Traduzione automatica e umana a confronto. Implicazioni per il post-editing

Simone Torsani

Abstract (italiano)

Il saggio indaga le differenze tra la traduzione umana e automatica al fine di individuare alcune caratteristiche di quest'ultima utili a definire meglio il lavoro di *post-editing*. La ricerca illustrata confronta, tramite un approccio per bigrammi, un corpus (N=36) di testi tradotti tramite il motore Google Translate con gli stessi testi tradotti da traduttori umani. I risultati mostrano che, rispetto alla traduzione umana, la traduzione automatica è generalmente più ridondante, tende a utilizzare maggiormente bigrammi più frequenti e a evitare bigrammi assenti dal corpus di riferimento. Il saggio discute le implicazioni dei dati ottenuti e le possibili applicazioni di un'analisi basata sui bigrammi per il lavoro di post-editing.

Parole chiave

Traduzione automatica; traduzione umana; post-editing; analisi per bigrammi;

Abstract (English)

The paper investigates the differences between human and machine translation to identify some characteristics of the latter that may be useful in better defining and understanding post-editing. The reported research compares, through a bigram-based approach, a corpus (*N*=36) of texts translated through the Google Translate engine with the same texts translated by human translators. The results show that, compared to human translation, machine translation is generally more redundant, tends to use more frequent bigrams and avoid bigrams absent from the reference corpus. The paper discusses the implications of the data obtained and possible applications of bigram-based analysis for post-editing work.

Keywords
Machine translation; human translation; post-editing; bigram-based analysis.

1. Introduzione

1.1 La conoscenza della traduzione automatica come competenza per il post-editing

La sviluppo e la diffusione sempre più capillare della Traduzione Automatica (d'ora in avanti, TA, si usa anche l'ing. *Machine Translation*, MT) hanno velocemente archiviato il dibattito sull'efficacia e sull'utilità di questa tecnologia, tanto che TA e *post-editing* (d'ora in avanti, PE) costituiscono oggi uno degli argomenti più importanti nella didattica della traduzione assistita. Il termine 'traduzione assistita' potrebbe forse apparire obsoleto visti sviluppi nel settore, come appunto la TA, ma non lo è, perché descrive ancora bene il rapporto che intercorre tra tecnologia e traduzione. Anche nel caso del PE, infatti, è sul traduttore che ricade la responsabilità del lavoro finale (do Carmo e Moorkens 2020). Si noti, infatti, che nella letteratura il testo prodotto da un sistema di TA è definito non 'traduzione', ma *output*, termine considerato più adatto a descrivere il materiale grezzo sul quale un *post-editor* deve lavorare. Non una revisione, insomma, ma una modifica più o meno approfondita (do Carmo e Moorkens 2020). È quindi sul rapporto tra tecnologia e traduttore che si fonda la ricerca sul PE e sulla sua didattica.

In una sintesi aggiornata della didattica della revisione e del PE (due attività simili e spesso confuse, ma che presentano differenze di rilievo, *ibidem*) Konntinen, Salmi e Koponen (2020) individuano tra le competenze strategiche necessarie a queste due attività (le altre sono le competenze interpersonali e strumentali), la conoscenza delle caratteristiche e degli errori tipici dei sistemi di TA. Un *post-editor*, infatti, dovrebbe avere coscienza delle aree di potenziale criticità dei sistemi, in modo da rendere più efficiente il lavoro di PE, ma anche per preparare i testi per la TA, operazione detta *pre-editing* (cfr. su questo Nitzke e Hansen-Schirra 2021). A tale riguardo Hansen-Schirra et al. (2019) identificano per la coppia inglese-tedesco alcuni problemi nell'output dei sistemi di TA, tra i quali le espressioni idiomatiche (spesso tradotte letteralmente) o la sintassi, in particolare la distinzione tra proposizioni principali e secondarie.

È proprio la profondità e la complessità del lavoro di PE, quindi, che rende la valutazione della qualità dell'output (cfr., tra gli altri, Läubli et al. 2018; Rivera-Trigueros 2022), e quindi la comprensione dei sistemi di TA, un requisito fondamentale per un *post-editor*, perché è la qualità dell'output a determinare la quantità di lavoro necessario al PE (Seewald-Heeg 2019). Quello di qualità è da

sempre un concetto fondamentale nel campo della traduzione assistita (Garcia 2023), e la riflessione intorno ad esso tocca oggi la TA. Viste le competenze necessarie per il PE, infatti, la valutazione della qualità di un output è funzionale anche alla conoscenza dei sistemi TA: anzi, il confine tra valutazione e conoscenza dei sistemi pare sfumare in una prospettiva didattica.

Il progetto, di cui si restituiranno di seguito i primi risultati, mira a costruire una base di conoscenza sulle caratteristiche della TA per la lingua italiana con l'obiettivo di migliorare la preparazione dei futuri traduttori rispetto a questa tecnologia. Il progetto, di medio-lungo termine, è incentrato sul confronto tra TA e traduzione umana (TU, cfr. infra) e prevede una prima fase dedicata ad analisi semi-automatiche su diversi corpora allineati di TA e TU per individuare aree di interesse sui cui lavorare.

1.2. Approcci nella valutazione della TA

Lommel e Burchardt (2019) illustrano quattro approcci alla valutazione della TA. Il primo raccoglie i diversi metodi basati su riferimenti, che confrontano l'output TA con una traduzione umana: la qualità dell'output è in questo caso valutata sulla base della somiglianza con la traduzione umana, per esempio, su quanti n-grammi sono presenti in entrambe. Questi metodi adottano, osservano gli autori, un approccio definito trascendente, cioè confrontano la TA con un modello ideale che coincide con la traduzione umana. Un secondo approccio è detto manuale ed è basato sul giudizio di un esperto che valuta una traduzione, in particolar modo scorrevolezza e adeguatezza, sulla base di criteri specifici. Un terzo approccio lavora sulla distanza: i metodi che a esso fanno capo misurano la quantità di lavoro necessaria da parte di un *post-editor* per modificare un output e renderlo accettabile. Ultima famiglia è quella dei sistemi automatizzati di valutazione, cioè sistemi che analizzano i diversi segmenti di partenza e di arrivo di un corpus di traduzioni per estrarne indicatori sulla base dei quali effettuano una stima della qualità dell'output, per esempio estraendo dal testo tradotto misure di fluidità, come il numero di token.

Bestgen (2021, 2022) utilizza un metodo ibrido, nel quale integra gli approcci basati sui riferimenti e quelli automatizzati, per indagare alcune caratteristiche dei sistemi TA. Più precisamente l'autore utilizza un metodo incentrato su bigrammi e misure di associazione per confrontare l'output di un sistema TA con la traduzione umana. Il metodo consiste nell'estrarre da un testo tutti i bigrammi, sequenze di due parole contigue (cioè, separate solo da spazio e non da punteggiatura), come per esempio (brano tratto dalla TU dell'intervento *A new way to help young people with their mental health* dal corpus utilizzato nella presente ricerca, cfr. infra):

> Tanto tempo fa, prima delle fattorie di canna da zucchero, prima delle capanne con il tetto di paglia (…)

che diventa

> tanto tempo / tempo fa / prima delle / delle fattorie / fattorie di / di canna ecc.

Si noti che la sequenza si interrompe in presenza della virgola perché come detto i bigrammi sono composti da parole non separate da punteggiatura. A ognuno dei bigrammi ottenuti sono quindi applicate diverse misure di associazione, cioè misure che quantificano il grado di attrazione tra due parole. L'autore ne utilizza due, t-score e *Mutual Information* (MI), qui descritte brevemente nella sezione dedicata al metodo. Le due misure restituiscono risultati diversi e valorizzano combinazioni di diverso tipo: il t-score assegna un valore alto a bigrammi con alta frequenza e composti da parole frequenti; al contrario bigrammi con frequenza osservata anche bassa, composti da parole poco frequenti, ma molto attratti tra loro hanno valori alti di MI.

Bestgen (2021), confronta TU e TA nella coppia inglese-francese e scopre che la TU presenta un numero maggiore di bigrammi con MI alto (MI>7), mentre la TA presenta un numero maggiore di bigrammi con t-score alto (t-score>10), un risultato che l'autore spiega con il fatto che la TA sarebbe più letterale. Ciò trova una spiegazione nel fatto che l'apprendimento di un sistema TA è influenzato dalla frequenza degli elementi durante l'addestramento (*ibidem*; cfr. anche Hansen-Schirra et al. 2019 sul ruolo della frequenza sull'apprendimento del linguaggio idiomatico e Koehn 2017 per un'introduzione alla traduzione automatica neurale) e quindi, questa dimensione dovrebbe essere privilegiata. Questi risultati, ottenuti su un corpus allineato di articoli di giornale (tradotti da traduttori umani e poi da sistemi TA) sono confermati da uno studio di replica dello stesso autore su trascrizioni di discorsi al Parlamento Europeo (2022). La ricerca illustrata nel presente contributo riprende, in parte, questo approccio.

2. Obiettivi e metodo

La prima parte del progetto, quella illustrata in questo saggio, consiste nel verificare un assunto base nella letteratura, e cioè che la Traduzione Automatica Neurale favorirebbe soluzioni (qui bigrammi) frequenti perché hanno un peso più forte nell'addestramento della rete. Ciò dovrebbe implicare che, rispetto a una TU, l'output di un sistema TA dovrebbe contenere:

1. Un maggior numero di bigrammi con t-score alto, che rappresentano bigrammi ad alta frequenza;

2. Un minor numero di bigrammi con MI alto, indicanti bigrammi più rari;

3. Un minor numero di bigrammi rari o assenti dal corpus di riferimento;

Per verificare l'ipotesi è stato raccolto un corpus parallelo di 36 presentazioni Ted Talk che sul sito della comunità[1] hanno disponibile una traduzione in lingua italiana. I testi selezionati sono di argomento diverso (tra cui economia, medicina e storia) e tradotti da traduttori diversi. I testi originali in inglese sono stati quindi tradotti con il sistema TA di Google Translate, uno tra i diversi sistemi attualmente disponibili (come DeepL o Reverso), già utilizzato in ricerche analoghe (Lee 2022). Al momento della ricerca il corpus era così composto:

	n. testi	n. token
Trad. Umana	36	34.000
Trad. Google	36	36.500

Tabella 1. Dettagli del corpus in esame.

Originale	Trad. Umana	Trad. Google
Globally, about 10% of people will experience an eating disorder during their lifetime. And yet, eating disorders are profoundly misunderstood. Misconceptions about everything from symptoms to treatment, make it difficult to navigate an eating disorder or support someone you love as they do so.	Circa il 10% della popolazione mondiale avrà a che fare con un disturbo alimentare nel corso della propria vita. Eppure, questi disturbi vengono profondamente fraintesi. Convinzioni errate su tutti gli aspetti, dai sintomi al trattamento, rendono difficile la gestione di un disturbo alimentare o sostenere una persona cara che ne è affetta.	A livello globale, circa il 10% delle persone sperimenterà un disturbo alimentare nel corso della propria vita. Eppure, i disturbi alimentari sono profondamente fraintesi. Idee sbagliate su tutto, dai sintomi al trattamento, rendono difficile affrontare un disturbo alimentare o supportare qualcuno che ami mentre lo fanno.

Tabella 2. Trad. umana e automatica a confronto (titolo originale *Why are eating disorders so hard to treat?*).

[1] Alla pagina https://www.ted.com/.

I testi dei due sotto corpora (TU e TA) sono stati elaborati da uno script che produce per ogni documento diversi indicatori. Quelli utilizzati nella presente ricerca sono:

- Numero di parole unità (*token*) per testo;
- Valore medio di t-score di tutti i bigrammi di un testo;
- Valore medio di MI di tutti i bigrammi di un testo;
- Frequenza osservata media di tutti i bigrammi di un testo;
- Percentuale sul totale dei bigrammi di un testo di bigrammi con t-score superiori alle soglie di 3, 5, 7, 9, 11;
- Percentuale sul totale dei bigrammi di un testo di bigrammi con MI superiori alle soglie di 3, 5, 7, 9, 11;
- Percentuale sul totale dei bigrammi di un testo di bigrammi assenti dal corpus di riferimento;
- Numero di bigrammi di un testo con MI maggiore o uguale a 7 e 10 e O minore di 1000, 500, 200 e 100, cioè bigrammi forti e rari in misure diverse;

T-score e MI confrontano, in maniera diversa, la frequenza osservata e la frequenza attesa in un corpus di riferimento. La frequenza attesa di un bigramma corrisponde alla probabilità di trovare due parole insieme se non ci fosse attrazione tra loro, data la loro frequenza e la dimensione del corpus di riferimento. Per esempio, il corpus Paisà, usato nella presente ricerca, consta circa 250.000.000 parole e le due parole *di* e *corsa* hanno una frequenza di, rispettivamente, 8388607 e 15561; la frequenza attesa del bigramma *di corsa* è quindi (8388607x15561)/250.000.000 = 522,14: in altre parole, se non ci fosse attrazione tra le due parole, queste si troverebbero insieme circa 522 volte. Le misure confrontano, in maniera diversa, frequenza osservata e attesa e indicano se le due si attraggono o respingono. Nel caso di *di corsa*, che ha frequenza osservata (indicata con O) 1417, le due parole si attraggono secondo la formula del t-score, e hanno infatti un valore (normalizzato su un corpus di riferimento di 100.000.000 di parole) di 15,03, molto alto perché superiore alla soglia di ≥10 che per convenzione indica i bigrammi ad alta frequenza. Al contrario, lo stesso bigramma ha un valore di MI molto basso, 1,44, che identifica il bigramma come non collocazionale (sempre secondo le convenzioni adottate). Riprendendo l'esempio del brano citato nell'introduzione (cfr. 1.3) limitatamente alla MI, ogni bigramma riceve un valore nella misura:

tanto tempo (MI=4,19) / tempo fa (MI=4,63) / prima delle (MI=1,28) / delle fattorie (MI=5,26) / fattorie di (MI=0,86) / di canna (MI=1,64) ecc.

La scelta dell'ultimo indicatore (numero di bigrammi di un testo con MI maggiore o uguale a 7 e O <1000 ecc.), non validato in letteratura, è dovuta al fatto che, almeno relativamente alla lingua italiana, le combinazioni con valori di MI alto (di norma si utilizza la soglia di ≥7) non hanno necessariamente bassa frequenza: il bigramma *ad esempio* ha O=46.354 e MI=7,46. Ciò rende meno significativo il confronto tra le fasce alte di MI perché contraddice il principio per cui i bigrammi con MI alto hanno frequenza bassa e sono formati da parole poco frequenti: le diverse declinazioni di MI e O, quindi, mirano ad identificare appunto i bigrammi autenticamente rari.

I valori delle due misure sono calcolati sul corpus Paisà (Lyding et al. 2014), un *web corpus* in lingua italiana, di circa 250.000.000 parole. I valori del t-score sono stati normalizzati, cioè calcolati come se il corpus di riferimento fosse 100.000.000: ciò è necessario per rendere i risultati confrontabili con altre ricerche dal momento che il valore di questa misura cambia a seconda della dimensione del corpus di riferimento[2].

La modalità di indagine utilizzata, quindi, adotta un metodo basato sui riferimenti per la valutazione della qualità della TA, cioè un metodo in cui «la qualità della MT [*machine translation*, n.d.A.] è […] vista come proporzionale alla sua similitudine con il testo di riferimento» (Lomme e Birchardt 2019: 89), cioè la traduzione umana. In ciò l'indagine adotta una prospettiva, come spiegato, trascendente, nella quale il testo umano è considerato ideale (ibidem).

I valori dei due sotto corpora sono stati confrontati con un test t accoppiato, cioè un test nel quale ogni documento in un sotto corpus è confrontato con il corrispettivo nell'altro, per esempio il documento 1 nel sotto corpus TU è confrontato con il documento 1 nel sotto corpus TA e così via. Il livello alfa per i test è stato stabilito in 0,05 ($p \leq 0{,}05^*$; $p \leq 0{,}01^{**}$; $p \leq 0{,}001^{***}$), la dimensione dell'effetto è calcolata con la d di Cohen utilizzando come soglie i valori di $d \geq 0{,}3$ dimensione piccola; $d \geq 0{,}5$ dimensione media e $d \geq 0{,}8$ dimensione grande.

[2] La formula del t-score è basata su una sottrazione e più grande è il corpus di riferimento, più cresce il valore di t-score di un bigramma. Per esempio *tenuto presente* ha t-score 11,52 se calcolato sulle dimensioni originali del corpus Paisà, ma 7,29 se normalizzato. Con ciò non si intende sostenere che il valore normalizzato sia più corretto, ma solo che è più facilmente comparabile. Inoltre, molte ricerche in questo ambito in lingua inglese usano di norma il British National Corpus che ha 100.000.000 parole unità come riferimento.

Sebbene costituisca un inizio e intenda indicare alcune aree di interesse per la ricerca in questo ambito, la ricerca presenta limiti che vanno tenuti presenti nella valutazione dei risultati. Il primo è legato alla natura dei testi in esame, limitati a un genere specifico, e pertanto è necessario lavorare anche su altri generi e tipologie di testi al fine di ottenere un quadro più ampio. Il secondo è che l'approccio per bigrammi non è in grado di indagare fenomeni importanti per la TA, come per esempio la natura letterale della traduzione, che andranno investigati con strumenti più adatti e, soprattutto, messi in relazione con gli aspetti qui emersi, come la ridondanza dell'output. Terzo, la ricerca ha preso in considerazione solo uno tra i diversi sistemi di TA e i suoi risultati devono essere quindi confermati da analisi sugli altri motori disponibili. La ricerca, infine, non considera l'intelligenza artificiale generativa, oggi sempre più spesso usata in connessione con la TA e il post-editing, e i suoi risultati vanno quindi considerati al netto dell'apporto di questo strumento.

3. Risultati

Il confronto tra i due sottocorpora nei diversi indicatori restituisce diversi risultati significativi che sono di seguito riportati.

I testi prodotti dalla TU sono in media più lunghi M=1040 (deviazione standard, ds=406) rispetto a quelli prodotti dalla TA, M=965 (ds=365), $t(35)=6{,}5$, $p<0{,}001^{***}$ e $d=1{,}09$ (dimensione dell'effetto alta).

Nella media del t-score per testo la TA presenta un valore medio M=24,2 (ds=3,2), mentre la TU ha M=22,2 (ds=3,2). La differenza è significativa, $t(35)=4{,}7$, $p<0{,}001^{***}$ e $d=0{,}79$ (dimensione dell'effetto medio-alta). Il risultato indica, quindi, che la TA produce bigrammi mediamente più collocazionali se analizzata con questa misura.

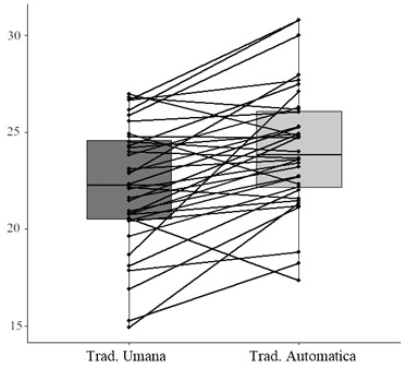

Figura 1. Valore medio di t-score per documento nei due sottocorpora. La TA mostra valori generalmente maggiori rispetto alla TU (d=1,09).

Poiché il t-score non è considerato una misura di frequenza, il risultato non implica che i bigrammi della TA siano mediamente più frequenti rispetto alla TU. Per questo si può invece confrontare la frequenza osservata media (O) per ogni testo. La media di O di tutti i bigrammi di un testo per la TA è M=9938,4, (ds=1906), leggermente superiore alla media della TU, che ha M=9267,4 (ds= 1766,7). La differenza è significativa $t(35)=2,6$, $p=0.013$* con $d=0,49$ (medio-bassa). Quindi la TA produce bigrammi mediamente, anche se di poco, più frequenti rispetto alla TU.

Sempre relativamente al t-score, la percentuale di bigrammi nelle diverse fasce mostra differenze significative tra TA e TU, in particolare nella fascia t-score≥10, con TA è M=46, 4, (ds=3,9) è superiore TU, che ha M=44,9 (ds= 4,4). La differenza è significativa $t(35)=3,3$, $p=0,002$** con $d=0,54$ (dim. media). Il risultato è analogo anche per valori superiori t-score, per esempio t-score ≥15, con TA che ha M=38, (ds=3,6), superiore TU, che ha M=36,5 (ds= 4). La differenza è significativa $t(35) = 3,5$, $p=0,001$** con $d=0,59$. La TA produce quindi più bigrammi ad alta frequenza. Questo risultato, insieme a quello relativo al valore medio di t-score, deve essere però contestualizzato. Se si considerano tutti i bigrammi di un corpus, i bigrammi con t-score alto, almeno relativamente alla lingua italiana, sono prevalentemente (nel sottocorpus della TA circa l'83%) combinazioni con, o composte da, parole grammaticali (pronomi, preposizioni, articoli ecc.), come mostra il seguente esempio di output TA, nel quale i bigrammi con t-score≥10 sono sottolineati (la doppia linea indica una sovrapposizione tra due bigrammi, come in *sulla ricerca* e *ricerca di* dove i due bigrammi che compongono il trigramma hanno entrambi t-score>10):

1. TA 'Alcuni astronomi concentrano il loro tempo e le loro energie sulla ricerca di pianeti a queste distanze dalle loro stelle. Quello che faccio riprende dove finisce il loro lavoro. Modello i possibili climi degli esopianeti'.

Come si può osservare tutti i bigrammi con t-score alto coinvolgono almeno una parola grammaticale. Lo stesso testo, nella traduzione umana, mostra una quantità di bigrammi con t-score alto sempre elevata, ma minore rispetto alla TA:

1. (bis) TU 'Alcuni astronomi dedicano il loro tempo, e le loro energie, a trovare pianeti alla giusta distanza dalla loro stella. Il mio lavoro comincia dove finisce il loro: elaboro modelli sui possibili tipi di clima degli esopianeti'.

Nell'esempio si può osservare, a parità di numero di parole nei due brani, il maggior numero di bigrammi con t-score ≥ 10 nella TA (13 in tutto, 41% sul to-

tale dei bigrammi contro 8, il 34% della TU). Gli output TA sarebbero quindi più ridondanti rispetto alla TU, perché sovrautilizzano combinazioni con parole grammaticali, più frequenti e quindi, nella logica dei sistemi TA, più affidabili.

Relativamente ai bigrammi assenti, la TA ne produce in media meno con M=7,6 (ds=2,2) rispetto alla TU M=8,5 (ds=2,2), $t(35)=2,5$, $p=0,002^{**}$ $d= 0,59$. Quella dei bigrammi assenti è una categoria particolare nei lavori su corpora di apprendenti (dove origina l'analisi per bigrammi), nei quali i bigrammi assenti possono essere indicativi di errori morfosintattici o ortografici. Ciò pone il problema di stabilire come questa categoria debba essere interpretata relativamente alla TA. Mentre, infatti, errori ortografici possono essere esclusi con una certa sicurezza, lo stesso non si può affermare di errori di sintassi, come dimostrerebbe l'esempio dal corpus:

2. TA 'È un classico racconto ammonito, giusto?' (orig. *It's a classic cautionary tale, right?*, trad. umana 'Proprio come nelle favole, no?')

Qui il bigramma assente è la spia di un inceppamento nel processo, con il sistema che, non trovando una formula adeguata, procede alla traduzione letterale, cioè parola per parola, di *cautionary tale*. Casi come il precedente pongono, allora, la questione se i bigrammi assenti possano essere indicatori di quelli che si possono, provvisoriamente, definire come punti problematici per la TA. Altri esempi dal corpus mostrano questa possibilità, come:

3. TA 'Non riesco a vedere questo pianeta a occhi nudi e nemmeno con i telescopi più potenti che attualmente possediamo.' (orig. *I can't see this planet with my naked eyes or even with the most powerful telescopes we currently possess*, trad. umana 'Oggi non riesco ancora a vedere questo pianeta a occhio nudo, né col più potente dei telescopi oggi a disposizione.')

Oppure

4. TA 'Le persone che pesano ciò che i medici potrebbero considerare un intervallo sano possono avere disturbi alimentari' (orig. *People who weigh what medical professionals might consider a healthy range can have eating disorders*, trad. umana 'Le persone che secondo i medici rientrano nel range normopeso possono avere disturbi anche gravi')

Il fatto è già osservato in letteratura (cfr. Schirra et al. 2019), anche se su unità di analisi più ampie, come le espressioni idiomatiche. Nel caso presente la traduzione letterale (e problematica) è segnalata appunto dai bigrammi assenti dal corpus di riferimento (*naked eye*s > occhi nudi; *weigh what* > pesano ciò; *healthy range* > intervallo sano).

Ultima analisi è quella relativa ai bigrammi forti e rari. Mentre non si osservano differenze significative né dimensioni dell'effetto rilevanti nel confronto tra le fasce di MI, il confronto tra il numero di bigrammi forti e rari è sempre significativo e con dimensioni dell'effetto medie e medio-basse, con la TA che produce più bigrammi di questo tipo:

	<1000	<500	<200	<100
TA	M=37,7; ds=17,3	M=31; ds=15,1	M=22,9; ds=12,1	M=17,2; ds=8,4
TU	M=33,3; ds=15,3	M=28,1; ds=13,8	M=20,7; ds=12,3	M=15,6; ds=7,2
t(35)	4 (*p*<0,001***)	3,1 (*p*=0,004**)	3 (*p*=0,005**)	2,6 (*p*=0,01*)
d	0,67	0,52	0,49	0,44

Si noti che anche questo fenomeno può essere, almeno in parte, interpretato in termini di ridondanza; per esempio nella traduzione dell'intervento *Why don't we cover the desert with solar panels?*, la TA ripete due volte 'pannelli solari' (MI=13,79 O=655) mentre la TU, per evitare la ripetizione, utilizza l'anafora (strategia di riduzione) per la seconda istanza di *solar panel* del testo di partenza:

5. TA: 'Quindi, coprire il deserto con pannelli solari potrebbe risolvere i nostri problemi energetici per sempre? I pannelli solari funzionano quando le particelle di luce colpiscono la loro superficie con energia sufficiente a far uscire gli elettroni dai loro legami stabili.' (orig. *So, could covering the desert with solar panels solve our energy problems for good? Solar panels work when light particles hit their surface with enough energy to knock electrons out of their stable bonds*, trad. umana 'Quindi, coprire il deserto con pannelli solari potrebbe risolvere il tutto? Ai pannelli serve che le particelle di luce li colpiscano con abbastanza energia da rompere i legami stabili degli elettroni.')

Un confronto sulla distribuzione di questo tipo di bigrammi (nelle diverse gradazioni di rarità) mostra, infine, che questi sono in parte sovrapponibili, cioè TU e TA usano lo stesso bigramma nello stesso punto della traduzione in circa il 66% di casi con O<1000, valore che scende a 58% con O<100.

	O<1000	O<500	O<200	O<100
TU	595	531	381	273
TU/TA	414	364	237	171
perc.	69,58%	68,55%	62,20%	62,64%

Tabella 3. Sovrapposizione tra TU e TA di bigrammi (sostantivo + aggettivo e aggettivo + sostantivo) con MI ≥7 e O <1000, 500, 200, 100. La seconda riga indica i casi in cui la TA ha prodotto lo stesso bigramma della TU nello stesso punto della traduzione.

Si noti che il dato è ancora grezzo, dal momento che lo status dei bigrammi in esame non è stato ancora indagato relativamente alla lingua italiana ed essi non possono pertanto essere automaticamente considerati collocazioni o termini (cfr. per es. nel corpus *ultimo decennio*: MI=8,46 o *proprie esperienze*: MI=7,55), veri elementi di interesse per la traduzione, in particolar modo quella specializzata. Né il calcolo considera l'impatto della ridondanza e della ripetitività nella TA. Alzando, arbitrariamente, la soglia minima di MI a 10, cioè lavorando su combinazioni con un maggior grado di attrazione (ma non per questo necessariamente più rare), le percentuali aumentano:
I risultati restituiti dalle analisi illustrate in questa ricerca sono quindi riassumibili in:

- La TA tende a prediligere la frequenza e produce testi con bigrammi mediamente più frequenti;
- La TA è più ridondante rispetto alla TU perché produce più i bigrammi t-score alto (indice di alta frequenza) che contengono in genere almeno una parola grammaticale: l'output TA è composto da testi più lunghi e ridondanti;
- La TA non ha problemi a produrre bigrammi rari e con forte attrazione, ma non è chiaro se questi costituiscano una traduzione corretta;
- I bigrammi forti e rari sono in parte (ma non è ancora chiaro in quale misura) sovrapponibili tra TU e TA;
- La TA tende a evitare i bigrammi assenti e quando ne produce c'è la possibilità che questi rappresentino traduzioni letterali errate (anche se non è chiaro in quale misura).

Le sezioni seguenti discuteranno le loro implicazioni per il post-editing e alcune possibili linee di ricerca future.

4. Discussione e implicazioni

Il primo risultato di interesse è la conferma che la TA produce testi con bigrammi mediamente più frequenti e, come già osservato da Bestgen (2021), un maggior numero di bigrammi ad alta frequenza, caratterizzati da t-score alto. Ciò è spiegabile con i meccanismi di allenamento delle reti neurali che favoriscono, cioè considerano più affidabili e quindi danno loro maggior peso, le soluzioni più frequenti. Il dato, come visto, va però interpretato: i risultati, infatti, mostrano che l'83% dei bigrammi con t-score alto contiene una parola grammaticale e il valore medio più alto di t-score è da imputarsi a una lingua generalmente più ridondante e meno sintetica, come suggerito anche dalla maggiore lunghezza media dei testi del sotto-corpus TA (circa il 7%, cfr. Tabella 1). Una prima implicazione, quindi, è che il PE può consistere in un lavoro di sintesi, operazione auspicabile in particolar modo nel caso di un PE completo (cfr. Nitzke e Hansn-Scirra, 2021 su PE leggero e completo), cioè un intervento mirato a migliorare anche la qualità linguistica della traduzione.

Non trova, invece, conferma la seconda ipotesi e cioè che la TU produca bigrammi con una media MI più alta o più bigrammi con MI alta, anzi, se si applica alla MI il filtro della frequenza osservata si scopre che la TA produce più bigrammi rari. Dal momento che il risultato è in contraddizione con ricerche precedenti su altre lingue e su altre tipologie di testo, ciò sembrerebbe indicare, in primo luogo, che l'analisi per bigrammi è influenzata da almeno una di queste due variabili, confermando l'importanza di replicare gli studi adattandoli a contesti diversi. Relativamente alla combinazione linguistica (inglese-italiano) e al tipo di testi indagati, in ogni caso, emerge un secondo aspetto importante e cioè che la TA non ha, in questo caso, difficoltà a produrre combinazioni rare e con forte attrazione, che possono consistere in vere e proprie collocazioni, come corteccia visiva: MI=12,31, O=38; *disturbi alimentari*: MI=9,90, O=56; *plastica biodegradabile*: MI=11,49, O=6. La produzione di bigrammi forti e rari, quindi, non è un problema per la TA: il sistema in esame, infatti, sembra riconoscere nella fase di addestramento anche le soluzioni poco frequenti e riesce quindi a produrle in fase di traduzione. Ciò non implica, si noti, una conclusione sull'efficacia della traduzione: in altre parole non si può affermare che le traduzioni sono sempre corrette. La conclusione, provvisoria, è quindi che la TA produce senza problemi bigrammi forti e rari, tra i quali trovano posto collocazioni e termini, anche se non è ancora chiaro in quale misura, fatto che rende necessario approfondire la questione attraverso analisi qualitative mirate. Il risultato è compatibile con quelli di Hansen-Schirra et al. (2019), i quali concludono, per la coppia inglese-tedesco, che «i termini specifici di un dominio spesso sono tradotti in maniera corretta perché il sistema tiene conto del contesto o è stato addestrato in maniera mirata»

(p. 125). Un lavoro di post-editing leggero (cioè orientato per lo più alla preservazione del significato) quindi dovrebbe sempre prestare attenzione al significato, ma la produzione di collocazioni analoghe a quelle della TU non pare, allo stato attuale, un problema per la TA.

Terzo risultato importante, anche questo caso collegato con la predilezione della TA per un linguaggio formulaico e frequente, è la minor percentuale negli output TA di bigrammi assenti dal corpus di riferimento rispetto alla traduzione umana. La TA, quindi, produce meno bigrammi assenti perché favorisce combinazioni frequenti, ma quando ne produce c'è la concreta possibilità che essi rappresentino produzioni creative. Mentre, però, nella TU queste sono traduzioni corrette c'è la possibilità, allo stato attuale non ancora quantificata, che le produzioni creative della TA siano traduzioni letterali potenzialmente errate, come mostrano gli esempi. Ciò implica quindi che il PE dovrebbe prestare particolare attenzione alle produzioni creative dei sistemi TA.

5. Conclusioni

Il saggio ha illustrato i primi risultati di un'indagine che intende individuare, tramite il confronto con la TU, le caratteristiche, ma anche i limiti della TA, elemento essenziale nella preparazione al post-editing. I primi risultati confermano, ma solo in parte, quelli di ricerche precedenti, confermando, come detto, l'importanza di replicare gli studi su combinazioni linguistiche e su tipologie di testi diversi. In particolare, relativamente alla combinazione inglese-italiano, la TA tende a produrre un output composto da bigrammi più frequenti e, sembrerebbe, più ridondante. L'implicazione più importante, quindi, è che la preparazione al post-editing, in particolare quello completo, dovrebbe concentrarsi sulla sintesi.

L'indagine illustrata, come detto, era preliminare e aveva soprattutto lo scopo di identificare alcune aree di interesse. Tra le aree emerse da questo primo confronto, due paiono promettenti. La prima è costituita dalle combinazioni forti e rare. Come visto la TA tende a produrre più combinazioni di questo tipo, ma il dato è ancora troppo generico. Non è stata infatti stabilita una soglia minima per identificare le collocazioni specializzate e, anzi, la soglia di 7, utilizzata in letteratura e originata da lavori sull'inglese, non pare adeguata a questo scopo perché, come visto negli esempi riportati, essa identifica anche bigrammi non di interesse, cioè collocazioni non specializzate. Il prossimo passaggio in questa direzione dovrebbe quindi consistere nell'identificare nella TU le combinazioni di interesse (come *fase preclinica* o *combustibili fossili*) per verificare il grado di sovrapponibilità tra TU e TA, cioè quante volte TA e TU usano nello stesso punto la stessa

combinazione. I risultati ottenuti alzando la soglia minima a 10, che sembrerebbe più efficace nel catturare le combinazioni specializzate, sono promettenti e non è da escludersi che la conclusione definitiva sia che la traduzione di collocazioni specializzate e terminologia non costituisca un problema per la TA, almeno nella combinazione linguistica inglese-italiano. Un secondo aspetto interessante è costituito dai bigrammi assenti. Se indagini future dovessero confermare che una quantità significativa di bigrammi assenti da un corpus di riferimento sono traduzioni letterali più o meno errate o problematiche, ciò avrebbe una grande implicazione per il PE perché significherebbe che l'analisi per bigrammi è in grado di individuare, appunto, aree problematiche nella traduzione. Sarebbe utile, in tal caso, realizzare un semplice strumento (da integrare, per esempio, negli applicativi di editing) in grado di rendere salienti questi bigrammi e aiutare il post-editor nell'individuare aree critiche nell'output della TA.

Rimane quindi, molto lavoro da fare per comprendere a fondo i sistemi di TA e demistificarne potenzialità e limiti. Anche i pochi risultati qui illustrati, infatti, sembrano indicare che i sistemi sono tutt'altro che infallibili, fatto che conferma l'importanza della preparazione dei post-editor.

Bibliografia

Bestgen, Y. 2021. Using CollGram to compare formulaic language in human and machine translation. In *Proceedings of the Translation and Interpreting Technology Online Conference*, R. Mitkov, V. Sosoni, J. C. Giguère, E. Murgolo, E. Deysel (a cura di), 174-180. Madrid/Sevilla: INCOMA.

Bestgen, Y. 2022. Comparing Formulaic Language in Human and Machine Translation: Insight from a Parliamentary Corpus. In *Proceedings of ParlaCLARIN III@ LREC2022*, D. Fiser, M. Eskevich, J. Lenardič & F.de Jong (a cura di), 101-106. Luxemburg: ELRA.

do Carmo, F. & Moorkens, J. 2020. Differentiating editing, post-editing and revision. In *Translation Revision and Post-editing*, M. Koponen, B. Mossop, I. S. Robert & G. Scocchera (a cura di), 35-49. London/New York: Routledge.

Garcia, I. 2023. Computer-aided translation systems. In *Routledge Encyclopedia of Translation Technology*, Chan S. W. (a cura di), 76-94. London/New York: Routledge.

Hansen-Schirra, S., Schaeffer, M. & Nitzke, J. 2019. Post-editing. Termini e definizioni. In *Machine Translation. Come usarla in modo professionale*, D. Lombardini, D. Cresceri & F. Mana (a cura di), 115-128. Roma: Aracne.

Koehn, P. 2017. Neural Machine Translation. <https://arxiv.org/pdf/1709.07809.pdf> (13.05.2023).

Konntinen, K., Salmi L., & Koponen, M. J. 2020. Revision and Post-Editing Competences in Translator Education. In *Translation Revision and Post-editing*, M. Koponen, B. Mossop, I. S. Robert & G. Scocchera (a cura di), 187-202. London/New York: Routledge.

Läubli, S., Sennrich, R., & Volk, M. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier & J. Tsujii (a cura di) 4791-4796. Brussels: Association for Computational Linguistics.

Lee, S. M. 2022. An investigation of machine translation output quality and the influencing factors of source texts. *ReCALL* 34(1): 81-94.

Lommel, A & Burchardt, A. 2019. La gestione della qualità nella traduzione. In *Machine Translation. Come usarla in modo professionale*, D. Lombardini, D. Cresceri D. & F. Mana (a cura di), 83-102. Roma: Aracne.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A. & Pirrelli, V. 2014. The PAISA corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, F. Bildhauer & R. Schäfer (a cura di), 36-43. Stroudsburg: The Association for Computational Linguistics.

Nitzke, J. & Hansen-Schirra, S. 2021. *A short Guide to Post-editing*. Berlin: Language Science Press.

Rivera-Trigueros, I. 2022. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation* 56(2): 593-619.

Seewald-Heeg, U. 2019. La formazione dei post-editor. In *Machine Translation. Come usarla in modo professionale*, D. Lombardini, D. Cresceri D. & F. Mana (a cura di), 109-114. Roma: Aracne.