ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO DI LAUREA MAGISTRALE IN
SPECIALIZED TRANSLATION

BIAS DI GENERE E TRADUZIONE AUTOMATICA

Esperimenti con il linguaggio non binario diretto nella traduzione dall'inglese all'italiano

Tesi di laurea magistrale in Machine translation

Relatore
Prof. Federico Garcea

Correlatrice
Prof. Rachele Raus

Correlatore
Prof. Alberto Barrón-Cedeño

Presentata da

Paolo Mainardi

Sessione luglio 2024

Anno Accademico 2023/2024

Alma Mater Studiorum Università di Bologna

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

Corso di Laurea magistrale Specialized Translation (classe LM – 94)

TESI DI LAUREA

in Machine translation

Bias di genere e traduzione automatica
Esperimenti con il linguaggio non binario diretto nella traduzione dall'inglese all'italiano

CANDIDATO:
Paolo Mainardi

RELATORE:
Federico Garcea

CORRELATRICE:
Rachele Raus

CORRELATORE:
Alberto Barrón-Cedeño

Anno Accademico 2023/2024

Primo Appello

Indice

Introduzione

Nel corso degli ultimi decenni, e con un'importante accelerazione negli anni più recenti, l'intelligenza artificiale (IA) è diventata un approccio centrale in moltissimi ambiti diversi, e di conseguenza, questa tecnologia è ormai utilizzata anche per prendere importanti decisioni che hanno un impatto a livello tanto sociale quanto individuale (vedi ad esempio Crawford, 2017).

Tuttavia, la ricerca affannosa di dataset e modelli sempre più grandi, a scapito di una considerazione più ampia dei presupposti su cui gli strumenti e i metodi dell'IA sono basati, ha portato allo sviluppo di una tecnologia profondamente ingiusta, che amplia le disuguaglianze sociali invece di ridurle. La riflessione critica sui danni causati dall'IA, quindi, mobilita una messa in discussione radicale del modello di conoscenza da cui è nata e su cui si basa questa tecnologia, anche tramite l'integrazione delle prospettive di altre discipline come la sociologia o l'etica (vedi ad esempio Birhane, 2021). Anche alla luce della rapida diffusione di queste tecnologie, negli ultimi anni ha preso sempre più importanza l'analisi degli assunti socioculturali (*bias*) alla base dei sistemi di IA e sui loro effetti negativi.

Un ambito dell'IA in rapida espansione è quello del trattamento automatico del linguaggio (TAL; anche NLP dall'inglese *natural language processing*), che comprende anche la traduzione automatica. Questa espansione è trainata in particolare dalla recente e ampia diffusione di applicazioni conversazionali (come ChatGPT[1]) basate su modelli linguistici di grandi dimensioni (LLM, dall'inglese *large language model*), che non sono naturalmente esenti dagli attuali problemi dell'IA.

Un aspetto specifico che sta ricevendo grande attenzione in questo ambito è quello del cosiddetto bias di genere (vedi ad esempio Blodgett et al., 2020; Savoldi et al., 2021). Come si vedrà meglio nel Capitolo 1, in particolare, le attuali pratiche linguistiche si basano sull'utilizzo del maschile come forma generica, una norma che comporta la cancellazione di tutte le altre identità, incluse le donne e le persone non binarie (ovvero tutte quelle che non si riconoscono nel binarismo di genere tipico della società occidentale, che consiste in una distinzione netta tra i generi maschile e femminile: vedi ad esempio Kendall, 2023).

I modelli di TAL, basati sulle pratiche linguistiche dominanti, riproducono questo comportamento. Per questo, negli ultimi anni si sta sviluppando sempre di più anche in questo ambito la riflessione sul bias di genere e sui possibili modi di integrare tale riflessione nei modelli e nei dati utilizzati per addestrarli, allo scopo di ridurre le disuguaglianze e garantire pari dignità e rappresentazione linguistica a qualunque persona in termini di identità di genere.

[1] https://openai.com/chatgpt/

Questo lavoro si inserisce in questa corrente, e si pone l'obiettivo di esplorare la possibilità di integrare l'espressione linguistica delle identità non binarie nella traduzione automatica dall'inglese all'italiano. È importante sottolineare che si tratta per lo più di uno studio, appunto, esplorativo: come si vedrà nel primo capitolo, la riflessione sugli effetti negativi dell'IA non può risolversi nei tentativi di aggiustare modelli esistenti o aggiungere dati più rappresentativi dei gruppi sociali dimenticati, ma deve stimolare una vera discussione sugli assunti socioculturali che hanno portato a questa situazione, e dev'essere quindi interdisciplinare e basata sulla comunità. Questo lavoro, necessariamente complesso e basato su sforzi collettivi, non è stato fatto per questa tesi, per ovvi limiti di tempo e risorse. Nonostante questo, speriamo che il contributo possa essere utile per la diffusione e l'avanzamento della riflessione su questi temi.

La tesi è divisa in tre capitoli principali: nel primo si discuterà di bias e IA, e in particolare di bias di genere nelle applicazioni legate al linguaggio, nonché delle pratiche linguistiche volte alla rappresentazione delle identità non binarie; nel secondo si illustreranno nel dettaglio gli esperimenti effettuati e i dati utilizzati; nell'ultimo verranno riportati e discussi i risultati ottenuti.

1. Bias (di genere) e linguaggio rispettoso del genere

In questo capitolo saranno introdotte alcune delle principali questioni legate all'etica dell'intelligenza artificiale (IA), in particolare in relazione al bias di genere nelle tecnologie linguistiche come la traduzione automatica e ai danni che ne conseguono, oltre a fornire una panoramica delle soluzioni proposte (§1.1). Si discuterà inoltre degli aspetti linguistici legati all'espressione dell'identità di genere (§1.2) e di linguaggio rispettoso del genere nelle sue diverse forme e della sua applicazione nelle tecnologie linguistiche (§1.3). Alcuni elementi legati all'approccio adottato in questo lavoro saranno introdotti in questo capitolo e poi approfonditi in quello successivo.

1.1. Bias e intelligenza artificiale

In scienze cognitive, il termine *bias* è utilizzato in relazione alle *euristiche*, ovvero processi mentali in cui alcune informazioni sono ignorate per arrivare velocemente a una decisione. In questo contesto, per bias si intende "the difference between human judgment and a 'rational' norm, often taken as a law of logic or probability" (Gigerenzer & Brighton, 2008: 117).

Con questa accezione, il concetto di bias è stato adottato nell'area dell'intelligenza artificiale e dell'apprendimento automatico (in inglese *machine learning*) (Savoldi et al., 2021), per affrontare le questioni etiche emerse in particolare con l'utilizzo di grandi collezioni di dati contenenti informazioni sensibili per l'addestramento dei sistemi (vedi Wallach, 2014). In questo ambito si parla in particolare di *bias predittivi*, ovvero bias propri di modelli predittivi in cui la distribuzione delle previsioni non rispetta quella ideale, e che danno risultati sistematicamente differenti per gruppi diversi (Shah et al., 2020).

Quando si tratta di sistemi basati su algoritmi, si parla inoltre di modelli e danni *sociotecnici*, un termine che sottolinea l'interazione tra questa tecnologia e il contesto sociale di riferimento: i potenziali danni causati da un sistema IA sono infatti tali in funzione di "existing power asymmetries and patterns of social inequality" (Shelby et al., 2023: 731). Questa prospettiva è fondamentale per comprendere, analizzare e affrontare i danni causati da sistemi IA.

Un caso tipico di applicazione di sistemi di IA predittiva con bias razziale è quello del *predictive policing*. Gebru (2020) porta l'esempio dei modelli utilizzati negli Stati Uniti per prevedere la probabilità che in una determinata zona si verifichi un crimine, con lo scopo di concentrare lì le attività delle forze di polizia. L'utilizzo dei dati relativi agli arresti precedenti, che colpiscono più frequentemente i gruppi marginalizzati (intesi come "communities that face structural forms of social exclusion": Shelby et al., 2023: 723), non farà che portare a un maggiore controllo sui quartieri abitati prevalentemente da tali comunità, che di conseguenza subiranno ancora più arresti, creando un

feedback loop che porta alla conferma e all'amplificazione delle disuguaglianze sociali (vedi anche Crawford, 2017).

Questi sistemi esibiscono quindi bias in termini di conferma di preconcetti, generalizzati dalla macchina a partire dai dati di addestramento (Lardelli & Gromann, 2022). Nell'ambito della ricerca delle fonti di bias in IA, una delle prime categorizzazioni è quella proposta da Friedman e Nissenbaum, ripresa ad esempio in Savoldi et al. (2021) in relazione al bias di genere in traduzione automatica (TA):

- i bias preesistenti sono propri del contesto socioculturale di riferimento e si riflettono nei dati utilizzati per addestrare i sistemi;
- i bias tecnici derivano dal modo in cui i dati sono manipolati per l'addestramento dei sistemi e in cui i sistemi stessi vengono costruiti;
- i bias emergenti diventano evidenti nel momento in cui i sistemi sono distribuiti, e derivano dall'amplificazione di bias preesistenti da parte dei modelli (anche in relazione a specifiche decisioni prese durante lo sviluppo di un sistema) o all'applicazione di un modello a determinati contesti di utilizzo.

In particolare, mentre i bias preesistenti ed emergenti sottolineano che "[h]arms from algorithmic systems emerge through the interplay of technical systems and social factors and can encode systemic inequalities" (Shelby et al., 2023: 724), i bias tecnici ed emergenti sono il risultato di specifiche scelte operate in diverse fasi dello sviluppo di un sistema, dalla raccolta dei dati al rilascio dei modelli (Suresh & Guttag, 2021; vedi anche Shah et al., 2020), tra cui:

- la definizione della popolazione e del campione (bias di rappresentazione) o delle etichette con cui annotare i dati (bias di misurazione);
- la (mancata) considerazione di specifiche differenze tra diversi sottogruppi che dovrebbero essere trattati diversamente affinché i risultati siano affidabili sull'intera popolazione (bias di aggregazione);
- la scelta dell'obiettivo statistico su cui modellare l'apprendimento del modello (ad esempio, la funzione obiettivo), che può portare a un'amplificazione di bias preesistenti (bias di apprendimento);
- la scelta delle *benchmark* utilizzate per la valutazione dei modelli (bias di valutazione): le scelte di chi sviluppa i modelli possono essere influenzate dalla volontà di raggiungere alte prestazioni sui dataset più comunemente utilizzati a questo scopo, senza considerare che questo potrebbe implicare, ad esempio, l'amplificazione di determinati stereotipi;
- la distribuzione del modello (bias di distribuzione): altri danni possono emergere nel momento della distribuzione del modello, in particolare se quest'ultimo viene utilizzato

da utenti o per scopi che non corrispondono a quelli presi in considerazione durante la fase di sviluppo.

In virtù del loro carattere sociotecnico, i danni causati dall'IA sono riconducibili a diversi aspetti dell'identità di un individuo o di un gruppo. Quando si tratta di tecnologie linguistiche, il bias di genere, in particolare, è diventato recentemente un tema di primo piano (vedi ad esempio Sun et al., 2019) alla luce di una rinnovata sensibilità riguardo al genere come costrutto sociale e la sua relazione con il genere come categoria linguistica, espressa diversamente in lingue diverse (vedi §1.2).

Nel prossimo paragrafo si chiarirà innanzitutto il ruolo e l'impatto del bias nell'IA in generale, per poi passare alla discussione specifica sul bias di genere in TAL e TA.

1.1.1. I danni causati da un'IA portatrice di bias

In generale, il concetto di bias è diventato fondamentale nella ricerca sull'etica dell'IA, in virtù della sempre maggiore importanza ricoperta da questa tecnologia nella società contemporanea. Infatti, anche se i bias non sono (solamente) un prodotto dei sistemi di IA, decisioni discriminatorie prese (anche) tramite questi sistemi possono avere importanti effetti negativi nella vita reale delle persone (vedi ad esempio Crawford, 2017).

Una delle tassonomie più complete dei potenziali danni dell'IA è quella compilata da Shelby et al. (2023) in base alla letteratura esistente, che comprende le seguenti macrocategorie:

- danni di rappresentazione: differenze nella rappresentazione di gruppi sociali diversi in base a tratti identitari;
- danni di allocazione: distribuzione iniqua di risorse a gruppi sociali diversi;
- danni relativi alla qualità del servizio: differenza nelle prestazioni di un sistema per utenti appartenenti a gruppi sociali diversi;
- danni interpersonali: danni alle relazioni tra individui o comunità;
- danni sociali: altri danni socioeconomici, politici e ambientali.

Gli effetti di questi danni seguono schemi normativi e gerarchie sociali esistenti, amplificando la discriminazione di gruppi già marginalizzati (vedi anche Gebru, 2020; D'Ignazio & Klein, 2020).

In questo lavoro si darà particolare importanza ai danni di rappresentazione. Nelle applicazioni di trattamento automatico del linguaggio (TAL), il bias di genere si manifesta infatti in modo evidente nella ripresa di ruoli di genere stereotipici (Bolukbasi et al., 2016) e nella cancellazione delle identità non binarie (Dev et al., 2021). Come sottolinea Crawford (2017), i danni di rappresentazione hanno un ruolo fondamentale nella propagazione di stereotipi preesistenti, e sono alla base di danni di altro tipo.

Nel caso specifico della TA, il bias di genere può manifestarsi, ad esempio, nell'assegnazione del genere maschile a utenti di un genere diverso (*misgendering*: vedi Dev et al., 2021; Lardelli &

Gromann, 2023b), come conseguenza dell'uso del maschile sovraesteso (vedi §1.2.1) nei dati di addestramento. La qualità del servizio è quindi compressa per determinati gruppi di utenti: se, in questo contesto, il sistema stesso e le traduzioni che produce sono la risorsa allocata (Savoldi et al., 2021), allora il misgendering da parte di un sistema di TA rappresenta un esempio di danno di allocazione che deriva dalla sottorappresentazione delle identità di genere diverse da quella maschile nei dati di addestramento.

Prima di parlare più nel dettaglio di bias di genere a livello linguistico, nel prossimo paragrafo saranno delineati alcuni dei principali approcci alla riduzione del bias.

1.1.2. La mitigazione del bias di genere in TAL

Una volta individuate e analizzate le principali questioni etiche legate al bias nell'IA, è naturale chiedersi quali siano le possibili soluzioni. In generale, la riduzione dei bias e dei danni causati dall'IA rappresenta un'operazione complessa, il che spiega la varietà di soluzioni proposte e analizzate nel tempo.

Per esempio, dal resoconto di Ferrara (2024) emerge che diverse tecniche di riduzione del bias possono essere messe in atto in tre fasi, ovvero al momento di preparare i dati (*data pre-processing*), nella scelta dei modelli da utilizzare, e una volta che sono disponibili gli output di tali modelli.

Nell'ambito del TAL, Sun et al. (2019) analizzano due modi di agire sui dati per ridurre specificamente il bias di genere: i dati di addestramento (corpora) e le rappresentazioni computazionali di tali dati (come le *word embeddings*: Bolukbasi et al., 2016; Guo & Caliskan, 2021). Lardelli & Gromann (2023a), invece, indagano il *post-editing* (ovvero la modifica, a posteriori, delle traduzioni fornite da un sistema di TA) come metodo di riduzione del bias maschile nelle traduzioni automatiche.

Una tecnica adottata in più lavori per ridurre il bias presente nei dati di addestramento è la *counterfactual data augmentation* (CDA; introdotta da Lu et al., 2020), che consiste nell'aggiungere nuovi dati che controbilanciano la rappresentazione esistente. Si tratta quindi di una tecnica mirata all'eliminazione della sottorappresentazione e della stereotipizzazione, che, come visto sopra, sono fondamentali nella riduzione dei danni causati dai sistemi IA. In TA, la CDA è stata applicata allo scopo di creare dataset bilanciati rispetto al genere, da utilizzare non solo come benchmark per la valutazione del bias di genere nei sistemi di TA (Stanovsky et al., 2019; Bentivogli et al., 2020; Vanmassenhove et al., 2021; Currey et al., 2022; Piergentili et al., 2023b), ma anche per ottimizzare modelli preesistenti (Costa-jussà & de Jorge, 2020; Saunders & Byrne, 2020) o per addestrare nuovi modelli di post-editing automatico (Zmigrod et al., 2019; Vanmassenhove et al., 2021; Sun et al., 2021).

Un esempio di progetto ad ampio respiro per la riduzione del bias (non solo di genere) in TA è quello di *Fairslator*[2] (Měchura, 2022). Il sito offre un plug-in per il post-editing automatico di tre sistemi di TA quando si traduce dall'inglese verso quattro lingue a genere grammaticale (ceco, tedesco, francese, irlandese; vedi §1.2), fornendo all'utente la possibilità di controllare in che modo l'output del sistema scelto viene modificato secondo le proprie esigenze. Fairslator permette di agire su diversi assi di ambiguità propri dell'inglese come lingua di partenza: oltre al genere dei referenti, il numero e il livello di formalità da usare nella traduzione del pronome inglese *you*. L'importanza di pensare soluzioni che permettano una maggiore personalizzazione dell'output dei sistemi di TA, in particolare per quanto riguarda il genere, è evidenziata ad esempio da Gromann et al. (2023), che si basano sui suggerimenti ricevuti da persone queer e non binarie.

Mentre Fairslator integra un modello basato su regole scritte a mano per modificare le traduzioni automatiche (in modo simile a Diesner-Mayer & Seidel, 2022), altri lavori puntano all'automazione di questo processo. I già citati lavori di Vanmassenhove et al. (2021) e Sun et al. (2021) si concentrano su una combinazione di sistemi basati su regole, che garantiscono risultati più robusti e interpretabili, e sistemi neurali, che restano meno ancorati ai dati di addestramento e offrono una migliore generalizzazione a dati nuovi (*unseen data*), seppur al costo di una minore trasparenza. Kostikova et al. (2023), invece, valutano la capacità di un sistema di TA adattivo di adattare le proprie traduzioni al linguaggio rispettoso del genere in tempo reale, dimostrando le potenzialità di questo approccio.

Infine, data la sempre maggiore popolarità dei modelli linguistici di grandi dimensioni (LLM) per un'ampia varietà di compiti di TAL, diversi lavori recenti esplorano la possibilità di sfruttarne l'adattabilità per ridurre il bias di genere tramite istruzioni ed esempi specifici, come si vedrà meglio nel prossimo capitolo.

1.1.3. Ripensare l'IA

Nonostante tutte le strategie esistenti volte alla mitigazione dei diversi tipi di bias, è fondamentale tenere a mente che la riduzione del bias in IA deve fare parte di uno sforzo interdisciplinare, e partire da un'attenta considerazione del rapporto di questa tecnologia con il contesto socioculturale ed economico: come visto all'inizio di questo capitolo, i sistemi di IA sono prodotti sociotecnici che non fanno altro che riflettere o esacerbare squilibri esistenti nella società di riferimento.

È quindi necessario ripensare a fondo gli approcci propri di questa scienza, a partire dal modo in cui l'ottimizzazione dei modelli e la raccolta dei dati sono state concepite fino a oggi (Wallach, 2014; Hardt, 2014; Barocas & Selbst, 2016), nonché rimettere al centro e dare voce alle persone che

[2] https://www.fairslator.com/

subiscono danni a causa dell'uso di questi sistemi (Blodgett et al., 2020; D'Ignazio & Klein, 2020; Gebru, 2020). Concretamente, è fondamentale porsi domande precise sulle dinamiche di potere che un sistema IA potrebbe riprodurre, fin dalla sua concezione: come ricordano D'Ignazio & Klein (2020), "what gets counted counts", e come sottolineano Shelby et al. (2023), lo sviluppo di un sistema basato su algoritmi deve essere accompagnato da una riflessione sulla semplificazione della realtà che si sta necessariamente operando e sui suoi effetti. Gebru (2020) spiega come la relazione dell'IA come scienza con i gruppi dominanti porti necessariamente ad approcci che non fanno altro che favorire questi gruppi e peggiorare la situazione di quelli già marginalizzati: "Who creates the technology determines whose values are embedded in it" (p. 264). L'autrice sottolinea infatti che questo problema continua a esistere anche nella ricerca sull'IA etica, che rimane legata agli stessi attori politico-economici, non riuscendo davvero a dare voce alle comunità che dovrebbe difendere.

Affinché la ricerca in questo campo raggiunga effettivamente il suo scopo di migliorare la rappresentazione e ridurre la discriminazione sociale dei gruppi marginalizzati, è fondamentale integrare le prospettive di altre discipline (ad esempio, la sociologia per comprendere i bias presenti nei dati [Wallach, 2014], o gli studi di genere e intersezionali per quanto riguarda il bias di genere [Leavy, 2018; Guo e Caliskan, 2021]), nonché coinvolgere le comunità impattate da una determinata tecnologia (come fanno, ad esempio, Gromann et al. [2023] per la TA): come ricorda Birhane (2021), "Given that harm is distributed disproportionately and that the most marginalized hold the epistemic privilege to recognize harm and injustice, […] for any solution that we seek, the starting point [must] be the individuals and groups that are impacted the most" (p. 5).

Alla luce di quanto visto fin qui, è importante sottolineare quindi che il presente lavoro è inteso semplicemente come uno studio esplorativo sulla possibilità di applicare strategie nuove a tecnologie di TA e TAL, e non come una soluzione a tutto tondo del problema del bias di genere.

I metodi e le strategie adottate a questo scopo saranno descritti nel dettaglio nel Capitolo 2. Prima, si discuterà dei modi in cui il bias di genere si manifesta nelle lingue e nella traduzione.

1.2. Genere e lingua in traduzione

La traduzione del genere pone problemi specifici presto identificati negli studi sulla traduzione (vedi Lardelli & Gromann, 2022). Infatti, il genere è espresso nelle varie lingue in modi diversi, e questo introduce delle asimmetrie che devono essere risolte quando si traduce tra due lingue che codificano il genere diversamente. Rispetto a queste differenze, si distinguono solitamente tre categorie (Savoldi et al., 2021; Lardelli, 2023; vedi anche Sczesny et al., 2016)[3]:

[3] Al di là di questa categorizzazione, con il termine "genere grammaticale" si intende in generale la categoria morfologica secondo cui nomi e altre parole sono divise in classi (Grandi, 2010).

- le lingue senza genere grammaticale (*genderless*, es. cinese mandarino, turco, lingue ugrofinniche) non hanno marche morfologiche del genere, ma esistono solitamente distinzioni lessicali (es. *fratello*, *sorella*);
- le lingue a genere naturale o nozionale (es. lingue germaniche a esclusione del tedesco) hanno un genere grammaticale, che però è raramente marcato, solitamente nel sistema pronominale (es. *lui*, *lei*), e in alcune distinzioni lessicali (in inglese, ad esempio, *mother* e *father*, ma anche *actor* e *actress*);
- le lingue a genere grammaticale (es. lingue romanze) sono quelle in cui il genere grammaticale ha l'impatto più importante: tutti i nomi (animati o inanimati), infatti, hanno un genere grammaticale, e anche gli aggettivi e alcune forme verbali sono solitamente marcate con lo stesso genere del nome o pronome a cui si riferiscono.

È importante sottolineare che il bias di genere (che, come visto, è un fatto socioculturale, più che linguistico) esiste in tutte queste categorie di lingue. Le strategie per l'eliminazione del bias di genere sono però diverse in base al sistema di genere grammaticale della lingua in questione: Sczesny et al. (2016) identificano le principali strategie adottate in lingue diverse.

In questo lavoro si parlerà in particolare della relazione tra due lingue rappresentative delle ultime due categorie, ovvero l'inglese e l'italiano. Nel caso specifico della traduzione tra queste due lingue, il problema del genere si pone perché in inglese la maggior parte dei nomi non ha un genere grammaticale, mentre in italiano tutti i nomi ce l'hanno, e le marche di genere si estendono a tutte le parti della frase collegate a un sostantivo (a causa della natura flessiva dell'italiano). Mentre il genere dei sostantivi riferiti a oggetti e concetti astratti non è legato ad alcuna categoria extralinguistica, il genere grammaticale dei referenti umani è tendenzialmente legato all'identità di genere della persona a cui ci si riferisce. Di conseguenza, quando si traduce una frase inglese in cui non c'è alcuna informazione sul genere dei referenti umani, bisogna necessariamente fare una scelta sul genere da usare in italiano (vedi Nissen, 2002).

Nella prossima sezione si vedrà perché questa scelta ricade quasi sempre sul maschile.

1.2.1. Il maschile sovraesteso

La situazione descritta sopra si applica, con le dovute differenze, a tutte le lingue a genere grammaticale (che includono non solo tutte le lingue romanze, ma anche quelle slave e semitiche e il tedesco, tra le altre), non solo in traduzione, ma anche quando si deve scrivere o parlare di persone di cui non si può o non si vuole esplicitare il genere.

In molte lingue si è affermata la regola del cosiddetto *maschile sovraesteso*, ovvero l'uso di forme al maschile per riferirsi a una o più persone di qualunque genere. Il maschile sovraesteso è particolarmente evidente nelle lingue a genere grammaticale e con una morfologia ricca come

l'italiano, ma si trova anche nelle lingue appartenenti alle altre categorie, come sottolineato da Sczesny et al. (2016), per esempio nell'utilizzo della parola per *uomo* con il senso di "essere umano".

Per capire l'impatto del maschile sovraesteso in lingue come l'italiano, è utile identificare i diversi contesti in cui viene utilizzato, in relazione al grado di specificità e di conoscenza del genere del referente da parte di chi scrive o parla (o traduce). Si possono distinguere infatti referenti specifici, che si riferiscono a una o più persone specifiche e fisse nel tempo, e referenti generici, che possono riferirsi a una o più persone diverse in base alla situazione; il genere di un referente, invece, può essere conosciuto, sconosciuto, misto (nel caso di gruppi di persone) o irrilevante (nel caso di referenti generici). In base a queste caratteristiche dei referenti, si possono categorizzare i diversi usi del maschile sovraesteso; per esempio, Raus (2015) distingue tra:

- maschile inclusivo: riferito a persone o gruppi di persone di genere sconosciuto o misto;
- maschile neutro o non marcato: utilizzato per referenti generici, il cui genere è quindi irrilevante;
- maschile estensivo: riferito a persone specifiche di genere non maschile.

Per esemplificare questi diversi usi del maschile sovraesteso, si considerino le seguenti frasi, tratte dal corpus GeNTE[4] (Piergentili et al., 2023b):

(1) Noi politici dobbiamo entrare a pieno titolo in questo dibattito e non lasciarlo esclusivamente alle burocrazie.

(2) Man mano che le Istituzioni europee ottengono sempre più potere sui cittadini, diviene importante porre dei limiti a questa influenza.

(3) Ringrazio la signora Commissario per il suo costante impegno nel campo della sicurezza marittima.

Nel primo esempio, "noi politici" rappresenta un maschile inclusivo con cui ci si riferisce a tutta la classe politica nel suo complesso; nel secondo, "i cittadini" è un maschile neutro, che si riferisce al gruppo eterogeneo della cittadinanza dell'Unione europea; l'ultima frase contiene invece un tipico esempio di maschile estensivo, dove si utilizza un nome professionale al maschile ("Commissario") per un referente il cui genere è conosciuto ed è femminile, come indicato dal titolo "signora". Su quest'ultimo punto, è importante notare che in Italia c'è ancora molta resistenza all'utilizzo delle forme femminili dei nomi riferiti a professioni e cariche istituzionali, soprattutto per quanto riguarda posizioni di alto prestigio socioeconomico, da cui le donne sono state e sono tuttora in gran parte escluse (vedi ad esempio Fusco, 2019 e Giusti, 2022; vedi anche Gheno, 2020b per una panoramica del dibattito sul tema).

[4] https://mt.fbk.eu/gente/

L'utilizzo del maschile sovraesteso comporta la sottorappresentazione delle persone che non vi si riconoscono, sia nei dati utilizzati per addestrare i modelli, sia, di conseguenza, nei loro output: alla luce di quanto visto su bias e IA, si può affermare quindi che si tratta di un esempio di bias preesistente, che si fa strada e viene amplificato dai modelli di TAL e TA (Savoldi et al., 2021).

In italiano, come in altre lingue in cui il genere grammaticale ha un ruolo importante, gli sforzi per la riduzione del bias di genere si concentrano in particolare sul superamento del maschile sovraesteso. In effetti, nonostante questa regola investa unicamente il genere grammaticale dei referenti, è stato ampiamente studiato il suo effetto sulle rappresentazioni mentali di chi legge o ascolta: infatti, a livello cognitivo, le forme grammaticali maschili, nonostante la regola, sono tendenzialmente percepite non in modo generico e inclusivo, ma esclusivo, ovvero come riferite unicamente o prevalentemente a uomini (Gygax et al., 2008). Il maschile sovraesteso, quindi, risulta essere un fattore che contribuisce al mantenimento della posizione dominante degli uomini nella società e alla cancellazione delle altre identità di genere, come dimostra tra l'altro la resistenza all'uso di forme femminili perfettamente grammaticali.

In generale, come già visto, quando si tratta di referenti umani, si instaura una corrispondenza tra genere grammaticale (binario in italiano) e genere sociale o identità di genere dell'individuo, che può rientrare in un sistema binario o meno (vedi Knisely, 2020). Nel secondo caso, gli strumenti linguistici standard forniti da una lingua con un sistema di genere grammaticale binario sono inadeguati all'espressione di identità che si situano al di fuori di tale binarismo.

Per questo, alla luce dell'importanza del linguaggio nella costruzione dell'identità individuale e tramite lo svelamento della matrice socioculturale alla base di norme come quella del maschile sovraesteso, i movimenti femministi cominciano dagli anni '70 a sovvertire questi usi (Pusterla, 2019). In Italia, uno dei primi e più famosi contributi scientifici e istituzionali è quello di Alma Sabatini (1987), che parla di "sessismo della lingua italiana" e propone soluzioni per evitare il maschile sovraesteso e garantire alle donne la stessa rappresentazione degli uomini.

I movimenti transfemministi che si sono diffusi nel ventunesimo secolo integrano le riflessioni sul linguaggio delle teorie queer, e spingono il dibattito verso una più radicale messa in discussione del sistema binario tanto a livello socioculturale quanto linguistico (Pusterla, 2019; vedi Non una di meno, 2017). Il dibattito sulla rappresentazione linguistica delle identità non binarie si è diffuso quindi in molte lingue del mondo, in particolare quelle in cui il genere grammaticale è più preponderante (vedi, ad esempio, Comandini, 2021 e Formato & Somma, 2023 per l'italiano; Lardelli & Gromann, 2023b per il tedesco e l'italiano; Ashley, 2019 e Knisely, 2020 per il francese; López,

2019 per lo spagnolo)[5]. Questi nuovi usi della lingua, corrispondenti a una nuova sensibilità rispetto al tema del genere, possono essere comprese sotto l'etichetta di linguaggio rispettoso del genere (GFL dall'inglese *gender-fair language*: Sczesny et al., 2016).

Questo lavoro si situa all'interno di un approccio non binario: nei prossimi paragrafi si discuterà più nel dettaglio del GFL e, in particolare, delle specifiche strategie elaborate in italiano e adottate in questo lavoro.

1.3. Il linguaggio rispettoso del genere (in italiano)

Come visto nel paragrafo precedente, il GFL comprende diverse strategie volte a una migliore rappresentazione delle diverse identità di genere e al superamento del maschile sovraesteso, anche allo scopo di evitare l'emergenza di fenomeni legati al bias di genere in IA.

Queste strategie possono essere suddivise in primo luogo in strategie binarie e non binarie. Come già accennato, l'approccio binario è stato il primo a emergere, nell'ambito delle prime ondate del femminismo; nei decenni successivi e con l'affermazione della prospettiva intersezionale e transfemminista, si sono diffusi gli approcci non binari (Pusterla, 2019; Ludbrook, 2022).

Inoltre, le strategie di GFL possono essere categorizzate in strategie *di visibilità* e *di oscuramento* (Robustelli, 2012; anche *gender-inclusive* e *gender-neutral*: Lardelli, 2023). Le prime hanno l'obiettivo di dare visibilità esplicita ai generi sottorappresentati, mentre le seconde permettono di non esprimere esplicitamente alcun genere specifico.

L'approccio binario è ancora quello dominante non solo in molti studi sul bias di genere in TAL per varie lingue (ad esempio: Zhao et al., 2018; Habash et al., 2019; Stanovsky et al., 2019; Zmigrod et al., 2019; Bentivogli et al., 2020; Costa-jussà & de Jorge, 2020; Gonen & Webster, 2020; Saunders & Byrne, 2020; Alhafni et al., 2022; vedi anche Dev et al., 2021; Piergentili et al., 2023a), ma anche nelle più recenti linee guida istituzionali sul GFL per l'italiano (ad esempio: MIUR, 2018; Parlamento Europeo, 2018; Thornton, 2020; Università di Bologna, 2020). Queste ultime, in particolare, adottano prevalentemente una visione binaria del genere, suggerendo ad esempio di utilizzare le forme maschili e femminili una accanto all'altra (il cosiddetto *sdoppiamento*, una strategia di visibilità), ma ammettendo anche l'uso del maschile sovraesteso in casi specifici (ad esempio, per riferirsi a cariche e ruoli in modo generico, quando non si fa riferimento alla persona che le ricopre: Parlamento Europeo, 2018) o a patto di adottare accorgimenti come l'aggiunta di un avviso sull'uso del maschile con valore generico e inclusivo (Università di Bologna, 2020).

[5] Il progetto Gender in Language (Papadopoulos, 2022) propone inoltre una panoramica dei sistemi di genere grammaticale e delle strategie di rappresentazione linguistica delle identità non binarie in diverse lingue.

Lo sdoppiamento è la principale strategia binaria di visibilità suggerita nelle linee guida per l'italiano, e consiste nell'affiancare le forme maschili e femminili, in forma estesa o contratta (vedi Raus, 2015), generalmente per riferirsi a gruppi misti o a referenti generici. Ad esempio:

(4) *gli alunni* > le alunne e gli alunni, le/gli alunne/i

(MIUR, 2018: 20)

Il principale problema legato a questa strategia è che può appesantire notevolmente la lettura, specialmente se ripetuta su tutti gli elementi della frase che concordano con il soggetto. Per questo motivo, le linee guida suggeriscono generalmente di mantenere l'accordo con la forma sdoppiata al plurale maschile, tornando di fatto a una forma di maschile sovraesteso.

Le strategie di oscuramento – come l'utilizzo di termini epiceni e collettivi o di frasi impersonali (5) – hanno generalmente un impatto minore sulla leggibilità:

(5) *i candidati* invieranno il curriculum > si invierà il curriculum

(Parlamento europeo, 2018: 12)

Come si vedrà meglio più sotto, tali strategie sono riconducibili a una forma di linguaggio non binario, pur non essendo riconosciute come tali in questi documenti.

Alcuni documenti recenti riconoscono esplicitamente l'esistenza di identità che si situano al di fuori del binarismo di genere: le linee guida pubblicate dall'Agenzia delle Entrate (2020) riconoscono la crescente diffusione delle strategie non binarie non standard (da cui, però, mettono in guardia), mentre altre, come quelle dell'Università di Padova (2017) suggeriscono di lasciare alcune forme aperte all'autodeterminazione, come *l'interessat_*, nel caso specifico dei moduli da compilare, dove "la necessità di assicurare visibilità a entrambi i generi è più importante della gradevolezza stilistica del testo" (Parlamento europeo, 2018:13).

L'approccio non binario al GFL può essere ulteriormente suddiviso in due gruppi di strategie (López, 2019): il linguaggio non binario diretto (LND) permette di mettere in evidenza le identità di genere non binarie (strategia di visibilità), mentre il linguaggio non binario indiretto (LNI) si basa sull'eliminazione di riferimenti espliciti a qualunque genere (strategia di oscuramento). Le strategie di LNI non escono dai confini della lingua standard e per questo compaiono anche nelle linee guida ufficiali, mentre quelle di LND si situano sul piano sperimentale; è importante però sottolineare che le etichette LNI e LND non identificano categorie discrete e opposte: i due approcci possono coesistere, ma investono parti diverse della lingua e hanno una portata sociolinguistica molto differente.

Nei prossimi paragrafi saranno descritte le principali strategie di LNI e LND applicabili in italiano, che si tratti di traduzione o meno. Prima, è utile specificare che il termine "neutro" sarà

utilizzato per riferirsi a frasi o espressioni linguistiche che non esplicitano il genere dei referenti umani, mantenendosi ambigue.

1.3.1. Linguaggio non binario indiretto

Come anticipato, al contrario del LND, che non ha status ufficiale, le strategie proprie del LNI compaiono anche nelle guida istituzionali sul GFL in italiano, pur senza essere riconosciute come forme di linguaggio non binario. Le strategie di LNI possono infatti essere sovrapposte alle strategie di oscuramento adottate anche all'interno di approcci binari, dato che consistono principalmente nell'evitare le forme linguistiche portatrici di informazioni sul genere dei referenti.

Questo è possibile, ad esempio, utilizzando nomi di genere promiscuo (vedi Gheno, 2020a), che si possono riferire a persone di qualunque genere indipendentemente dal genere grammaticale del sostantivo (6), oppure di riformulazioni in frasi passive, relative (7) o impersonali (vedi anche Raus, 2015):

> (6) *Il responsabile* del procedimento amministrativo > La persona responsabile del procedimento amministrativo
>
> (7) L'assicurazione contro le malattie è a carico *del fruitore* della borsa > L'assicurazione contro le malattie è a carico di chi fruisce della borsa
>
> (Università di Padova, 2017: 20)

L'utilizzo di nomi epiceni – ovvero che mantengono la stessa forma al maschile e al femminile (Gheno, 2020a) – può non essere una soluzione definitiva, dal momento che il genere della parola può comunque essere indicato dall'articolo o da altri elementi che concordano con il soggetto. Nell'esempio (6), *responsabile* è un nome epiceno, ma il genere è indicato dall'articolo *il*; per eliminare ogni marca di genere è quindi necessario utilizzare un nome di genere promiscuo, come *la persona*.

Un'altra difficoltà legata al LNI riguarda la sua applicazione a testi specialistici, ad esempio di tipo legale:

> (8) [...] *i creditori*, *obbligazionisti* o no, ed *i portatori* di altri titoli delle società partecipanti alla scissione devono essere *tutelati* onde evitare che la realizzazione della fusione *li* leda;
>
> (8a) [...] le persone titolari di credito, tra cui chi detiene titoli obbligazionari, e chi detiene altri titoli delle società partecipanti alla scissione devono avere una

tutela onde evitare che la realizzazione della fusione risulti in una lesione dei loro interessi;

JRC-Acquis, EN-IT[6]

L'esempio (8) riporta una frase presa dal corpus parallelo di leggi europee (JRC-Acquis: Steinberger et al., 2006), dove la traduzione italiana contiene diversi maschili sovraestesi (in corsivo); la proposta di riformulazione di questa frase in LNI (8a) comporta modifiche complesse alla sintassi della frase e, soprattutto, la sostituzione di termini specifici come *creditori obbligazionisti*, che potrebbe non essere accettabile. L'applicazione di questa strategia a testi legali e altri testi specialistici, dove la chiarezza è una necessità di primo piano e il linguaggio è particolarmente rigido, dovrebbe quindi essere idealmente supervisionata da una figura esperta di quel campo, e potrebbe non essere possibile in alcuni casi, soprattutto a causa della necessità di mantenere una terminologia specifica e condivisa.

In questi casi, l'unica strategia possibile sarebbe quella dello sdoppiamento (del tipo *i/le creditori/e obbligazionisti/e*), che però riporta a un sistema binario. Come si vedrà nel prossimo paragrafo, il LND, pur non essendo compatibile con la grammatica standard, può permettere di adottare un linguaggio neutro rispetto al genere rispettando la terminologia specifica ed evitando di appesantire eccessivamente la sintassi.

1.3.2. Linguaggio non binario diretto

Come dice il nome, il LND si basa sulla rappresentazione diretta, esplicita, delle identità non binarie, spesso tramite soluzioni linguistiche innovative che possono agire sul vocabolario o sulla morfologia della lingua. Per questo motivo, le strategie di LND, così come le istanze della comunità queer e non binaria, non sono menzionate nella maggior parte delle linee guida ufficiali sull'uso del GFL in italiano.

Le strategie di LND nascono infatti negli ambienti transfemministi e queer, e sono proposte dal basso volte a rispondere a un'esigenza delle persone direttamente interessate da questi aspetti (Pusterla, 2019; Gheno, 2022; Acanfora, 2022). Per questo motivo, diverse strategie sono emerse nel tempo nelle varie lingue, e, allo stato attuale, coesistono a vari livelli.

Per quanto riguarda l'italiano, Comandini (2021) passa in rassegna le diverse strategie di LND che si sono diffuse in Italia negli ultimi anni. Dalla sua analisi emerge che le due strategie più utilizzate online erano quelle dell'asterisco (*) e dello schwa (ə). Quest'ultima strategia in particolare ha avuto ampio spazio nel dibattito pubblico italiano negli ultimi anni (Gheno, 2022; Sulis & Gheno, 2022).

[6] https://opus.nlpl.eu/JRC-Acquis/en&it/v3.0/JRC-Acquis

La proposta dello schwa nasce in particolare per sopperire all'impossibilità di pronuncia delle precedenti proposte grafiche come l'asterisco (*) o la chiocciola (@): si tratta infatti di un grafema mutuato da un simbolo dell'alfabeto fonetico internazionale (IPA), /ə/, usato per indicare la vocale centrale media non arrotondata, pronunciata con il tratto fonatorio rilassato (Gheno, 2022). Non è un suono proprio dell'italiano standard, ma esiste in molte altre lingue tra cui l'inglese e il francese, oltre che in diverse lingue regionali italiane. Negli ultimi anni, il suo uso si è diffuso soprattutto negli ambienti transfemministi e vicini alla comunità queer, sia allo scritto che al parlato, e si sta facendo strada anche in ambiti istituzionali (Sulis & Gheno, 2022; Proto, 2021).

La diffusione dello schwa come strategia di LND è passata anche attraverso la sua sistematizzazione da parte del progetto Italiano inclusivo[7] già dal 2015 e di case editrici come effequ[8], che nel 2020 l'ha adottata proprio per risolvere un problema di traduzione, ovvero per tradurre un testo in portoghese che conteneva desinenze non binarie; la casa editrice ha quindi esteso l'uso dello schwa alla sua intera collana di saggistica, proponendo una strategia diversa da quella di Italiano inclusivo, principalmente per il fatto che la distinzione tra singolare e plurale nella declinazione dei nomi è stata abbandonata (vedi anche Gheno, 2020b; Cavallo et al., 2021; Papadopoulos et al., 2022; Lardelli & Gromann, 2023b).

L'uso di strategie di LND come lo schwa è in certi casi utile quando il LNI appesantirebbe troppo il discorso; inoltre, come visto in §1.3.1, talvolta non è possibile sostituire un termine specialistico con un'alternativa epicena, perciò il LND può rendere un termine neutro rispetto al genere, senza cambiarne la radice lessicale. La frase in (8) può quindi essere riscritta anche in questo modo:

> (8b) […] ə creditorə, obbligazionistə o no, ed ə portatorə di altri titoli delle società partecipanti alla scissione devono essere tutelatə onde evitare che la realizzazione della fusione lə leda.
>
> JRC-Acquis, EN-IT[6]

La riformulazione in (8b) richiede in effetti meno modifiche, e permette appunto di mantenere la stessa sintassi e terminologia della frase originale.

In questo lavoro, lo schwa sarà utilizzato come morfema che indica un genere non specificato o non binario, tanto per referenti generici quanto per referenti specifici di genere sconosciuto o non binario (vedi §1.2.1), come strategia di LND volta alla diffusione del GFL in italiano e secondo il principio per cui "sarebbe forse corretto identificare questi tentativi come la ricerca non di *neutro* o di un *terzo genere*, ma di una forma *priva di genere*" (Gheno, 2022). Le norme specifiche che hanno

[7] https://italianoinclusivo.it/
[8] https://www.effequ.it/schwa/

guidato la riscrittura delle frasi in questo lavoro saranno descritte nel dettaglio nel prossimo capitolo (§2.6).

2. Metodo

Gli esperimenti effettuati sono volti all'adattamento di due modelli, utilizzati per la traduzione automatica, ModernMT (MMT) e Llama 3 (da qui in poi anche solo Llama), al linguaggio non binario diretto (LND: vedi §1.3.2) nella traduzione verso l'italiano di frasi inglesi ambigue rispetto al genere dei referenti umani, con l'obiettivo principale di evitare l'uso del maschile sovraesteso (definito in §1.2.1). Come contributo aggiuntivo, sono stati inoltre addestrati due classificatori automatici che identificano frasi inglesi e italiane come marcate o neutre (§2.5).

Dunque, in base a quanto visto nello scorso capitolo, è evidente che non si tratta di soluzioni soddisfacenti per affrontare il problema del bias di genere in TAL nel suo complesso, ma piuttosto di soluzioni più immediate e legate a un problema specifico. Queste soluzioni richiedono infatti una quantità limitata di dati e sono relativamente semplici da implementare, dal momento che si basano su modelli già esistenti. Per affrontare il problema in modo più completo sarebbe necessario pensare alla creazione di un diverso tipo di IA, un lavoro al di fuori degli scopi e delle ambizioni di questa tesi.

Entrambi i modelli testati sono basati sull'architettura *transformer* (Vaswani et al., 2017) – oggi alla base di numerose applicazioni di TAL (vedi ad esempio Patwardhan et al., 2023) – ma sono fondamentalmente diversi: da un lato, ModernMT include due componenti principali, un *encoder* e un *decoder*; dall'altro, LLaMa 3 è un cosiddetto *decoder-only*. I modelli delle famiglie Llama, GPT o T5, tra le altre, sono conosciuti anche come *large language models* (LLM) perché contengono milioni o miliardi di parametri e sono addestrati per avere una comprensione generale di una o più lingue, in modo da poter svolgere una grande varietà di compiti in TAL; al contrario, ModernMT è pensato unicamente come sistema di TA. Infine, i modelli *encoder-decoder* sono addestrati sulla generazione di testo condizionata a un input specifico, mentre i *decoder-only* sono spesso ottimizzati sulla conversazione (*instruction-tuned*), quindi sulla continuazione dell'input ricevuto.

Nonostante siano addestrati principalmente come strumenti di generazione di testo, i LLM, date le loro grandi dimensioni, possono acquisire diverse *capacità emergenti* (vedi ad esempio Wei et al., 2022), ovvero che non rientrano nei compiti su cui sono stati originariamente addestrati, tra cui la traduzione. Da questa proprietà deriva la possibilità di controllarne l'output, fornendo direttamente al modello delle istruzioni specifiche in linguaggio naturale e/o degli esempi del compito da svolgere, senza ottimizzazione dei parametri (una tecnica chiamata *few-shot prompting*: Brown et al., 2020), cosa che non è possibile con sistemi di TA come ModernMT. Come visto nel precedente capitolo, in effetti, la possibilità di controllare le traduzioni di un sistema di TA è un passaggio fondamentale

nella mitigazione del bias di genere. Per questo motivo, recentemente si sono cominciate a esplorare le possibilità offerte da questi modelli in questo ambito.

Per esempio, Sánchez et al. (2023) e Vanmassenhove (2024) testano la capacità di due LLM conversazionali, rispettivamente Llama 2 e ChatGPT 3.5, di fornire più traduzioni alternative, rispetto al genere dei referenti, per un'unica frase di partenza ambigua; Savoldi et al. (2024) e Piergentili et al. (2024), invece, valutano diversi modelli sulla generazione di traduzioni in linguaggio non binario indiretto e diretto, rispettivamente, confrontando diverse strategie di adattamento.

Questi approcci non sono possibili con sistemi come ModernMT, per i quali i tentativi di riduzione del bias di genere si concentrano solitamente sull'adattamento tramite memorie di traduzione (MT) (ad esempio, Kostikova et al., 2023) o sul post-editing (manuale in Lardelli & Gromann, 2023a; automatico in Jain et al., 2021; Sun et al., 2021; Vanmassenhove et al., 2021).

Alla luce della letteratura esistente, quindi, per questo lavoro sono state previste tre configurazioni per ciascun modello: nella prima, chiamata *baseline*, abbiamo valutato le traduzioni fornite da ModernMT e Llama 3 senza istruzioni di alcun tipo; in seguito, per ciascuno abbiamo implementato due diverse strategie di adattamento, volte alla riduzione del bias di genere e in particolare dell'uso del maschile sovraesteso.

- Per ModernMT:
    1. Adattamento tramite memoria di traduzione;
    2. Post-editing automatico.
- Per Llama 3:
    1. Utilizzo di istruzioni esplicite per dirigere l'output verso l'utilizzo del LND;
    2. Aggiunta di esempi contrastivi che consistono, per ogni frase di partenza, in due traduzioni: una marcata e una neutra, differenziate tramite un prefisso come illustrato in Tabella 8.

Da questa impostazione risultano quindi sei esperimenti in totale, considerando le tre diverse configurazioni di ciascun modello; tutti gli esperimenti sono stati effettuati tra marzo e giugno 2024. Nel paragrafo successivo (§2.1) sono riportati i dati utilizzati per tali esperimenti, descritti nel dettaglio nei paragrafi §2.2 (ModernMT) e §2.3 (Llama 3). In §2.4 è descritto il procedimento seguito per la valutazione dei risultati, mentre in §2.6 sono presentate le linee guida formulate per la riscrittura delle frasi in schwa. Gli script utilizzati per l'addestramento e la valutazione dei classificatori e degli altri modelli sono disponibili al link https://github.com/paolo-mainardi/tratec-tesi.

2.1. Dati

La Tabella 1 fornisce una panoramica dei dati utilizzati, con il numero di frasi da ogni dataset suddivise in dati di addestramento (Train) e di valutazione (Test) per ogni modello.

	ModernMT		**Llama 3**		**IT5**		**Cls_EN**		**Cls_IT**	
Dataset	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
GeNTE	0	750	0	750	0	750	1274	226	2231	394
NeuTralRewriter	0	0	0	0	0	0	5778	1023	0	0
MT-GenEval	1528	0	1528	0	1528	0	0	0	0	0
Totale parziale	**1528**	**750**	**1528**	**750**	**1528**	**750**	**7052**	**1249**	**2231**	**394**
Totale	2278		2278		2278		8331		2625	

Tabella 1. Numero di frasi utilizzate per l'addestramento e la valutazione di ogni modello, suddivise per dataset di provenienza. Gli 0 indicano che un dataset non è stato utilizzato in quel contesto per quel modello.

La mancanza di dataset adatti a esperimenti con il linguaggio non binario è nota (vedi ad esempio Dev et al., 2021), nonostante quest'area di ricerca sia in espansione; per questo motivo è stato necessario manipolare i dati raccolti, e non è stato possibile prestare particolare attenzione a questioni come i domini di appartenenza. I dataset utilizzati – descritti nel dettaglio nei paragrafi che seguono (§2.1.1-2.1.3) – appartengono quindi ad aree specifiche e lontane tra loro, oltre a essere piuttosto limitati in quantità, e questi aspetti potrebbero aver influenzato i risultati.

2.1.1. GeNTE

Come illustrato in Tabella 2, il dataset GeNTE[9] (Piergentili et al., 2023b) contiene 1500 triplette contenenti una frase di partenza inglese (colonna "SRC") , la sua traduzione marcata (colonna "REF-G"), raccolte dal corpus Europarl (Koehn, 2005)[10], e una traduzione alternativa in linguaggio non binario indiretto (LNI) (colonna "REF-N"). Le coppie di frasi raccolte da Europarl contengono uno o più referenti umani, che hanno tutti lo stesso genere, e sono annotate in base a se tali referenti nella frase inglese portano marche di genere esplicite (maschili, femminili) o se il loro genere è ambiguo (N) (colonna "GENDER"). Alcune frasi di partenza marcate ("Set-G") sono state duplicate e riscritte con un genere diverso rispetto a quello originale (da maschile a femminile o viceversa), e ne esistono quindi due riformulazioni identiche in LNI, mentre le frasi di partenza neutre ("Set-N") hanno una sola traduzione marcata e una sola traduzione neutra.

[9] Scaricato da https://mt.fbk.eu/gente/.
[10] https://www.statmt.org/europarl/

ID	Europarl_ID	SET	SRC	REF-G	REF-N	GENDER
163	ep-en-it-18662	Set-G	We have not succeeded in convincing her.	Abbiamo fallito nel tentativo di convincerla.	Abbiamo fallito nel tentativo di convincere questa persona.	F
606	ep-en-it-22552	Set-G	He needs to receive care and be released immediately.	Deve essere immediatamente rilasciato e curato.	Occorre rilasciare e curare immediatamente questa persona.	M
1465	ep-en-it-9431	Set-N	Rights of language teachers	Diritti degli insegnanti di lingue	Diritti delle persone che insegnano le lingue	N

Tabella 2. Esempi di voci estratte casualmente da GeNTE. La colonna "SET" distingue le frasi di partenza (colonna "SRC") in marcate (Set-G) e neutre (Set-N); la colonna "GENDER" identifica il genere del referente nella frase inglese.

Questo corpus è stato utilizzato per l'addestramento e la valutazione dei classificatori, in una versione ridotta chiamata GeNTE_CLS (Tabella 3):

- per il classificatore inglese abbiamo usato tutte le frasi inglesi originali (colonna "SRC"), che abbiamo etichettato come marcate (femminili/maschili) (1) o neutre (0);
- per il classificatore italiano abbiamo invece usato tutte le traduzioni originali (colonna "REF-G") (maschili/femminili), che abbiamo taggato come marcate (1); delle traduzioni in LNI (colonna "REF-N"), taggate come neutre (0), abbiamo usato soltanto quelle corrispondenti a frasi di partenza maschili o neutre, rimuovendo così 375 frasi neutre dal dataset originale, in modo da non avere nessun duplicato.

Nei dati utilizzati per questo esperimento sono presenti quindi tutte le frasi di partenza inglesi di GeNTE con le loro traduzioni marcate (ovvero le coppie originariamente contenute in Europarl), più un subset delle traduzioni neutre. In Tabella 3 è riportato il numero di frasi utilizzate per l'addestramento dei classificatori, suddivise in marcate e neutre e in base alle marche di genere originali nelle frasi di partenza inglesi.

GeNTE_CLS			
en		**it**	
Marcate	Neutre	Marcate	Neutre
750	750	1500	1125
1500		**1125**	

Tabella 3. Distribuzione delle frasi per lingua e marche di genere in GeNTE_CLS, la versione di GeNTE utilizzata per l'addestramento dei classificatori italiano e inglese.

A partire dallo stesso dataset abbiamo creato anche un test set per la valutazione e la comparazione dei risultati dei sei esperimenti con MMT e Llama. Per fare questo abbiamo scritto manualmente nuove riformulazioni, stavolta in LND, per le frasi di partenza neutre ("Set-N"), secondo le linee guida presentate in §2.6. Nel test set sono presenti anche le traduzioni marcate contenute nel dataset originale (colonna "REF-G"), utilizzate solo per creare gli esempi contrastivi forniti a Llama 3. Il test set, chiamato GeNTE – Set-N_Schwa e illustrato in Tabella 4, contiene quindi 750 frasi inglesi ambigue rispetto al genere, ognuna accompagnata da due traduzioni: una marcata, presa dal dataset GeNTE originale, e l'altra, in schwa, creata per questi esperimenti.

GeNTE – Set-N_Schwa

en	**it**	
Neutre	Marcate	Schwa
750	**750**	**750**

Tabella 4. Distribuzione delle frasi in GeNTE - Set-N_Schwa. Ciascuna frase inglese è accompagnata da due traduzioni italiane: una marcata e una neutra (in schwa).

2.1.2. NeuTralRewriter

Il dataset NeuTralRewriter[11] (Vanmassenhove et al., 2021) contiene frasi inglesi suddivise in tre subset appartenenti a domini diversi:

- 500 frasi dal corpus OpenSubtitles[12];
- 500 frasi da Reddit[13];
- WinoBias+ (dati sintetici da Zhao et al., 2018).

Le frasi marcate contenute nel corpus sono bilanciate tra maschili e femminili; per questo motivo alcune frasi nei subset di dati naturali (OpenSubtitles e Reddit) sono il risultato di una duplicazione e conversione automatica dal maschile al femminile o viceversa. Ogni frase marcata è allineata a una sua riformulazione neutra creata automaticamente; per le frasi convertite al genere diverso, quindi, la riformulazione neutra si ripete due volte, come in GeNTE.

Come si vede in Tabella 1, abbiamo utilizzato questi dati unicamente per l'addestramento del classificatore inglese, aggiungendoli a quelli descritti sopra. Una volta rimosse le frasi ripetute, abbiamo taggato quelle rimaste come marcate (maschili/femminili) (1) o neutre (0). La distribuzione delle frasi in questa versione ridotta del dataset è riportata in Tabella 5.

[11] File scaricati da https://github.com/vnmssnhv/NeuTralRewriter.
[12] https://opus.nlpl.eu/OpenSubtitles/corpus/version/OpenSubtitles
[13] https://www.reddit.com/

NeuTralRewriter		
Subset	**Marcate**	**Neutre**
OpenSubtitles	495	495
Reddit	500	500
WinoBias+	3161	1650
Totale	**4156**	**2645**

Tabella 5. Distribuzione delle frasi nella versione di NeuTralRewriter utilizzata in questo studio, in base al subset e alle marche di genere.

2.1.3. MT-GenEval

Il dataset MT-GenEval[14] (Currey et al., 2022) è una benchmark multilingue per la valutazione di sistemi di TA rispetto ai fenomeni legati alla traduzione del genere (vedi §1.2). Contiene frasi inglesi raccolte da Wikipedia[15] con le loro traduzioni in otto lingue, ed è suddiviso in due sezioni:

- *counterfactual set*: per il primo subset sono state raccolte circa 3000 frasi inglesi in base a liste di parole portatrici di informazioni sul genere dei referenti (tratte da Zhao et al., 2018); per ogni frase sono fornite due traduzioni: una che mantiene le marche di genere corrette, l'altra che usa un genere diverso (da femminile a maschile o viceversa);
- *contextual set*: la seconda sezione contiene circa 1500 frasi inglesi raccolte tramite parole riferite a professioni che in inglese non specificano il genere; alle frasi contenute in questa sezione è stata aggiunta nel dataset una frase precedente nel contesto del documento, che disambigua il genere di uno dei referenti contenuti nella frase ambigua; sono state quindi aggiunte due traduzioni in base allo stesso principio seguito per il counterfactual set: una con il genere suggerito dalla frase di contesto, l'altra con un genere diverso (da maschile a femminile o viceversa).

Per questo lavoro abbiamo utilizzato soltanto la seconda sezione (la cui struttura è illustrata in Tabella 6), nella direzione inglese-italiano, selezionando le frasi di partenza neutre (colonna "Source") e le loro traduzioni marcate con il genere corretto ("Reference_Original").

[14] File scaricati da https://github.com/amazon-science/machine-translation-gender-eval.
[15] https://www.wikipedia.org/

Context	Source	Reference_Original	Reference_Flipped
She grew up and attended school in Gaborone, where her parents were working.	Masisi is an accountant.	Masisi è una contabile.	Masisi è un contabile.
He named Gutfeld his successor.	Zinczenko became editor-in-chief in 2000.	Zinczenko divenne caporedattore nel 2000.	Zinczenko divenne caporedattrice nel 2000.

Tabella 6. Voci di esempio da MT-GenEval - *contextual*. Il genere del referente nella prima traduzione (colonna "Reference_Original") è determinato in base alla frase nella colonna "Context", mentre nella seconda ("Reference_Flipped") è convertito dal maschile al femminile o viceversa.

Per svolgere gli esperimenti, per ogni frase abbiamo aggiunto una riformulazione in LND della traduzione esistente; come per GeNTE, le riformulazioni sono state scritte a mano secondo le linee guida presentate in §2.6.

Abbiamo utilizzato questi dati (versione MT-GenEval – Context_Schwa) per l'adattamento di ModernMT e Llama 3, con le frasi di partenza abbinate alle loro traduzioni riformulate in schwa. Per l'addestramento del modello di riformulazione, invece, abbiamo utilizzato le coppie di traduzioni italiane (di cui una marcata, contenuta in "Reference_Original", e la nostra riscritta in schwa).

Durante il processo di aggiunta delle riformulazioni in LND, 31 frasi del dataset originale sono state rimosse perché non adeguate alla riformulazione. Il dataset finale contiene 1528 triplette: in Tabella 7 è riportato il numero di frasi utilizzate per gli esperimenti, divise in base alla lingua e alle marche di genere.

MT-GenEval – Context_Schwa			
en	**it**		
Neutre	M	F	Schwa
1528	**812**	**716**	**1528**

Tabella 7. Distribuzione delle frasi nella versione di MT-GenEval con riformulazioni italiane in LND, utilizzata per questo studio.

2.2. ModernMT

Come anticipato, ModernMT è un sistema commerciale di TA neurale accessibile a pagamento tramite API[16]. Utilizzando quest'ultima abbiamo tradotto le 750 frasi inglesi neutre del test set (Tabella 4). Al sistema è stata poi aggiunta una memoria di traduzione contenente le frasi inglesi di

[16] https://www.modernmt.com/api#introduction

partenza dal set di addestramento insieme alle loro traduzioni riscritte con schwa (Tabella 7); abbiamo quindi ritradotto le frasi del test set, utilizzando stavolta il *context vector* fornito dall'API, per verificare la capacità del modello di adattare le traduzioni agli esempi forniti (come in Kostikova et al., 2023). Il *context vector* è una stringa contente un identificatore delle memorie di traduzione da utilizzare e un peso che ne determina l'importanza: per il secondo esperimento abbiamo utilizzato soltanto la memoria di traduzione descritta sopra, con il peso massimo, ovvero 1. Per il terzo esperimento con ModernMT, le traduzioni ottenute nella prima fase sono state riscritte automaticamente tramite il modello descritto più sotto (§2.2.1).

I tre set di traduzioni ottenute sono stati quindi valutati manualmente e automaticamente, confrontando gli output di ModernMT con le frasi di riferimento (ovvero, le riformulazioni in schwa contenute nel test set), come descritto in §2.4.

2.2.1. Modello di riformulazione

Per il post-editing automatico delle traduzioni ottenute da ModernMT nella configurazione `baseline` abbiamo appositamente addestrato un modello sulla riscrittura in LND di frasi italiane maschili o femminili.

Per questo compito abbiamo utilizzato IT5 (Sarti & Nissim, 2022)[17], la versione di T5 ottimizzata sulla lingua italiana. T5 (Raffel et al., 2020) è un transformer *encoder-decoder*, addestrato sulla generazione di testo condizionata da un input (*text-to-text*). Abbiamo scelto questo modello perché può svolgere diversi compiti – tra cui tradurre, riassumere o riscrivere (ad esempio, con uno stile diverso) testi – in base alle istruzioni che riceve, ed è l'unico di questo tipo addestrato specificamente sull'italiano. Per questo esperimento, abbiamo quindi ottimizzato IT5 nella versione base (da 220 milioni di parametri) sul compito di riformulazione di frasi italiane da marcate a neutre, inteso come una forma di traduzione intralinguistica.

Per l'ottimizzazione di IT5, effettuata tramite il `Trainer` di Hugging Face in PyTorch[18], abbiamo sfruttato lo stesso dataset (MT-GenEval – Context_Schwa) utilizzato per l'adattamento di ModernMT e Llama 3, con la differenza che, in questo caso, le frasi di partenza non sono quelle inglesi, ma le traduzioni italiane marcate contenute nel dataset originale, affiancate sempre dalle riformulazioni in schwa create per questo studio.

Per l'addestramento abbiamo utilizzato *batch* di 64 esempi, applicando il *padding* per far sì che ogni frase avesse la stessa lunghezza, in questo caso quella della frase più lunga nella batch; come ottimizzatore abbiamo utilizzato AdamW, con un tasso di apprendimento iniziale di 0.0005 e *linear*

[17] Checkpoint ottenuti tramite Hugging Face da https://huggingface.co/gsarti/it5-base.
[18] https://huggingface.co/docs/transformers/main_classes/trainer#trainer

decay, senza *warmup*; il modello è stato addestrato per 10 epoche, monitorando la *loss* sui dati di validazione (2% del set di addestramento).

Con il modello ottimizzato abbiamo quindi riscritto automaticamente le traduzioni ottenute da ModernMT (baseline). In questa fase, implementata sempre tramite Hugging Face, per la generazione delle frasi abbiamo applicato una tecnica di decodifica che combina *beam search* (con dimensione dei beam b=5) e *multinomial sampling* (con il valore predefinito k=50). Come discusso, ad esempio, da Holtzman et al. (2020), tecniche basate sulla maggiore probabilità per la selezione dei token con cui continuare il testo prodotto fino a quel momento (come la beam search) tendono a produrre testi fluenti ma anche poco creativi; la combinazione di questa tecnica con una di *sampling* permette di ridurre il problema della generazione di testi ripetitivi, mantenendo comunque un'alta accuratezza, necessaria per il compito in questione.

2.3. Llama 3

Come anticipato, Llama 3[19] è un transformer di tipo decoder-only, addestrato sulla continuazione di un input di testo.

Per testare l'adattabilità di questo modello alla traduzione neutra rispetto al genere, abbiamo utilizzato l'API GroqCloud[20], che permette di interrogare la versione *Instruct* (ottimizzata sulla conversazione) gratuitamente tramite una struttura a turni, dove, per ogni turno, l'utente formula una richiesta (*prompt*) e il modello fornisce una risposta. Per questi esperimenti abbiamo utilizzato la versione del modello con 70 miliardi (70b) di parametri, quella più grande disponibile.

I tre esperimenti effettuati con Llama sono i seguenti:

- *baseline*: al modello viene chiesto di tradurre le 750 frasi del test set, senza ulteriori indicazioni;
- *istruzioni*: seguendo la stessa struttura, vengono richieste le traduzioni delle stesse frasi, ma stavolta aggiungendo la richiesta esplicita di utilizzare lo schwa nelle traduzioni;
- *esempi contrastivi*: la stessa richiesta esplicita del secondo esperimento viene arricchita da esempi del compito da svolgere; in questa configurazione, forniamo al modello le frasi di partenza del set di addestramento (MT-GenEval – Context_Schwa) accompagnate dalle due traduzioni italiane (quella originale e quella in schwa), nonché le frasi di partenza del test set accompagnate dalle traduzioni originali (marcate); chiediamo quindi di fornire le traduzioni neutre corrispondenti a queste ultime, seguendo gli esempi forniti.

[19] L'articolo che ha accompagnato l'uscita del modello si trova nel blog di Meta AI, al link https://ai.meta.com/blog/meta-llama-3/. Al momento (giugno 2024) non è ancora disponibile una pubblicazione scientifica; è disponibile invece un preprint relativo al primo rilascio di questa famiglia di modelli (Touvron et al., 2023).
[20] https://console.groq.com/docs/quickstart

Dato il limite di 6000 token al minuto nelle richieste inviate tramite l'API, sono state usate *batch* di 15 traduzioni per prompt, in tutte le configurazioni. Per la configurazione con gli esempi contrastivi, ogni prompt contiene 30 esempi; per semplicità abbiamo usato solo le prime 1500 righe di MT-GenEval – Context_Schwa, escludendo le ultime 28 frasi in eccesso. In tutto abbiamo quindi inviato al modello 50 richieste per ogni configurazione.

I prompt per le tre configurazioni sono stati strutturati come illustrato in Tabella 8 (sul modello di quelli utilizzati in Savoldi et al., 2024 e Piergentili et al., 2024), dove `user` è seguito dalle richieste dell'utente, mentre `system` serve a identificare gli esempi di output desiderati.

Configurazione	Template
baseline	**user**: "Translate the following English sentences into Italian: English: <frase di partenza inglese dal test set> Italian: " …
istruzioni	**user**: "Translate the following English sentences to Italian using the schwa (ə) neomorpheme to replace masculine and feminine morphemes for human beings: English: <frase di partenza inglese dal test set> Italian: " …
esempi contrastivi	**user**: "Translate the following English sentences to Italian using the schwa (ə) neomorpheme to replace masculine and feminine morphemes for human beings: English: <frase di partenza inglese dal training set> Italian, gendered: <traduzione italiana marcata>" … **system**: "Italian, neutral: <riformulazione italiana in schwa>" … **user**: "English: <frase di partenza inglese dal test set> Italian, gendered: <traduzione italiana marcata> Italian, neutral: " …

Tabella 8. Struttura dei prompt inviati a Llama 3 nelle tre configurazioni. Ogni prompt contiene 15 richieste di traduzioni e, nell'ultima configurazione, anche 30 esempi.

Abbiamo quindi valutato i tre set di traduzioni manualmente e automaticamente, confrontando gli output di Llama 3 con le frasi di riferimento (le riformulazioni con schwa dal test set), come descritto nella prossima sezione.

2.4. Valutazione

La valutazione è basata sulle frasi contenute in GeNTE – Set-N_Schwa (Tabella 4). Con gli stessi dati abbiamo eseguito sia una valutazione automatica, utilizzando le quattro metriche descritte in §2.4.1, sia una valutazione manuale su due livelli, descritta in §2.4.2.

2.4.1. Valutazione automatica

La valutazione automatica comprende due metriche basate su *n-gram* (gruppi di parole o caratteri di lunghezza variabile): **BLEU**, che opera a livello di token (Papineni et al., 2002), e **chrF** a livello di caratteri (Popović, 2015); e due basate sulla distanza di Levenshtein, che rappresenta il numero di modifiche necessarie per passare da un testo all'altro: **TER** (Snover et al., 2006) e la sua variante per i caratteri, **characTER** (Wang et al., 2016). Tutte le metriche sono state implementate tramite la libreria `evaluate` di Hugging Face[21].

La combinazione di queste metriche dovrebbe permettere di cogliere diversi aspetti legati alla differenza tra gli output del modello e le frasi di riferimento, sia a livello di token che di singoli caratteri: infatti, data la morfologia dell'italiano, la riformulazione delle frasi in LND si limita spesso alla sostituzione di un singolo carattere su una o più parole.

2.4.2. Valutazione manuale

Per la valutazione manuale abbiamo implementato un protocollo a due livelli ispirato a quello di Savoldi et al. (2024), basato sull'effettiva neutralizzazione dei referenti umani e sull'accettabilità linguistica della traduzione. La valutazione manuale, quindi, si basa non solo sulle frasi di riferimento, ma anche su quelle di partenza. Una persona ha annotato tutte le frasi rispetto a entrambi gli aspetti su due scale Likert, illustrate nelle Tabelle 9 e 10, rispettivamente per la neutralità e l'accettabilità delle traduzioni.

[21] https://huggingface.co/docs/evaluate/index

Neutralità

1	Completamente neutra. *Tutti i referenti umani, comprese tutte le parole che concordano con il nome o pronome, non sono marcati come maschili né come femminili; questo a prescindere dalla presenza di schwa nella frase.*
2	Parzialmente neutra. *Alcuni referenti sono marcati come maschili o femminili, per intero o in parte (ad esempio, all'interno di uno stesso sintagma nominale alcuni elementi potrebbero avere lo schwa e altri no).*
3	Completamente marcata. *Tutti i referenti umani presenti nella frase sono marcati. Se compaiono delle schwa, sono applicate a parole che non si riferiscono a esseri umani (es. oggetti inanimati, participi passati dove l'accordo non è necessario, ecc.).*

Tabella 9. Scala Likert utilizzata per la valutazione manuale rispetto alla effettiva neutralità di genere delle traduzioni.

Accettabilità

1	Accettabile. *La frase è adeguata sia dal punto di vista traduttivo, rispetto alla frase di partenza inglese, sia nell'utilizzo dello schwa rispetto alle linee guida presentate in §2.6.*
2	Per lo più accettabile. *La frase è per lo più adeguata; contiene (pochi) errori linguistici o di traduzione e/o nell'uso dello schwa, ma la comprensibilità non è compromessa.*
3	Per lo più inaccettabile. *La frase è per lo più inadeguata: contiene errori linguistici o di traduzione e nell'uso dello schwa, e può essere poco comprensibile in alcuni punti.*
4	Inaccettabile. *La frase è incompleta e/o inadeguata dal punto di vista linguistico/traduttivo, non è comprensibile o contiene segmenti ripetuti più volte; può contenere errori nell'uso dello schwa.*

Tabella 10. Scala Likert utilizzata per la valutazione manuale rispetto all'accettabilità linguistica delle traduzioni.

In particolare, nella valutazione della neutralità di una traduzione vengono presi in considerazione tutti i referenti umani presenti nella frase: se tutti questi sono marcati al maschile e/o al femminile, la valutazione di neutralità è 3; all'opposto, se tutti i referenti sono ambigui rispetto al genere, la frase è considerata completamente neutra e valutata con 1. Per i casi intermedi, dove almeno un referente è neutro e almeno uno è marcato, la valutazione è 2, come nell'esempio (11).

Perché un referente sia considerato neutro non è necessario che abbia la desinenza in schwa: è sufficiente che sia ambiguo rispetto al genere, il che è possibile anche utilizzando, ad esempio, nomi epiceni e promiscui (vedi anche §1.3.1). Per esempio, la frase (9) non contiene schwa, ma è comunque completamente neutra (punteggio 1):

(9) Non ignorare le tue stesse parole negando alle persone sorde, alle persone con difficoltà di apprendimento e alle organizzazioni autogestite di persone disabili la possibilità di sedersi almeno al tavolo.

`Llama_baseline` [526]

Allo stesso modo, la frase (10) è completamente neutra, grazie all'uso dello schwa per entrambi i referenti (la terza persona plurale sottintesa a cui si riferisce la forma verbale "sono salitə" e "l'autorə"):

(10) sono salitə al di sopra della politica di partito e hanno dimostrato che, quando un emendamento è giusto, lo sosterranno, chiunque ne sia l'autorə.

`MMT_post-editing` [3]

Al contrario, nella (11), un referente è marcato al maschile ("cittadini europei"), mentre "custodi" e "migliori", al plurale e senza articoli, sono entrambe forme epicene, quindi valide per referenti di qualunque genere. A questa frase, quindi, è stata assegnata una valutazione di 2 rispetto alla neutralità (parzialmente neutra).

(11) Non ci sono custodi dei Trattati migliori dei cittadini europei.

`MMT_baseline` [237]

Lo stesso vale per la frase (12): mentre "dei rifugiati" rappresenta un caso di maschile sovraesteso, "del personale" è una forma collettiva che può riferirsi a un gruppo di referenti di qualunque genere, pur essendo declinata al maschile.

(12) […] il destino dei rifugiati viene messo nelle mani […] del personale delle compagnie aeree e degli aeroporti.

`MMT_adapted` [205]

Molti di questi casi intermedi si trovano in particolare negli esperimenti di adattamento dove i modelli inseriscono alcune forme in schwa: in questi casi, spesso i referenti sono neutralizzati soltanto in parte, ad esempio applicando lo schwa soltanto ad alcune delle parole che vi si riferiscono. Anche queste frasi, come la (13), dove la preposizione articolata "negli" è lasciata al maschile nonostante l'uso di schwa sul nome cui si riferisce ("ispettorə"), sono considerate solo parzialmente neutre (punteggio 2).

(13) Abbiamo anche bisogno di investire negli ispettorə, […].

`Llama_istruzioni` [135]

Per quanto riguarda l'accettabilità, invece, la valutazione si basa sulla qualità linguistica e traduttiva, anche rispetto alla frase di partenza, nonché sull'utilizzo dello schwa in relazione alle linee guida presentate più avanti (§2.6).

L'esempio (14) è esemplificativo di uno degli errori di accettabilità più frequenti in tutti gli esperimenti: la traduzione è molto aderente alla struttura della frase di partenza, e per questo non è del tutto comprensibile.

(14) Representatives of the governments, MEPs and national MPs, together with the Commission, reaching an agreement on the Charter of Fundamental Rights of the European Union, in ten months, seemed an impossible objective.
> I rappresentanti dei governi, europarlamentariə e parlamentari nazionali, insieme alla Commissione, raggiungere un accordo sulla Carta dei Diritti Fondamentali dell'Unione Europea in dieci mesi, sembrava un obiettivo impossibile.

`Llama_istruzioni` [319]

Il punteggio assegnato a questa frase per l'accettabilità è 3, visto anche l'errore nell'uso dello schwa su "europarlamentariə", dove non era necessaria ed è usata nel modo sbagliato.

In effetti, per quanto riguarda l'uso dello schwa, uno degli errori più frequenti è la sua errata generalizzazione a parole che non devono essere neutralizzate, per esempio perché sono già ambigue rispetto al genere. In (15), l'articolo "aə" è ridondante, dal momento che "membri" è sì grammaticalmente maschile, ma è un nome promiscuo, quindi già adatto a referenti di qualunque genere:

(15) "[...] spetta in particolare aə membri del parlamento divulgare maggiormente le decisioni critiche, [...]."

`MMT_post-editing` [92]

In altri casi, lo schwa è utilizzata in modo diverso da quanto previsto nelle linee guida e, quindi, dagli esempi forniti al modello:

(16) Commissarə Byrne, siamo molto soddisfattə.

`Llama_contrastiva` [126]

Nell'esempio (16), entrambi i referenti sono resi ambigui (neutralità: 1), ma nel caso di "Commissarə", la forma è grammaticalmente scorretta (accettabilità: 2): come stabilito nelle linee guida (§2.6), le parole in -io/-ia prendono lo schwa dopo la -i-, sia al singolare che al plurale. La forma corretta sarebbe stata quindi *Commissariə*.

Nella prossima sezione sarà descritto il metodo adottato per l'addestramento dei classificatori, per poi presentare le linee guida in quella successiva.

2.5. Classificatori

Come anticipato, abbiamo addestrato anche due classificatori per identificare frasi potenzialmente contenenti linguaggio non binario, rispettivamente in inglese e in italiano. Questi modelli potrebbero essere utilizzati per filtrare frasi inglesi e italiane che siano ambigue rispetto al genere dei referenti umani, permettendo di creare dataset come quelli utilizzati in questo lavoro, utili per l'addestramento e la valutazione di strumenti volti alla identificazione e alla riduzione di fenomeni legati al bias di genere in TAL.

Classificatori come questi potrebbero essere usati anche in combinazione con un metodo per riscrivere automaticamente le traduzioni ottenute da sistemi di TA, filtrando quelle che devono effettivamente essere riscritte e lasciando come sono quelle già neutre. Un approccio simile è adottato ad esempio in Habash et al. (2019) e Jain et al. (2021), che abbiamo seguito per l'addestramento dei classificatori.

Il problema è stato impostato come un compito intralinguistico di classificazione binaria, considerando come marcate (etichetta 1) tutte le frasi che contengono almeno uno o più referenti marcati al maschile o al femminile, e come neutre (etichetta 0) quelle che contengono solo referenti il cui genere è ambiguo.

Per l'italiano abbiamo utilizzato le coppie di traduzioni marcate ("REF-G") e neutre ("REF-N") contenute in GeNTE_CLS (Tabella 3); per l'inglese, oltre alle frasi di partenza di GeNTE_CLS, suddivise in marcate ("Set-G") e neutre ("Set-N"), abbiamo aggiunto i dati di NeuTralRewriter come presentati in Tabella 5.

Per quanto riguarda le rappresentazioni dei dati, abbiamo adottato due tipi di word embeddings. Nelle rappresentazioni statiche, più classiche, il vettore associato a ogni token rimane lo stesso a prescindere dal contesto: i dati utilizzati per questo esperimento sono stati tokenizzati, e il vettore associato a ogni token, se presente, è stato preso dal vocabolario scaricato per Word2Vec, fastText e GloVe. Dall'altro lato, le rappresentazioni contestuali tengono più conto del comportamento semantico delle parole, che possono assumere significati diversi, e sono quindi rappresentate diversamente, in base al contesto. Abbiamo quindi utilizzato (solo lato *encoder*) due transformer della famiglia BERT (Devlin et al., 2019), pre-addestrati per la creazione di rappresentazioni contestuali: RoBERTa e UmBERTo.

In particolare, per Word2Vec abbiamo utilizzato due vocabolari diversi, quello originale rilasciato da Google (Mikolov et al., 2013) per l'inglese[22] e quello per l'italiano pubblicato da Di Gennaro et al. (2020)[23]; il progetto fastText, invece, fornisce rappresentazioni pre-addestrate per 157 lingue, tra cui l'inglese e l'italiano (Grave et al., 2018)[24]; GloVe (Pennington et al., 2014)[25] è disponibile solo per l'inglese, quindi è stato utilizzato solo per questa lingua; infine, RoBERTa (Conneau et al., 2020)[26] è un modello multilingue, quindi è stato utilizzato per entrambe le lingue, mentre UmBERTo[27] è una versione di RoBERTa ottimizzata per l'italiano, quindi è stato utilizzato solo per questa lingua.

Per ciascuna lingua, abbiamo utilizzato un CNN e un LSTM, addestrati su rappresentazioni statiche (Word2Vec, fastText, GloVe) e contestuali (ottenute utilizzando UmBERTo e RoBERTa). Abbiamo quindi ottenuto quattro modelli per lingua, tra cui abbiamo selezionato il migliore in base alla valutazione sul test set.

Durante l'addestramento di ciascuno degli otto classificatori sono stati ottimizzati automaticamente alcuni parametri tramite il tuner di `Keras`[28]; la migliore configurazione di parametri trovata è stata poi utilizzata per comparare quel modello agli altri.

Abbiamo quindi valutato automaticamente le prestazioni di ogni modello confrontandone gli output con le annotazioni manuali, e in base ai risultati (riportati nel capitolo successivo), i modelli selezionati sono due CNN addestrati su embedding da RoBERTa per l'inglese e da UmBERTo per l'italiano.

2.6. Linee guida per la riformulazione

Per la riscrittura delle frasi necessarie ad addestrare e testare i modelli abbiamo adottato un approccio non binario basato sul LND, secondo le linee guida riportate di seguito.

I principi generali che guidano l'approccio alla riformulazione delle frasi sono ispirati a quelli proposti da Piergentili et al. (2023a) per la traduzione neutra rispetto al genere (GNT dall'inglese *gender-neutral translation*). Lo scopo della GNT è quello di tradurre da una lingua a un'altra applicando il GFL (*gender-fair language*: vedi §1.3), evitando ad esempio di marcare il genere dei referenti umani nel testo di arrivo quando questo non è necessario. In particolare:

[22] Scaricati da https://code.google.com/archive/p/word2vec/.
[23] Scaricati da https://mlunicampania.gitlab.io/italian-word2vec/.
[24] Vettori scaricati da https://fasttext.cc/docs/en/crawl-vectors.html.
[25] Scaricati tramite gensim nella versione glove-wiki-gigaword-300 da https://github.com/piskvorky/gensim-data.
[26] Checkpoint ottenuti tramite Hugging Face da https://huggingface.co/FacebookAI/xlm-roberta-large.
[27] Checkpoint ottenuti tramite Hugging Face da https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1.
[28] https://keras.io/keras_tuner/

1. il genere di un referente non dovrebbe essere espresso se non è possibile ottenere informazioni a riguardo dalla frase di partenza (ovvero, se la frase di partenza non contiene marche di genere esplicite per quel referente);
2. se, invece, la frase di partenza contiene informazioni specifiche circa il genere di uno o più referenti umani, questi referenti devono essere marcati con lo stesso genere nella frase di arrivo;
3. eventuali maschili sovraestesi (§1.2.1) contenuti nella frase di partenza non dovrebbero essere propagati alla frase di arrivo; in caso di dubbi o ambiguità sul fatto che una forma maschile possa essere propria (ovvero che si possa effettivamente riferire a un referente di genere maschile), una traduzione neutra è sempre preferibile.

Gli esperimenti effettuati (§2) sono basati su frasi di partenza che non contengono referenti marcati: per questo, nella riscrittura delle traduzioni marcate, tutti i referenti sono stati resi neutri, a prescindere dai controsensi e dal genere reale delle persone menzionate, al solo scopo di addestrare i modelli. Eventuali maschili neutri (ovvero forme marcate al maschile usate per referenti generici: §1.2.1) sono stati rimossi, come nell'esempio (17), dove si usa "man" o "uomo" per intendere *persona* o *essere umano* (le espressioni corrispondenti sono sottolineate nell'esempio).

(17) At the same time we would easily destroy the belief the entrepreneur and the man in the street have in the common sense of EU decision-makers. / Rottameremmo così altrettanto facilmente anche la fiducia dell'imprenditore e dell'uomo della strada nel buon senso di quanti sono preposti ad adottare le decisioni a livello comunitario.

> Rottameremmo così altrettanto facilmente anche la fiducia dell'imprenditorə e della persona comune nel buon senso di quantə sono prepostə ad adottare le decisioni a livello comunitario.

`GeNTE - Set-N_Schwa` [325]

Nello specifico, la strategia di LND adottata è quella dello schwa (vedi §1.3.2). Le linee guida che seguono sono state formulate principalmente in base ai suggerimenti di effequ[29] e alla grammatica proposta da Italiano inclusivo[30], nonché alla panoramica fornita da Gender in Language[31].

1. Il **carattere** usato per scrivere lo schwa minuscolo (*ə*) ha il codice Unicode 0259, mentre la forma maiuscola *Ə* corrisponde al codice Unicode 018F.

[29] https://www.effequ.it/schwa/
[30] https://italianoinclusivo.it/scrittura/
[31] https://www.genderinlanguage.com/italian

2. Gli **articoli determinativi** sono *lə* per il singolare e *ə* per il plurale; se il nome inizia per vocale, l'articolo singolare subisce elisione (*l'*) per qualunque genere.
3. L'**articolo indeterminativo singolare** è *unə*; anche se il nome inizia per vocale, l'articolo non subisce né elisione né troncamento. La forma **plurale** corrisponde alla preposizione articolata *deə*.
4. Le **preposizioni articolate** si formano dalla base de-, a-, da-, su- a cui viene aggiunto lo schwa, con elisione nei casi previsti; esempio: *del/dello* / *della* / *dellə* / *dell'*, *dei/degli* / *delle* / *deə*; *sul/sullo* / *sulla* / *sullə* / *sull'*, *sui/sugli* / *sulle* / *suə*.
5. Il **pronome personale** di terza persona singolare è *ləi* in funzione di soggetto, *lə* in funzione di oggetto diretto e indiretto singolare; al plurale, il pronome di oggetto diretto è *lə*, quello indiretto è *loro* (con eventuale accordo in -ə). Il pronome soggetto *ləi* rappresenta l'unico caso in cui lo schwa si trova in posizione tonica e intrasillabica, cosa che ne potrebbe rendere più difficile la pronuncia, come nota anche Gheno (2022). Ricordiamo inoltre che il pronome formale di terza persona singolare *lei* si può riferire indipendentemente a qualunque genere (con accordo in schwa). I pronomi *egli*/*ella*, *essi*/*esse* (quando riferiti a persone) diventano *ellə* al singolare e *essə* al plurale; possono anche essere sostituiti con *ləi* o *loro* o con un pronome zero in quasi tutti i contesti.
6. I **dimostrativi** sono *questə* e *quellə* sia al singolare che al plurale, con elisione al singolare davanti a nomi che iniziano per vocale (*quest'*, *quell'*). La distinzione di numero è data dagli altri elementi della frase; ad esempio: "Se le relazioni si guastano con uno di questi, l'utilità dell'assistente […] è […] compromessa"[32] à "Se le relazioni si guastano con unə di questə, l'utilità dell'assistente […] è […] compromessa"; in alcuni casi il contesto frasale non è però sufficiente a distinguere tra singolare e plurale: "Si è quasi tentati di invitare questi anziani leader ad aprire la porta e a lasciare entrare il successo."[33] à "Si è quasi tentatə di invitare questə anzianə leader ad aprire la porta e a lasciare entrare il successo."
7. I **nomi** vengono trattati diversamente in base alla classe (vedi Gheno, 2020a), ma in tutti i casi non ci sono differenze tra la forma singolare e quella plurale (la distinzione è data dalle altre parole che concordano con il nome, in particolare gli articoli):
    a. Nomi di genere mobile
        i. Nomi in -o/-a à -ə: *il maestro* / *la maestra* / *lə maestrə*, *i maestri* / *le maestre* / *ə maestre*. Casi particolari:

[32] Frase tratta da MT-GenEval
[33] GeNTE – Set-N_Schwa [208]

1. Nomi in -co/-ca, -ci/-che e -go/-ga, -gi/-ghe: la consonante ha suono duro davanti a schwa, sia al singolare che al plurale, senza bisogno di scrivere la *h*: *l'amico* / *l'amica* / *l'amicə*, *gli amici* / *le amiche* / *ə amicə* (in IPA [amikə]); *lo psicologo / la psicologa / lə psicologə, gli psicologi / le psicologhe / ə psicologə* (in IPA [psikologə]);
2. Nomi in -cio/-cia, -ci/-cie e -gio/-gia, -gi/-ge: la pronuncia della consonante rimane dolce davanti a schwa, mantenendo la -i- nella grafia: *il saggio* / *la saggia* / *lə saggiə*, *i saggi* / *le sagge* / *ə saggiə* (in IPA [sadʒːə];
3. Nomi in -io/-ia/, -i(i)/-ie: la forma con schwa finisce sempre in -iə: *il segretario* / *la segretaria* / *lə segretariə*, *i segretari* / *le segretarie* / *ə segretariə*.

ii. Nomi in -e/-a à -ə, esempio: *pompiere* / *pompiera* / *pompierə*. Casi particolari:
1. Nomi in -tore/-trice/-tora o -sore/-ditrice/-sora: per la forma con schwa si utilizza il modello del maschile o del femminile analogico à -torə, -sorə: *l'autore* / *l'autrice/autora* / *l'autorə*, *gli autori* / *le autrici/autore* / *ə autorə*; *il difensore* / *la difenditrice/difensora* / *lə difensorə*, *i difensori* / *le difenditrici/difensore* / *ə difensorə*; *l'assessore* / *l'assessora* / *l'assessorə*, *gli assessori* / *le assessore* / *ə assessorə*.

iii. Femminili in -essa:
1. Se il nome di partenza è modellato su un participio presente in -ente, -enti o -ante, -anti, si utilizzerà questa come forma epicena (vedi punto b.) valida per qualunque genere: *il presidente* / *la presidentessa/presidente* / *lə presidente*; *i*/*le*/*ə parlanti*;
2. Se al femminile in -essa corrisponde un maschile in -sore, si utilizza la stessa strategia applicata ai femminili in -trice/-tora, -ditrice/-sora: *il professore* / *la professoressa/professora* / *lə professorə*, *i professori* / *le professoresse/professore* / *ə professorə*.

b. <u>Nomi di genere comune (epiceni)</u>
i. Nomi epiceni sia al singolare che al plurale (per la maggior parte modellati su participi presenti): si utilizza la stessa forma per qualunque genere: *lo/la/lə studente*, *gli/le/ə studenti*; *il*/*la*/*lə giudice*, *i*/*le*/*ə giudici*.

ii. Nomi epiceni al singolare ma mobili al plurale (es. nomi in -eta, -ista, -iatra): *l'atleta*, *gli atleti* / *le atlete* / *ə atletə*; *il*/*la*/*lə dentista*, *i dentisti* / *le dentiste* / *ə dentistə*. Casi particolari:
    1. Nomi in -ga, -ghi/-ghe: il suono rimane duro anche davanti a schwa, sia al singolare che al plurale, senza bisogno di scrivere la *h*: *il*/*la*/*lə collega*, *i colleghi* / *le colleghe* / *ə collegə* (in IPA [kol:egə]).

c. Nomi di genere promiscuo: nessun intervento sui nomi di questa classe, per cui l'accordo segue il genere grammaticale del nome a prescindere dal genere della persona; esempio: *la persona*, *il membro*, *la guida*, *la spia*, *l'individuo*, ecc.

d. Nomi di genere fisso: non è stata trovata una soluzione per questi nomi, che per quanto riguarda i referenti umani comprendono principalmente i nomi di parentela. Non sono infatti state trovate soluzioni condivise per superare la distinzione maschile-femminile in questi nomi (es. *madre/padre, fratello/sorella*, ecc.).

8. Per gli **aggettivi** – qualificativi e non – si applicano generalmente le stesse regole dei nomi con la stessa struttura morfologica. Ad esempio, i possessivi si formano sul modello dei nomi di genere mobile in -o/-a, quindi: *(il) mio* / (*la) mia* / (*lə) miə*, *(i) miei* / *(le) mie* / *(ə) miə*; (il) nostro / (la) nostra / (lə) nostrə; per la terza persona plurale, il possessivo *loro* si applica a qualunque genere, ma la distinzione può comunque essere data dall'articolo: ad esempio, *i loro* / *ə loro*.
9. L'accordo dei **participi** si fa come per i nomi e gli aggettivi. I participi presenti possono essere trattati come nomi epiceni (*presidente*), mentre quelli passati come nomi di genere mobile. In italiano contemporaneo, il participio passato si accorda obbligatoriamente con il soggetto solo se il verbo è un intransitivo con ausiliare *essere*: "*No one has been able* to explain to me yet […]"[34] à "Finora nessunə è riuscitə a spiegarmi […]"; oppure con l'oggetto, se questo è un pronome personale di terza persona: "[…] we are too dubious […] not to refrain from *putting them on their guard*."[35] à "[…] nutriamo troppi dubbi […] per astenerci dal metterlə in guardia." (Telve, 2011).

Chiaramente, le forme in schwa interessano soltanto le parole che possono riferirsi a esseri umani, compresi i nomi e tutte le parole influenzate dall'accordo (es. aggettivi, pronomi, forme participiali, ecc.).

In particolare, lo schwa viene usato in tutti i casi in cui in italiano standard si usa frequentemente il maschile sovraesteso (§1.2.1; vedi anche Proto, 2021), ovvero per:

[34] GeNTE – Set-N_Schwa [49]
[35] GeNTE – Set-N_Schwa [84]

- singoli referenti specifici di genere non binario o sconosciuto:

After retiring from teching, Cook became a novelist. > Dopo aver lasciato l'insegnamento, Cook è diventatə unə scrittorə di romanzi.[36]

- gruppi di referenti di genere (presumibilmente) misto:

[The novel] left the literary critics divided […]. > [Il romanzo] lasciò […] ə criticə letterariə divisə […].[36]

- referenti generici, al singolare o al plurale:

A company secretary routinely enters into contracts in the company's name […]. > Lə segretariə di una società subentra nei contratti a nome della società […].[36]

In questo lavoro, tutte le parole riferite a esseri umani e marcate rispetto al genere sono state riscritte in LND anche dove il LNI sarebbe stato appropriato, al solo scopo di addestrare i modelli. Come visto in §1.3, tuttavia, le strategie di LNI e LND possono coesistere, e anzi, come sottolineato anche nella linea editoriale di effequ, in una situazione reale, la strategia migliore sarebbe limitare l'uso dello schwa ai casi in cui il LNI non sarebbe una soluzione possibile o ideale; nel caso in cui ci si riferisca specificamente a persone non binarie, il LND rimane però l'approccio più raccomandabile, soprattutto se si sta traducendo un testo dove si utilizzano strategie dello stesso tipo. Le strategie di LNI non sono infatti sufficienti a rappresentare esplicitamente le identità non binarie, e López (2020) evidenzia come il LNI rischia di essere un altro modo di cancellarle.

Attualmente, invece, molte traduzioni di testi e prodotti audiovisivi che comprendono personaggi non binari tendono a cancellarne l'identità, riportandoli a forme binarie (vedi ad esempio Zanfabro, 2019; Misiek, 2020), così come succede anche negli articoli di informazione (Lardelli & Gromann, 2023b); la diffusione del LND è quindi importante anche perché un'adozione più generalizzata di queste forme può contribuire a migliorare la rappresentazione di tali identità anche in questi contesti.

Nel prossimo capitolo sono riportati i risultati delle valutazioni automatica e manuale di tutti gli esperimenti descritti fin qui; la presentazione dei risultati è seguita da una discussione contenente alcuni esempi.

[36] Frase tratta da MT-GenEval

3. Risultati

In questo capitolo sono riportati i dati relativi alle prestazioni dei classificatori (§3.1) e i risultati della valutazione automatica (§3.2) e manuale (§3.3) dei tre set di traduzioni ottenute da ModernMT (MMT) e Llama 3, seguiti da una loro interpretazione e discussione.

I risultati integrali, con gli output di tutti i modelli e la valutazione manuale per tutti gli esperimenti, si trovano al link https://github.com/paolo-mainardi/tratec-tesi.

3.1. Classificatori

Le Tabelle 11 e 12 contengono i risultati della valutazione automatica delle previsioni dei classificatori sul test set per ogni tipo di rappresentazione dei dati, rispettivamente per l'inglese e l'italiano. Tutte le metriche sono state calcolate tramite `scikit-learn`[37], utilizzando per ogni modello la configurazione migliore, e rappresentano la differenza tra le previsioni del modello e l'annotazione manuale. Tutti i punteggi sono riportati come media delle prestazioni sulle due classi (0 per le frasi neutre e 1 per le frasi marcate).

	W2V		**FT**		**GloVe**		**RoBERTa**	
	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM
Precision	**1.00**	0.94	0.99	0.35	0.99	0.35	**1.00**	0.99
Recall	**1.00**	0.94	0.99	0.59	0.99	0.59	**1.00**	0.99
F1	**1.00**	0.94	0.99	0.44	0.99	0.44	**1.00**	0.99
Accuracy	0.9967	0.9399	0.9935	0.5908	0.9903	0.5908	**0.9983**	0.9919

Tabella 11. Valutazione automatica dei classificatori per l'inglese. I valori in grassetto rappresentano le prestazioni migliori.

	W2V		**FT**		**RoBERTa**		**UmBERTo**	
	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM
Precision	0.84	0.33	0.83	0.33	0.87	0.60	**0.88**	0.56
Recall	0.85	0.57	0.83	0.57	0.87	0.60	**0.88**	0.55
F1	0.84	0.42	0.83	0.42	0.87	0.57	**0.88**	0.55
Accuracy	0.8451	0.5710	0.8299	0.5710	0.8680	0.6040	**0.8807**	0.5507

Tabella 12. Valutazione automatica dei classificatori per l'italiano. I valori in grassetto rappresentano le prestazioni migliori.

In base a questi risultati si possono fare due considerazioni principali. Da un lato, per una stessa architettura, i risultati migliori si ottengono sempre con delle rappresentazioni contestuali:

[37] https://scikit-learn.org/stable/about.html

come visto in §2.5, infatti, esse sono più dinamiche e, quindi, adeguate a rappresentare il linguaggio rispetto a quelle statiche. Dall'altro, utilizzando una stessa rappresentazione dei dati, i CNN raggiungono sempre prestazioni migliori rispetto agli LSTM: data la loro architettura, infatti, i primi permettono di dare più importanza a caratteristiche localizzate, a livello di n-gram, con un raggio di pochi token, mentre i secondi prendono in considerazione l'intera sequenza. Nel nostro caso, per entrambe le lingue, il modello migliore utilizza una finestra di 3 token. Questo suggerisce che, per risolvere il compito in questione, è a questo livello che si trovano le informazioni più importanti: in effetti, data l'impostazione dell'esperimento, una frase è considerata marcata se anche solo uno dei suoi referenti è marcato al maschile o al femminile; l'intero contesto frasale può quindi contenere informazioni contrastanti, che possono portare a confusione, e questo potrebbe spiegare l'inferiorità degli LSTM, seppur minima in alcuni casi.

La disparità tra l'inglese e l'italiano, invece, può essere spiegata alla luce delle differenze morfologiche tra le due lingue. Come visto in §1.2, la morfologia dell'italiano, anche ma non solo in termini di codifica del genere grammaticale, è molto più complessa rispetto a quella dell'inglese, che per altro contiene molte meno forme marcate rispetto all'italiano. Per questo motivo, nelle tre configurazioni comuni alle due lingue (ovvero quelle con Word2Vec, fastText e RoBERTa) abbiamo ottenuto risultati migliori sull'inglese rispetto all'italiano.

Nonostante l'approccio risulti promettente nel complesso, l'impostazione dell'esperimento non permette di giungere a conclusioni definitive riguardo alle possibili prestazioni in condizioni di uso effettivo, in particolare perché le frasi utilizzate per addestrarli e valutarli contengono ognuna un solo set di marche di genere: ovvero, ogni frase può avere uno o più referenti, ma tutti i referenti in un'unica frase saranno tutti o maschili, o femminili, o ambigui. Per assicurare risultati più affidabili, in futuro sarebbe quindi necessario testare – ed eventualmente addestrare – i classificatori su dati più realistici e naturali da questo punto di vista.

3.2. Valutazione automatica

Le metriche utilizzate per la valutazione automatica degli output di ModernMT e Llama, presentate nello scorso capitolo (§2.4.1), sono quattro. Da un lato, BLEU e chrF indicano il grado di sovrapposizione tra la risposta del modello e la frase di riferimento, rispettivamente a livello di token e di caratteri: più alto è il punteggio, maggiore è la corrispondenza; dall'altra parte, TER e characTER (cTER) danno un'idea della quantità di modifiche necessarie per passare dalla risposta del modello alla frase di riferimento , la prima a livello di token e la seconda di caratteri: in questo caso, a un punteggio più alto corrisponde una maggiore distanza tra le due frasi.

In Tabella 13 sono riportati i punteggi relativi a tutte queste metriche, per ciascuno dei sei esperimenti. I punteggi originali relativi alle metriche (charac)TER sono stati convertiti sottraendoli a 1, in modo da renderle più facilmente comparabili con le altre due. Per BLEU, il punteggio globale rappresenta la media tra i valori specifici calcolati sugli n-gram da 1 a 4; per il calcolo del punteggio chrF, invece, è stata utilizzata una finestra di 6 caratteri.

Modello	Configurazione	BLEU	chrF	TER	cTER
ModernMT	baseline	0.2255	0.5580	0.3195	0.5093
	adapted	0.2179	0.5577	0.3028	0.5047
	post-editing	0.2320	0.5573	0.3027	0.5046
Llama 3	baseline	0.1899	0.5336	0.2927	0.4847
	istruzioni	0.1863	0.5311	0.2916	0.4821
	contrastiva	**0.8101**	**0.9312**	**0.8858**	**0.9650**

Tabella 13. Risultati della valutazione automatica dei sei esperimenti con ModernMT e Llama 3. I valori sottolineati indicano la prestazione migliore su una metrica tra le tre configurazioni di uno stesso modello, mentre i valori in grassetto indicano la prestazione migliore su una metrica in assoluto.

Considerando solo le prime due configurazioni per ciascun modello, i risultati rivelano che la qualità delle traduzioni di Llama (`Llama_baseline`, `Llama_istruzioni`) non raggiunge quella di ModernMT (`MMT_baseline`, `MMT_adapted`). Inoltre, sia l'adattamento tramite memoria di traduzione (MT) (`MMT_adapted`) che l'aggiunta delle istruzioni (`Llama_istruzioni`) hanno prodotto un leggero peggioramento nelle traduzioni prodotte da entrambi i modelli: sembra, quindi, che questi due approcci, applicati al compito in questione, creino ambiguità difficilmente risolvibili dai sistemi testati, come emerge anche dall'analisi manuale, in particolare per Llama (§3.3).

Al contrario, gli esempi contrastivi sembrano aver avuto un impatto importante per Llama, visto il grande distacco nei punteggi di questa configurazione (`Llama_contrastiva`) rispetto a tutte le altre, compresa quella migliore per ModernMT (`MMT_post-editing`). Questi risultati rivelano una quasi totale sovrapposizione delle traduzioni di Llama in questa configurazione con le traduzioni di riferimento. A questo proposito, è importante ricordare che gli esempi contrastivi costituiscono una grande quantità di dati in più su cui il modello si può basare per produrre le traduzioni, il che spiega in parte questo grande distacco: fornendo due traduzioni per ogni frase del set di addestramento, e una traduzione marcata per quelle di valutazione, il modello ha imparato a rimanere molto aderente alle traduzioni marcate fornite come esempi contrastivi, distanziandosi significativamente anche dalle traduzioni prodotte nelle altre due configurazioni (`Llama_baseline` e `Llama_istruzioni`).

ModernMT non ha avuto accesso a tali dati in nessuna configurazione, compresa quella con post-editing, che comporta semplicemente la modifica delle marche di genere delle traduzioni già ottenute in `MMT_baseline`: gli eventuali errori linguistici o di traduzione presenti in quest'ultime non possono essere corretti con questa strategia, e questo si riflette quindi nei risultati relativi a questa configurazione.

In ogni caso, queste metriche forniscono un'idea generale della qualità delle traduzioni ottenute rispetto alle frasi di riferimento, ma, da sole, non permettono di verificare quale configurazione raggiunge meglio l'obiettivo di questo studio, ovvero la traduzione neutra rispetto al genere (GNT), e non premiano eventuali traduzioni perfettamente accettabili ma semplicemente formulate in modo diverso rispetto a quelle di riferimento. Per poter valutare i risultati di questi esperimenti in modo più dettagliato, soprattutto in termini di riduzione dei fenomeni legati al bias di genere, è stato quindi necessario eseguire anche un'analisi manuale, i cui risultati sono presentati e discussi nella sezione seguente.

3.3. Valutazione manuale

Le Figure 1 e 2 riportano il numero di frasi a cui è stato assegnato ogni punteggio rispettivamente per accettabilità e neutralità, in ognuno dei sei esperimenti.

Ricordiamo che per entrambe le dimensioni, un punteggio più basso significa che la frase ha raggiunto le aspettative: 1 in neutralità significa che la frase è completamente neutra, e allo stesso modo, 1 in accettabilità significa che la frase è completamente accettabile dal punto di vista linguistico e traduttivo, come spiegato nello scorso capitolo (§2.4).

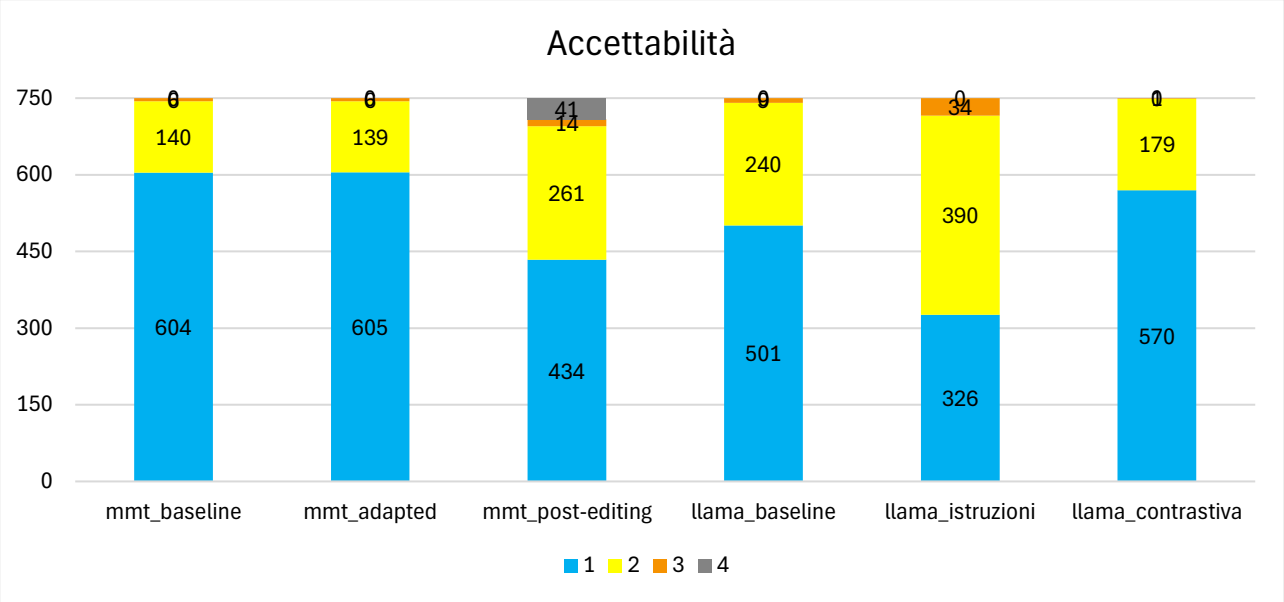

Figura 1. Distribuzione dei punteggi di accettabilità nei sei esperimenti con ModernMT e Llama 3.

In primo luogo, la dimensione dell'accettabilità misura la qualità di una frase dal punto di vista traduttivo, quindi rispetto alla frase di partenza, e linguistico o grammaticale rispetto alla lingua di arrivo; in questo caso, questo include anche l'accuratezza nell'uso dello schwa, dove presente, in base alle linee guida presentate nello scorso capitolo (§2.6).

Dalla Figura 1 si può notare immediatamente che le traduzioni di ModernMT, nelle prime due configurazioni (`MMT_baseline`, `MMT_adapted`), hanno il numero più alto di frasi a cui è stato assegnato un punteggio di 1, corrispondente a una frase del tutto accettabile, in linea con i risultati della valutazione automatica, che suggeriscono un'alta qualità delle traduzioni. Lo scarso impatto dell'adattamento con MT è invece confermato anche dall'analisi manuale.

D'altra parte, si può notare una variazione più ampia nella qualità delle traduzioni di Llama: in particolare, in `Llama_istruzioni`, la qualità delle traduzioni è significativamente peggiore rispetto alle altre due configurazioni per lo stesso modello. In effetti, l'analisi manuale ha permesso di rilevare che, diversamente da `Llama_baseline`, lo schwa compare nella maggior parte delle traduzioni, ma viene usato spesso in modo errato:

(18) Mi sono fermatoə davanti al carcere a Tunisi il 9 aprile e ho visto centinaia di poliziottiə che scioglievano una manifestazione di solidarietà organizzata da democraticiə.

`Llama_istruzioni` [502]

L'esempio (18) è rappresentativo del comportamento di Llama in questa configurazione, con lo schwa attaccato alla fine delle parole, ma senza sostituire le marche maschili. Tale comportamento sembra confermare l'ipotesi secondo cui questa strategia introduce più ambiguità che altro per il modello, in linea anche con Vanmassenhove (2024), che adotta questa strategia per richiedere a un altro LLM conversazionale di fornire nelle traduzioni tutte le alternative possibili in termini di genere, quindi senza l'introduzione di strategie non-standard come quella utilizzata qui.

Sempre in linea con i risultati dell'analisi automatica, anche dalla valutazione dell'accettabilità delle traduzioni di Llama aumenta significativamente nella configurazione `Llama_contrastiva`, verosimilmente per i motivi esposti sopra. Questa strategia, inizialmente pensata prevalentemente per spingere il modello a utilizzare lo schwa, si è rivelata quindi utile anche per una migliore qualità delle traduzioni, innescando un comportamento più simile alla riformulazione delle traduzioni marcate che alla traduzione diretta delle frasi di partenza.

Infine, dalla Figura 1 si può notare che `MMT_post-editing` è l'unica configurazione in cui a un certo numero di frasi (41) è stato assegnato un punteggio di 4, che denota una frase completamente inaccettabile. Tali errori non sono da imputare a ModernMT, ma al modello di

riformulazione, che in questi casi ha generato frasi incomplete (19) o con segmenti ripetuti più volte (20):

(19) Sono fiducioso che con questo insieme sistematico di regole uniformi stiamo gettando le basi per colmare le lacune nella legislazione esistente e migliorare la sicurezza alimentare lungo tutta la catena alimentare.
> fiduciosə che con questo insieme sistematico di regole uniformi stiamo gettando le basi per colmare le lacune nella legislazione esistente e migliorare la sicurezza alimentare lungo tutta la catena alimentare.

`MMT_post-editing` [212]

(20) Ma abbiamo bisogno di un testo costituzionale per l'Europa il più rapidamente possibile, un testo per i nostri cittadini che contenga dichiarazioni chiare e comprensibili sull'Europa, che dica loro quali diritti hanno, dove l'Unione si è assunta la responsabilità, quali poteri dovrebbero esercitare gli Stati membri, anche in futuro, e come l'Europa intende fare il suo lavoro, con ogni garanzia che lo Stato di diritto proteggerà i loro diritti umani.
> ma abbiamo bisogno di un testo costituzionale per l'europa il più rapidamente possibile, un testo per i nostri cittadini che contenga dichiarazioni chiare e comprensibili sull'europa, che dica loro quali diritti hanno, dove l'unione si è assunta la responsabilità, quali poterə dovrebbero esercitarə gli stati membri, anche in futuro, e come l'europa intende fare l'europa into bisogno di un testo costitul'europa l'europa l'europa l'europa l'europa l'europa l'europa, un testo per l'europa intul'europa, un testo costitul'europa intul'europa intə l'europa intə l'europa intə l'europa intə

`MMT_post-editing` [72]

In (19), l'aggettivo "fiducioso" è correttamente convertito in "fiduciosə", ma manca la prima parola: "Sono", e per questo la frase, altrimenti del tutto accettabile, non lo è più. La frase in (20), invece, è esemplificativa della *degenerazione del testo*, un problema noto nei modelli generativi (vedi ad esempio Holtzman et al., 2020). Nel caso di questo esperimento, il problema è stato contenuto grazie alla tecnica di generazione applicata (descritta nello scorso capitolo: §2.2.1), ma è rimasto in queste frasi.

I risultati relativi all'accettabilità, che, come visto, prendono in conto anche l'uso dello schwa, sono meglio compresi se osservati in relazione a quelli sulla neutralità, presentati in Figura 2.

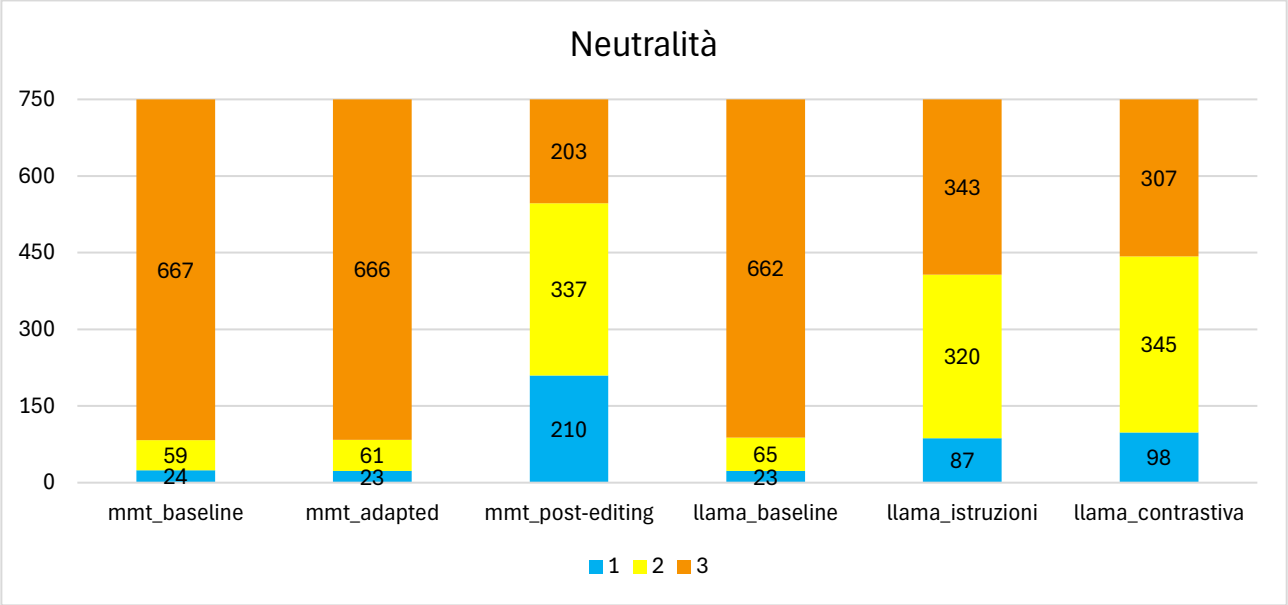

Figura 2. Distribuzione dei punteggi di neutralità nei sei esperimenti con ModernMT e Llama 3.

I punteggi di neutralità indicano quanti, tra i referenti umani contenuti in una frase, sono neutri rispetto al genere, ovvero non marcati al maschile o al femminile. Dalla distribuzione dei punteggi assegnati alle traduzioni ottenute nei sei esperimenti si configura una situazione diversa rispetto a quella restituita dai punteggi di accettabilità.

Stando all'accettabilità, infatti, le traduzioni migliori sono quelle di `MMT_baseline` e `MMT_adapted`; tuttavia, da uno sguardo rapido al grafico sulla neutralità si può notare immediatamente che, da questo punto di vista, quelle due configurazioni sono ampiamente insufficienti a raggiungere gli obiettivi della GNT (con solo 22 e 24 frasi completamente neutre, rispettivamente), insieme anche a `Llama_baseline` (con solo 23 frasi completamente neutre), un'altra delle configurazioni con i punteggi di accettabilità migliori.

Tramite l'analisi manuale, abbiamo notato che, effettivamente, in queste tre configurazioni non compare nemmeno una volta lo schwa; questo risultato era previsto per le due baseline, ma non per `MMT_adapted`, il che conferma l'osservazione, basata sulla valutazione automatica e sui punteggi di accettabilità, che l'adattamento tramite MT non ha influenzato significativamente l'output di ModernMT.

In queste tre configurazioni, le frasi parzialmente o completamente neutre non lo sono quindi grazie all'utilizzo dello schwa, ma, come visto in §2.4.2, sono date dall'utilizzo di forme già ambigue rispetto al genere, in base anche al lessico utilizzato nella frase di partenza corrispondente, come esemplificato dalle tre traduzioni della frase (21):

(21) We Liberals, along with certain other members of this House, would like to take this opportunity to make it absolutely crystal clear that we are in favour of improving this openness regulation.

> (21a) Noi liberali, insieme ad alcuni altri membri di questo Parlamento, vorremmo cogliere l'occasione per chiarire in modo assolutamente cristallino che siamo a favore del miglioramento di questo regolamento sull'apertura.

`MMT_baseline/adapted` [25]

> (21b) Noi Liberali, insieme ad alcuni altri membri di questa Camera, vorremmo cogliere questa occasione per rendere assolutamente chiaro che siamo a favore di migliorare questa regolamentazione sulla trasparenza.

`Llama_baseline` [25]

La traduzione di ModernMT (21a) è la stessa sia con che senza MT e, così come quella di Llama (21b), è completamente neutra pur senza utilizzare lo schwa: "liberali" è una forma epicena, e "membri", come già visto, è un nome promiscuo; entrambe le scelte possono essere state influenzate dalla frase di partenza, che contiene "Liberals" e "members".

Al contrario, Llama è stato in grado di adottare lo schwa nelle sue traduzioni con entrambe le strategie di adattamento; tuttavia, se, come emerso dalla valutazione di accettabilità, con le sole istruzioni il modello non è in grado di applicarla sempre correttamente e in modo sistematico, con gli esempi contrastivi si ottengono risultati più accurati: la maggiore aderenza alle frasi di riferimento, come visto sopra, risulta in un importante miglioramento rispetto alle prime due configurazioni. Un esempio è dato dalle traduzioni (22a) e (22b):

(22a) Come ex-negociatore sindacale, sono abituatə a difendere l'esito delle trattative.

`Llama_istruzioni` [336]

(22b) In qualità di ex negoziatorə sindacale, sono abituatə a difendere i risultati delle negoziazioni.

`Llama_contrastiva` [336]

Nella prima (`Llama_istruzioni`), al di là dell'errore di ortografia, il nome "negociatore" rimane al maschile, nonostante la corretta neutralizzazione dell'aggettivo "abituatə" ad esso collegato; al contrario, nella seconda (`Llama_contrastiva`) il modello neutralizza correttamente l'intera unità: "negoziatorə […] abituatə".

Le traduzioni di ModernMT con post-editing, nonostante il peggioramento rispetto alla baseline emerso dalla valutazione dell'accettabilità, rappresentano anche la configurazione migliore in quanto a neutralità, ovvero quella con il numero più alto di traduzioni completamente neutre (210, contro le 98 di Llama con esempi contrastivi, la seconda migliore sotto questo aspetto). In questo caso, rispetto alla baseline, molte più frasi sono effettivamente neutre. Dall'analisi manuale risulta in effetti che `MMT_post-editing` è l'unica configurazione in cui compaiono forme in schwa non solo sui nomi, ma anche sugli articoli e le preposizioni, cosa che non accade nelle due versioni adattate di Llama:

(23) What is significant is that the majority of Russian citizens supported this move.

> (23a) Ciò che è significativo è che la maggioranza dei cittadinə russi hanno sostenuto questa mossa.

`Llama_istruzioni` [596]

> (23b) ciò che è significativo è che la maggior parte deə cittadinə russə ha sostenuto questa mossa.

`MMT_post-editing` [596]

Per la traduzione della frase in (23), Llama con le istruzioni (23a) applica giustamente lo schwa per la traduzione di *citizens* ("cittadinə"), ma non a tutte le altre parole che dovrebbero concordare con il nome ("dei", "russi"); al contrario, il modello di post-editing applicato alla traduzione di MMT (23b) utilizza correttamente lo schwa su tutto il sintagma nominale ("deə cittadinə russə").

In base a quanto visto fin qui, nella prossima e ultima sezione si trarranno alcune conclusioni su questo lavoro e se ne discuteranno le limitazioni principali.

Conclusione

In questa tesi abbiamo discusso di bias di genere in traduzione automatica, parte del più ampio problema del bias in intelligenza artificiale, da una prospettiva non binaria (Capitolo 1). Abbiamo inoltre effettuato sei esperimenti esplorativi (descritti nel Capitolo 2) sulla possibilità di adottare una forma di linguaggio non binario diretto (LND: vedi §1.3.2) nella traduzione automatica di frasi inglesi ambigue rispetto al genere verso l'italiano, mantenendo questa ambiguità tramite l'uso dello schwa (ə). Viste le difficoltà riscontrate nel trovare dati per l'addestramento e la valutazione di modelli che possano svolgere il compito della traduzione neutra rispetto al genere (GNT: vedi §2.6), abbiamo anche addestrato due classificatori per l'identificazione automatica di frasi marcate o ambigue rispetto al genere, rispettivamente in inglese e in italiano (§2.5), con lo scopo di automatizzare in parte questo processo in futuro.

I sei esperimenti effettuati hanno coinvolto due modelli principali: ModernMT, un sistema di traduzione automatica neurale, e Llama 3, uno dei *large language model* rilasciati più di recente, utilizzato per la traduzione automatica. Entrambi i modelli, usati senza alcun tipo di adattamento, non raggiungono gli obiettivi della GNT, dimostrando un ampio utilizzo del maschile sovraesteso (§1.2.1) nella traduzione di frasi inglesi ambigue rispetto al genere, in linea con quanto emerso da lavori precedenti. Per questo motivo, abbiamo adottato, per ciascun modello, due strategie volte a promuovere l'uso dello schwa, un neomorfema introdotto recentemente in italiano per indicare un genere indistinto, al posto del maschile sovraesteso; per fare questo una persona ha riscritto manualmente le traduzioni marcate contenute nei dataset di addestramento (1528) e di valutazione (750), sostituendo tutte le marche maschili e femminili con lo schwa, in base a delle linee guida elaborate appositamente per questo studio (§2.6).

Da un lato, per ModernMT (§2.2), i due esperimenti consistevano nell'adattamento tramite una memoria di traduzione – in cui le frasi di partenza erano allineate alle traduzioni riformulate in schwa – e nel post-editing automatico delle traduzioni ottenute, in seguito all'ottimizzazione (*fine-tuning*) del modello di riscrittura IT5 su questo compito (§2.2.1). Per Llama 3 (§2.3), invece, abbiamo sperimentato i risultati dell'adattamento in due fasi progressive: prima abbiamo fornito al modello una semplice richiesta esplicita di tradurre le frasi utilizzando lo schwa al posto dei morfemi maschili e femminili per i referenti umani; poi abbiamo aggiunto alle frasi da tradurre due esempi di traduzione – una marcata e una con schwa – spingendo il modello a seguire il secondo esempio per le frasi da tradurre.

Le traduzioni ottenute nei sei esperimenti sono state valutate sia tramite metriche automatiche – volte a cogliere la differenza tra l'output del modello e il risultato atteso, a livello sia di token che

di singoli caratteri – sia tramite una valutazione manuale basata su due aspetti, ovvero l'effettivo mantenimento dell'ambiguità di genere e l'accettabilità generale delle traduzioni dal punto di vista linguistico.

I risultati di questa valutazione hanno confermato l'inadeguatezza di entrambi i modelli, così come sono, di fornire traduzioni neutre rispetto al genere. Gli esperimenti di adattamento hanno invece fornito risultati diversi.

Da un lato, per ModernMT, l'uso della memoria di traduzione non ha permesso di raggiungere i risultati sperati, con un impatto praticamente nullo sulle traduzioni fornite dal modello; l'applicazione del post-editing automatico alle traduzioni fornite da ModernMT, invece, ha permesso di ottenere frasi con un utilizzo molto più ampio e sistematico dello schwa, risultando come il miglior approccio da questo punto di vista, nonostante alcuni problemi nel testo generato (vedi §3.3).

Llama 3 si è rivelato invece più sensibile all'adattamento, inserendo alcune forme in schwa già solo con una semplice richiesta esplicita e senza nessun esempio; in questa configurazione, però, il modello non sembra in grado di applicare tale strategia in modo sistematico né coerente, non esistendo regole condivise per il suo uso in italiano (vedi §1.3.2). Al contrario, gli esempi contrastivi hanno avuto un impatto più ampio del previsto, spingendo il modello a svolgere sostanzialmente un compito di riscrittura delle traduzioni marcate fornite insieme alle frasi da tradurre, applicandovi le regole dedotte dagli esempi di traduzioni con schwa forniti come dati di addestramento. In questa configurazione, quindi, anche la qualità delle traduzioni di Llama 3 è aumentata significativamente, un effetto che non era stato previsto. Una lacuna di questa configurazione risiede tuttavia nell'uso di schwa su articoli e preposizioni, soddisfacente soltanto nel caso del post-editing, che infatti raggiunge una valutazione migliore per quanto riguarda la neutralità delle traduzioni.

Nel complesso, gli esperimenti confermano l'inadeguatezza dei sistemi testati per quanto riguarda il linguaggio non binario. I risultati ottenuti suggeriscono inoltre che soluzioni linguistiche come il LND, che implicano l'utilizzo di strategie non-standard come lo schwa, siano integrabili con più successo tramite strumenti più flessibili, di tipo generativo (come i LLM). L'adattamento di ModernMT tramite memoria di traduzione, in particolare, si è rivelato insoddisfacente, probabilmente a causa della differenza tra questo tipo di esperimento e l'uso tipico delle memorie di traduzione, ovvero l'adattamento a un dominio specifico, solitamente dal punto di vista di terminologia e stile. Parallelamente, le semplici istruzioni non sono state sufficienti per adattare Llama 3, e sembrano creare ambiguità risolvibili soltanto con l'aggiunta di esempi; l'adattamento degli LLM conversazionali è un'area di ricerca ancora poco sviluppata, ma l'insufficienza delle sole istruzioni è rilevata, ad esempio, anche nel già citato lavoro di Vanmassenhove (2024), e in questo caso potrebbe essere aggravata dalla natura innovativa della soluzione utilizzata.

A questo proposito, era stato ipotizzato (§2.1) che i diversi domini di appartenenza dei dati utilizzati potessero influenzare negativamente i risultati; tuttavia, dato il maggiore successo degli esperimenti con il modello di riscrittura (IT5) per ModernMT e con gli esempi contrastivi per Llama 3, questa seconda spiegazione, relativa alla natura del compito da svolgere, sembra più convincente.

A prescindere dai risultati ottenuti, questo studio presenta ovviamente diverse limitazioni, a cui si è accennato brevemente in diverse sezioni e che abbiamo raccolto sistematicamente nella prossima.

Limitazioni e sviluppi futuri

Una prima questione da prendere in considerazione è relativa alle regole formulate e utilizzate per riscrivere le traduzioni italiane usando lo schwa. Come spiegato in §1.3, le strategie di linguaggio non binario diretto (come lo schwa) non hanno in questo momento status ufficiale e non sono accettate in modo uniforme nelle comunità di parlanti. In questa fase, quindi, diverse strategie coesistono, e le linee guida formulate e adottate qui sono solo una proposta di sistematizzazione, che potrebbe non corrispondere all'uso che si potrebbe affermare in futuro.

Un'altra questione fondamentale riguarda i modelli utilizzati negli esperimenti. A parte IT5 (così come T5, di cui è una versione), di cui sono disponibili e documentati i dati, il codice e i metodi di addestramento, per ModernMT e Llama 3 queste informazioni non sono disponibili, trattandosi di sistemi commerciali. Per questi motivi, non è possibile garantire la riproducibilità degli esperimenti con ModernMT e con Llama 3 (via GroqCloud). In effetti, Meta, azienda che ha creato i modelli della famiglia Llama, ha dichiarato in diverse occasioni che tali modelli sono *open-source*[19]; tuttavia, come dimostrato ad esempio in Liesenfeld & Dingemanse (2024), questa affermazione non corrisponde a una vera trasparenza e il rilascio dei modelli non è mai stato supportato da una pubblicazione sottoposta a revisione paritaria.

Per quanto riguarda, invece, i classificatori, abbiamo già discusso in §2.5 la loro possibile integrazione con un metodo di post-editing automatico (come quello adottato qui per ModernMT); tuttavia, per questa tesi non è stato possibile combinare le due risorse, e abbiamo lasciato un eventuale esperimento in questo senso per lavori futuri.

Un altro aspetto che potrebbe essere interessante esplorare è relativo al maggiore successo degli esperimenti con modelli generativi rispetto all'adattamento tramite memoria di traduzione. Per esempio, si potrebbe verificare empiricamente se, in fase di valutazione, i modelli testati applicano meglio lo schwa per le espressioni presenti anche nei dati di addestramento: i risultati di un tale esperimento potrebbero infatti fornire più spunti sulla possibilità o meno di trasferire le regole apprese dai dati di addestramento anche a testi appartenenti a domini lontani.

Inoltre, questo studio è limitato dal punto di vista delle lingue considerate: nonostante la traduzione dall'inglese all'italiano sia utile per l'analisi dei fenomeni legati al bias di genere in TA e generalizzabile a coppie di lingue simili, concentrarsi su un'unica direzione non permette di tenere conto dei fenomeni che potrebbero emergere, ad esempio, nella traduzione tra due diverse lingue a genere grammaticale o verso una lingua senza genere grammaticale. Lo studio è inoltre incentrato in particolare sulla riduzione dell'uso del maschile sovraesteso, ma altri tipi di bias di genere, come la stereotipizzazione, non sono stati indagati né affrontati in profondità.

Infine, come già sottolineato in diverse occasioni, è bene tenere a mente che questo progetto non aveva l'ambizione di affrontare in modo soddisfacente il problema del bias di genere in traduzione automatica, legato a ben più complesse dinamiche che si manifestano anche in altri ambiti (discusse nel Capitolo 1), ma, piuttosto, di esplorare le possibili soluzioni alle istanze linguistiche della comunità non binaria italofona, sempre più accolte in ambito scientifico ma comunque senza un vero riscontro dal punto di vista delle tecnologie disponibili (vedi ad esempio Piergentili et al., 2024). Il fatto di basarsi su sistemi esistenti, quindi, fa sì che queste soluzioni non siano adatte ad affrontare alla radice il problema del bias; tuttavia, possono essere utili come spunto di riflessione e come contributo alla ricerca in questo ambito.

In futuro, quindi, sarebbe auspicabile farsi carico della responsabilità degli strumenti di IA nella conferma ed esacerbazione dell'oppressione delle comunità marginalizzate, non solo in ambito linguistico. Questa consapevolezza non può che nascere dal dialogo con le comunità in questione, che, nel caso della traduzione e non solo, riflettono da decenni sul superamento di pratiche linguistiche dannose.

Riferimenti bibliografici [38]

Acanfora, F. (2022), 'Schwa: una questione identitaria', *Lingua italiana*. online: https://www.treccani.it/magazine/lingua_italiana/speciali/Schwa/1_Acanfora.html.

Alhafni, B., N. Habash, H. Bouamor, O. Obeid, S. Alrowili, D. AlZeer, K.M. Shnqiti, A. Elbakry, M. ElNokrashy, M. Gabr, A. Issam, A. Qaddoumi, V. Shanker, M. Zyate (2022), 'The Shared Task on Gender Rewriting', in *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp. 98-107. https://aclanthology.org/2022.wanlp-1.10.

Ashley, F. (2019), 'Les personnes non-binaires en français : une perspective concernée et militante', *H-France Salon*, 11(14), pp. 1-15. https://h-france.net/Salon/SalonVol11no14.5.Ashley.pdf.

Barocas, S. & A. D. Selbst (2016), 'Big Data's Disparate Impact', *California Law Review*, 104(3), pp. 671-732. https://www.ssrn.com/abstract=2477899.

Bentivogli, L., B. Savoldi, M. Negri, M.A. Di Gangi, R. Cattoni, M. Turchi (2020), 'Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6923-6933. https://www.aclweb.org/anthology/2020.acl-main.619.

Birhane, A. (2021), 'Algoritmic injustice: a relational ethics approach', *Patterns*, 2(2), pp. 1-9. https://www.cell.com/patterns/fulltext/S2666-3899(21)00015-5.

Blodgett, S. L., S. Barocas, H. Daumé III, H. Wallach (2020), 'Language (Technology) is Power: A Critical Survey of "Bias" in NLP', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454-5476. https://www.aclweb.org/anthology/2020.acl-main.485.

Bolukbasi, T., K. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai (2016), 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings', in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4356-4364. https://dl.acm.org/doi/10.5555/3157382.3157584.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herber-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S.Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei (2020), 'Language Models are Few-Shot Learners', in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 1877-1901. https://dl.acm.org/doi/10.5555/3495724.3495883.

[38] Tutti i collegamenti esterni riportati in questa sezione sono stati visitati l'ultima volta il 30 giugno 2024.

Cavallo, A., L. Lugli, M. Prearo (2021), *Cose, spiegate bene: Questioni di un certo genere*. Milano: Iperborea.

Comandini, G. (2021), 'Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. Ricerca sul corpus CoGeNSI', *Testo e Senso*, 23, pp. 43-64. https://testoesenso.it/index.php/testoesenso/article/view/524.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov (2020), 'Unsupervised Cross-lingual Representation Learning at Scale', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440-8451. https://www.aclweb.org/anthology/2020.acl-main.747.

Costa-jussà, M. & A. de Jorge (2020), 'Fine-tuning Neural Machine Translation on Gender-Balanced Datasets', in *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 26-34. https://aclanthology.org/2020.gebnlp-1.3/.

Crawford, K. (2017), 'The trouble with bias', Keynote speech, 31st International Conference on Neural Information Processing Systems. online: https://www.youtube.com/watch?v=ggzWIipKraM.

Currey, A., M. Nadejde, R. Pappagari, M. Mayer, S. Lauly, X. Niu, B. Hsu, G. Dinu (2022), 'MT-GenEval: A Counterfactual and Contextual Dataset for Evaluating Gender Accuracy in Machine Translation', in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4287–4299. https://aclanthology.org/2022.emnlp-main.288/.

Di Gennaro, G., A. Buonanno, A. Di Girolamo, A. Ospedale, F. Palmieri, G. Fedele (2020), 'An Analysis of Word2Vec for the Italian Language', arXiv preprint arXiv:2001.09332v1. online: http://arxiv.org/abs/2001.09332.

Dev, S., M. Monajatipoor, A. Ovalle, A. Subramonian, J. Phillips, K. Chang (2021), 'Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1968-1994. https://aclanthology.org/2021.emnlp-main.150.

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1: *Long and Short Papers*, pp. 4171-4186. https://aclanthology.org/N19-1423/.

Diesner-Mayer, T. & N. Seidel (2022), 'Supporting Gender-Neutral Writing in German', in *Proceedings of Mensch und Computer 2022*, pp. 509-512. https://dl.acm.org/doi/10.1145/3543758.3547566.

D'Ignazio, C. & L. F. Klein (2020), '"What gets counted counts"', in C. D'Ignazio & L. F. Klein (a cura di), *Data feminism*. MIT Press. https://direct.mit.edu/books/oa-monograph/4660/Data-Feminism.

Ferrara, E. (2024), 'Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies', *Sci*, 6(1), Articolo 3. https://www.mdpi.com/2413-4155/6/1/3.

Formato, F. & A. L. Somma (2023), 'Gender Inclusive Language in Italy: A Sociolinguistic Overview', *The journal of Mediterranean and European Linguistic Anthropology*, 5(1), pp. 22-40. https://jomela.pub/v5-i1-a3/.

Fusco, F. (2019), 'Il genere femminile tra norma e uso nella lingua italiana: come e perché', in S. Adamo, G. Zanfabro, E. Tignani Sava (a cura di), *Non esiste solo il maschile. Teorie e pratiche per un linguaggio non discriminatorio da un punto di vista di genere*, pp. 27-49. Edizioni Università di Trieste. https://www.openstarts.units.it/handle/10077/27061.

Gebru, T. (2020), 'Race and Gender', in M. D. Dubber, F. Pasquale, S. Das (a cura di), *The Oxford Handbook of Ethics of AI*. Oxford University Press. https://academic.oup.com/edited-volume/34287.

Gheno, V. (2020a), 'Ministra, portiera, architetta: le ricadute sociali, politiche e culturali dei nomi professionali femminili (prima parte)'. *Linguisticamente*. online: https://www.linguisticamente.org/nomi-femminili/.

Gheno, V. (2020b), 'La questione dei nomi delle professioni al femminile una volta per tutte'. *Valigia Blu*. online: https://www.valigiablu.it/professioni-nomi-femminili/.

Gheno, V. (2022), 'Schwa: Storia, motivi e obiettivi di una proposta'. *Lingua italiana*. online: https://www.treccani.it/magazine/lingua_italiana/speciali/Schwa/4_Gheno.html.

Gigerenzer, G. & H. Brighton (2009), 'Homo Heuristicus: Why Biased Minds Make Better Inferences', *Topics in Cognitive Science*, 1(1), pp. 107-143. https://onlinelibrary.wiley.com/doi/10.1111/j.1756-8765.2008.01006.x.

Giusti, G. (2022), 'Inclusività della lingua italiana, nella lingua italiana: come e perché', *Deportate, esuli, profughe*, 48, pp. 1-19. https://www.unive.it/pag/44259/?L=0.

Gonen, H. & K. Webster (2020), 'Automatically Identifying Gender Issues in Machine Translation using Perturbations', in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1991-1995. https://www.aclweb.org/anthology/2020.findings-emnlp.180.

Grandi, N. (2010), 'Genere', in *Enciclopedia dell'italiano*. Istituto della Enciclopedia Italiana. online: https://www.treccani.it/enciclopedia/genere_(Enciclopedia-dell'Italiano)/.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov (2018), 'Learning Word Vectors for 157 Languages', in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 3483-3487. https://aclanthology.org/L18-1550/.

Gromann, D., M. Lardelli, K. Spiel, S. Burtscher, L. Klausner, A. Mettinger, I. Miladinovic, S. Schefer-Wenzl, D. Duh, K. Bühn (2023), 'Participatory Research as a Path to Community-Informed, Gender-Fair Machine Translation', in *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pp. 49-59. https://aclanthology.org/2023.gitt-1.5/.

Guo, W. & A. Caliskan (2021), 'Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases', in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 122-133. https://dl.acm.org/doi/10.1145/3461702.3462536.

Gygax, P., U. Gabriel, O. Sarrasin, J. Oakhill, A. Garnham (2008), 'Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men', *Language and Cognitive Processes*, 23(3), pp. 464-485. http://www.tandfonline.com/doi/abs/10.1080/01690960701702035.

Habash, N., H. Bouamor, C. Chung (2019), 'Automatic Gender Identification and Reinflection in Arabic', in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 155-165. https://www.aclweb.org/anthology/W19-3822.

Hardt, M. (2014), 'How big data is unfair'. *Medium.* online: https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

Holtzman, A., J. Buys, L. Du, M. Forbes, Y. Choi (2020), 'The Curious Case of Neural Text *De*Generation', *8th International Conference on Learning Representations*. https://iclr.cc/virtual_2020/poster_rygGQyrFvH.html.

Jain, N., M. Popović, D. Groves, E. Vanmassenhove (2021), 'Generating Gender Augmented Data for NLP', in *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pp. 93-102. https://aclanthology.org/2021.gebnlp-1.11.

Kendall, E. (2023), 'gender binary', in *Encyclopaedia Britannica*. online: https://www.britannica.com/topic/gender-binary.

Knisely, K. A. (2020), 'Le français non-binaire: Linguistic forms used by non-binary speakers of French', *Foreign Language Annals*, 53(4), pp. 850-876. https://onlinelibrary.wiley.com/doi/10.1111/flan.12500.

Koehn, P. (2005), 'Europarl: A Parallel Corpus for Statistical Machine Translation', in *Proceedings of Machine Translation Summit X: Papers*, pp. 79-86. https://aclanthology.org/2005.mtsummit-papers.11/.

Kostikova, A., J. Daems, T. Lazarov (2023), 'How adaptive is adaptive machine translation, really? A gender-neutral language use case', in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 95-97. https://aclanthology.org/2023.gitt-1.9/.

Lardelli, M. & D. Gromann (2022), 'Gender-Fair (Machine) Translation', in *Proceedings of the New Trends in Translation and Technology conference*, pp. 166-177. online: https://www.researchgate.net/publication/369948882_Gender-Fair_Machine_Translation.

Lardelli, M. & D. Gromann (2023a), 'Gender-Fair Post-Editing: A Case Study Beyond the Binary', in *Proceedings of the 24th Annual Meeting of the European Association for Machine Translation*, pp. 251-260. https://aclanthology.org/2023.eamt-1.24/.

Lardelli, M. & D. Gromann (2023b), 'Translating non-binary coming-out reports: Gender-fair language strategies and use in news articles', *The Journal of Specialised Translation*, 40, pp. 213-240. https://jostrans.soap2.ch/issue40/issue40_toc.php.

Lardelli, M. (2023), 'Gender-fair translation: a case study beyond the binary', *Perspectives: Studies in Translation Theory and Practice*. online: https://www.tandfonline.com/doi/full/10.1080/0907676X.2023.2268654.

Leavy, S. (2018), 'Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning', in *Proceedings of the First International Workshop on Gender Equality in Software Engineering*, pp. 14-16. https://dl.acm.org/doi/10.1145/3195570.3195580.

Liesenfeld, A. & M. Dingemanse (2024), 'Rethinking open source generative AI: open washing and the EU AI Act', in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1774-1787. https://dl.acm.org/doi/10.1145/3630106.3659005.

López, Á. (2019), 'Tú, yo, elle y el lenguaje no binario', *La Linterna del Traductor*, 19, pp. 142-150. https://lalinternadeltraductor.org/pdf/lalinterna_n19.pdf.

López, Á. (2020), 'Cuando el lenguaje excluye: consideraciones sobre el lenguaje no binario indirecto', *Cuarenta naipes*, 3, pp. 295-312. https://fh.mdp.edu.ar/revistas/index.php/cuarentanaipes/article/view/4891.

Lu, K., P. Mardziel, F. Wu, P. Amancharla, A. Datta (2020), 'Gender Bias in Neural Natural Language Processing', in V. Nigam et al. (a cura di), *Logic, Language, and Security*, pp. 189-202. Springer. https://link.springer.com/chapter/10.1007/978-3-030-62077-6_14.

Ludbrook, G. (2022), 'From Gender-Neutral to Gender-Inclusive English. The Search for Gender-Fair Language', *Deportate, esuli, profughe*, 48, pp. 20-30. https://www.unive.it/pag/44259/?L=0.

Mikolov, T., K. Chen, G. Corrado, J. Dean (2013), 'Efficient Estimation of Word Representations in Vector Space', arXiv preprint arXiv:1301.3781. online: http://arxiv.org/abs/1301.3781.

Měchura, M. (2022), 'Introducing Fairslator: A machine translation bias removal tool', in *Proceedings of Translating and the Computer 44*, pp. 90-95. https://www.tradulex.com/varia/TC44-luxembourg2022.pdf.

Misiek, S. (2020), 'Misgendered in Translation?: Genderqueerness in Polish Translations of English-language Television Series', *Anglica*, 29(2), pp. 165-185. https://anglica-journal.com/resources/html/article/details?id=207730.

Nissen, U. K. (2002), 'Aspects of translating gender', *Linguistik Online*, 11(2), pp. 25-37. https://bop.unibe.ch/linguistik-online/article/view/914.

Non una di meno (2017), *Piano femminista contro la violenza maschile sulle donne e la violenza di genere*. online: https://nonunadimeno.files.wordpress.com/2017/11/abbiamo_un_piano.pdf.

Papadopoulos, B. (2022, a cura di), *Gender in Language Project*. online: www.genderinlanguage.com.

Papadopoulos, B., S. Cintrón, C. Hartman, D. Rusignuolo (2022), 'Italian', in B. Papadopoulos (a cura di), *Gender in Language Project*. online: www.genderinlanguage.com/italian/.

Papineni, K., S. Roukos, T. Ward, W. Zhu (2002), 'BLEU: a method for automatic evaluation of machine translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318. https://aclanthology.org/P02-1040/.

Patwardhan, N., S. Marrone, C. Sansone (2023), 'Transformers in the Real World: A Survey on NLP Applications', *Information*, 14(4), Articolo 242. https://www.mdpi.com/2078-2489/14/4/242.

Pennington, J., R. Socher, C. Manning (2014), 'GloVe: Global Vectors for Word Representations', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532-1543. http://aclweb.org/anthology/D14-1162.

Piergentili, A., D. Fucci, B. Savoldi, M. Negri, L. Bentivogli (2023a), 'Gender Neutralization for an Inclusive Machine Translation: from Theoretical Foundations to Open Challenges', in *Proceedings of the first Workshop on Gender-Inclusive Translation Technologies*, pp. 71-83. https://aclanthology.org/2023.gitt-1.7/.

Piergentili, A., B. Savoldi, D. Fucci, M. Negri, L. Bentivogli (2023b), 'Hi Guys or Hi Folks? Benchmarking Gender-Neutral Machine Translation with the GeNTE Corpus', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14124-14140. https://aclanthology.org/2023.emnlp-main.873/.

Piergentili, A., B. Savoldi, M. Negri, L. Bentivogli (2024), 'Enhancing Gender-Inclusive Machine Translation with Neomorphemes and Large Language Models', arXiv preprint arXiv:2405.08477. online: https://arxiv.org/abs/2405.08477.

Popović, M. (2015), 'CHRF: character n-gram F-score for automatic MT evaluation', in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392-395. https://aclanthology.org/W15-3049/.

Proto, D. (2021), *Ləi e tuttə. Differenti usi dello schwa nell'italiano contemporaneo*. Tesi di Laurea in Lettere, Università di Bologna. online: https://drive.google.com/file/d/1mrdSH0UXRehqTaeCVstYIbooS_4db_xW/view.

Pusterla, M. (2019), 'Parlare femminista: La lingua di Non una di meno', in S. Adamo, G. Zanfabro, E. Tignani Sava (a cura di), *Non esiste solo il maschile. Teorie e pratiche per un linguaggio non discriminatorio da un punto di vista di genere*, pp. 27-49. Edizioni Università di Trieste. https://www.openstarts.units.it/handle/10077/27061.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu (2020), 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', *Journal of Machine Learning Research*, 21, Articolo 140, pp. 1-67. https://jmlr.csail.mit.edu/papers/v21/.

Sánchez, E., P. Andrews, P. Stenetorp, M. Artetxe, M.R. Costa-jussà (2023), 'Gender-specific Machine Translation with Large Language Models', arXiv preprint arXiv:2309.03175. online: http://arxiv.org/abs/2309.03175.

Sarti, G. & M. Nissim (2022) 'IT5: Text-to-text Pretraining for Italian Language Understanding and Generation', in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 9422-9433. https://aclanthology.org/2024.lrec-main.823/.

Saunders, D. & B. Byrne (2020), 'Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7724-7736. https://www.aclweb.org/anthology/2020.acl-main.690.

Savoldi, B., A. Piergentili, D. Fucci, M. Negri, L. Bentivogli (2024), 'A Prompt Response to the Demand for Automatic Gender-Neutral Translation', in *Proceedings of the 18th Chapter of the Association for Computational Linguistics*, Volume 2: *Short papers*, pp. 256-267. https://aclanthology.org/2024.eacl-short.23/.

Savoldi, B., M. Gaido, L. Bentivogli, M. Negri, M. Turchi (2021), 'Gender Bias in Machine Translation', *Transactions of the Association for Computational Linguistics*, 9, pp. 845-874. https://aclanthology.org/2021.tacl-1.51/.

Sczesny, S., M. Formanowicz, F. Moser (2016), 'Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?', *Frontiers in Psychology*, 7, Articolo 25. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.00025/full.

Shah, D., H. A. Schwartz, D. Hovy (2020), 'Predictive Biases in Natural Language Processing: A Conceptual Framework and Overview', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5248-5264. https://aclanthology.org/2020.acl-main.468/.

Shelby, R., S. Rismani, K. Henne, AJ. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, G. Virk (2023), 'Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction', in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723-741. https://dl.acm.org/doi/10.1145/3600211.3604673.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul (2006), 'A Study of Translation Edit Rate with Targeted Human Annotation', in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pp. 223-231. https://aclanthology.org/2006.amta-papers.25/.

Stanovsky, G., N.A. Smith, L. Zettlemoyer (2019), 'Evaluating Gender Bias in Machine Translation', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679-1684. https://www.aclweb.org/anthology/P19-1164.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, D. Varga (2006), 'The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages', in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 2142-2147. https://aclanthology.org/L06-1196/.

Sulis, G. & V. Gheno (2022), 'The Debate on Language and Gender in Italy, from the Visibility of Women to Inclusive Language (1980s-2020s)', *The Italianist*, 42(1), pp. 153-183. https://www.tandfonline.com/doi/full/10.1080/02614340.2022.2125707.

Sun, T., A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K. Chang, W. Y. Wang (2019), 'Mitigating Gender Bias in Natural Language Processing: Literature Review', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630-1640. https://www.aclweb.org/anthology/P19-1159.

Sun, T., K. Webster, A. Shah, W. Wang, M. Johnson (2021), 'They, Them, Theirs: Rewriting with Gender-Neutral English', arXiv preprint arXiv:2102.06788. online: http://arxiv.org/abs/2102.06788.

Suresh, H. & J. Guttag (2021), 'A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle', in *Proceedings of the 1st ACM Conference on Equity and Access*

in Algorithms, Mechanisms, and Optimization, Articolo 17. https://dl.acm.org/doi/10.1145/3465416.3483305.

Telve, S. (2011), 'Accordo [prontuario]', in *Enciclopedia dell'Italiano*. Istituto della Enciclopedia Italiana. online: https://www.treccani.it/enciclopedia/accordo-prontuario_(Enciclopedia-dell'Italiano)/.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample (2023), 'LLaMA: Open and Efficient Foundation Language Models', arXiv preprint arXiv:2302.13971. online: http://arxiv.org/abs/2302.13971.

Vanmassenhove, E. (2024), 'Gender Bias in Machine Translation and the Era of Large Language Models', arXiv preprint arXiv:2401.10016. online: http://arxiv.org/abs/2401.10016.

Vanmassenhove, E., C. Emmery, D. Shterionov (2021), 'NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8940–8948. https://aclanthology.org/2021.emnlp-main.704/.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin (2017), 'Attention Is All You Need', in *Advances in Neural Information Processing Systems 30*, pp. 5998-6008. https://papers.nips.cc/paper_files/paper/2017.

Wallach, H. (2014), 'Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency'. *Medium.* online: https://hannawallach.medium.com/big-data-machine-learning-and-the-social-sciences-927a8e20460d.

Wang, W., J.T. Peter, H. Rosendahl, H. Ney (2016), 'CharacTer: Translation Edit Rate on Character Level', in *Proceedings of the First Conference on Machine Translation*, Volume 2: *Shared Task Papers*, pp. 505-510. http://aclweb.org/anthology/W16-2342.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus (2022), 'Emergent Abilities of Large Language Models', in *Transactions on Machine Learning Research*. https://openreview.net/pdf?id=yzkSU5zdwD.

Zanfabro, G. (2019), 'Translation Trouble: a proposito di Tyke Tiler, A. e George', in S. Adamo, G. Zanfabro, E. Tignani Sava (a cura di), *Non esiste solo il maschile. Teorie e pratiche per un linguaggio non discriminatorio da un punto di vista di genere*, pp. 121-146. Edizioni Università di Trieste. https://www.openstarts.units.it/handle/10077/27061.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, K. Chang (2018), 'Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods', in *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2: *Short papers*, pp. 15-20. http://aclweb.org/anthology/N18-2003.

Zmigrod, R., S. Mielke, H. Wallach, R. Cotterell (2019), 'Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651-1661. https://www.aclweb.org/anthology/P19-1161.

Linee guida istituzionali

Agenzia delle Entrate (2020), *Linee guida per l'uso di un linguaggio rispettoso delle differenze di genere*. online: https://www.agenziaentrate.gov.it/portale/documents/20143/1742359/Linee_guida_linguaggio_genere_2020.pdf/.

Ministero dell'Istruzione, dell'Università e della Ricerca (2018), *Linee guida per l'uso del genere nel linguaggio amministrativo del MIUR*. online: https://www.miur.gov.it/documents/20182/0/Linee_Guida_+per_l_uso_del_genere_nel_linguaggio_amministrativo_del_MIUR_2018.pdf/3c8dfbef-4dfd-475a-8a29-5adc0d7376d8?version=1.0.

Parlamento europeo (2018), *La neutralità di genere nel linguaggio usato al Parlamento europeo*. online: https://www.europarl.europa.eu/cmsdata/187102/GNL_Guidelines_IT-original.pdf.

Raus, R. (2015), 'Le questioni non risolte dal punto di vista linguistico', in S. Giorcelli, M. Spanò, R. Raus, M. Abouyaala, I. Catrano, V. Patti (a cura di), *Un approccio di genere al linguaggio amministrativo. Linee Guida - Una proposta del CUG e della Consigliera di Fiducia dell'Università degli Studi di Torino*, pp. 18-28. online: https://www.unito.it/sites/default/files/linee_guida_approccio_genere.pdf.

Robustelli, C. (2012), *Linee guida per l'uso del genere nel linguaggio amministrativo*. online: https://www.uniss.it/sites/default/files/documentazione/c._robustelli_linee_guida_uso_del_genere_nel_linguaggio_amministrativo.pdf.

Sabatini, A. (1987), *Raccomandazioni per un uso non sessista della lingua italiana*. online: https://www.funzionepubblica.gov.it/sites/funzionepubblica.gov.it/files/documenti/Normativa%20e%20Documentazione/Dossier%20Pari%20opportunit%C3%A0/linguaggio_non_sessista.pdf.

Thornton, A.M. (2020), *Per un uso della lingua italiana rispettoso dei generi*. online: https://www.univaq.it/include/utilities/blob.php?item=file&table=allegato&id=4925.

Università degli Studi di Padova (2017), *Generi e linguaggi: Linee guida per un linguaggio amministrativo e istituzionale attento alle differenze di genere*. online: https://www.unipd.it/sites/unipd.it/files/2017/Generi%20e%20linguaggi.pdf.

Università di Bologna (2020), *Linee guida per la visibilità di genere nella comunicazione istituzionale dell'Università di Bologna*. online: https://www.unibo.it/it/ateneo/chi-siamo/linee-guida-per-la-visibilita-del-genere-nella-comunicazione-istituzionale-universita-di-bologna.