Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano

Mirko Tavosanis
Dipartimento di Filologia, letteratura e linguistica
Università di Pisa
Via Santa Maria 36 – 56126 Pisa PI
mirko.tavosanis@unipi.it

Riassunto[1]

Il contributo presenta una valutazione delle prestazioni di Google Traduttore e di DeepL attraverso le interfacce web disponibili al pubblico. Per la valutazione è stato usato un campione di 100 frasi tratto da testi giornalistici in lingua inglese tradotti in italiano. Le traduzioni prodotte sono state valutate da esseri umani e i risultati della valutazione sono stati confrontati con il calcolo del punteggio BLEU. La valutazione umana dei sistemi automatici ha mostrato livelli di qualità vicini a quelli della traduzione umana, mentre il punteggio BLEU non ha mostrato una stretta corrispondenza con la valutazione umana.

Abstract

The paper describes an assessment of the performance of Google Translator and DeepL when the systems are used through their public web interfaces. The assessment was carried on a sample of 100 sentences from English-language journalistic texts translated into Italian. The translation outputs were evaluated by humans and the results of the evaluation were compared with the calculation of the BLEU score. Human evaluation of machine translation has shown quality levels very close to those of human translation, while the BLEU score has not shown a strict connection with human evaluation.

1 Introduzione

I sistemi di traduzione automatica stanno assumendo un ruolo sempre più importante nella vita quotidiana, da soli o integrati in altre pratiche (Bersani Berselli 2011). La loro diffusione potrebbe anche produrre innovazioni strutturali e trasformare in profondità alcuni settori lavorativi, a cominciare dall'insegnamento delle lingue straniere (Ostler 2010; Tavosanis 2018).

Tuttavia, la valutazione delle effettive prestazioni di questi sistemi rimane un problema complesso sia dal punto di vista teorico sia dal punto di vista pratico. Inoltre, la difficoltà di valutazione è considerata da tempo uno dei vincoli principali anche per lo sviluppo dei sistemi di traduzione (Pieraccini 2012, p. 275; Hajič 2008, p. 85).

Per la valutazione sono state sviluppate numerose metriche di tipo automatico o semiautomatico; la più usata in tempi recenti è stata BLEU (Papinieni e altri, 2002). Il lavoro sulle metriche è costante e, in particolare, alla valutazione delle metriche è dedicato uno degli degli *shared tasks* delle conferenze WMT (i risultati della più recente sono presentati in Fourth Conference on Machine Translation 2019, pp. 494-525).

Tuttavia, nel corso degli ultimi anni è diventato evidente che le metriche più usate non sono in realtà in grado di descrivere adeguatamente le differenze e i miglioramenti di prestazioni dei sistemi più recenti di traduzione automatica, e in particolare di quelli basati su reti neurali. Il pro-

blema può essere descritto in generale come problema di scarsa correlazione tra le metriche e il giudizio umano. Usare come punto di riferimento il giudizio umano sembra d'altra parte del tutto corretto dal punto di vista metodologico: l'obiettivo dei sistemi di traduzione è principalmente quello di fornire traduzioni che gli esseri umani considerino di buon livello.

In particolare, la non perfetta correlazione tra BLEU e il giudizio umano è stata notata da tempo (per esempio: Callison-Burch, Osborne e Koehn 2006) e diversi valutatori hanno ribadito la necessità di considerare la valutazione umana come primaria (Callison-Burch e altri 2008, p. 72). La situazione è stata probabilmente resa meno evidente anche dall'abitudine frequente di valutare sistemi diversi confrontando le prestazioni tra di loro e non su una scala assoluta; tuttavia, questa prassi non è mai stata l'unica e i sistemi presentati nella principale campagna di valutazione sulla traduzione automatica, i task WMT, sono valutati solo con giudizi assoluti, non con giudizi relativi.

Il problema si è mostrato con particolare evidenza negli ultimi anni, in seguito alla rapida introduzione dei sistemi di traduzione basati su reti neurali. BLEU, come i sistemi di traduzione statistica (PB-SMT), basa il proprio funzionamento sugli *n*-grammi. Si ritiene però che questo meccanismo mostri un "inherent bias" contro i sistemi che adottano meccanismi di traduzione non basati su *n*-grammi, quali appunto i sistemi basati su reti neurali (Way 2018, p. 170).

Diverse verifiche hanno mostrato che in pratica BLEU sottovaluta fortemente i risultati dei sistemi di traduzione a reti neurali (Bentivogli e altri 2018a; Shterionov e altri 2018). Naturalmente, la validità di queste verifiche può essere relativizzata alle caratteristiche di specifici campioni. Tuttavia, i dati oggi disponibili giustificano l'idea che BLEU non possa essere usato come indicatore generale di qualità di questi sistemi.

In questo contesto non mancano dichiarazioni in cui si rivendica il raggiungimento della "parità" tra traduzione automatica e traduzione umana per alcuni sistemi commerciali (Hassan e altri 2018). Le verifiche indipendenti non hanno però al momento confermato questi risultati; al contrario, hanno evidenziato differenze significative (Toral e altri 2018).

Dichiarazioni del genere mostrano comunque l'utilità di una valutazione esterna delle prestazioni dei sistemi più usati. Anche il presente contributo concorre a questa attività, documentando lo stato delle cose per prodotti di ampia diffusione e in un contesto d'uso reale per una lingua su cui le valutazioni sono state finora piuttosto ridotte. Alcuni testi generati con traduzione automatica sono stati quindi sottoposti a valutazione umana, assieme a prodotti di traduttori umani, con l'obiettivo di:

1. fornire una valutazione umana delle prestazioni (assolute e non relative) di due diversi sistemi
2. confrontare i risultati della valutazione umana con quelli della valutazione ottenuta attraverso BLEU

2 Il contesto della traduzione

Le verifiche descritte di seguito sono state compiute usando due sistemi liberamente accessibili al pubblico e spesso indicati come i migliori nel loro genere: Google Traduttore e DeepL.

I due sistemi non sono forse i più utilizzati su scala mondiale. Si può pensare che Google Traduttore sia il sistema più comunemente usato, ma in assenza di indicazioni ufficiali è possibile che questo primato vada in realtà assegnato al sistema di traduzione automatica di Facebook.

DeepL non solo è sicuramente meno noto di Google Traduttore, ma è probabilmente meno usato anche di un quarto sistema di traduzione, Microsoft Translator. Tuttavia, DeepL è frequentemente segnalato come uno dei migliori prodotti della sua categoria e nelle valutazioni con BLEU ha ottenuto negli ultimi anni punteggi spesso superiori a quelli di Google Traduttore (Heiss e Soffritti 2018).

3 Google Traduttore

Le origini di Google Traduttore risalgono al 2003, quando il servizio venne lanciato con il nome di Google Translate. In seguito il servizio è stato rinominato, per l'italiano, come Google Traduttore.

Alle origini, il sistema si basava su prodotti SYSTRAN. Già nel 2006 Google iniziò comunque a usare un sistema di traduzione statistica sviluppato in proprio, GSMT (Google Statistical Machine Translation). Caratteristica di questo sistema è l'uso dell'inglese come lingua ponte, per cui le traduzioni tra lingue diverse dall'inglese vengono fatte passando comunque da una versione in lingua inglese e poi ritradotte – con un peggioramento significativo della qualità rispetto alle traduzioni dirette da e verso l'inglese (una sintesi delle fasi di sviluppo è presentata in Tavosanis

2018, pp. 95-96). Le lingue coperte sono aumentate rapidamente e, anche se nell'ultimo anno non ne sono state aggiunte di nuove, nel luglio del 2019 risultavano in tutto 103 (la lista completa è disponibile sul sito https://translate.google.com/), traducibili reciprocamente per un totale di poco più di 10.000 diverse combinazioni.

Nel frattempo, Google ha sviluppato il prodotto inserendovi caratteristiche di intelligenza artificiale basate sull'apprendimento automatico e sulle reti neurali. Il 15 novembre 2016 è stato quindi annunciato il passaggio di una parte dei servizi di Google Traduttore dal sistema GSMT a quello GNMT (Google Neural Machine Translation). Rispetto al precedente, GNMT ha il vantaggio di tradurre, secondo gli sviluppatori, frase intere e non spezzoni di frase, curando in particolare la coesione grammaticale, che nei sistemi precedenti non sempre veniva rispettata (Turovsky 2016). Nel marzo 2017, il sistema GNMT era già disponibile per traduzioni tra otto lingue: inglese, cinese, francese, tedesco, giapponese, coreano, portoghese, spagnolo e turco. Nell'aprile dello stesso anno è stato esteso ad altre lingue europee, tra cui l'italiano (Google 2017).

4 DeepL

Realizzato dall'azienda tedesca DeepL GmbH, il sistema di traduzione DeepL è stato reso disponibile al pubblico nell'agosto del 2017 (sito: https://www.deepl.com/). Rispetto a Google, copre un numero relativamente ridotto di lingue, tutte di origine indoeuropea: italiano, inglese, tedesco, francese, spagnolo, portoghese, olandese, polacco e russo. Dal punto di vista tecnico, l'azienda ha dichiarato che il sistema di traduzione si basa su reti neurali, ma non ha fornito altre informazioni.

5 Procedura di valutazione

Per la valutazione del lavoro è stato usato un corpus di articoli di quotidiani e periodici. Tale scelta è stata fatta in base a diversi fattori:

- Importanza, in quanto l'italiano giornalistico è centrale nell'architettura dell'italiano contemporaneo (Bonomi 2002, Berruto 2012)
- Verosimiglianza, in quanto la traduzione di articoli di questo tipo è un impiego realistico dei sistemi descritti, nella loro versione rivolta all'utente generico e resa disponibile attraverso un'interfaccia web
- Disponibilità, in quanto è facile ottenere ragionevoli quantitativi di articoli in doppia versione, originali e tradotti
- Praticità, in quanto le traduzioni degli articoli spesso hanno una corrispondenza 1:1 tra le frasi del testo originale e quelle del testo tradotto.

Il lavoro è stato condotto su un campione di 100 frasi, valutate separatamente (da valutatori diversi) per l'adeguatezza (*adequacy*) e per la fluenza (*fluency*). Anche se i risultati delle verifiche WMT hanno confermato la maggior rilevanza dell'adeguatezza (Bentivogli e altri 2018b: 62), le due valutazioni diverse sono state conservate per verificare l'esistenza di differenze nei prodotti commerciali. Va comunque notato che dal punto di vista dell'adeguatezza, nonostante sia teoricamente possibile che una frase tradotta con sistemi a reti neurali non abbia nulla a che fare contenutisticamente con il testo di partenza, nella pratica non si è prodotto nessun caso di questo genere.

Le scale utilizzate sono state:

Adeguatezza

1. Il contenuto informativo dell'originale è stato completamente alterato
2. È stata trasmessa una parte del contenuto informativo, ma non la più importante
3. Circa metà del contenuto informativo è stata trasmessa
4. La parte più importante del contenuto informativo originale è stata trasmessa
5. Il contenuto informativo è stato tradotto completamente

Fluenza

1. Impossibile da ricondurre alla norma
2. Con più di due errori morfosintattici
3. Con non più di due errori morfosintattici e/o molti usi insoliti di collocazioni
4. Con non più di un errore morfosintattico e/o un uso insolito di collocazioni
5. Del tutto corretta

All'interno del campione sono state inserite casualmente frasi provenienti da un corpus di 15 articoli di quotidiani e periodici, scelti casualmente sulla base della disponibilità online sia del testo originale sia di una traduzione in lingua italiana. In alcuni casi, le traduzioni umane prese in esame sono opera di volontari ma sono comunque di buon livello qualitativo. I testi originali in inglese

sono stati ripuliti e sottoposti alle interfacce web di Google Traduttore e DeepL. Poiché queste interfacce, nella versione liberamente accessibile, accettano testi di una lunghezza massima di 5000 caratteri, i testi più lunghi sono stati scomposti in blocchi di lunghezza inferiore, rispettando i confini di frase (e spesso di capoverso). I blocchi stessi sono stati poi sottoposti individualmente ai sistemi.

Al termine della procedura, per ogni articolo erano quindi disponibili:

1. Il testo originale in lingua inglese
2. La traduzione umana
3. La traduzione prodotta da Google
4. La traduzione prodotta da DeepL

Le frasi da esaminare sono state selezionate in modo causale. Sono poi state sottoposte ai valutatori in ordine casuale e senza indicazioni sulla loro origine: i valutatori non avevano quindi elementi esterni per decidere se l'origine di una singola frase era un traduttore umano, DeepL o Google. Nella valutazione per adeguatezza le frasi erano accompagnate dal testo originale in lingua inglese, secondo l'orientamento *DA-src* (Bentivogli e altri 2018b: 62), mentre nella valutazione per fluenza era disponibile solo il testo italiano. La valutazione è stata eseguita su carta, in condizioni controllate, per un tempo medio di un'ora per ogni campione.

I valutatori sono stati complessivamente 14: 6 hanno valutato l'adeguatezza, 8 la fluenza. La valutazione della fluenza è stata condotta su un campione più esteso di 147 frasi, per rendere la lunghezza dell'attività paragonabile a quella della valutazione dell'adeguatezza. Ai fini della valutazione sono state tuttavia usate solo le 100 frasi coincidenti con frasi valutate per adeguatezza.

Il gruppo dei valutatori era interamente formato da studenti del corso di laurea magistrale in Informatica umanistica dell'Università di Pisa. Tutti i valutatori avevano l'italiano come lingua madre e disponevano di una conoscenza della lingua inglese di livello B2 o superiore. Nessuno di loro aveva esperienza di attività redazionale o di revisione di traduzioni e nessuno è stato coinvolto nella fase di scelta e preparazione degli articoli.

Per migliorare l'omogeneità del risultato, una settimana prima della valutazione vera e propria è stata fatta una sessione di prova con i valutatori interessati. In questa sessione sono state valutate frasi diverse da quelle esaminate in seguito. I punteggi assegnati sono stati discussi sulla base dei testi, cercando di arrivare quanto più possibile alla condivisione di parametri per il lavoro effettivo.

6 Esito della valutazione

Nel giudizio finale la varianza dei giudizi è stata piuttosto ridotta. Le medie della varianza calcolata su ogni singola frase sono state infatti:

	Adeguatezza	Fluenza
Google	0,3982	0,4631
DeepL	0,4312	0,4375
Umano	0,4320	0,3432
Totale	0,4192	0,4243

Tabella 1: Varianza media nei giudizi per frasi.

Per quanto riguarda la fluenza, la varianza massima (0,1728) si è avuta nei giudizi per questa traduzione, con sei punteggi 4 e due punteggi 3:

Originale: As Rahme served a frugal dish of rice in vine leaves, her son unspooled a familiar Palestinian narrative.

Traduzione DeepL: Mentre Rahme serviva un frugale piatto di riso in foglie di vite, suo figlio ha sboccato un racconto familiare palestinese.

Più consistente è stata la varianza massima per l'adeguatezza, con due frasi che hanno ottenuto il livello di 1,9592:

Originale: And though people can be induced to use social media addictively, while ordering Deliveroo night after night, pausing only to take an Uber to the pub, wedding addiction remains a rarity.

Traduzione Google: E anche se le persone possono essere indotte a usare i social media in modo assopito, mentre ordinano Deliveroo notte dopo notte, facendo una pausa solo per portare un Uber al pub, la dipendenza da matrimonio rimane una rarità.

Originale: And now the Trump administration, having failed to repeal the ACA when Republicans controlled Congress, is suing to have the whole thing declared unconstitutional in court – because what could be a better way to start off the 2020 campaign than taking insurance away from 20 million Americans?

Traduzione umana: E ora l'amministrazione Trump, non essendo riuscita ad abrogare l'ACA quando i repubblicani controllavano il Congresso, sta facendo causa per far dichiarare l'intera cosa

incostituzionale in tribunale - perché quale modo migliore di togliere l'assicurazione a 20 milioni di americani per iniziare la campagna del 2020?

Le frasi oggetto di valutazione sono state poi riassemblate in tre diversi documenti, a seconda dell'origine, ed è stato calcolato il punteggio BLEU per i prodotti della traduzione automatica, confrontati con la traduzione umana. La valutazione risultante è stata:

Traduttore	N. frasi	Media adeguatezza	Media fluenza	BLEU
Google	37	4,15	3,90	0,2538
DeepL	39	4,30	3,94	0,3254
Umano	24	4,60	4,46	n. a.

Tabella 2: Valutazione complessiva delle traduzioni.

Per la fluenza, va notato che il punteggio 5 è stato assegnato all'unanimità solo a pochissime frasi. Tuttavia, alcune frasi sia di Google sia di DeepL hanno ottenuto questo punteggio massimo, cosa che viceversa non è successa per le traduzioni umane. Questo giudizio è stato assegnato soprattutto a frasi brevi, ma non solo a esse. Per esempio, sono state valutate 5 queste traduzioni:

Originale: Which is weird, because the truth is that everyone's judging everyone else's relationships all the time.

Traduzione DeepL: Il che è strano, perché la verità è che tutti giudicano sempre le relazioni altrui.

Originale: In an attempt to avert this awful fate, the American Medical Association launched what it called Operation Coffee Cup, a pioneering attempt at viral marketing.

Traduzione Google: Nel tentativo di scongiurare questo terribile destino, l'American Medical Association lanciò quella che chiamò Operation Coffee Cup, un tentativo pionieristico di marketing virale.

7 Esame dei risultati

In risposta alle domande presentate nel paragrafo 1 è innanzitutto notevole l'alto livello raggiunto da entrambi i sistemi. Nessuno dei due può essere considerato all'altezza della traduzione umana, e non mancano i casi di frasi tradotte in modo molto insoddisfacente, come questa (valutazione media 1,43):

Originale: If you are used to the boil-them-whole, admire, tug-leaf-by-leaf, scape-with-bottom-teeth school of artichoke preparation and eating, it comes as a shock when you first see Romans deal, in typically direct style, with their favourite vegetable.

Traduzione Google: Se sei abituato a bollire tutto, ammira, rimorchia la foglia per pianta, scolpisci i denti di fondo con la preparazione e il consumo di carciofo, diventa un vero shock quando vedi per la prima volta i romani, in genere stile diretto, con il loro vegetale preferito.

Tuttavia, nel complesso, colpisce che per esempio per l'adeguatezza la distanza relativa tra la traduzione umana e DeepL sia pari solo al 6,5%. Il dislivello per quanto riguarda la fluenza è maggiore, ma rimane comunque molto contenuto.

I dati confermano inoltre la superiorità delle prestazioni di DeepL già segnalata da diverse fonti, anche se la differenza con Google è molto contenuta. Il margine relativo di vantaggio di DeepL è infatti solo del 3,5% per l'adeguatezza e dell'1% per la fluenza.

Va notato che la differenza nella composizione del campione potrebbe spiegare parte dei risultati; all'interno di eventuali prove future sarebbe sicuramente opportuno sottoporre alla valutazione campioni omogenei. Tuttavia, per esempio, la lunghezza media delle frasi, che influenza in negativo la qualità della traduzione automatica, non solo è molto simile nei due campioni, ma è superiore nel caso del sistema che ha ottenuto la valutazione più alta. Il campione usato per DeepL ha infatti una lunghezza media di 25,79 token per frase, mentre in quello usato per Google il valore equivalente è di 25,03.

Per quanto riguarda BLEU, la correlazione con la valutazione umana risulta davvero debole. Il ridotto scarto tra Google e DeepL nella valutazione umana diventa infatti una differenza relativa del 22% con BLEU.

Soprattutto, però, è notevole la differenza rispetto ai punteggi BLEU per la traduzione umana spesso indicati in bibliografia (Papinieni e altri 2002), che si aggirano attorno a 0,6. Per DeepL questo corrisponderebbe a una differenza relativa del 45,8%, difficile da considerare rappresentativa della differenza tra i risultati su una scala di giudizio assoluta.

Va inoltre notato che negli ultimi anni i punteggi BLEU di sistemi come Google o Microsoft Translator si sono spesso collocati tra 0,2 e 0,4 (Tavosanis 2018). In questo contesto, se il punteggio di DeepL è piuttosto elevato, quello di Google si avvicina alla media.

8 Conclusioni e sviluppi futuri

Il lavoro descritto qui rappresenta una delle prime concretizzazioni di un progetto più ampio, dedicato a studiare le possibilità di inserimento strutturale dei traduttori automatici nella pratica didattica delle lingue partendo dall'analisi delle prestazioni e della possibilità di integrare facilmente i prodotti nel percorso di un traduttore in formazione. Nel giro di pochi mesi dovrebbero essere quindi disponibili valutazioni più estese. Per la traduzione italiana, queste valutazioni potrebbero essere di particolare interesse, considerando non solo la rapidità dei miglioramenti recenti ma anche il fatto che l'italiano è stato relativamente poco rappresentato nelle analisi condotte finora.

Per gli sviluppi futuri, l'aver preso in esame un unico genere testuale, per quanto variato, è un limite evidente dell'analisi (Burchardt e altri 2017: 159-160): l'estensione della valutazione a tipologie diverse rispetto all'articolo di quotidiano o periodico potrebbe facilmente portare a risultati molto diversi da quelli descritti qui. L'inclusione di altri generi testuali rappresenta quindi senz'altro il requisito più importante nella progettazione di un lavoro di valutazione su scala più estesa. In quest'ottica, sembra particolarmente interessante l'estensione del lavoro a testi specialistici.

Bibliografia

Bentivogli, Luisa, e altri (2018a). *Neural versus phrase-based mt quality: An in-depth analysis on english–german and english–french*. In *Computer Speech & Language*, 49, pp. 52-70.

Bentivogli, Luisa, e altri (2018b). *Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment*. In Proceedings of the 15th International Workshop on Spoken Language Translation, Iwslt, pp. 62-69.

Berruto, Gaetano (2012). *Sociolinguistica dell'italiano contemporaneo. Nuova edizione.* Roma: Carocci.

Bersani Berselli, Gabriele (a cura di, 2011), *Usare la traduzione automatica*. Bologna: CLUEB.

Bonomi, Ilaria (2002). *L'italiano giornalistico. Dall'inizio del '900 ai quotidiani online*. Firenze: Cesati.

Burchardt, Aljoscha, e altri (2017). *A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engine*s. In *The Prague Bulletin of Mathematical Linguistics*, 108, pp. 159-70.

Callison-Burch, Chris, e altri (2008). *Further meta-evaluation of machine translation*. In *Proceedings of the third workshop on statistical machine translation*, Association for Computational Linguistics, pp. 70-106.

Callison-Burch, Chris, Miles Osborne e Philipp Koehn (2006). *Re-evaluation the role of BLEU in machine translation research*. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 249-256.

Fourth Conference on Machine Translation (2019), *Proceedings of the Conference, Volume 2: Shared Task Papers, Day 1*, Stroudsburg, ACL <http://www.statmt.org/wmt19/pdf/53/WMT-2019-2.pdf>.

Google (2017). *Translation API Language Support*. Sito Google <https://cloud.google.com/translate/docs/languages>.

Hajič, Jan (2008). *Linguistics Meets the Exact Sciences*. In *A companion to digital humanities*, a cura di Susan Schreibman, Ray Siemens e John Unsworth, Hoboken, John Wiley & Sons, pp. 79-87.

Hassan, Hany, e altri (2018). *Achieving human parity on automatic Chinese to English news translation.* arXiv preprint arXiv:1803.05567 (2018).

Heiss, Christine e Marcello Soffritti (2018). *DeepL Traduttore e didattica della traduzione dall'italiano in tedesco–alcune valutazioni preliminari*. In *Translation and Interpreting for Language Learners (TAIL). Lessons in honour of Guy Aston, Anna Ciliberti, Daniela Zorzi*, a cura di Laurie Anderson, Laura Gavioli e Federico Zanettin, Milano, AItLA, pp. 241-258.

Ostler, Nicholas (2010). *The Last Lingua Franca. English until the Return of Babel*. Londra: Allen Lane.

Papinieni, Kishore, e altri (2002). *BLEU: A Method for Automatic Evaluation of Machine Translation*. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, ACL, Stroudsburg, pp. 311-8.

Pieraccini, Roberto (2012). *The Voice in the Machine. Building Computers that Understand Speech*. Boston: MIT Press.

Shterionov, Dimitar, e altri (2018). *Human versus automatic quality evaluation of NMT and PBSMT*. In *Machine Translation*, 32, 3, pp. 217-235.

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

Toral, Antonio, e altri (2018). *Attaining the unattainable? Reassessing claims of human parity in neural machine translation*. arXiv preprint arXiv:1808.10432.

Turovsky, Barak (2016). *Found in translation: More accurate, fluent sentences in Google Translate*. Google Blog <https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/>.

Way, Andy (2018). *Quality expectations of machine translation.* In *Translation Quality Assessment*, Springer, Cham, pp. 159-178.