Studi Italiani di Linguistica Teorica e Applicata, anno LII, 2023, numero 3

Rachele Raus Alma Mater Studiorum - *Università di Bologna*

DEEP LEARNING E TRADUZIONE INTRALINGUISTICA: RIFORMULARE I TESTI DELLA PUBBLICA AMMINISTRAZIONE IN MODO INCLUSIVO

Abstract

In this article, we present the experimentation of the E-MIMIC device (*Empowering Multilingual Inclusive Communication*), an application based on neural networks that aims at making administrative texts more inclusive. As result of the collaboration of data scientists and linguists, this device is giving particularly significant results for the Italian language of Italy and for the French language of France.

E-MIMIC was designed as an intralingual translation device. This allows us to exploit, by analogy, previous knowledge related to interlinguistic automatic translation and the problems it raises (e.g. the "smoothing" of linguistic diatopia and biases produced by this type of translation).

The novelty of this device lies on the one hand in the AI models adopted, which provide for supervised learning through the "humans in the analytics loop" system, and on the other in the linguistic-discursive criteria used to improve and optimize the deep learning of the networks.

From this point of view, E-MIMIC also becomes an experiment to propose new professional profiles in the translation and linguistic fields, straddling linguistics and computational or computer sciences.

Keywords: Intralingual Translation, Deep Learning, Language of Public Administration, Inclusive Language.

1. Introduzione

Le evoluzioni nell'apprendimento profondo, che permette a un dispositivo informatico basato su reti neurali di apprendere su grandi quantità di dati, imitando il cervello umano (Le Cun, 2019), hanno consentito all'intelligenza artificiale (IA) di svilupparsi recentemente in tutti i settori delle attività umane e di ricerca (cfr. Zouinar, 2020; von Braun *et al.*, 2021, ecc.). Nel settore linguisti-

co (Tavosanis, 2018), l'IA è stata impiegata nell'industria linguistica con finalità diverse e in particolare nel settore della traduzione automatica (Kohen, 2020).

Tuttavia, questo successo solleva anche diverse problematiche. In primo luogo, porta a riflettere sulla maggior presenza di dati e di strumenti informatici in inglese a discapito di altre lingue meno rappresentate (Kim *et al.*, 2019; Vetere, 2023). Inoltre, solleva la questione dell'impatto che alcune lingue veicolari o controllate (Ryan, 2009), a partire dalle quali le reti neurali apprendono per il tramite dell'uso di corpora multilingui messi a disposizione dalle organizzazioni internazionali o da multinazionali, hanno sulla diatopia delle lingue nazionali (Caliskan *et al.*, 2017; Rastier, 2021), contribuendo a generare disparità linguistiche (lingua nazionale "fagocitata" dall'omologa internazionale). In ultimo, ma non per importanza, la generazione e la diffusione massiva di *bias* (Bartoletti, 2020), ovvero di veri e propri errori che distorcono la realtà creando pregiudizi di genere e non solo nella fase di generazione del testo da parte di un dispositivo basato su reti neurali.

Partendo da un caso sperimentale, cioè dall'applicativo E-MIMIC (*Empowering Multilingual Inclusive Communication*) basato su reti neurali capace di riformulare i testi amministrativi in chiave inclusiva, intendiamo valutare se l'apporto della linguistica in fase di *pre-editing* nell'apprendimento profondo possa contribuire a preservare la diatopia linguistica e a realizzare più generalmente una comunicazione inclusiva.

Il dispositivo, attualmente sviluppato quasi integralmente nella versione italiana (lingua italiana d'Italia)[1], e che è in fase di sviluppo per il francese di Francia e lo spagnolo di Spagna, ha permesso infatti di ottenere risultati promettenti.

In questa sede, dopo aver presentato la metodologia su cui si basa l'applicativo, riportiamo alcune riflessioni in merito ai test condotti sulla versione francese di E-MIMIC per mostrare in che modo il paradigma alla base permetta di salvaguardare la diatopia linguistica; successivamente, faremo il caso dell'applicativo italiano per valutare quanto i criteri adottati da E-MIMIC consentano d'individuare con un alto livello di accuratezza gli elementi che producono non inclusione ed eventuali *bias* e che quindi necessitano di essere riformulati. Trarremo quindi delle conclusioni in merito all'innovatività del dispositivo.

2. La metodologia di E-Mimic

Frutto della collaborazione di un'équipe del Politecnico di Torino e dell'Università di Bologna[2], l'applicativo E-MIMIC è stato progettato come un dispositivo di traduzione intralinguistica, sfruttando per analogia le conoscenze pregresse legate alla traduzione automatica interlinguistica e alle problemati-

che da essa sollevata, in particolare riguardo alla diatopia linguistica – a causa dell'adeguamento delle lingue nazionali alle lingue veicolari e/o controllate utilizzate nella comunicazione internazionali –, e riguardo al rischio di *bias* sociolinguistici, come la perdita dell'accordo di genere o di tratti sia morfosintattici sia semantici d'inclusività (Marzi, 2021).

Il modello informatico adottato da E-MIMIC prevede l'addestramento sorvegliato da personale umano per testare e migliorare le *performance* traduttive. Proprio su questi due elementi di attenzione (diatopia e *bias*) legati all'intervento di personale competente si basa l'innovazione maggiore di E-MIMIC.

In merito alle fasi di test, il sistema adottato è stato quello definito *humans in the analitycs loop* (Cerquitelli, Raus, Molino, in corso di pubblicazione), nel quale l'essere umano è coinvolto nel ciclo di analisi e si utilizzano metriche di valutazione basate sulla modellizzazione del giudizio umano, in modo da quantificare in modo corretto e imparziale la solidità e la completezza delle operazioni linguistiche (*tasks*) che sono state svolte con metodi di apprendimento profondo.

Riguardo al miglioramento delle *performance* del dispositivo, in particolare rispetto alla qualità della traduzione in termini di preservazione della diatopia linguistica, con conseguente miglioramento della leggibilità del testo, e resa dell'inclusività, il personale linguistico coinvolto nel progetto ha operato alcune scelte previsionali al fine di orientare al meglio l'addestramento della macchina già in fase di *pre-editing*, ovvero di preparazione dei corpora da utilizzare e da annotare per l'addestramento delle reti neurali. Questa scelta è piuttosto rara nel panorama dell'industria linguistica supportata da algoritmi d'IA, dal momento che normalmente si predilige il *post-editing* – per lo più automatizzato o semiautomatizzato – funzionalmente al miglioramento della qualità del testo tradotto (O'Brien *et al.*, 2018; Monti, 2019: 13), scelta che non sempre si rivela efficace (Toral, 2019).

Precisiamo che al momento, per le tre lingue considerate da E-MIMIC (IT, FR, ES), non esistono applicativi di riformulazione inclusiva dei testi amministrativi, considerando peraltro che per inclusione non intendiamo solo quella di genere ma di qualsiasi tipo di "minoranza" (Raus *et al.*, 2022). Solamente per lo spagnolo è disponibile online un prototipo meno avanzato di quello che stiamo elaborando[3]. La riformulazione di testo inclusivo tramite *chatbot* come *ChatGPT*[4] produce risultati non proprio soddisfacenti. In un test condotto il 10 giugno 2023, il risultato per la richiesta di riformulare un testo italiano non rispettoso dell'accordo di genere (utilizzo del maschile di professione in presenza di nome propri di genere femminile, utilizzo del maschile generico sovraesteso, ecc.) ha dato luogo dapprima a una riformulazione identica e poi, a seguito della richiesta di rispondere nuovamente, alla riformulazione delle forme maschili con la doppia forma "maschile/femminile", che ha reso il testo illeggibile.

Per effettuare tale tipo di operazione, l'architettura informatica dell'applicativo E-MIMIC, nelle tre versioni linguistiche proposte, integra due modelli di *deep learning* in cascata. Il primo effettua il task di classificazione della frase "inclusiva" o "non inclusiva". Il secondo propone una riformulazione inclusiva ove necessario. (Cerquitelli *et al.*, 2023).

In merito alla riformulazione, e perciò nella traduzione intralinguistica del testo iniziale in chiave inclusiva, essa avviene addestrando il dispositivo su criteri d'inclusione che rispettano quanto riportato per le lingue selezionate dal Segretariato del Consiglio dell'Unione europea in merito alla *Comunicazione inclusiva* (2018) e che tengono conto delle indicazioni degli enti e delle istituzioni che, su basi più o meno volontaristiche, si occupano di normare la lingua a livello nazionale[5].

Il preaddestramento avviene anzitutto su dati, o meglio corpora[6], che presenteremo successivamente per le versioni linguistiche prese in considerazione in questa sede, poi specializzando il dispositivo in due modi:

1. tramite l'utilizzo di dati sintetici predisposti dal personale linguistico per le varie lingue;
2. tramite materiale autentico opportunamente annotato[7].

Dal punto di vista informatico, le componenti di E-MIMIC possono essere rappresentate come in Figura 1, in cui sono facilmente reperibili:

- le forme di preaddestramento (*pre-training*) sui corpora iniziali non annotati;
- le forme di addestramento per effettuare un *task* (*fine-tuning*), cioè un compito specifico, sui paradigmi linguistici, ossia sui dati sintetici creati dal personale linguistico, e sui dati annotati[8];
- il compito o i compiti (*tasks*) attesi alla fine, riportati a destra della figura, ovvero, nel nostro caso, la classificazione delle frasi in inclusive e non inclusive, la loro riformulazione inclusiva e la scrittura inclusiva finale.

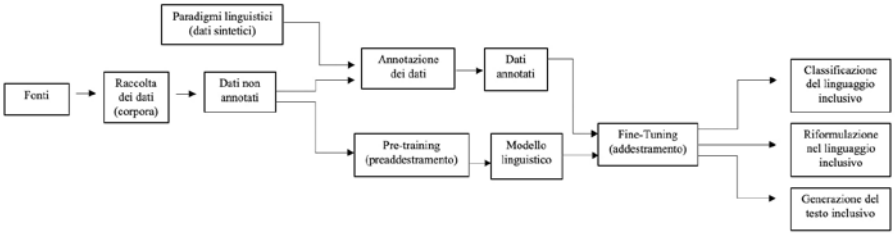

Figura 1: Rappresentazione delle componenti di E-MIMIC (fonte: Attanasio *et al.* 2021: 4231[33])

3. La versione francese di E-Mimic: la selezione dei corpora per salvaguardare la diatopia linguistica

Nella traduzione interlinguistica, le reti neurali dei dispositivi più noti di traduzione automatica (es. *DeepL*, *Google Translate*…)[9] tendono, soprattutto durante le prime fasi di preaddestramento, a privilegiare i corpora delle organizzazioni internazionali perché fonti "autorevoli" (International Organization for Standardization, 2009: 7) di grandi banche di dati multilingui[10]. Ciò ha comportato problemi legati all'apprendimento linguistico sia per ciò che concerne l'inclusività, i corpora istituzionali internazionali privilegiano infatti il maschile generico che impedisce ai dispositivi di traduzione automatica d'imparare l'accordo al femminile quando si tratta di persona (Marzi, 2021; Raus *et al.*, 2022), sia rispetto alla diatopia linguistica, le lingue veicolari e/o controllate usate nei contesti internazionali, finendo per avere maggior peso statisticamente rispetto alle lingue equivalenti nazionali in fase di addestramento delle reti neurali. I dispositivi di traduzione neurale automatica, in altre parole, finirebbero per imparare le lingue veicolari internazionali rispetto alle loro omologhe nazionali, con rare eccezioni[11]. Non esistono studi che mostrino fino a che punto queste lingue siano utilizzate e quanto impattino sulle lingue nazionali omologhe, dal momento che generano testo ulteriore sul web proprio tramite i dispositivi di traduzione automatica. Quello che, però, sta emergendo è la presenza di una "convergenza lessicale" (Hermand, 2014) tra lingue romanze e germaniche, a partire dall'utilizzo di termini che normalmente derivano dall'inglese veicolare.

In relazione a tale questione, la versione francese di E-MIMIC è preaddestrata su corpora generici indifferenziati tratti dal web e impara quindi il francese senza tener conto della diatopia. In tal senso, il camemBERT-base[12] che è stato utilizzato, è addestrato sui dati del 2018 del progetto OSCAR[13], ovvero da corpora *Common Crawl*[14] scaricati da Internet e ad accesso libero (Ortiz Suárez *et al.*, 2019). OSCAR, infatti, è una piattaforma di risorse e di set di dati multilingui utili proprio per addestrare le reti neurali, che in questo modo imparano le lingue naturali. Tuttavia, imparando sulla base del criterio statistico di maggior frequenza, non siamo in grado di capire se, una volta preaddestrate su questi corpora, le reti neurali di E-MIMIC siano capaci di privilegiare le forme lessicali del francese di Francia o quelle delle varianti francesi utilizzate a livello internazionale, spesso usato proprio per i siti web multilingui dai quali i corpora monolingui di OSCAR sono scaricati.

Per questo motivo, abbiamo provveduto a specializzare E-MIMIC su un corpus di dati nazionali, selezionando testi tratti dal Parlamento francese[15].

Il test che presentiamo verte a verificare la presenza di elementi dei francesi utilizzati a livello internazionale, che sono veicolari rispetto alla comunicazione istituzionale internazionale e *in primis* rispetto alla comunicazione

delle organizzazioni internazionali[16]. Tali elementi variano tra canadismi, calcati sulle forme inglesi equivalenti anche rispetto al contenuto concettuale (Paquet-Gauthier, 2018), e concetti generalmente iperonimici tipici delle varianti linguistiche utilizzate dalle organizzazioni internazionali, la cui finalità è la produzione di testi giuridici sovrannazionali che saranno trasposti e adattati nelle lingue nazionali[17]. In questa sede, ci occuperemo dei secondi, quindi di termini generalmente usati a livello del francese internazionale ma meno usati, almeno ad oggi, nel francese nazionale.

Il test consiste nel valutare l'apprendimento di E-MIMIC della lingua francese, comparando la situazione dopo il preaddestramento su OSCAR a quella successiva ottenuta con la specializzazione sui corpora nazionali, così da vedere quanto la specializzazione possa incidere sulla presenza dei termini delle varianti veicolari del francese, dato che i corpora nazionali sono esigui rispetto alla quantità di dati iniziali tratti da OSCAR.

Per far ciò, è stata predisposta una demo di *encoding*[18] che permette alle reti neurali di "prevedere" all'interno delle frasi l'uso di parole apprese su base statistica dai corpora usati per il preaddestramento (OSCAR e poi Parlamento francese). L'applicativo si presenta con una maschera utente (Figura 2) nella quale possono essere inserite frasi predisposte dal linguista per testare le reti neurali.

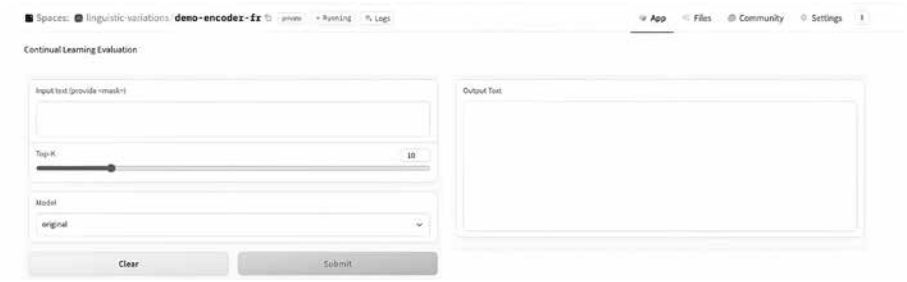

Figura 2: Maschera utente della demo di *encoding* utilizzata per testare la versione francese di E-MIMIC

Nell'ambito di queste frasi, la presenza di *<mask>* nel testo digitato come input (*Input text*) serve a dare la possibilità d'inserire la parola o il sintagma che il dispositivo prevede come il più idoneo, su basi statistiche, in quel contesto frastico, a seguito dell'apprendimento (preaddestramento o addestramento). Tale parola comparirà nella frase come testo di output (*Output text*). La demo permette anche di selezionare:

- il modello (*Model*) che si vuole testare, nello specifico il modello "originale" preaddestrato su OSCAR (*original*), oppure quello specializzato sui corpora del Parlamento francese (*trained*);
- il numero di risultati che vogliamo visualizzare (indicatore "*Top K*"). Abbiamo selezionato i primi 10 risultati più significativi dal punto di vista statistico.

Riportiamo di seguito due delle dieci frasi sottoposte al software con risultati analoghi, per illustrare quanto vogliamo dimostrare:

(1) La violence <*mask*> est un crime que les femmes ne peuvent plus tolérer.
(2) L'égalité des <*mask*> est un principe fondamental.

Abbiamo cercato di ricreare contesti che potessero indurre la comparsa di parole specifiche, ovvero in (1) l'aggettivo "*domestique*" e, in (2), il termine "*genre*", sapendo che a livello internazionale, i sintagmi "*violence domestique*" (Nugara, 2011a, 2011b) ed "égalité des *genres*" (Raus, 2013b, 2020) sono frequenti, diversamente dal contesto nazionale che privilegia adattamenti (es. *violence conjugale*, égalité des femmes[19]…). Per operare tale scelta, ci siamo basate sulla nozione di "profilo lessico-discorsivo" introdotta da Marie Veniard (2021: 109), per la quale si sceglie l'utilizzo di una determinata parola contestuale rispetto a un'altra per ragioni sintattiche, semantiche, sintagmatiche, testuali, enunciative e interdiscorsive.

I risultati del test, condotto in data 19 giugno 2023, sono riportati nella Tabella 1:

Modello *original* della demo	**Modello *trained* della demo**
Risultati per l'esempio (1)	*Risultati per l'esempio (1)*
0.541 - conjugale	0.404 - conjugale
0.177 - sexuelle	0.140 - sexuelle
0.057 - masculine	0.062 - physique
0.028 - féminine	0.052 - systémique
0.016 - **domestique**	0.046 - masculine
0.013 - qui	0.037 - quotidienne
0.006 - extrême	0.033 - sex
0.006 - physique	0.030 - verbale
0.006 - corporelle	0.023 - qui
0.005 - sociale	0.017 - féminine
Risultati per l'esempio (2)	*Risultati per l'esempio (2)*
0.495 - chances	0.248 - femmes

0.167 - droits	0.239 - chances
0.092 - **genres**	0.178 - droits
0.018 - citoyens	0.030 - hommes
0.016 - sexe	0.021 - citoyens
0.013 - revenus	0.018 - enfants
0.010 - races	0.010 - lois
0.010 - traitements	0.008 - forces
0.009 - personnes	0.008 - filles
0.008 - territoires	0.007 - **genres**

Tabella 1: Risultati del test condotto sulla versione francese di E-MIMIC rispetto all'utilizzo di termini internazionali (a sinistra) o nazionali (a destra) in frasi inserite nella demo per testare le reti neurali di E-MIMIC preaddestrate su corpora generici (OSCAR, modello "*original*" a sinistra) e poi specializzate sul corpus nazionale (modello "*trained*" a destra)

Il numero a sinistra delle parole in tabella indica la percentuale di possibilità di occorrenza del termine (es. 0.541 equivale al 54,1% di possibilità) stando alla previsione condotta dalle reti neurali sulla base dei calcoli statistici desunti durante le fasi di preaddestramento su corpora internazionali generici nei casi a sinistra della tabella, e di specializzazione successiva su corpus nazionale del Parlamento francese per i casi a destra.

Tra gli elementi che potremmo evincere dalla tabella, è chiaro che l'addestramento su quantità di dati non ingenti ma specializzati impatta sulla conoscenza linguistica delle reti neurali, al punto che si possono segnalare molte differenze tra i risultati presenti nelle due colonne.

Ai fini della nostra analisi, ci limitiamo a segnalare alcune variazioni specifiche. Rispetto all'esempio (1), sebbene la prima previsione resti la stessa ("*conjugale*"), è interessante notare che la variante internazionale "*domestique*" sparisce a seguito della specializzazione delle reti neurali sul corpus nazionale. La stessa cosa avviene per l'esempio (2), dato che il termine "*genre*" diventa raro a livello nazionale, dove è "*femme*" a essere privilegiato.

Viene confermata la tendenza del lessico internazionale all'iperonimia, come vediamo dagli attanti eufemizzati tramite metonimia ("*genres, sexe, races*") o generalizzati con iperonimi ("*citoyens, personnes*") nella versione preaddestrata su corpora internazionali (OSCAR), rispetto alle categorie specifiche ("*femmes, hommes, enfants, filles*") presenti nel modello specializzato sul corpus nazionale.

Il test condotto mostra come dati quantitativamente non significativi rispetto a quelli generici iniziali del preaddestramento possono impattare sensibilmente sui secondi laddove siano opportunamente selezionati rispetto alla tematica oggetto della ricerca. I sintagmi "*violence* + X" ed "égalité des + X" sono, infatti, sufficientemente frequenti nel corpus del Parlamento francese al punto che la loro frequenza relativa al corpus specifico diventa significativa, per-

mettendo di avere un impatto significativo sul preaddestramento generico. Ciò consente una maggior garanzia di rispetto della variante diatopica del francese di Francia rispetto al modello originale preaddestrato su corpora scaricati dal web, più generici e caratterizzati maggiormente dalla presenza di elementi tipici dei francesi internazionali.

Questo primo risultato è fondamentale rispetto alla tutela del multilinguismo.

4. La versione italiana di E-Mimic: paradigmi linguistici e criteri discorsivi per produrre testo inclusivo

La versione italiana di E-MIMIC permette di effettuare altri test significativi ai fini della riformulazione di un testo in chiave inclusiva. Per questa versione, abbiamo usato il modello BERT preaddestrato per l'italiano su dati messi a disposizione dalla Bavarian State Library, tratti da *Wikipedia* del 2019 e dal corpus OPUS (Tiedemann, 2012)[20].

L'addestramento funzionale a effettuare dei *task* specifici (*fine-tuning*), che sono tre nel nostro caso (classificare le frasi come inclusive o meno, la loro riformulazione inclusiva e la scrittura inclusiva finale), è stato effettuato su materiale autentico tratto da documenti messi a disposizione da Città metropolitana di Torino o pubblicati sul web da Università ed enti pubblici nazionali[21]. A differenza della versione francese, per quella italiana abbiamo già avviato l'addestramento per rinforzo tramite l'inserimento di dati sintetici (paradigmi linguistici) e la specializzazione del modello tramite l'annotazione di materiali autentici effettuata dal team linguistico[22]. Questi due interventi vengono dettagliati nei sottoparagrafi seguenti.

4.1 Dati sintetici utili all'apprendimento per rinforzo: i paradigmi linguistici

In merito ai paradigmi linguistici per l'addestramento per rinforzo, l'attività di riformulazione di testi in chiave inclusiva da noi condotta dal 2016 presso enti piemontesi[23] ha portato alla redazione di diverse guide relative alle attività di riformulazione[24] che, assieme a vari testi della letteratura di settore[25], hanno funto da ispirazione per i paradigmi usati per il progetto E-MIMIC. I dati sintetici prodotti sono stati raccolti in fogli excel che riportano *template* sotto forma di frasi, nelle quali dei <*seed*> permettono le riformulazioni sull'asse paradigmatico. Il *seed* consente di saturare la frase con formulazioni non inclusive e di riformularle con le corrispettive inclusive che permettono alle reti neurali di astrarre la regola virtuosa d'inclusione.

In Tabella 2, riportiamo l'esempio di dieci *template*[26] relativi a riformulazioni inclusive rispetto al genere, che utilizzano elementi "neutri" ("persona", "personale", "componente", astrazioni metonomiche varie).

Template	**Non Inclusivo**	**Inclusivo**
Occorre fornire i dati [seed]	dei clienti	della clientela
È necessario conoscere in anticipo il numero [seed]	degli studenti	della componente studentesca
Sono stati analizzati i riscontri [seed]	dei docenti	del personale docente
Il parere [seed] deve essere acquisito entro l'anno	degli insegnanti	del personale insegnante
Il riscontro [seed] è necessario	dei direttori tecnici	della direzione tecnica
L'avviso [seed] è fortemente richiesto	dei dirigenti	del personale dirigente
Occorre richiedere la firma [seed]	dei tecnici	del personale tecnico-amministrativo
Richiedere la firma [seed] è opportuno	dei lettori	del personale con contratto CEL
Sostenere le proposte [seed] aiuta a migliorare le performance	degli esperti linguistici	del personale con contratto CEL
Acquisire il parere [seed] aiuta a migliorare le performance	dei dipendenti	del personale dipendente
	dei richiedenti	delle persone richiedenti
	dei docenti	delle persone che svolgono docenza
	dei sottoposti a indagine	delle persone sottoposte a indagine
	dei verbalizzanti	delle persone verbalizzanti
	dei candidati interessati	delle persone candidate
	degli interessati	delle persone interessate
	dei vaccinati	delle persone vaccinate
	dei vigili	delle persone della polizia municipale
	dei tesserati	delle persone iscritte con tessera
	dei laureati	delle persone laureate
	dei dottori di ricerca	delle persone addottorate
	dei delegati	delle persone delegate
	dei malati	delle persone malate
	dei manifestanti	delle persone manifestanti

Tabella 2: Esempio di dati sintetici (paradigmi linguistici) utilizzati per l'apprendimento per rinforzo di E-MIMIC

Per questi test è stato usato un modello preaddestrato, specializzandolo sul 90% dei *template*, mentre il restante 10% è stato usato per valutarne l'accuratezza su dati mai visti prima dal dispositivo. In questo caso specifico, i *template* sono stati divisi in modo tale che il modello non vedesse il *template* di base (ad esempio: [MASK] vanno al mare) durante l'addestramento se questo stesso *template* fosse stato usato per il test.

I test condotti sui paradigmi[27] dimostrano che, addestrando il modello solo su dati sintetici, si raggiunge un grado di accuratezza (*accuracy*) dell'85%, ovvero che il modello impara a riconoscere frasi inclusive con quel risultato di *accuracy*. Questo risultato mostra la capacità delle moderne architetture neurali di modellare efficacemente un linguaggio inclusivo già solo utilizzando dei paradigmi linguistici.

I paradigmi sono serviti non solo per il *task* di classificazione dell'inclusività ma anche per quello di traduzione intralinguistica in fase di riformulazione inclusiva. Un test condotto sempre da personale umano ha mirato a comparare le *performance* del modello utilizzato per il *task* riformulazione, addestrato solo su dati annotati nel primo caso e addestrato sugli stessi dati ma aggiungendo il rinforzo dei paradigmi linguistici nel secondo (Tabella 3).

Modello usato	**Riformulazioni corrette**	**Riformulazioni parzialmente corrette**	**Riformulazioni non corrette**
IT5 senza dati sintetici	64,76%	15,71	19,52
IT5 con dati sintetici	*69,52%*	*17,14*	*13,33*

Tabella 3: Risultati dei test condotti sul *task* di riformulazione inclusiva con o senza l'utilizzo di dati sintetici

Nella tabella, abbiamo segnato in corsivo i risultati migliori rispetto all'inclusione, in modo da cogliere quanto l'uso dei paradigmi si sia rivelato utile.

Con questo test, inoltre, è stato possibile individuare elementi nocivi all'apprendimento, come ad esempio il paradigma concernente la riformulazione dal plurale non inclusivo al singolare inclusivo di parole epicene inizianti per vocale (es. "gli atleti" *vs* "l'atleta"; "gli asceti" *vs* "l'asceta", ecc.). In questo caso, il dispositivo ha astratto la regola della riformulazione generale dell'epiceno senza tener conto dell'elemento vocalico iniziale. Il plurale non inclusivo "le visite guidate sono pensate per *i turisti*" è stato perciò riformulato con "le visite guidate sono pensate per *il turista*", che però resta non inclusivo dato che la consonante iniziale dell'epiceno non permette l'elisione inclusiva dell'articolo determinativo.

4.2 Criteri discorsivi per migliorare le performance di apprendimento

Per quanto concerne i criteri di annotazione, sono stati ispirati dall'analisi francese del discorso[28]. Chi deve annotare le frasi ottenute dallo *splitting* in frasi del documento iniziale ha a disposizione una maschera (Figura 3), creata sulla piattaforma *Label Studio*[29], per indicare quando la frase è inclusiva o meno, ma soprattutto il tipo di sequenza discorsiva e l'utilizzo o meno di linguaggio chiaro.

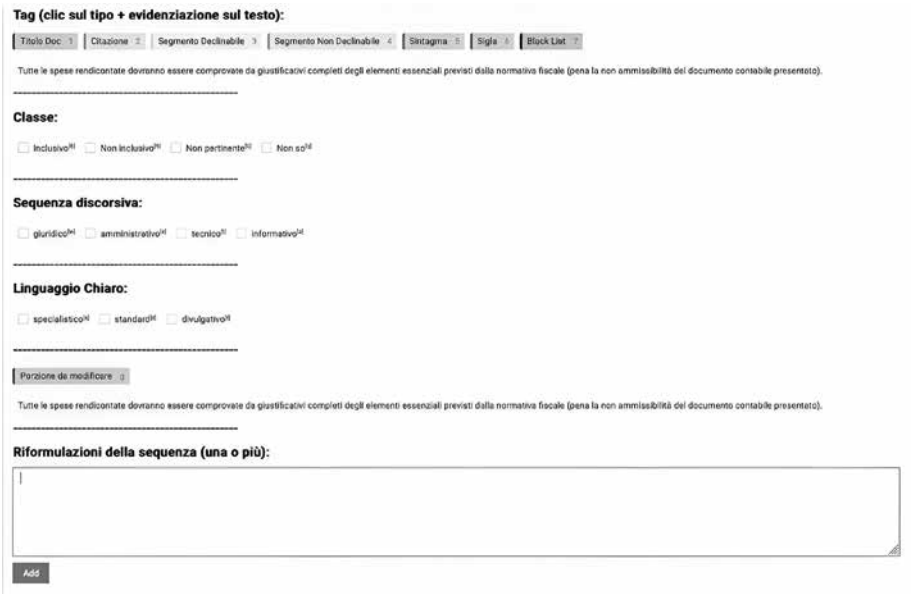

Figura 3: Maschera di annotazione nella versione italiana di E-MIMIC

In alto, sono presenti dei *Tag* che permettono di annotare elementi della frase come declinabili o meno[30], come facenti parte di sintagmi o titoli o citazioni, la presenza di sigle ed elementi opachi dal punto di vista dell'accessibilità del testo ed elementi che possono produrre *bias* semantici e che quindi sono inseriti in una *black list*. La maschera permette poi d'inserire la riformulazione inclusiva di frasi iniziali non inclusive e viceversa.

I test condotti sulle schede annotate hanno anzitutto permesso di ottenere ottimi punteggi (95,43 di accuratezza) rispetto alla capacità del dispositivo di classificare le frasi come inclusive, non inclusive o non pertinenti (versione *Single-task*).

Ulteriori test sono stati effettuati[31] creando una versione *Multi-task* del modello che, rispetto a quella *Single-task*, aggiunge due *task* specifici, uno in

merito al riconoscimento della sequenza discorsiva (*discourse classification*) e un secondo sull'identificazione delle porzioni evidenziate tramite i *tag*, come sintagma, *black list*, ecc. (*specific portion identification*). Abbiamo, quindi, comparato le *performance* dei due modelli (*Single-task* e *Multi-task*) su un sottocorpus di 21 testi amministrativi.

Riportiamo i risultati del test in Tabella 4:

Versione single-task	**Versione Multi-Task**		
Task *inclusive_classification*	Task (1) *inclusive_classification*	Task (2) *discourse_classification*	Task (3) *specific_portion_identification*
Training: 762	Training: 762	Training: 300	Training: 266
Validation: 95	Validation: 95	Validation: 37	Validation: 33
Test: 96	Test: 96	Test: 38	Test: 34
Il modello raggiunge un'accuratezza dell'**81.25%** sul test set	Sul task di classificazione di inclusività il modello raggiunge un'accuratezza dell'**85.41%** sul test set.		

Tabella 4: Stima sul numero di frasi per i vari *task* delle due versioni (*Single* e *Multi-task*) con i risultati rispetto all'accuratezza della capacità dei modelli di classificare la frase come inclusiva o non inclusiva

Le diciture *Training*, *Validation* e *Test* si riferiscono rispettivamente ai dati usati per allenare, scegliere il modello migliore e calcolare le metriche. Il test dimostra che l'ausilio di criteri discorsivi migliora l'accuratezza del modello, che quindi diventa più capace di classificare la frase come inclusiva, non inclusiva o non pertinente. Segnaliamo ancora margini di miglioramento delle *performance* che potrebbero essere ottenute da un lato, con l'aggiunta di sequenze discorsive più variegate e dall'altro, con l'incremento della numerosità dei dati concernenti i *tag* di annotazione.

Detto ciò, i test condotti in questo e nel precedente sottoparagrafo dimostrano che la generazione di paradigmi linguistici e l'utilizzo di criteri discorsivi per l'annotazione dei corpora migliorano sensibilmente le *performance* del dispositivo, agevolando l'individuazione di segmenti non inclusivi e la loro riformulazione in chiave inclusiva.

5. Conclusioni

I test condotti sulle versioni italiana e francese di E-MIMIC hanno permesso di dimostrare che, diversamente dai *large language models* d'intelligenza artificiale, che stanno ormai divenendo lo standard attuale nei prodotti di generazione di testo e nella riformulazione o traduzione di testi (cfr., ad esempio,

ChatGPT, che oltre a generare testi, ne traduce basandosi sui traduttori automatici più noti o ne riformula su richiesta dell'utente) e che sono modelli che necessitano di grandi dati e si fondano sull'apprendimento automatico per lo più non sorvegliato da personale umano con forti rischi ai danni del multilinguismo e dell'inclusione nel linguaggio, l'alternativa di utilizzare quantità di dati più esigue, ma opportunamente selezionate o annotate da personale esperto in fase di *pre-editing* e non a valle del processo di traduzione, permette di risolvere le questioni legate da un lato al non rispetto della diatopia linguistica, come abbiamo avuto modo di vedere in merito al caso francese, e dall'altro rispetto all'individuazione di segmenti non inclusivi, che peraltro possono produrre *bias*, come abbiamo visto nella versione italiana dove il sistema è messo in condizione di riconoscere tali segmenti per poi riformularli. La presenza dell'intervento umano previsto, inoltre, durante tutte le fasi di test permette d'inserire correttivi eventuali e di seguire quanto il dispositivo è in grado di fare e come.

Infine, tale alternativa consente al personale esperto nelle scienze del linguaggio di far parte di equipe di lavoro interdisciplinari, lasciando configurare la possibile creazione di nuovi profili professionali transdisciplinari, all'intersezione tra le scienze del linguaggio e quelle computazionali e informatiche, come attualmente richiesto dal mercato del lavoro (Ferraresi *et al.*, 2021; Miličević Petrović *et al.*, 2021)[32].

NOTE

[1] La versione italiana, in fase di finalizzazione, è stata denominata "*Inclusively*".

[2] Ringraziamo Tania Cerquitelli (coordinatrice), Luca Cagliero, Moreno La Quatra, Salvatore Greco e Micaela Tonti della collaborazione. Cfr. https://dbdmg.polito.it/e-mimic/index.php/research-team.

[3] https://huggingface.co/spaces/hackathon-pln-es/es_nlp_gender_neutralizer presentato all'Hackaton di PLN organizzato dal gruppo Somos NLP (https://somosnlp.org/hackathon). Per una presentazione del gruppo Somos, cfr. https://somosnlp.org/.

[4] https://openai.com/blog/chatgpt.

[5] Intendiamo l'Accademia della Crusca per l'Italia, l'*Académie française* per la Francia e la *Real Academia Española* in Spagna. Per la trattazione di tali criteri per l'italiano e il francese, cfr. anche Attanasio *et al.* (2021) e Raus *et al.* (2022).

[6] Seppur nella consapevolezza che i due concetti, l'uno informatico (dati) e l'altro linguistico (corpora), non possano sovrapporsi (cfr. in tal senso Rastier, 2021), in questo contesto li useremo indifferentemente.

[7] In merito ai team di annotazione per le tre lingue, cfr. https://www.jmcoe.unito.it/content/e-mimic-empowering-multilingual-inclusive-communication.

[8] Va precisato che sia per l'annotazione sia per l'addestramento l'unità di base scelta è la frase.

[9] Consultabili ai link: https://www.deepl.com/translator e https://translate.google.it.

[10] Le fonti per l'addestramento di *DeepL* sono le stesse utilizzate per il software di concordanze *Linguee* e, come facilmente reperibile dai risultati di ricerche bilingui in questo dizionario, sono essenzialmente siti di organizzazioni internazionali (ONU, UE, ecc.). *Google Translate* è stato addestrato sulla banca dati dell'ONU (Organizzazione delle Nazioni Unite, 2017: 120).

[11] Il traduttore francese *Reverso* (https://www.reverso.net/), ad esempio, è più "virtuoso" dei concorrenti rispetto alla diatopia linguistica perché si basa anche su fonti nazionali.

[12] L'espressione "camemBERT" nel *Natural Language Processing* rinvia al modello linguistico BERT francese. Il modello "base" consta di 138 GB di testo. Cfr. Martin *et al.* (2020) e il link https://huggingface.co/camembert/camembert-base.

[13] https://oscar-project.org/.

[14] https://commoncrawl.org/.

[15] Sono stati scaricati i dossier legislativi del Senato e dell'Assemblea nazionale dal gennaio 2019 al dicembre 2022 per un totale di c.a. 80MB. Ringraziamo Micaela Rossi dell'Università di Genova e Danio Maldussi dell'Università di Bergamo per aver coordinato i lavori su tali corpora.

[16] Queste lingue veicolari variano molto anche sulla base delle persone addette alla traduzione che possono essere più o meno madrelingua rispetto al paese rappresentato (es. presenza di personale canadese nei servizi di traduzione dell'ONU, presenza di madrelingua presso il Consiglio d'Europa rispetto alle istituzioni dell'Unione europea che non richiedono per forza personale madrelingua…). L'eurogergo o eurocratese, ovvero il gergo istituzionale dell'UE che varia nelle diverse lingue (Raus, 2013a), è stato analizzato da diversi anni (Goffin, 1994; Mori 2018); diversamente, le varianti linguistiche dell'ONU e di altre istituzioni sono meno conosciute.

[17] Ad esempio, per l'italiano utilizzato nell'UE, cfr. Nydstet (1999) ma anche, per le ventiquattro lingue ufficiali dell'UE, il lavoro coordinato da Mori (2018).

[18] È stata usata la piattaforma *Gradio* (https://gradio.app/) ospitata su *HuggingFace Hub* (https://huggingface.co/) sotto forma di *space* (https://huggingface.co/spaces).

[19] Sulla base del contesto, infatti, lo stesso termine inglese *gender* è stato adattato in Francia con "*sexe*, *sexiste*, *femme*"… privilegiando la chiarezza (cfr. https://www.culture.fr/FranceTerme/Recommandations-d-usage/GENDER).

[20] Il modello è disponibile al link https://huggingface.co/dbmdz/bert-base-italian-xxl-cased; il corpus consta di 13 GB e di circa 2 miliardi di parole.

[21] Si tratta di 68 testi per un totale di 3.661 frasi.

[22] Per la presentazione del team linguistico, cfr. https://www.jmcoe.unito.it/content/e-mimic-empowering-multilingual-inclusive-communication.

[23] L'iniziativa, ancora in corso, è seguita alla firma della carta d'intenti *Io Parlo e non Discrimino* e coinvolge la testata regionale della RAI, l'Università di Torino, il Politecnico di Torino, il Consiglio Regionale del Piemonte, la Città e la Città metropolitana di Torino, enti associativi sul territorio. http://www.regione.piemonte.it/pinforma/images/DOCUMENTI/cartaintenti.pdf.

[24] Segnaliamo A.A..VV. (2015), *Un approccio di genere al linguaggio amministrativo. Linee guida*. Università degli Studi di Torino, https://www.unito.it/sites/default/files/linee_guida_approccio_genere.pdf e *Decalogo delle buone pratiche per un linguaggio amministrativo non discriminatorio*, Torino, Città metropolitana di Torino. http://www.cittametropolitana.torino.it/speciali/2021/linguaggio_genere/dwd/BastaunclicDecalogo.pdf.

[25] In particolare, i testi di Cecilia Robustelli, soprattutto le *Linee guida per l'uso del* genere *nel linguaggio amministrativo* (2012) (https://www.uniss.it/sites/default/files/documentazione/c._robustelli_linee_guida_uso_del_genere_nel_linguaggio_amministrativo.pdf).

[26] Tali template hanno permesso di produrre 480 record.

[27] Il progetto prevede l'utilizzo di paradigmi anche per l'inclusione delle persone ipovedenti o con cecità e per rendere il linguaggio maggiormente accessibile (es. riformulazione delle formule abbreviate, rimozione di elementi di disturbo come le barre oblique, ecc.). Tali paradigmi sono stati elaborati dalla componente studentesca dell'Università di Bologna, dell'Università IULM di Milano e dell'Università degli Studi di Trieste nell'ambito del corso *Professional Skills – Traduzione, inclusione e intelligenza artificiale* erogato all'Università di Bologna (campus di Forlì) nell'A.A. 2022-2023. Si ringrazia la componente studentesca che ha partecipato e le colleghe Elena Liverano e Helena Lozano Miralles che hanno coadiuvato l'iniziativa.

[28] Per una trattazione esaustiva di tali criteri, cfr. Raus *et al.* (2002).

[29] https://labelstud.io/.

[30] Questo *tag* è peraltro fondamentale per forzare l'apprendimento di elementi necessari come, ad esempio, il femminile dei nomi di professione.

[31] Per questi test, ringraziamo in particolare il collega Moreno La Quatra.

[32] Cfr. il progetto *UPgrading the SKIlls of Linguistics and Language Students – UPSKILLS.* https://upskillsproject.eu.

[33] Abbiamo tradotto noi dall'inglese.

BIBLIOGRAFIA

Attanasio G., Cagliero L., Cerquitelli Greco S., La Quatra M., Raus R., Tonti M. (2021), "E-MIMIC: Empowering Multilingual Inclusive Communication" in: *2021 IEEE International Conference on Big Data (Big Data)*, Orlando (USA), 15-18 dicembre 2021, pp. 4227-4234.

Bartoletti, I. (2020), *An Artificial Revolution. On Power, Politics and AI*, Edinburgh, Indigo.

Braun, J. (von), Archer, M. S., Reichberg, G. M., Sánchez Sorondo, M. (2021), *Robotics, AI, and Humanity. Science, Ethics, and Politics*, Svizzera, Springer.

Caliskan, A., Bryson, J.J., Narayan, A. (2017), "Semantics derived automatically from language corpora contain human-like biases", *Sciences*, 356(6334), pp. 183-186.

Cerquitelli, T., Cagliero, L., Greco, S., La Quatra, M., Raus, R., Tonti, M. (2023), "Riformulazione automatica di test in lingua italiana in forma inclusiva mediante Intelligenza Artificiale", *Poster* presentato al convegno *Just the Woman I Am – JTWIA*, Torino, 8 marzo 2023, https://jtwia.org (ultimo accesso 23/06/2023).

Cerquitelli, T., Raus, R., Molino, A. (in corso di pubblicazione), "Artificial Intelligence and Neural Machine Translation", in: Baumgartner, S., Tieber, M. (eds.) *Handbook of Translation Technology and Society*, New York e Londra, Routledge.

Consiglio dell'Unione europea (2018), *Comunicazione inclusiva*. https://www.consilium.europa.eu/it/documents-publications/publications/inclusive-comm-gsc/ (ultimo accesso 23/06/2023).

Ferraresi, A., Aragrande, G., Barrón-Cedeño, A., Bernardini, S., Miličević Petrović, M. (2021), "Competences, skills and tasks in today's jobs for linguists: Evidence from a corpus of job advertisements". UPSKILLS Intellectual output 1.3. https://zenodo.org/record/5030879 (ultimo accesso 23/06/2023).

Goffin, R. (1994), "L'eurolecte : Oui, jargon communautaire: Non", *Meta*, XXXIX(4), pp. 636-642.

Hermand, M.-H. (2014), "Le discours eurorégional. Indices convergents de légitimation d'un espace institutionnel", *Mots. Les langages du politique*, 106, pp. 71-86.

International Organization for Standardization (2009), *Assessment and benchmarking of terminological resources – General concepts, principles and requirements*, ISO 23185: 2009, Ginevra.

Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S., Ney, H. (2019), "Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages", in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, November 3-7, pp. 866-876.

Kohen, P. (2020), *Neural Machine Translation*, Cambridge, Cambridge University Press.

Le Cun, Y. (2019), *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*, Parigi, Odile Jacob.

Martin, L. *et al.* (2020), "CamemBERT: a Tasty French Language Model", in: *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, July 2020, https://arxiv.org/abs/1911.03894 (ultimo accesso 23/06/2023).

Marzi, E. (2021), "La traduction automatique neuronale et les biais de genre : le cas des noms de métiers entre l'italien et le français", *Synergies Italie*, 17, pp. 19-36. http://gerflint.fr/Base/Italie17/marzi.pdf (ultimo accesso 23/06/2023).

Miličević Petrović, M., Bernardini, S., Ferraresi, A., Aragrande, G., Barrón-Cedeño, A. (2021), "Language data and project specialist: A new modular profile for graduates in language-related disciplines". *UPSKILLS Intellectual output 1.6*. Zenodo. https://dx.doi.org/10.5281/zenodo.5030929 (ultimo accesso 23/06/2023).

Monti, J. (2019), *Dalla Zairia alla traduzione automatica. Riflessioni sulla traduzione nell'era digitale*, Napoli, Paolo Loffredo Editore.

Mori, L. (2018), *Observing Eurolects. Corpus Analysis of linguistic variation in EU law*, Amsterdam, John Benjamins Publishing Company.

Nugara, S. (2011a), "De l'anglais onusien au français européen : L'émergence de la dénomination violence domestique à l'égard des femmes dans le discours du conseil de l'Europe", in: Calvo, A., *et al.* (a cura di), *World Wide Women / Mondialisation, genres, langages*, Torino, CIRSDe. https://iris.unito.it/retrieve/handle/2318/143354/24029/WWWsilvia.nugara.pdf (ultimo accesso 23/06/2023).

Nugara, S. (2011b), "Féminisme et universalisme dans le Conseil de l'Europe : Le cas de la dénomination *violence domestique à l'égard des femmes*". *Synergies Italie*, 7, pp. 39-49.

Nydstet, J. (1999), "L'italiano che si scrive a Bruxelles", *Italiano e oltre*, XIV, pp. 198-206.

O'Brien, S., Simard, M., Goulet, M.-J. (2018), "Machine translation and self-post-editing for academic writing support: Quality explorations" in: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.), *Translation quality assessment: From principles to practice*, Svizzera, Springer, pp. 237-262.

Organizzazione delle Nazioni Unite (2017), *Change. Rapport annuel 2017*, Ginevra, ONU.

Ortiz Suárez, P. J., Sagot, B., Romary, L. (2019), "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures", in: *Challenges in the Management of Large Corpora (CMLC-7) 2019*, https://corpora.ids-mannheim.de/CMLC7-final/CMLC-7_2019-Oritz_et_al.pdf (ultimo accesso 23/06/2023).

Paquet-Gauthier, M. (2018), "Changements sémantiques sous l'influence de l'anglais : Le cas de quatre "emprunts de sens" en français au Québec (1992–2012)", in: Jacquet-Pfau, C., Napieralski, A., Sablayrolles, J.-F. (éds.), *Emprunts néologiques et équivalents autochtones: Études interlangues*, Łódź, University of Łódź, pp. 177-200.

Rastier, F. (2021), "Data vs Corpora", in: Mayaffre, D., Vanni, L. (éds.) *L'intelligence artificielle des textes : des algorithmes à l'interprétation*, Parigi, Champion, pp. 203-245.

Raus, R. (2013a), "L''eurojargon' et sa variante française", *Argotica* 1(2), pp. 383-394.

Raus, R. (2013b), *La terminologie multilingue : la traduction de l'égalité homme/femme dans les organisations internationales*, Bruxelles, De Boeck.

Raus, R. (2020), "Mémoires 'féministes' et déconstruction des différences : la traduction française de *gender* et d'*intersectionality* dans les Organisations internationales", *De genere. Journal of Literary, Postcolonial and Gender Studies*, 5, pp. 43-57. http://www.degenere-journal.it (ultimo accesso 23/06/2023).

Raus, R., Tonti, M., Cerquitelli Cagliero, L. T., Attanasio, G., La Quatra, M., Greco, S. (2022), "L'analyse du discours et l'intelligence artificielle pour réaliser une écriture inclusive : le projet E-MIMIC", in : *Congrès mondial de Linguistique française – CMLF 2022*, Orléans, 4-9 luglio 2022, https://www.shs-conferences.org/articles/shsconf/pdf/2022/08/shsconf_cmlf2022_01007.pdf (ultimo accesso 23/06/2023).

Ryan, R. (2009), "Les langues contrôlées, une valeur ajoutée pour le traducteur", *Revue française de la traduction*, 22, pp. 57-67.

Tavosanis, M. (2018), *Lingue e intelligenza artificiale*, Roma, Carocci.

Tiedemann, J. (2012), "Parallel Data, Tools and Interfaces in OPUS", in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association, pp. 2214-2218.

Toral, A. (2019), "Post-editese: An Exacerbated Translationese", in: *Proceedings of MT Summit XVII*, Dublino, August, 19-23, 1, pp. 273-281.

Veniard, M. (2021), *La nominazione degli eventi nella stampa. Saggio di semantica discorsiva*, Roma, Tab edizioni.

Vetere, G. (2023), "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive", *De Europa Special issue 2022*, Milano, Ledizioni, pp. 69-87.

Zouinar, M., 2020, "Évolutions de l'Intelligence artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail", *Activités*, 17(1). https://journals.openedition.org/activites/4941#ftn1 (ultimo accesso 23/06/2023).