UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in
Lingue Moderne per la Comunicazione e Cooperazione Internazionale
Classe LM-38

Tesi di Laurea

Traduzione umana e automatica a confronto: un'analisi quantitativa di articoli tradotti dal russo all'italiano

Relatrice
Prof.ssa Arjuna Tuzzi
Correlatrice
Prof.ssa Linda Torresin

Laureanda
Francesca Ragogna
n° matr.2005679 / LMLCC

Anno Accademico 2021/2022

Indice

Introduzione

La traduzione è da secoli l'attività umana che permette l'incontro fra lingue e culture diverse. Si tratta essenzialmente della produzione di un testo di arrivo (o metatesto) a partire da un testo di partenza (o prototesto). Come afferma il linguista, semiologo e traduttore italo-americano Roman Jakobson (Bertazzoli 2020, p.26), una traduzione non avviene solo tra lingue con sistemi linguistici differenti, ma anche tra testi appartenenti allo stesso codice. Il linguista individua tre diversi tipi di traduzione: *intralinguistica*, *intersemiotica*, *interlinguistica*. La prima indica la riformulazione di un testo utilizzando altre parole appartenenti allo stesso sistema linguistico. La seconda individua la trasposizione di un testo da un mezzo ad un altro, ad esempio l'adattamento cinematografico di un romanzo. La terza tipologia, invece, indica ciò che tutti noi comunemente intendiamo per traduzione, ovvero la produzione di un testo di arrivo a partire dall'interpretazione di un testo di partenza elaborato con segni linguistici differenti. Negli ultimi decenni, con lo sviluppo di programmi sempre più sofisticati, l'interpretazione di questi segni linguistici è stata delegata a *software* di traduzione automatica, che traducono in pochi secondi testi scritti od orali da una lingua ad un'altra. La domanda allora sorge spontanea, la traduzione è la semplice transcodifica di segni linguistici o qualcosa di più?

Secondo Raffaella Bertazzoli, la traduzione è molto di più di una semplice trasposizione di segni. Si tratta di "uno scambio con l'altro non semplicemente strumentale, ma [di un incontro] espressivo e comunicativo tra lingue e sistemi culturali [diversi]" (Bertazzoli 2020, p. 7). Se questo fosse vero, il traduttore umano potrebbe essere ancora considerato una figura indispensabile che non verrà completamente sostituita in futuro? O in alcune tipologie di testo, come l'articolo di giornale, la traduzione automatica riesce a raggiungere i livelli della traduzione umana? A queste domande si cercherà di rispondere all'interno della presente tesi.

Lo scopo della ricerca svolta per questa tesi è di confrontare e analizzare testi originali scritti in russo con le relative traduzioni italiane, realizzate da esseri umani e da due *software* di traduzione automatica: *Google Translate* e *DeepL*. Il genere testuale analizzato è l'articolo di giornale pubblicato online. Verranno effettuate una serie di analisi quantitative dei testi per verificare se è possibile studiare

determinati aspetti traduttivi solo attraverso l'uso di *software* di analisi testuale quantitativa.

In primo luogo, verrà analizzata la correttezza delle traduzioni di acronimi, abbreviazioni, prestiti stranieri, forestierismi e traslitterazioni dall'alfabeto cirillico a quello latino e viceversa. Inoltre, si studierà la traduzione di termini, che almeno in teoria, prevedono una sola traduzione possibile (es. Кремль > Cremlino), controllando se il numero delle occorrenze di tali parole sia lo stesso sia nei testi originali che nelle traduzioni. Si confronterà, quindi, la capacità dei programmi di traduzione automatica di tradurre correttamente diverse tipologie di parole, considerando come variante corretta la traduzione umana. Infine, oltre alle differenze tra HT (*human translation* o traduzione umana) e MT (*machine translation* o traduzione automatica), si osserveranno anche le differenze tra i due programmi di traduzione automatica. Il *software* utilizzato per questa analisi sarà *AntConc*. In questa prima fase ci si aspetta che determinate parole, come forestierismi, prestiti stranieri o acronimi, non vengano sempre tradotti correttamente dai programmi di *machine translation*, perché la traduzione automatica non è precisa nel tradurre costruzioni lessicali, semantiche e pragmatiche complesse (Li et al. 2014, p. 190). Ad esempio, ci si può aspettare che il forestierismo "телеграм канал[1]" non sia sempre riconosciuto come nome proprio di un *social network*, ma venga tradotto a volte con "canale telegrafico". D'altra parte, ci si aspetta che termini come "Кремль"[2] o "Коронавирус"[3], i quali teoricamente dovrebbero avere una traduzione univoca ed essere presenti tanto nel testo russo quanto nelle traduzioni, abbiano in realtà un numero di occorrenze diverso nella traduzione umana. Una delle cause può essere ricondotta alla tipologia di traduzione che viene effettuata. Nella presente tesi, infatti, si analizzerà la traduzione per il web, nello specifico la traduzione di articoli pubblicati su un sito web. In questo caso l'obiettivo del traduttore è quello di "adattare un testo dalla

[1] trad. Canale Telegram

[2] trad. Cremlino

[3] trad. Coronavirus

lingua di origine a quella di destinazione prendendo in considerazione la cultura del Paese ricevente, il contesto sociale […] oltre che il medium utilizzato" (Torresin 2022, p. 20). Un esempio di adattamento può essere la traduzione di "коронавирус" in "pandemia", anche se la traduzione letterale è "coronavirus".

In secondo luogo, per confrontare la semplicità lessicale delle traduzioni umane e quelle automatiche, verranno confrontati due parametri: il *Type Token Ratio* (cfr. cap. 2.2) e la lunghezza media di una frase nelle singole traduzioni umane e automatiche (cfr. Ondelli & Viale 2010, pp. 3-5). Questa analisi è presente nella ricerca di Kunilovskaya et al. (2018), in cui sono state confrontate traduzioni di traduttori professionisti con quelle di studenti di traduzione. Nello studio è stato scoperto che tendenzialmente gli studenti tendono a produrre traduzioni meno complesse lessicalmente e ad essere più influenzati dai testi originali rispetto ai traduttori professionisti. Date le caratteristiche della traduzione automatica, che tende a tradurre in modo letterale ed è altamente influenzata dal testo di partenza (Ibanez 2021), si ipotizza che i risultati di Kunilovskaya et al. (2018) sulle traduzioni di studenti siano simili a quelle delle traduzioni automatiche. Ci si aspetta, quindi, che le traduzioni automatiche abbiano una minor varietà lessicale e sintattica rispetto alle traduzioni umane (cfr. Kunilovskaya et al. 2018).

Successivamente, si proseguirà con un'analisi degli argomenti (*topic*) contenuti negli articoli originali russi, per confrontarli con il numero e la tipologia di *topic* presenti nelle traduzioni umane e automatiche. Si verificherà se il numero di *topic* è rispettato nelle *human translation* e *machine translation* o se sussistono delle differenze tra gli originali e le traduzioni e tra HT e MT. In questo modo si potrà capire se i traduttori umani si prendono delle libertà allontanandosi dal testo originale o meno. In questa fase di lavoro si utilizzerà il *software Iramuteq*. Ci si aspetta che il numero e la tipologia di *topic* delle traduzioni automatiche siano le stesse degli originali, perché traducono in modo più letterale rispetto a un essere umano (Ibanez 2021). Al contrario, le traduzioni umane potrebbero avere categorie leggermente diverse dovute alle tecniche di traduzione di adattamento, messe in atto da un traduttore umano quando traduce per un pubblico che non conosce le tradizioni e la cultura di quel paese straniero (cfr. Torresin 2022, p. 20).

In seguito, verranno studiate le similitudini e le differenze stilistiche tra i testi originali russi e le relative traduzioni umane e automatiche. Per questa analisi verrà utilizzato il pacchetto *stylo* all'interno del programma *R* e si farà uso della *cluster analysis*. Si ipotizza che i traduttori umani siano meno influenzati dallo stile di scrittura degli autori originali dei testi, perché non traducono sempre in modo letterale come i traduttori automatici. Al contrario, si suppone che un traduttore automatico produca traduzioni con uno stile più simile all'autore originale del testo in russo. Si sottolinea che, in questa analisi, con il termine "stile" si intende solo l'insieme di bigrammi, trigrammi o delle prime *n* parole più frequenti all'interno di un testo (Eder et al. 2016, pp. 107-106).

Nell'ultima fase di ricerca, verrà fatto uso del metodo *machine learning* per verificare se e in che misura un algoritmo sia in grado di distinguere la mano del traduttore umano da quella del traduttore automatico (Google Translate o DeepL). Per questo scopo si utilizzeranno due algoritmi di *Authorship Attribution* (*Support Vector Machine* e *Random Forest*), che verranno eseguiti all'interno del programma *R*. Sulla base degli studi precedenti (cfr. Li et al. 2015; Fu et al. 2021) è stato confermato che, attraverso il *machine learning*, un algoritmo sia in grado di distinguere la traduzione automatica da quella umana. Di conseguenza ci si aspetta che anche gli algoritmi sopracitati possano individuare correttamente le traduzioni umane.

Infine, si ipotizza che in tutte le analisi sarà evidente la superiorità del *software* di traduzione DeepL rispetto a Google Translate. Entrambi i programmi sono in grado di elaborare bilioni di parole sia nel testo di partenza che nel testo di arrivo (Costa-jussa et al. 2012, p. 254), ma solo in DeepL vengono effettuati regolarmente dei test qualitativi da traduttori professionisti, che giudicano la resa migliore tra un gruppo di traduzioni automatiche eseguite da DeepL e altri programmi concorrenti. È stato verificato che DeepL supera la concorrenza in un rapporto di 3:1. (cfr. https://www.deepl.com/en/whydeepl).

1. Contesto

1.1. Traduzione Automatica: definizione e tipologie

Con il termine Traduzione Automatica o Machine Translation (MT) ci si riferisce a una branca della linguistica computazionale che studia l'uso di programmi informatici per tradurre testi da una lingua di partenza a una lingua d'arrivo (Costa-jussa et al. 2012, p. 246). I *software* di traduzione automatica non vanno confusi con i *software* di traduzione assistita, che hanno il compito di aiutare il traduttore nel suo lavoro senza, però, fornirgli la traduzione del testo (Hutchins 2001, p. 6). Esistono diverse tipologie di MT e un modo per classificarle è analizzare la loro metodologia di base. Si possono individuare due approcci differenti di MT: i *rule-based machine translation* (RBMT) e i *corpus-based approach*.

La RBTM si fonda su una serie di regole linguistiche individuate da un gruppo di esperti linguistici umani. L'algoritmo alla base dei sistemi RBTM è costituito da diversi step, che all'interno di questa tesi riassumeremo nelle seguenti fasi (cfr. Costa-jussa et al. 2012):

1. Analisi del testo: il programma segmenta il testo in gruppi di frasi seguendo delle indicazioni preimpostate dagli esperti linguistici, controlla i segmenti in un dizionario e fornisce dei risultati sulle possibili traduzioni;
2. Transfer da un codice linguistico all'altro: il *software* trova la traduzione più adatta tra le porzioni di testo che prevedono più di una traduzione e risolve eventuali problemi di disambiguazione nel testo di partenza per trovare l'equivalente corretto nella lingua target;
3. Produzione della traduzione: il programma trova le divergenze grammaticali tra la lingua di partenza e la lingua di arrivo (e.g. concordanze di genere e numero), crea sequenze di *chunk* (porzioni di testo) ordinate e sostituisce i *chunk* della lingua di partenza nella lingua di arrivo. Infine, viene effettuato un controllo ortografico per correggere eventuali refusi.

Questo approccio venne utilizzato per creare i primi *software* di traduzione automatica, ed è possibile trovarlo ancora oggi in programmi come *Apertium* o *Translendium*. Il problema di questo tipo di traduzione automatica è che richiede

un grande investimento di tempo e di personale specializzato, per questo motivo non è più molto utilizzato (Costa-jussa et al. 2012, p. 249).

La maggior parte dei programmi di traduzione automatica, infatti, si basavano fino a pochi anni fa sul *corpus-based approach*. Quest'ultimo estrae in modo automatico la propria 'conoscenza' partendo dall'analisi di esempi di traduzione presi da *corpus* di testi paralleli creati da esperti umani (Costa-jussa et al. 2012, p. 247). All'interno del *corpus-based approach* possono essere distinte due tipologie di traduzione automatica: *example-based machine translation* (EBMT) e *statistical-based machine translation* (SMT).

EBTM utilizza come database *corpus* bilingui di testi paralleli e traduce per analogia nuovi testi (Costa-jussa et al. 2012, p. 247). Anche SMT parte da testi paralleli, ma fa uso del *machine learning* per analizzare i testi di partenza e le traduzioni umane per creare soluzioni nuove e non previste dal *corpus* di testi iniziale. Questa tipologia di MT si basa sulla traduzione di una stringa[4] del testo di partenza (*source string*) in una stringa nella lingua di arrivo (*target string*) cercando tra tutte le possibili *target string* quella che è ha la probabilità maggiore di essere la traduzione più fedele all'originale. Questo processo può essere sintetizzato con la seguente formula:

$$\tilde{t}_1^I = \underset{t_1^I}{\operatorname{argmax}} P(t_1^I | s_1^J)$$

Fig. 1 (da Costa-jussa, et al. 2012, p. 249)

Dove *t* e *s* indicano rispettivamente *target string* e *source string*, mentre *I* e *J* sono il numero delle parole delle frasi *target* e *source*. Con *argmax* si intende un'operazione specifica del machine learning, che permette in generale di trovare un argomento (qui una stringa) con il grado di probabilità più alto possibile. I primi programmi SMT moderni hanno implementato tale approccio utilizzando un modello lineare-logaritmico per ottenere la traduzione di una stringa con una

[4] Stringa: "sequenza finita di caratteri alfanumerici registrata in memoria o in un altro supporto (nastro, disco, ecc.), che rappresenta dati in forma codificata" (https://www.treccani.it/vocabolario/stringa2/)

probabilità di correttezza maggiore (Li, et al. 2015, p.1). Questo sistema può essere sintetizzato nel seguente schema:

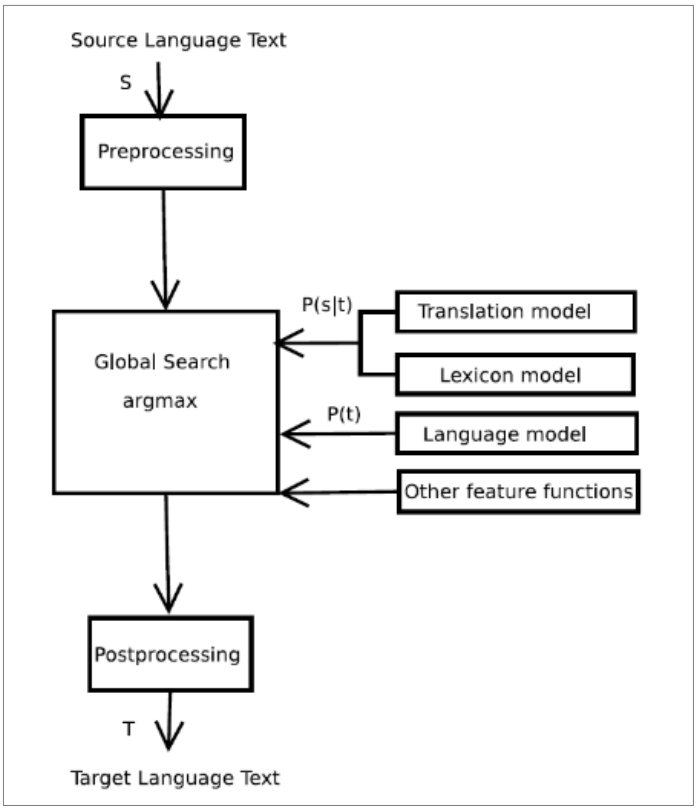

Fig. 2 (da Costa-jussa, et al. 2012, p. 250)

Con il termine *source language text* si intende il testo di partenza che, una volta inserito nel *software* di MT, viene diviso in unità fondamentali di traduzione (fase di *preprocessing*), che possono essere sia parole (per i vecchi sistemi *word-based*) sia frasi (nei sistemi più recenti *phrase-based*). Le unità di traduzione vengono riconosciute dal sistema come stringhe in base alla lunghezza media delle frasi (o parole nei sistemi *word-based*) stimata sui testi bilingui allineati, usati per 'allenare' il programma alla traduzione. Nella fase di *global search argmax* vengono utilizzati diversi modelli. Il primo è il *translation model* costituito da tutte le traduzioni possibili di una stringa contenute nel *training corpus*. Inoltre, i sistemi SMT utilizzano un modello lessicale e uno linguistico (*lexicon model* e *language model*), che permettono rispettivamente di calcolare la probabilità di tradurre una stringa parola per parola e la probabilità che una stringa si presenti all'interno di un'altra. Infine, i programmatori possono decidere di aggiungere delle funzioni aggiuntive

(*other feature functions*) per migliorare la performance del *software* (Costa-jussa et al. 2012, p. 250-251). Il problema di questo approccio è che risulta difficile da applicare nei casi di lingue meno utilizzate, come l'islandese, perché potrebbe non esistere una quantità di testi sufficiente per un corretto funzionamento della SMT. In questi casi la *rule-based machine translation* potrebbe essere più proficua (Ibid., p. 249). Data la grandezza del *corpus* con cui i programmi SMT lavorano, lo studio condotto da Costa-jussa et al. (2012), ha confermato che la *statistical-based machine translation* riesce a lavorare meglio sull'aspetto semantico rispetto ai *rule-based approaches*, probabilmente grazie all'enorme quantità di dati contenuti nel *training corpus* che permettono di avere un contesto d'uso delle parole più elevato.

Nonostante la diffusione e la precisione dei *software* di traduzione SMT, negli ultimi anni si è sviluppato un ulteriore tipologia di traduzione automatica: la *neural machine translation* (NMT). Il decennio scorso molte aziende, tra cui *Google* a partire dal 2016, hanno iniziato ad abbandonare gradualmente la traduzione *statistical-based* per utilizzare la NMT. I sistemi NMT si fondando in generale su network neurali, detti *encoder* e *decoder*. Il primo legge e codifica una frase del testo di partenza (*source sentence*) in un vettore di lunghezza fissa, il secondo produce la traduzione effettiva a partire dalla decodifica del vettore (Bahdanau et al. 2015, p. 1). Il grande vantaggio della NMT è che non ha bisogno necessariamente dell'intervento umano e può funzionare autonomamente. Inoltre, riesce, almeno in teoria, ad affrontare meglio la traduzione di periodi complicati con frasi dipendenti molto distanti dalla principale (*long-distance dependencies*), che comportano dei problemi nella fase di allineamento e rappresentano una sfida per i *software* SMT (cfr. Wang et al. 2017). Purtroppo, questa tipologia di traduzione automatica non è infallibile e, come scrivono Wang et al. (2017) nel loro studio, presenta tre problemi principali:

1. *Coverage problem*: il *software* pecca nel capire quali parole sono già state tradotte e quali no, creando problemi di *over-translation* o *under-translation*;
2. *Imprecise translation problem*: NMT tende a proporre parole che sembrano naturali nella lingua target, ma non riflettono il significato originale della lingua di partenza;

3. *UNK problem*: i programmi basati sulla NMT usano un vocabolario fisso relativamente grande caratterizzato dalle parole più frequenti in una lingua (*most frequent words*) e sostituisce le parole più rare con un *UNK word* (parole sconosciute). Di conseguenza la qualità della traduzione diminuisce.

Si è pensato che una combinazione della *statistical-based machine translation* con la *neural machine translation* possa permettere di creare programmi di traduzione automatica più precisi e ancora più indipendenti dall'intervento umano. Ricerche in questo campo sono state effettuate, ad esempio, dal già citato studio condotto da Wang et al. (2017) presso la Soochow University a Suzhou (Cina). I ricercatori hanno creato un *training model* end-to-end contenente sia NMT che SMT. In questo modo sono riusciti a diminuire le probabilità di errore della NMT e hanno dimostrato l'efficacia della combinazione delle due tipologie di MT.

Nel caso specifico di questa tesi, i *software* di traduzione automatica che verranno utilizzati sono *Google Translate* e *DeepL*. Entrambi utilizzano la NMT, ma *DeepL*, come si può leggere nel loro sito ufficiale (https://www.deepl.com/en/whydeepl), si avvale del controllo qualità delle traduzioni da parte di traduttori professionisti. Maggiori informazioni sul funzionamento dei due *software* e sulle loro differenze verranno presentati nel capitolo 2.4.1.

1.2. Differenze tra traduzione e testo originale: il *traduttese*

All'interno di questo lavoro verranno analizzate una serie di traduzioni sia in relazione al testo originale sia autonomamente. Il dubbio che può sorgere è se sia possibile considerare una traduzione come un testo indipendente dall'originale. Negli ultimi decenni si è scoperto che la traduzione possiede effettivamente delle caratteristiche uniche e diverse dal testo di partenza, che possono essere sintetizzate all'interno di una parola: il *traduttese* (cfr. Gellerstam, 1986). Con questo termine si designa una serie di caratteristiche lessicali, sintattiche e/o testuali che distinguono le traduzioni dai testi originali e vengono per questo definiti *universali traduttivi* (cfr. Rubino et al. 2016; Ondelli & Viale 2010). Usando le parole di Ondelli e Viale (2010), "gli universali traduttivi si concretizzerebbero in costanti che caratterizzano l'agire del traduttore e che si ripercuotono sull'assetto del testo

di arrivo" (p. 3). Studi sul *traduttese* hanno individuato in particolare quattro universali traduttivi, che verranno esposti sulla base dello studio di Ondelli e Viale (2010, pp. 3-5):

1. *Semplificazione*, ovvero la tendenza del traduttore a semplificare la lingua durante la produzione del testo di arrivo. Gli indicatori che vengono considerati maggiormente per verificare questo universale sono la lunghezza media delle frasi e la ricchezza lessicale, calcolata in base al rapporto tra *word tokens* (N) e *word types* (V) (questi due concetti verranno approfonditi nel capitolo 2.2). In caso di semplificazione, entrambi i calcoli daranno valori inferiori nelle traduzioni rispetto ai testi originali. Un altro parametro da poter utilizzare è la densità lessicale, ovvero il rapporto tra parole piene e vuote all'interno di un testo, che dovrebbe risultare minore nelle traduzioni;
2. *Esplicitazione*, ovvero la tendenza del traduttore di spiegare concetti, invece di lasciarli impliciti. Questo può comportare una disparità di *word token* tra il testo di arrivo e quello di partenza, ovvero la produzione di traduzioni più lunghe rispetto al testo originale. Dal punto di vista sintattico, invece, è possibile che nella traduzione vi sia una maggior frequenza d'uso di congiunzioni, connettivi e ripetizioni, per rendere il contenuto del testo più esplicito;
3. *Normalizzazione (o Conservatorismo)*, ovvero la tendenza a esagerare le caratteristiche linguistiche della lingua target e conformare la traduzione alle regole della lingua di arrivo. Da un punto visto pratico si è notato che nella traduzione orale gli interpreti preferiscono eliminare dalla traduzione fenomeni legati all'oralità, come le false partenze, le autocorrezioni o i periodi incompleti. Nella traduzione scritta, invece, il traduttore sceglie se utilizzare un linguaggio più elevato o più colloquiale in base sia al testo di partenza sia alla sua tendenza più innovatrice o conservatrice. Nel caso specifico

dell'italiano, questa scelta può essere notata nell'uso o meno di parole appartenenti all'*italiano dell'uso medio*[5];

4. *Levelling out (o Convergenza)*, ovvero un livello di omogeneità maggiore tra un *corpus* di traduzioni rispetto a un *corpus* di testi non tradotti. Si è notato che i testi tradotti presentano meno differenze dal punto di vista della densità, della ricchezza lessicale e della lunghezza media dei periodi, rispetto a un gruppo di testi non tradotti.

Come affermano Ondelli e Viale (2010), "non sempre risulta agevole distinguere tra fenomeni ascrivibili a un universale traduttivo piuttosto che all'altro" (p. 5). Nel caso della traduzione inglese-italiano, ad esempio, l'uso esagerato del pronome soggetto potrebbe essere causato sia dall'*esplicitazione* sia da un'interferenza con la lingua inglese, che non prevede il soggetto sottointeso. Lo stesso studio condotto dai ricercatori sopracitati ha fatto sorgere dei dubbi sulla possibilità di riconoscere gli universali traduttivi. Ondelli e Viale (2010) hanno analizzato un *corpus* costituito da articoli giornalistici esplicitamente dichiarati come traduzioni dall'inglese all'italiano, mettendolo a confronto con un secondo *corpus* di articoli giornalistici scritti originariamente in italiano. Il loro obiettivo era quello di riconoscere la presenza di universali traduttivi direttamente nelle traduzioni, concentrandosi "esclusivamente sulle eventuali divergenze tra articoli tradotti e articoli [originali]" (Ondelli & Viale 2010, p. 2). In questo modo, i due ricercatori hanno saltato la fase di confronto con i testi di partenza che viene di solito effettuata negli studi che si occupano di *traduttese*. I risultati della ricerca, però, non sono stati soddisfacenti come si aspettavano i ricercatori. Si è notato che molti dei presunti universali traduttivi potevano essere giustificati anche solo dall'interferenza con la lingua inglese. Ad esempio, i ricercatori hanno notato che la minore ricchezza lessicale dei testi tradotti poteva essere attribuita sia all'universale della *semplificazione* sia alle caratteristiche della lingua inglese, che tollera maggiormente le ripetizioni a breve distanza nel testo e utilizza un registro

[5] Con questo termine si indica un "italiano caratterizzato da una serie di tratti che, un tempo esclusi dallo standard, appaiono ora ampiamente diffusi e accettati da tutti i parlanti" essendo diminuito il confine tra italiano scritto e parlato (Berruto, 2010 in Enciclopedia Treccani)

meno elevato (Ondelli & Viale 2010, p. 56). I due ricercatori hanno, inoltre, osservato che i dati sulla densità lessicale e sulla lunghezza dei periodi non erano conformi agli universali traduttivi. Ad esempio, la percentuale di parole piene risultava simile sia nei testi tradotti sia in quelli non-tradotti, "segno che in quest'ambito non emergono differenze significative nell'assetto dei testi frutto di traduzione rispetto a quelli originali" (Ondelli & Viale 2010, p. 16).

Da questo studio si evince la difficoltà di riconoscere gli universali traduttivi senza effettuare un confronto con i testi di partenza, che al contrario può essere molto utile, come si può notare da uno studio più recente condotto da Rubino, et al. (2016) presso l'Università di Saarland (Germania). Gli studiosi sono partiti dall'analisi dei seguenti tre *corpora*:

1. il primo *corpus* è costituito da testi originali inglesi di vario tipo (es. sia di stampo politico che letterario);
2. il secondo *corpus* si compone dalla traduzione inglese-tedesco prodotta da professionisti;
3. il terzo *corpus* è formato dalla traduzione inglese-tedesco realizzata da studenti.

Uno degli scopi della loro ricerca era quello di provare che l'*information density* possa essere un ulteriore universale traduttivo valido per studiare il *traduttese*. Con il termine *information density* ci si riferisce alla decisione di un parlante di modulare l'ordine, la densità e la specificità delle parole usate per comunicare (cfr. Rubino et al. 2016). Un parlante in una situazione comunicativa tende ad evitare l'utilizzo di variazioni linguistiche (es. sinonimi) per semplificare la comunicazione (*uniform information density*). Da un punto di vista statistico, per calcolare la prevedibilità di un'espressione all'interno di un determinato contesto si utilizza una misura chiamata *surprisal* (Rubino et al. 2016, p. 962). All'interno di questo studio, i ricercatori hanno calcolato il *surprisal* a livello lessicale, all'interno di una parte del discorso e a livello sintattico.

Al termine della ricerca si è potuto concludere che il *surprisal* può essere considerato un criterio attendibile per determinare se una frase è stata tradotta o appartiene a un testo originale, ma risulta meno affidabile se si desidera distinguere testi tradotti da professionisti o da studenti (Rubino et al. 2016, p. 966).

Possiamo quindi concludere che l'*information density* possa essere considerata una caratteristica del *traduttese* al pari delle altre e che studi sul *traduttese* sono facilitati se si analizzano anche i testi di partenza delle traduzioni.

Nella presente tesi verrà preso in considerazione l'universale traduttivo della *semplificazione* nella seconda fase di analisi dei dati, che ha lo scopo di verificare la semplicità lessicale dei *corpora* di traduzione attraverso il calcolo del *Type Token Ratio* e della lunghezza media della frase (cfr. Ondelli & Viale 2010, pp. 3-5).

1.3. Traduzione umana e automatica a confronto

1.3.1. Attendibilità della traduzione automatica

I primi esempi di traduzione automatica sono comparsi negli anni '50 del Novecento e fin dal principio ricercatori e sviluppatori si erano posti l'obiettivo di creare un programma in grado di produrre traduzioni 'perfette' (cfr. Hutchins 2001). Lo scopo non è stato evidentemente ancora raggiunto, dal momento che le traduzioni automatiche non sono ancora in grado di sostituire completamente le traduzioni umane. Negli ultimi anni, però, si è assistito ad un grande miglioramento dei *software* di traduzione automatica, che in alcuni casi, come vedremo, sono in grado di equiparare una traduzione umana.

Nel 2014, Li et al. hanno condotto uno studio nell'Università di Memphis (USA) analizzando un *corpus* costituito da una collezione di 289 testi, sia scritti che parlati, tratti dall'opera *Selected Works of Mao Zedong*. I testi raccolti contenevano sia gli originali in cinese sia le relative traduzioni in inglese e i testi di partenza sono stati poi tradotti con Google Translate. Attraverso l'uso dei programmi LIWC[6] e Coh-Metrix[7], hanno analizzato il grado di formalità e coesione delle traduzioni automatiche mettendole a confronto prima con le traduzioni eseguite da professionisti, poi con i testi originali.

[6] LIWC: è un programma di analisi testuale che conta la percentuale di parole di un testo appartenenti a una stessa categoria linguistica o psicologica (Li et al. 2012, p.191).
[7] Coh-Metrix: è uno strumento computazionale che permette di analizzare la coesione, la difficoltà lessicale e sintattica di un *corpus* di testi (Ibid., p. 191)

Dai risultati è emerso che la traduzione automatica di Google Translate si avvicina alla traduzione umana a livello semantico e pragmatico. Dal punto di vista sintattico e grammaticale, invece, è stato registrato che il *software* necessitava di miglioramenti. Si è concluso che Google Translate sia in grado di produrre traduzioni leggibili e comprensibili nonostante la possibile presenza di errori grammaticali e che risulti uno strumento molto utile per coloro che hanno bisogno di una traduzione veloce per acquisire determinate informazioni (Li et al. 2014, p. 194). È importante porre l'attenzione all'anno di pubblicazione di questo studio, quando Google Translate utilizzava ancora una *statistical-based machine translation*. Come già accennato nel capitolo 1.1, nel 2016 l'azienda ha deciso di iniziare ad abbandonare gradualmente l'SMT per utilizzare esclusivamente la traduzione neurale (NMT). Nello stesso anno Yonghui Wu et al. (2016) hanno pubblicato uno studio in cui analizzavano il *neural machine translation system* di Google e lo confrontavano con la versione precedente SMT. Gli autori erano consapevoli dei seguenti limiti della traduzione neurale:

1. I sistemi NMT sono *computationally expensive*, ovvero richiedono molti procedimenti sia nella fase di *training* che di *translation inference* (qui: proposte traduttive del *software*). Questo può rallentare il processo traduttivo, rendendolo talvolta proibitivo nel caso in cui si utilizzino *data set* e *corpora* di grandi dimensioni;
2. Inoltre, i sistemi NMT peccano in *robustness*[8], soprattutto quando i testi da tradurre contengono parole a bassa frequenza d'uso (Wu et al. 2016, p. 1).

Per cercare di risolvere queste problematiche, i ricercatori hanno utilizzato un algoritmo *sequence-to-sequence learning framework* costituito da tre componenti: *encoder network*, *decoder network* e *attention module*. L'encoder trasforma una frase del testo di partenza (input) in una lista di vettori, in cui un vettore sostituisce un simbolo dell'input (es. lettera o segno di interpunzione). Da questa lista di vettori, il decoder produce un simbolo alla volta fino al simbolo finale *end-of-sentence* (EOS). L'encoder e il decoder sono collegati dall'*attention module* che

[8] Robustness: abilità di un *software* di gestire gli errori

permette al decoder di focalizzarsi su diverse parti della frase di partenza nel corso della decodifica, migliorando il parallelismo tra testo di partenza e di arrivo, e diminuendo il tempo necessario nella fase di *training*[9] (cfr. Wu et al., 2016). Per accelerare la velocità di produzione della traduzione finale, è stato utilizzato un *low-precision arithmetic* nella fase di inferenza (qui: il procedimento di traduzione che permette di trovare il termine equivalente nella lingua di arrivo). Inoltre, per migliorare la gestione dei *low frequency words*, le parole sono state divise in una lista di *common sub-word units* sia nell'input (testo di partenza) che nell'output (testo di arrivo) (cfr. Wu et al., 2016). L'utilizzo di questo metodo ha permesso di migliorare l'accuratezza del *software* e la velocità di traduzione.

Dopo aver eseguito questi miglioramenti nel sistema neurale di Google Translate, i ricercatori hanno comparato le traduzioni automatiche prodotte dal *software* con quelle umane, utilizzando come punto di riferimento due *corpora* pubblici utilizzati durante il WMT[10] del 2014, costituito il primo da traduzioni dall'inglese al francese e il secondo da traduzioni dall'inglese al tedesco. Le traduzioni umane e automatiche sono state analizzate da un gruppo di essere umani detti *raters*, che hanno comparato le traduzioni *side-by-side* di una frase di partenza e ne hanno valutato la qualità in una scala da 1-6. Al termine del confronto è stato verificato che in alcune frasi la qualità della traduzione di Google Translate aveva raggiunto l'accuratezza di un traduttore bilingue medio. Inoltre, è stato concluso che la traduzione neurale, con le accortezze descritte all'inizio, è in grado di diminuire del 60% la probabilità di errore su diverse coppie linguistiche (es. inglese-tedesco; inglese-francese) rispetto alla versione precedente di Google Translate basata su un sistema SMT (cfr. Wu et al., 2016).

Le conclusioni di questo studio portano a pensare che la traduzione automatica sia effettivamente in grado di avvicinarsi e talvolta raggiungere l'accuratezza della traduzione umana. In realtà il confronto è stato svolto a livello frasale e non è stato effettuato un 'controllo qualità' mettendo in relazione le singole frasi con il contesto

[9] La fase di *training* è una delle fasi di cui è composto il *machine learning*, che verrà approfondito nel capitolo 3.

[10] Workshop on Machine Translation: una conferenza tenuta annualmente in cui si discute di traduzione automatica e ricerche in tale ambito.

in cui erano inserite. Di conseguenza non è stato possibile valutare la coerenza e la coesione delle frasi inserite all'interno di un testo.

Proprio su questo punto si è concentrato uno studio successivo condotto da Läubli, et al. (2018) presso l'Università di Zurigo in collaborazione con l'Università di Edimburgo. I ricercatori hanno testato empiricamente la veridicità delle seguenti conclusioni:

1. "[*Machine Translation*] approaches the accuracy achieved by average bilingual human translators [on some test sets]" (Wu et al., 2016 p.20)[11];
2. "[*Machine*] translation quality is at human parity when compared to professional human translators" (Hassan et al., 2018 p.1)[12].

Il *corpus* di testi analizzato è costituito da 123 articoli scritti in cinese presi dal test set del WMT 2017. Le traduzioni dal cinese all'inglese sia automatiche che umane provengono dai dati raccolti da Hassan et al. (2018). Per valutare correttamente l'adeguatezza delle traduzioni sono stati reclutati traduttori professionisti (*raters*) con almeno tre anni di esperienza e recensioni positive dei clienti nel sito ProZ[13]. In tutto sono stati selezionati 4 parlanti nativi (due in cinese, uno in inglese e uno in entrambe le lingue) e 4 parlanti di madrelingua inglese. I primi avevano il compito di valutare l'adeguatezza[14] delle traduzioni rispetto all'originale, i secondi la scorrevolezza[15] dei testi tradotti senza vedere l'originale. Tutti i *raters* hanno valutato sia testi interi sia singole frasi in ordine casuale (Läubli et al. 2018, p. 3). Grazie a questo tipo di analisi (sia testuale sia frasale) i ricercatori sono stati in grado di ottenere dati più attendibili riguardanti la qualità delle traduzioni automatiche rispetto agli studi precedenti. Al termine della ricerca è stato concluso che dal punto di vista dell'adeguatezza a livello frasale MT e HT non sono particolarmente differenti, ma a livello testuale i *raters* hanno dimostrato di preferire la traduzione umana (Läubli et al. 2018, pp. 3-4). Dal punto di vista della

[11] trad. La traduzione automatica si avvicina all'accuratezza di un traduttore umano bilingue medio.
[12] trad. La qualità della traduzione automatica ha raggiunto la parità delle traduzioni umane di traduttori professionisti.
[13] ProZ: network di traduttori e agenzie di traduzione che collega traduttori e aziende e/o clienti singoli.
[14] i.e. quanto il significato della frase di partenza viene mantenuto nella traduzione (cfr. Popel et al. 2020).
[15] i.e. quanto una frase suona scorrevole nella lingua di arrivo (cfr. Popel et al. 2020).

scorrevolezza del testo, invece, i *raters* hanno sempre preferito la traduzione umana, sia a livello frasale sia testuale (Läubli et al. 2018, p. 4).

Grazie a questo studio, è stato dimostrato che per valutare l'accuratezza di una traduzione automatica non basta analizzare la correttezza di una singola frase, ma è importante vedere il contesto in cui si trova. Questa conclusione è coerente con quello che già nel 2014 Li et al. avevano dedotto, ovvero che un traduttore automatico è in grado di produrre testi leggibili e comprensibili, ma non riesce sempre a evitare errori grammaticali o sintattici che rendono la traduzione meno scorrevole da leggere o ascoltare.

Questa consapevolezza ha cambiato il modo di analizzare le MT e ha portato i ricercatori successivi a studiare il buon funzionamento delle MT non solo a livello frasale, ma anche testuale. Interessante in questo senso è lo studio condotto da Popel et al. (2020), che hanno analizzato i sistemi neurali di traduzione automatica per mettere nuovamente in discussione la superiorità della traduzione umana rispetto a quella automatica (cfr. Läubli et al., 2018).

Popel et al. (2020) hanno studiato il sistema *neural-based* CUBBIT[16], che ha superato le prestazioni di traduttori professionisti nella traduzione di frasi singole dall'inglese al ceco durante il WMT 2018. La tipologia di testo al quale appartenevano le frasi era la cronaca giornalistica e il test è stato effettuato solo nella coppia di lingue inglese-ceco. Il grande difetto di questa 'competizione uomo-macchina' è stato il fatto che sono state prese nuovamente in considerazione solo frasi fuori contesto. Per questo motivo, dopo la pubblicazione dello studio di Läubli et al. (2018), Popel at al. hanno deciso di approfondire le capacità traduttive di CUBBIT valutando le frasi all'interno di un contesto (Popel et al. 2020 p. 4).

Prima di parlare dell'analisi vera e propria, è utile spiegare cos'è CUBBIT e come si differenzia da altri sistemi neurali. Il sistema utilizza di base lo stesso algoritmo *sequence-to-sequence learning framework* descritto nello studio di Wu et al. (2016) riguardante Google Translate. Una delle differenze è che CUBBIT utilizza un *attention module* perfezionato, detto *multi-head attention*, in cui più

[16] CUBBIT: Charles University Block-Backtranslation-Improved Transformer Translation

funzioni indipendenti (*independent attention functions*) vengono addestrate contemporaneamente permettendo all'algoritmo di rappresentare più fenomeni linguistici in contemporanea. In questo modo si facilita la traduzione di parole ambigue e coreferenti (Popel et al. 2020 p. 2).

Il *corpus* di traduzioni analizzato nella ricerca era costituito dalle stesse traduzioni umane e automatiche utilizzate durante il WMT 2018, analizzando, però, le singole frasi all'interno dei testi. Il giudizio delle traduzioni è stato affidato a un gruppo di valutatori, costituito da 6 traduttori professionisti madrelingua cechi, 7 non professionisti madrelingua cechi con un'ottima conoscenza dell'inglese e 3 teorici della traduzione. Come in Läubli et al. (2018), anche in questa ricerca il compito dei valutatori era quello di giudicare la traduzione dal punto di vista dell'adeguatezza e della scorrevolezza. Inoltre, a differenza dello studio precedente, avevano anche il compito di valutare la qualità dell'intera traduzione.

Al termine della ricerca è stato confermato che il sistema CUBBIT è effettivamente in grado di diminuire la distanza tra la traduzione automatica e quella umana sia dal punto di vista della adeguatezza che della scorrevolezza, tanto che molti partecipanti non sono riusciti a distinguere le traduzioni umane da quelle automatiche. Inoltre, è stato notato che i traduttori professionisti sono più sensibili agli errori di scorrevolezza rispetto ai non professionisti, tanto da preferire una traduzione scorrevole, rispetto a una semplicemente adeguata. Questa preferenza non è stata riscontrata nei non professionisti, che invece hanno dato più importanza all'adeguatezza del testo, nel momento in cui dovevano valutare la traduzione nella sua interezza. Di conseguenza, si può concludere che all'interno di domini, come l'articolo di giornale o scientifico, il mantenimento del significato originale di una frase è più importante per il lettore rispetto alla scorrevolezza della traduzione (Popel et al. 2020, p. 10). Ovviamente questa conclusione non può valere per testi di scrittura creativa, in cui la scorrevolezza rimane fondamentale e la traduzione automatica non riesce a eguagliare quella umana.

Si capisce, allora, che la traduzione automatica potrà sostituire la traduzione umana in alcuni ambiti, perché permette a molti clienti di ottenere una traduzione adeguata velocemente a un costo irrisorio se non nullo. Un esempio può essere trovato all'interno dell'articolo pubblicato da Takakusagi et al. (2021) in cui è stata

valutata l'accuratezza del *software* DeepL nella traduzione di un articolo medico dal giapponese all'inglese. L'articolo in questione si intitola *Dosimetric Comparison between Carbon-Ion Radiotherapy and Photon Radiotherapy for Stage I Esophageal Cancer*. La valutazione della traduzione è avvenuta in modo diverso rispetto agli studi presentati precedentemente. Il testo è stato tradotto automaticamente dal giapponese all'inglese, dopodiché tre traduttori hanno eseguito individualmente una *back translation* in giapponese a partire dalla traduzione in inglese di DeepL. Infine, sono stati confrontanti il testo originale giapponese con il testo giapponese tradotto a partire dalla traduzione in inglese di DeepL. I traduttori dovevano essere madrelingua giapponesi, esperti di radio-oncologia certificati dalla JASTRO[17] e aver pubblicato come primo autore 5 o più articoli in riviste scientifiche inglesi *peer reviewed*. I giudici dovevano essere radio-oncologi madrelingua giapponesi e non traduttori. Al termine dello studio si è notato che il testo *back translated* in giapponese aveva un alto tasso di corrispondenza sia di significati sia di strutture grammaticali con il testo originale giapponese. Al contrario, le frasi lunghe non sono state tradotte molto bene dal *software* di traduzione automatica (Takakusagi et al., 2021).

Al termine di questo breve excursus sulle principali scoperte riguardanti le traduzioni automatiche negli ultimi anni, si può evincere che ci sono ancora pareri discordanti sulla capacità dei *software* di MT di sostituire completamente la traduzione umana. Si è notato come la MT abbia fatto dei grandi passi in avanti negli ultimi anni, ma rimane imprecisa nella traduzione di frasi lunghe (cfr. Takakusagi et al. 2021). D'altro canto, si è notato che nella traduzione frasale le MT abbiano effettivamente superato le HT (cfr. Popel et al. 2020), ma questa precisione dipende dal tipo di testo che viene tradotto. Articoli di giornale e scientifici, ad esempio, (cfr. Läubli et al. 2018; Popel et al. 2020; Takakusagi et al. 2021) vengono tradotti con molta precisione dai sistemi di traduzione automatica, ma la traduzione di un romanzo o una poesia non potrà probabilmente mai essere affidata completamente alla MT (cfr. Popel et al. 2020). Infine, è stato osservato

[17] Japanese Society for Radiation Oncology

che la MT sta migliorando nell'adeguatezza della traduzione peccando, però, nella scorrevolezza. Secondo Popel et al. (2020) sembrerebbe che la scorrevolezza e la fluidità di un testo passino in secondo piano e che sia molto più importante per un lettore capire il messaggio dell'autore originale del testo, rispetto a leggere un testo coeso. D'altro canto, però, una traduzione poco scorrevole con qualche errore grammaticale non può essere pubblicata all'interno di siti aziendali o testate giornalistiche perché ne risentirebbe la qualità del servizio prestato al cliente. Sembrerebbe, quindi, che la conclusione di Li et al. (2014) sia vera ancora oggi. Secondo i loro studi una MT è uno strumento utile e veloce per comprendere le informazioni di base di un testo in lingua straniera, ma per ottenere la redazione di un testo scorrevole e adeguato in grado di essere pubblicato senza problemi, sembra che ci sia ancora bisogno del traduttore umano.

1.3.2. Possibili nuovi metodi di valutazione della MT

Negli articoli presentati nel paragrafo precedente, le traduzioni automatiche venivano giudicate buone o meno buone in base alla loro vicinanza alla traduzione umana. Come scrivevano Li et al. (2014), questo è il classico punto di vista che viene adottato in letteratura per valutare la prestazione di una traduzione automatica. Nel 2021, però, uno studio condotto da Yingxue Fu e Mark-Jan Nederhof presso l'Università di St. Andrews (UK) ha messo in dubbio questo metodo di valutazione. All'interno della loro ricerca hanno utilizzato un *corpus* di testi tratto da *News commentary parallel corpus v 13* fornito dal WMT 2018. Le coppie linguistiche di studio erano: ceco-inglese; tedesco-inglese; russo-inglese. I testi originali e le traduzioni umane erano contenuti nel *corpus* del WMT 2018, mentre le traduzioni automatiche sono state ottenute da Google Translate. Uno degli obiettivi dello studio era capire se la MT e la HT possano essere distinte da un algoritmo di machine learning. In questo caso è stato utilizzato il BERT[18] model. Al termine della ricerca è stato osservato che la traduzione automatica e quella umana possono essere distinte da un *software* con un alto livello di precisione (cfr.

[18] Bidirectional Encoder Represantations from Transformers

Fu & Nederhof 2021). Questo ha fatto concludere ai ricercatori che esistano delle diversità lessicali tra la traduzione automatica e quella umana, tali da far supporre l'esistenza di caratteristiche linguistiche specifiche della MT, soprattutto a livello lessicale. Come le *human translation* si distinguono dai testi originali per il *traduttese*, è probabile in futuro che caratteristiche analoghe possano essere trovate anche nelle traduzioni automatiche. Una simile scoperta mette in dubbio le analisi condotte finora, che partivano dal presupposto che più una traduzione automatica è simile alla traduzione umana, più è corretta (cfr. Fu & Nederhof, 2021). Ci si può aspettare che in futuro verranno pubblicati studi che approfondiranno questi aspetti e li prenderanno in considerazione per studiare meglio le differenze tra MT e HT. Forse la scoperta di caratteristiche simili al *traduttese* per la MT sarà utile anche per il miglioramento degli stessi programmi di traduzione automatica. Al momento sembra che questo sia un campo di ricerca ancora inesplorato e che forse permetterà di diminuire ancora di più la distanza tra la traduzione umana e quella automatica.

2. Analisi dei dati testuali: lavorare su un *corpus*

2.1. Cosa si intende per analisi dei dati testuali

All'interno della presente tesi si è deciso di utilizzare l'*analisi dei dati testuali* come metodo di analisi per confrontare le traduzioni umane e automatiche e i testi originali russi. Con questo termine si designa, in senso ampio, un'attività che permette di acquisire, sintetizzare e restituire delle informazioni contenute in un insieme di testi (cfr. Tuzzi 2003, p. 17). I metodi possono essere quantitativi o qualitativi, possono prevedere attività svolte manualmente o con il supporto del computer, possono avere come finalità quella di analizzare il contenuto o quella di riconoscere altre caratteristiche rilevanti presenti nei testi.

Nelle scienze sociali l'analisi dei dati testuali è stata per lungo tempo soprattutto analisi del contenuto. Un esempio di analisi del contenuto può essere l'esposizione orale della trama di un film: anche se si tratta di materiale audiovisivo e non di un testo scritto, durante la fase di acquisizione si guarda attentamente il film, dopodiché si elabora ciò è stato visto cercando di riassumerne e sintetizzarne il contenuto (fase di sintesi), infine si è in grado di esporre oralmente i punti salienti del film a qualcuno (fase di restituzione).

Come scrive Tuzzi (2003), anche con un computer oggi si è in grado di eseguire un'analisi dei dati testuali. Un computer, infatti, riesce a leggere e ad acquisire le informazioni di un testo digitando le lettere all'interno di un word processor, con l'ausilio di una tastiera, o dettando le parole attraverso un sistema di riconoscimento vocale. Successivamente il computer è in grado di organizzare, elaborare e sintetizzare le informazioni acquisite attraverso l'uso di uno o più programmi per l'analisi statistica di dati testuali (cfr. Tuzzi 2003, p. 18). L'intero processo viene supervisionato da uno o più ricercatori che hanno poi il compito di analizzare i risultati dei *software* di analisi dei dati testuali e comunicarli in altra forma (fase di restituzione). Come ribadisce Tuzzi (2003, pp. 18-19), la bontà dei risultati di una ricerca dipende non solo dalla precisione degli strumenti utilizzati, ma anche dalla sensibilità e conoscenza in materia del ricercatore stesso. Questo ci fa capire che, per quanto un *software* riesca ad analizzare autonomamente dei dati, c'è sempre bisogno dell'intervento umano per sintetizzarli, comprenderli e comunicarli correttamente.

Per quanto già nel XVII secolo siano stati rinvenuti i primi abbozzi di analisi testuale con obiettivi di analisi del contenuto con lo studio dei *Canti di Sion* in Svezia, il primo studio rilevante scientificamente venne condotto dagli americani William I. Thomas e Florian Zaniecki nel 1918. I due "svolsero uno studio qualitativo di 754 lettere, scambiate da emigranti polacchi con parenti e amici rimasti in Polonia, allo scopo di comprendere le condizioni di vita dell'emigrante polacco in America" (Tuzzi 2003, p. 21). Dopo la Prima Guerra Mondiale molti studiosi iniziarono ad interessarsi alle tecniche di propaganda politica dei primi del '900 e, soprattutto nell'università di Chicago, si chiesero quali metodi fossero i più adatti per analizzarla. Il primo a coniare il termine *content analysis* (qui trad. analisi del contenuto) fu il ricercatore Harold D. Lasswell, che partendo dalla sua tesi di dottorato intitolata *Propaganda Technique in the World War* (1927), propose un nuovo sistema di analisi di tipo quantitativo, nel tentativo di risolvere le carenze delle ricerche sulla propaganda politica dell'epoca, che peccavano, ad esempio, di "sufficiente rigore [nei] criteri di campionamento, di selezione dei materiali, di costruzione degli indicatori" (Tuzzi 2003, p. 21). All'epoca con il termine *content analysis* ci si riferiva solo a metodologie di analisi quantitativa di testi di propaganda politica, successivamente questo termine venne esteso a tutte le ricerche che analizzavano il contenuto di un messaggio scritto od orale (cfr. Tuzzi 2003, p. 21). Non tutti gli studiosi dell'epoca, però, si ritrovarono d'accordo nell'analizzare le parole attraverso numeri. Secondo lo storicista tedesco Wilhelm Dilthey, *comprendere* un testo non significa semplicemente *spiegarlo*, bensì "risalire dalla espressione dello spirito alla sua interiorità" (Giuliano & La Rocca 2008, p. 8). In queste parole si può riassumere il punto di vista dell'ermeneutica, ovvero l'arte di interpretazione di testi, leggi e documenti che trae le sue origini dall'Antica Grecia (cfr. Enciclopedia Treccani; Giuliano & La Rocca 2008, p. 8). Il conflitto tra queste due prospettive, una eccessivamente quantitativa l'altra eccessivamente 'romantica', ha portato a una stasi nello sviluppo di procedure di analisi rigorose, controllabili e condivise per interpretare un testo (cfr. Giuliano & La Rocca 2008, p. 8).

Dal punto di vista dei metodi statistici, una svolta si ha negli anni '60 del Novecento quando linguisti e matematici iniziarono ad utilizzare strumenti di

analisi quantitativa dei testi per studiare le corrispondenze in ambito fonetico, e successivamente lessicale, di un *corpus* di testi sufficientemente ampio da considerarlo rappresentativo. I primi studi in questo senso vennero effettuati in Francia da Jean P. Benzécri e la sua scuola (cfr. Tuzzi 2003, p. 23) e portarono allo sviluppo di nuove tecniche, come l'analisi delle corrispondenze, che poi sono diventate uno standard per l'analisi esplorativa dei dati testuali.

In questi anni, parallelamente a uno sviluppo di *software* specifici per l'analisi dei dati, si è arrivati ad un'integrazione degli approcci quantitativo ed ermeneutico sintetizzabili nel seguente modo:

1. Analisi dei dati qualitativi assistita dal computer (*Computer Assisted Qualitative Data Analysis Softwares – CAQDAS*), ovvero l'analisi semi-automatica di testi. In questo caso il ricercatore utilizza *software* che facilitano la lettura e interrogazione di documenti per trarre delle risposte su specifiche domande di ricerca o aiutano a costruire delle ipotesi. Programmi utili per svolgere questo tipo analisi sono Nvivo o AtlasT;
2. Analisi statistica dei dati testuali (*Analyse statistique des données textuelles*), ovvero un approccio di tipo lessicometrico che utilizza *software* per confrontare i profili lessicali dei testi, senza che il ricercatore li legga direttamente. Per questo motivo viene definito "automatico" (cfr. Giuliano & La Rocca 2008, p. 8). Quest'ultimo metodo permette di analizzare grandi quantità di dati testuali più rapidamente rispetto al primo approccio, perché il ricercatore non deve necessariamente leggere i testi per analizzarli. All'interno della presente tesi verranno utilizzati programmi appartenenti a questa seconda tipologia.

Inoltre, gli antenati dell'analisi dei dati testuali sono da cercare non solo nelle scienze sociali ma, come è ovvio, anche in ambito umanistico (cfr. De Mauro & Chiari 2005). Tullio De Mauro (1995) fa risalire la nascita degli studi quantitativi del linguaggio addirittura alle riflessioni di Orazio.

Anche se esulano dagli obiettivi di questa tesi, vale la pena ricordare che gli studi quantitativi delle regolarità matematiche della lingua e delle leggi universali del linguaggio vantano una lunga tradizione e hanno impegnato linguisti e matematici di grande spessore. La legge più classica e famosa è quella di Zipf, ma

sono numerosi i nomi di grandi studiosi del passato che, con i loro contributi, hanno datto grande impulso alla produzione di modelli e misure quantitative (Mandelbrot, Markov, Shannon, Yule, Simpson, Guiraud, Herdan, Sichel, Honoré, per citarne solo alcuni) (cfr. Tuzzi 2003, pp. 115-116).

Infine, come scrive Tuzzi (2003, p. 27), anche oggi l'analisi dei dati testuali non è l'unico campo di studi a utilizzare strumenti di tipo statistico per analizzare testi. Nell'ambito della ricerca ne ricordiamo almeno uno, ovvero il *Natural Language Processing* (NLP), che "si occupa di intelligenza artificiale, riconoscimento vocale, traduttori multilingue, allineamento degli idiomi [e altro]" (Tuzzi 2003, p. 27). Proprio questa metodologia è quella utilizzata per creare *software* di traduzione automatica. Inoltre, anche attraverso lo studio del NLP, si conducono i principali studi sul confronto tra traduzione umana e automatica. Per l'approfondimento di queste ricerche si rimanda al capitolo 1, in particolare al sottocapitolo 1.3.

2.2. Oggetto dell'analisi dei dati testuali: il *corpus*

L'oggetto dell'analisi dei dati testuali è il *corpus*, ovvero una collezione di testi "coerente con gli scopi perseguiti dalla ricerca" (Tuzzi, 2003, p. 29). Da questa definizione si potrebbe pensare che qualsiasi collezione di testi possa essere considerata un *corpus*. Per quanto questo sia vero in senso lato, è interessante il ragionamento di Ondelli (2018) che invece cita la definizione di Barbera et al. (2007) per distinguere la semplice collezione di testi dal corpus *linguistics*, la disciplina che studia i contenuti del *corpus*. Secondo Barbera et al. (2007):

> A *corpus* is a finite collection of (written, oral or multimodal) texts or parts thereof in electronic format, consistently processed (i.e. tokenised and added with adequate mark-up) so as to be treated and investigated automatically by means of *software*.[19]
>
> Barbera et al. 2007, p. 70

Come la descrive Ondelli (2018, p. 134), questa definizione potrebbe essere considerata "dura", in quanto sottintende l'uso di procedure specifiche per

[19] trad. Un *corpus* è una collezione finita di testi (scritti, orali o multimodali) o parti di essi in formato elettronico, elaborata coerentemente (ovvero tokenizzata e adattata con un mark-up adeguati) in modo da poter essere trattata e analizzata automaticamente da un *software*.

l'identificazione di unità di testuali, come la *tokenisation* e il *mark-up*. Secondo Barbera et al. (2007), sembrerebbe che questi due parametri dovrebbero essere i primi ad essere presi in considerazione nella preparazione di un *corpus*, visto non solo come una semplice collezione di testi, ma come un insieme di dati da inserire e analizzare in un *software*. In realtà, sostiene Ondelli (2018, p. 135), ci sono anche altri fattori (socio)linguistici che entrano in gioco nella fase di preparazione del *corpus*, come le cinque dimensioni di variazione linguistica definite da G. Berruto (1987, pp. 19-279):

1. Variazione diacronica, ovvero la variazione della lingua nel corso del tempo (es. anni, secoli o millenni);
2. Variazione diatopica, ovvero la variazione della lingua in base all'area geografica in cui viene utilizzata (es. uso di un italiano neostandard con influenze dialettali diverse in Veneto o in Sicilia);
3. Variazione diafasica, ovvero la variazione della lingua in base alla situazione comunicativa (es. uso di un linguaggio formale in una conferenza e informale in una conversazione tra amici);
4. Variazione diastratica, ovvero la variazione della lingua nelle diverse classi sociali (es. l'uso di vocaboli diversi in una conversazione condotta da muratori o studenti universitari);
5. Variazione diamesica, ovvero la variazione linguistica in base al mezzo di comunicazione utilizzato (es. l'uso di una sintassi diversa in un romanzo, rispetto alla sua trasposizione cinematografica).

Se le variazioni linguistiche sono elementi fondamentali da prendere in considerazione in fase di raccolta del *corpus*, la *tokenisation* e i *mark-up* nominati da Barbera et al. (2007) sono importanti per una corretta analisi dei dati testuali.

Con il termine *tokenisation* si intende un processo che identifica le unità minime di un *corpus*, detti *token*, ovvero le parole di cui sono composti i testi (cfr. Ondelli 2018, p. 135). Se un essere umano automaticamente distingue le singole parole di un testo, una macchina ha bisogno di essere "istruita" nel riconoscere quali caratteri dovrebbero essere considerati lettere che formano il *token* e quali dovrebbero essere considerati dei distanziatori, come i numeri, gli apostrofi o i segni di interpunzione. I *mark-up*, invece, sono strumenti utili che permettono al *software* di escludere dal

calcolo dei *token* informazioni soprasegmentali, come il titolo o l'autore di un articolo di giornale (cfr. Ondelli 2018, p. 135).

Prima di iniziare la fase di analisi dei dati testuali vera e propria, si effettua un'analisi preliminare del *corpus*, chiamata in inglese *pre-processing*. Prima di tutto si deve decidere se si vuole lavorare per *word tokens*, *lemmi* o *stem* (tema o radice di una parola). Il primo, come già accennato precedentemente, indica una qualsiasi unità linguistica che di solito coincide con una parola (Baker et al 2006, p. 159). Se si decide di lavorare con le parole intere, è importante allora distinguere il *word token* dal *word type*, ovvero le forme grafiche distinte di un *corpus* (Tuzzi 2003, p. 72). Ad esempio, nella frase "occhio per occhio, dente per dente" ci sono 6 *word token* (numero di parole totali) e 3 *word type* ("occhio", "per", "dente"). Conoscere la somma dei *word token* (N) ci permette di capire la grandezza generale del *corpus*, mentre con la somma dei *word type* (V) possiamo capire nello specifico la grandezza del vocabolario del *corpus*. Attraverso il rapporto tra V/N possiamo calcolare il TTR (*Type Token Ratio*), che ci permette di capire la ricchezza lessicale di un *corpus*. Generalmente un alto valore di TTR corrisponde ad una grande quantità di variazione lessicale, mentre un valore basso indica una minor variazione lessicale (Ali & Hussein 2014, p. 113). L'alto valore di TTR dipende anche dal numero di *hapax* (o *hapax*-legomena) di un testo, ovvero "una parola che compare una sola volta nel *corpus*" (Tuzzi 2003, p. 73). Ne consegue che un numero troppo elevato di TTR impedirebbe una buona analisi quantitativa del *corpus*, in quanto per trovare dati significativi attraverso metodi statistici c'è bisogno di ridondanza. D'altra parte, un *corpus* con un numero minimo di *hapax* e V può influire altrettanto negativamente in fase di analisi. Empiricamente è stato dimostrato che se il rapporto tra *word type* e *word token* è maggiore del 20%:

$$\frac{V(N)}{N} > 0{,}20$$

il *corpus* possiede un vocabolario troppo poco ridondante e non è sufficientemente esteso dal punto di vista statistico. Infatti, se la ricchezza lessicale è elevata, c'è bisogno di raccogliere una collezione di testi più estesa per poter analizzare le caratteristiche del *corpus* attraverso strumenti statistici (cfr. Tuzzi 2003, p. 76). Un altro parametro da considerare per capire la trattabilità di un *corpus* in via preliminare è il calcolo della percentuale di *hapax* presente in esso, perché troppe

parole ripetute solo una volta indicano, nuovamente, un vocabolario poco ridondante e potrebbe compromettere una corretta analisi del *corpus*. Empiricamente è stato valutato che se il rapporto tra il numero di *hapax* e *word types* è maggiore del 50% del vocabolario:

$$\frac{V_1}{V} > 0{,}50$$

il *corpus* presenta un numero troppo elevato di parole usate un'unica volta e non risulta trattabile da un punto di vista statistico (Tuzzi 2003, p. 76).

È necessario puntualizzare che questi parametri empirici sono stati sperimentati e possono valere per lingue come il francese, l'inglese o l'italiano, ma non per tutte. Ogni lingua, infatti, possiede delle regole proprie che la rendono più o meno tendente a creare parole diverse, che un programma può considerare come *word type* o *hapax* diversi. Nel caso specifico della lingua russa (lingua flessiva) si ha la presenza di 6 casi diversi (nominativo, genitivo, dativo, accusativo, strumentale, prepositivo), 3 generi (maschile, femminile, neutro), due forme diverse per ogni verbo (aspetto imperfettivo e perfettivo) e vengono declinati non solo i nomi comuni, ma anche nomi propri e numeri (cfr. Cevese et al. 2000). Ne consegue che un tale sistema linguistico porti alla proliferazione di numerose parole apparentemente differenti, ma che in realtà possiedono la stessa radice e sono solo declinate in modo diverso. Per questi motivi, è tollerabile nella lingua russa un rapporto V/T maggiore del 20% e un rapporto V_1/V maggiore del 50%. Come si può notare, ad esempio, nello studio di Kelih (2010), che verrà descritto più nello specifico nel sottocapitolo 2.3, il rapporto tra N/V del testo analizzato è pari al 30% (cfr. Kelih 2010, p. 2). Un altro esempio di TTR maggiore del 20% si trova nello studio condotto da Dmitrieva e Tiedmann (2021), che avevano lo scopo di creare un *aligned corpus*, costituito da testi letterari originali russi con i relativi adattamenti in una lingua russa semplificata per studenti che imparano il russo come seconda lingua. Prima di effettuare l'allineamento vero e proprio, sono stati analizzati i due *corpora* secondo diversi parametri, tra cui il TTR, il cui valore per i testi originali corrispondeva al 42%, mentre per i testi adattati al 43% (cfr. Dmitrieva & Tiedemann 2021, p. 75).

In alcuni studi, per risolvere il "problema" dell'alto grado di flessione della lingua russa, è stata utilizzata la lemmatizzazione (cfr. Kunilovskaya & Kutuzov

2015; Kunilovskaya et al. 2018), ovvero un procedimento che associa ciascuna parola alla sua forma base (o lemma) e alla relativa categoria grammaticale (si tratta di una lemmatizzazione basata sulle parti del discorso o *part of speech* – POS). Ad esempio, nella frase "Ieri sono andato a scuola, mentre mia mamma andava al lavoro e mia sorella andava all'università"[20] risultano i seguenti lemmi verbali: sono → **essere**_Verbo; andato → **andare**_Verbo; andava → **andare**_Verbo. Come si può notare la lemmatizzazione può essere utile nel caso di lingue flessive perché permette di lavorare sul lemma, ovvero "[un'] unità grafica che costituisce l'intestazione di un articolo o voce di dizionario o di enciclopedia" (Faloppa, 2010). In questo modo si riesce a calcolare meglio la ricchezza lessicale e il numero di *hapax*, ma quando il *corpus* presenta ambiguità semantica o un uso di un linguaggio figurato, diventa difficile capire a quale parte del discorso appartiene una determinata parola (cfr. Ondelli 2018, p. 147). Per questi motivi la lemmatizzazione, dal punto di vista metodologico, rimane un passaggio molto delicato e non adottato sistematicamente da tutti i linguisti e all'interno di questa tesi non verrà utilizzata, perché la tipologia di testi che verrà analizzata presenta figure retoriche e in alcuni casi una semantica ambigua.

2.3. Esempi di analisi quantitativa dei testi nel confronto tra traduzioni

In letteratura i maggiori studi riguardanti il confronto tra HT (*human translation*) e MT (*machine translation*) non utilizzano sistematicamente i parametri di base dell'*analisi dei dati testuali*, citati nel paragrafo precedente. Come si è visto nel sottocapitolo 1.3., le ricerche si basano soprattutto su un sistema di valutazione effettuato da *raters* scelti in precedenza (cfr. Wu et al., 2016; Läubli et al. 2018; Popel et al. 2020; Takakusagi et al. 2021), a volte integrato da una modifica dell'algoritmo del traduttore automatico per migliorarne la performance (cfr. Wu et al., 2016; Popel et al. 2020). Solo in uno degli studi precedentemente citati (cfr. Li et al. 2014) sono stati utilizzati due *software* di analisi statistica dei

[20] La lemmatizzazione di questa frase è stata effettuata con TreeTagger.

dati testuali, ovvero il Coh-Metrix e l'LIWC[21]. All'interno di questo studio, però, non sono stati eseguiti dei calcoli, come il TTR o il CCR[22], che di solito vengono inseriti all'interno di un'analisi dei dati testuali. Di conseguenza, nonostante l'analisi dei dati testuali non sia la metodologia maggiormente utilizzata negli studi di confronto tra HT e MT, la ricerca di Li et al. (2014) ci fa capire che anche i *software* tipici dell'analisi dei dati testuali possono essere utili per trovare le differenze e convergenze tra HT e MT. A sostegno di questa tesi troviamo diverse ricerche che, invece, utilizzando l'analisi dei dati testuali come metodologia per studiare le differenze tra testi originali e traduzioni umane e, in alcuni casi, anche le differenze tra traduzioni umane e automatiche.

Nello studio condotto da E. Kelih (2010), ad esempio, è stato utilizzato il numero di *Word types* (V) e *Word tokens* (N) come metro di misura per confrontare la distanza tra varie lingue slave. Nello studio è stata confrontata la traduzione in sloveno, croato, serbo, bulgaro, macedone, slovacco, ceco, polacco, serbo superiore, bielorusso e ucraino del romanzo russo *Come fu temprato l'acciaio* (Как закалялась сталь) scritto da N.A. Ostrovskij tra il 1932-1934. Il ricercatore ha dimostrato che il calcolo di N e V può essere utile per capire la vicinanza linguistica di due lingue. Infatti, dai risultati è stata confermata, ad esempio, la vicinanza tra le lingue russa, bielorussa e ucraina (lingue slave orientali) o tra il croato e il serbo (lingue slave meridionali), provando così l'attendibilità di V e N anche nello studio delle traduzioni.

Successivamente, nel 2018, una ricerca condotta da M. Kunilovskaya et al., ha utilizzato altri criteri dell'analisi dei dati testuali per comparare le traduzioni dall'inglese al russo eseguite da studenti di traduzione con quelle eseguite da professionisti, tenendo sempre conto del confronto con i testi originali in inglese. L'obiettivo era analizzare le caratteristiche lessicali e testuali di base delle traduzioni per trovare le differenze tra le performance traduttive degli studenti e dei professionisti, prendendo in considerazione le caratteristiche del *traduttese*. Per rispondere alle domande di ricerca sono stati effettuati dei procedimenti dell'analisi

[21] cfr. note p. 14 per ulteriori informazioni sul funzionamento di questi programmi
[22] maggiori informazioni sul ccr (*corpus coverage rate*) verranno fornite nel paragrafo 2.4.2.

dei dati testuali, tra cui la lemmatizzazione, il calcolo del TTR, del HTR e la distribuzione degli *high frequency words*[23] per verificare la lunghezza media delle frasi, la varietà e densità lessicale. Questi risultati hanno confermato la teoria degli universali del *traduttese* e le differenze qualitative tra le traduzioni prodotte da studenti e professionisti. In questo modo si conferma nuovamente l'attendibilità dell'analisi dei dati testuali in studi di traduttologia.

Per quanto riguarda il confronto tra traduzione automatica e traduzione umana, anche se in minor quantità, sono stati trovati degli studi che utilizzano parametri simili a quelli che verranno utilizzati nella presente tesi. Un esempio di questi studi è la ricerca condotta da C.S. Lee nel 2019. Lo studio aveva lo scopo di analizzare le differenze stilistiche tra la traduzione umana e automatica per quanto riguarda la coppia linguistica coreano-inglese nell'ambito della traduzione letteraria. Tra le varie domande di ricerca, l'autore si è chiesto se la traduzione automatica ha un proprio stile e se è chiaramente distinguibile dallo stile delle traduzioni umane. Per rispondere a queste domande sono stati utilizzati due metodi di rappresentazione della distanza tra i testi, ovvero il *multidimensional scaling* e la *hierarchical cluster analysis*. In entrambi è stata utilizzata la misura *Burrow's Delta*. I due metodi hanno permesso di rispondere con successo alle domande di ricerca ed è stato concluso che i traduttori automatici hanno degli stili interdipendenti e chiaramente distinguibili dalle traduzioni umane. Metodi stilometrici analoghi sono stati utilizzati anche all'interno di due studi recenti contenuti in due tesi di laurea magistrale supervisionate da J. Rybicki presso l'Università Jagielloński in Cracovia (Polonia). Il primo (Cembrzyńska et al. 2021) si è occupato del confronto delle traduzioni in inglese delle opere di Stanisław Lem prodotte sia da traduttori umani sia dal *software* di traduzione automatica DeepL. Nell'analisi sono stati utilizzati metodi stilometrici basanti sulla frequenza delle parole (*word frequency*). Lo stesso metodo è stato utilizzato anche da A. Żak et al. (2021) per confrontare le traduzioni umane ed automatiche, sempre realizzate da DeepL, dall'inglese al polacco delle

[23] maggiori informazioni su *high frequency words* (parole con una frequenza alta) verranno fornite nel paragrafo 2.4.2.

opere di J.R.R. Tolkien, Christopher Tolkien, C.S. Lewis e Charles Williams. Lo scopo della tesi era di testare l'attendibilità del *software* package *stylo* di R nel determinare la paternità, il genere e la cronologia delle opere e associarle alle relative traduzioni. Inoltre, studi approfonditi sulle traduzioni di DeepL, hanno portato alla conclusione che la traduzione automatica ha molte similitudini con la traduzione umana, ma, utilizzando l'*oppose() function* del pacchetto *stylo* ci si è accorti che la traduzione automatica ha dei problemi nell'utilizzo corretto degli articoli nella traduzione in polacco, perché semplifica eccessivamente certe parole e ha difficoltà nel tradurre elementi letterari complessi, come l'uso di forme linguistiche arcaiche. I risultati di queste analisi hanno portato gli autori a concludere che la MT di DeepL è progredita a tal punto dall'essere in grado di tradurre anche testi letterari, seppur con delle imperfezioni stilistiche.

Al termine di questo breve excursus, si può quindi confermare che i metodi e gli strumenti dell'analisi quantitativa di testi che verranno utilizzati all'interno di questa tesi sono attendibili e potranno essere utili non solo nell'ambito traduttologico, ma anche nel confronto tra traduzioni automatiche e umane.

2.4. *Corpus* oggetto di analisi

2.4.1. Descrizione delle fonti dei testi

Il *corpus* oggetto di analisi in questa tesi è costituito quattro *corpora*:

- "Articoli russi": articoli di giornale originali scritti in russo;
- "Traduzioni umane": traduzioni degli articoli russi dal russo all'italiano realizzate da traduttori (intesi come "esseri umani" in questo contesto);
- "DeepL": traduzioni automatiche dal russo all'italiano degli articoli russi effettuate da *DeepL*
- "Google Translate": traduzioni automatiche dal russo all'italiano degli articoli russi effettuate da *Google Translate*.

Ogni *corpus* è costituito da 34 testi per un totale di 136. Si è deciso di collezionare solo articoli di giornale pubblicati online in diverse testate russe, le cui traduzioni in lingua italiana fossero presenti nel sito russiaintranslation.com. Come si può leggere nel sito ufficiale, *Russia in Translation* "è un progetto online [non a scopo di lucro] che si prefigge di tradurre in maniera fedele ed imparziale articoli

dalle principali testate giornalistiche della Federazione Russa". Le traduzioni vengono prodotte solo dal russo all'italiano e lo scopo del progetto è quello di diffondere le notizie dei principali media russi al pubblico italiano senza schierarsi politicamente. Gli articoli non riguardano solo la Federazione russa, ma anche i paesi che appartenevano all'ex Unione Sovietica. Le traduzioni degli articoli toccano temi diversi, che vengono raggruppati in quattro sezioni principali: "Cucina", "Cultura", "Politica" e "Società". Per creare nel migliore dei modi un *corpus* omogeneo e ridurre l'influenza esercitata dei diversi argomenti sui dati testuali, si è deciso di collezionare solo le traduzioni degli articoli appartenenti alle sezioni Politica e Società. Inoltre, si è cercato il più possibile di raccogliere più traduzioni dello stesso autore e articoli russi della stessa redazione giornalistica. L'ideale sarebbe stato raccogliere anche più articoli russi scritti dallo stesso autore e tradotti dallo stesso traduttore, ma data la vastità di fonti russe utilizzate da *Russia in Translation* e l'ampio gruppo di traduttori che partecipano al progetto, ciò non è stato possibile (cfr. cap 6.2). Sempre per ridurre le potenziali fonti di variazione e creare un coprus di testi omogeneo, anche il periodo di tempo è stato ridotto a due anni di osservazione (2020-2021).

Per essere certi della presenza della traduzione italiana, gli articoli russi sono stati selezionati a partire dagli articoli presenti nel sito *Russia in Translation*, che contengono sempre un collegamento ipertestuale all'articolo originale russo. La collezione dei testi e delle loro traduzioni è iniziata a dicembre 2021 ed è terminata a giugno 2022.

Nella raccolta dei testi sono stati tenuti in considerazione il TTR e la percentuale di hapax. Inizialmente, infatti, erano stati raccolti 21 testi per ogni *corpus*, ma i dati di TTR e %*hapax* non erano soddisfacenti; quindi, è stato ritenuto opportuno ampliare i *corpus* fino a 34 testi per *corpus*, raggiungendo misure lessicometriche più soddisfacenti e adeguate a un approccio statistico.

Durante la raccolta dei testi è scoppiata la crisi russo-ucraina, che ha portato alla censura e/o chiusura e/o sospensione del lavoro di molte testate giornalistiche. Inoltre, la pagina *Russia in Translation* si è apertamente dichiarata contro la guerra e ha interrotto momentaneamente la traduzione di articoli a tema "Politica" e "Società". Data la situazione, si è deciso, quindi, di non utilizzare articoli pubblicati

nel 2022 in Russia per evitare la collezione di testi propagandistici, in grado di alterare il vocabolario del *corpus*, dove per "vocabolario" si intende una lista (la *word list*) di *word type* corredata dalla loro occorrenza nel *corpus* (cfr. paragrafo 2.4.2). Un approfondimento della crisi russo-ucraina dal punto di vista dell'analisi testuale di testi propagandistici merita una ricerca a parte, che in questa sede non verrà trattata.

Il testo originale più recente presente nel *corpus* "Articoli Russi" è stato pubblicato il 13 dicembre 2021 nel giornale Ekspert dal titolo *Protesti, turisti, ustojčivost' i soobšestva – četyre slova dlja ponimaija 2021 goda*[24] (V. Kozlov), mentre l'articolo più datato è stato pubblicato il 30 marzo 2020 nel giornale Tass e si intitola *"Nas b'jut, potomušto mi ne sašišaemsja". Trener rossijskich velogohšikov ob iske k WADA*[25] (A. Kyznezov). Le relative traduzioni di questi articoli sono *Proteste, turisti, sostenibilità e comunità: quattro parole per comprendere il 2021* (pubblicato il 30/01/2022, di E. Groppi) e *Ci colpiscono perché non ci difendiamo". L'allenatore dei ciclisti russi sulla causa contro la WADA* (pubblicato il 01/05/2020, di A. Lazzari), e sono rispettivamente la traduzione più recente e più datata del *corpus* "Traduzioni Umane".

Al contrario delle traduzioni provenienti tutte dalla stessa fonte, gli articoli sono stati pubblicati nelle seguenti testate giornalistiche (fonti dei siti presenti al cap. 6.4):

1. The Insider: testata giornalistica indipendente online specializzata in giornalismo investigativo e analisi politica. La sua redazione ha sede a Riga in Lettonia e pubblica articoli in russo con delle parti del sito disponibili in traduzione inglese. Il focus principale del giornale è la Russia;
2. Novaya Gazeta: periodico russo indipendente che pubblica articoli sia online sia in versione cartacea. La sede principale si trova a Mosca in Russia e pubblica principalmente in lingua russa, eccetto per una newsletter dal

[24] Russo: *Протесты, туристы, устойчивость и сообщества — четыре слова для понимания 2021 года*

[25] Russo: *"Нас бьют, потому что мы не защищаемся". Тренер российских велогонщиков об иске к WADA*

titolo *Russia, Explained* in cui vengono pubblicati articoli in lingua inglese. Il caporedattore è Dmitry Muratov, che ha ricevuto il premio Nobel per la Pace nel 2021. Il 28 marzo 2022 la *Novaya Gazeta* ha pubblicato un post nel sito, in cui dichiarava di dover sospendere la pubblicazione di articoli sul sito, sulle reti e sulla carta a causa di due avvertimenti ricevuti dal *Roskomnadzor* (Servizio federale per la supervisione delle comunicazioni, della tecnologia, dell'informazione e dei mass media) e di riprendere l'attività al termine della' "operazione speciale in territorio ucraino". Come si può vedere nel sito ufficiale, non hanno completamente interrotto la pubblicazione e pare continuino a lavorare da Riga, ma sicuramente la produzione è stata ridotta drasticamente;

3. Lenta.ru: testata giornalistica online specializzata in attualità, sport e pubblicità con sede a Mosca;
4. Expert: testata giornalistica nata nel 1995 a Mosca, che dal 2005 pubblica articoli anche online nel sito Expert.ru. I temi maggiormente trattati sono l'economia e la politica sia in ambito russo che internazionale. La testata si autodefinisce neutrale ideologicamente e dichiara di esporre i fatti in modo oggettivo. Il giornale possiede anche quattro redazioni locali, che si occupano di temi specifici di determinate aree russe, ovvero la Siberia, gli Urali, il Nord-Ovest e il Sud della Federazione Russa;
5. Meduza: testata giornalistica online fondata nel 2014, con sede in Riga (Lettonia). Il gruppo di giornalisti è costituito da professionisti, sia di nazionalità russa che no, esperti del mondo russo e dei Paesi dell'ex Unione Sovietica. Gli articoli vengono pubblicati sia in inglese che in russo e non sempre viene specificato l'autore del testo; quindi, può essere che in questi casi l'articolo sia una traduzione in russo del testo originale pubblicato precedentemente in inglese. La redazione si dichiara neutrale e di voler mettere in luce anche gli aspetti più scomodi del mondo russo e dell'ex Unione Sovietica. Al momento dell'attuale crisi in Ucraina, il giornale scrive di essere bandito in Russia, ma di poter continuare la propria attività al di fuori del Paese;

6. RIA Novosti: testata giornalistica online fondata nel 2014 a Mosca dall'azienda statale specializzata in media *Rossiya Segodnya*. Tratta di temi principalmente relativi alla Russia, ma possiede una parte chiamata *mir* (mondo) in cui tratta di temi internazionali. L'unica lingua in cui vengono pubblicati gli articoli è il russo.
7. Gazeta.Ru: testata giornalistica online con sede a Mosca e fondata nel 2016. Il giornale possiede anche delle redazioni più piccole, tra cui il giornale "Lenta.ru" nominato precedentemente.
8. Izvestija (o MIZ Izvestija): testata giornalistica fondata nel 2017 a Mosca. Oltre a pubblicare articoli online, mette in onda servizi giornalistici nei canali televisivi REN e 5TV.
9. Vedomosti: testata giornalistica fondata nel 1999 a Mosca. Il giornale, pubblicato sia online sia su carta, si occupa di economia e di fornire notizie in modo onesto e indipendente da correnti politiche. L'unica lingua di pubblicazione è il russo.
10. Moskvič: testata giornalistica russa online con sede a Mosca fondata dalla compagnia pubblicitaria russa *Urbanmedia*.
11. Diletant: giornale russo fondato nel 2015, che ha lo scopo di pubblicare articoli riguardanti la storia della Russia e dell'Unione Sovietica, utilizzando un linguaggio semplice e accessibile a lettori di tutte le età ed educazione.
12. Argumenti i Facti: testata giornalistica russa ufficialmente registrata presso il *Roskomnadzor* a partire da aprile 2020. Il sito web della redazione viene finanziato dal Ministero dello Sviluppo digitale, delle Comunicazioni e dei Mass Media della Federazione Russa.
13. Deutsche Welle (DW): compagnia mediatica tedesca che si occupa anche della stesura di articoli online non solo a tema tedesco, ma anche internazionale. La redazione pubblica articoli in 32 lingue diverse, tra cui il russo. Nel *corpus* della presente tesi l'unico articolo proveniente dalla DW è stato scritto dal giornalista russo K. Eggert in lingua russa.
14. TASS: agenzia di stampa statale russa fondata nel 1904 e attiva anche durante l'Unione Sovietica. Si occupa di diffondere tempestivamente notizie riguardanti la Russia nel modo più oggettivo possibile ed è presente nelle

principali piattaforme web, come VK e Telegram. Il sito web è disponibile anche in lingua inglese.

15. Pravda.ru: redazione online registrata per la prima volta nel 1999, mentre la versione cartacea esiste dal 1912. L'obiettivo dichiarato dall'azienda è quello di pubblicare articoli scritti da autori con opinioni diverse, in modo da dare la possibilità al lettore di sviluppare il proprio punto di vista a partire da una pluralità di opinioni.

Al capitolo 6 è possibile trovare le indicazioni bibliografiche degli articoli presenti nei *corpora* "Articoli Russi" (paragrafo 6.1) e "Traduzioni Umane" (paragrafo 6.2).

Gli ultimi due *corpora* oggetto di analisi sono costituiti da traduzioni automatiche effettuate da due strumenti di traduzione automatica: Google Translate e DeepL. Come già accennato precedentemente, Google Translate è un *software* elaborato dall'azienda Google e lanciato online nel 2006. Inizialmente utilizzava un algoritmo *statistical-based* (cfr. Costa-jussa et al. 2012, p. 254), successivamente a partire dal 2016 gli sviluppatori hanno iniziato ad utilizzare algoritmi *neural-based* (cfr. cap. 1.1 e 1.3). Attualmente il programma è in grado di tradurre in 133 lingue diverse, è gratuito e riesce a tradurre file con le seguenti estensioni: .docx, .pdf, .pptx, .xlsx (cfr. https://translate.google.com/). Anche DeepL utilizza network neurali e traduce in 28 lingue diverse. Il *software* offre la possibilità di tradurre frasi gratuitamente fino a un massimo di 5000 caratteri. Inoltre, è possibile tradurre file con estensione .pdf, .docx, .pptx senza limiti di caratteri, ma nella versione gratuita non è possibile modificare, copiare o incollare la traduzione prodotta dal *software*. Ai fini di questa tesi, si è deciso di utilizzare una delle versioni a pagamento di DeepL Pro per avere accesso a più funzioni del programma, tra cui la possibilità di copiare e incollare le traduzioni. In base alla tipologia di abbonamento che si decide di fare, c'è sempre un numero massimo di traduzioni che possono essere effettuate in un mese. Nella versione *Starter* possono essere effettuate massimo 5 traduzioni, in quella *Advanced* massimo 20, mentre nell'abbonamento *Ultimate* massimo 100. Sia *Starter* che *Advanced* permettono di usufruire di un periodo di prova gratuito fino a 30 giorni (cfr. https://www.deepl.com/pro?cta=header-pro).

Entrambi i *software* di traduzione possiedono un riconoscimento automatico della lingua da tradurre e l'utente deve solo inserire la lingua di arrivo in cui desidera tradurre il testo. All'interno di questa tesi si è deciso di utilizzare due *software* di traduzione automatica per capire se esistono delle differenze qualitative tra *software* di traduzione automatica. Entrambi utilizzano reti neurali e dovrebbero produrre traduzioni di qualità simile. In realtà il *software* DeepL è molto più utilizzato da aziende e traduttori professionisti, facendo pensare quindi che sia più preciso rispetto a Google Translate. Inoltre, per accertarsi della qualità del *software*, l'azienda conduce regolarmente dei *blind test*, in cui traduttori professionisti devono decidere la qualità della traduzione automatica senza sapere quale *software* l'ha generata. Secondo il sito ufficiale dell'azienda, DeepL produce traduzioni migliori rispetto alle traduzioni prodotte da altri programmi di traduzione automatica con un rapporto di 3:1 (cfr. https://www.deepl.com/en/whydeepl). I risultati di questi test e le critiche dei traduttori professionisti vengono utilizzati per verificare e migliorare le *performance* del programma.

2.4.2. Tassi di copertura e struttura interna del *corpus*

Come già scritto nel paragrafo 2.2, prima di iniziare ad analizzare i *corpus* è necessario capire se i *corpus* collezionati siano buoni o meno. In fase preliminare sono stati calcolati i TTR e la quantità di *hapax* per ogni *corpus* oggetti di analisi. I risultati di queste analisi sono riassunti nella tabella sottostante:

Corpus	**N**	**Mean Size**	**V**	**TTR %**	**V_1**	**hapax%**
Articoli Russi	41419	1218,2	13837	33,4	9379	67,8
Traduzioni Umane	56525	1662,5	9628	17,0	5416	56,3
DeepL	52945	1557,2	8856	16,7	4851	54,8
Google Translate	53370	1569,7	8920	16,7	4898	54,9

Tab. 1- *Valutazione preliminare delle dimensioni dei corpora*

dove "N" indica il numero di *word token* presenti nel singolo *corpus*; "Mean Size" è la media del numero di *word token* per singolo articolo (N/№ testi); "V" rappresenta il numero di *word type* presenti in ogni *corpus*; il "TTR" è il rapporto

tra il numero di *word type* e il numero di *word token* (V/N*100) espresso in percentuale; "V1" indica il numero di *hapax* presenti nei singoli *corpora*; "hapaxes%" è il rapporto rapporto tra il numero di *hapax* e il numero di *word type* (V_1/V*100) espresso in percentuale. Per calcolare il numero di *word token* e *word type* è stato utilizzato il *software AntConc* sia per i *corpora* in italiano che per quello in russo.

Per rendere il calcolo comparabile, il programma è stato settato in modo da individuare come *token* le parole costitute da lettere dell'alfabeto, considerare separatori i segni di interpunzione e ignorare i numeri (che non vengono quindi inclusi nel calcolo di *token*). Inoltre, sono stati utilizzati dei *tag* per escludere dal vocabolario dei *corpora* informazioni non rilevanti, come il nome dell'autore dell'articolo, la data di pubblicazione e il giornale di appartenenza.

Come si può notare, il TTR è sotto la soglia limite del 20% per tutti i *corpora* tradotti in italiano, mentre raggiunge il 33,4 % nel *corpus* in lingua russa. Questo risultato è conforme con i risultati ottenuti negli studi precedenti (cfr. Kelih 2010, p. 2; Dmitrieva & Tiedemann 2021, p. 75). Come scritto nel paragrafo 2.2, il russo è una lingua molto flessiva e il programma *AntConc* riconosce come diversi *word token* la forma coniugata di uno stesso verbo o la forma declinata di una stessa parola in quanto è settato per riconoscere le parole diverse (intese come forme grafiche, cioè *word type*). Per questo motivo ci si aspettava un superamento della soglia del 20%, anzi il TTR ottenuto è più basso di alcuni TTR che si trovano in letteratura (cfr. Dmitrieva & Tiedemann 2021; Kunilovskaya & Kutuzov 2015; Kunilovskaya et al. 2018). Il dato meno positivo è dato dalla percentuale di *hapax* presente nei *corpora*. Se per il *corpus* in russo ci si aspettava che superasse la soglia del 50% per i motivi sopraesposti, per quanto riguarda i *corpora* in italiano i risultati ottenuti non sono ottimali. Come scritto nel paragrafo 2.2, se il *corpus* presenta un numero di *hapax* superiore al 50%, significa che c'è un numero troppo elevato di parole originali e non risulta trattabile da un punto di vista statistico (Tuzzi 2003, p. 76).

Con la collezione iniziale di 21 testi, si era raggiunta una percentuale di *hapax* del 57% nel *corpus* di Google Translate e del 58% nei *corpora* di DeepL e degli articoli tradotti da esseri umani. Inizialmente anche i TTR erano molto elevati (19%

per tutti i *corpora* in italiano). Si è cercato di diminuire questi valori aggiungendo testi, fino al raggiungimento di 34 articoli. Il TTR è diminuito, raggiungendo dei valori quasi ottimali, mentre la percentuale *hapax* è diminuita, ma non ha raggiunto valori minori del 50%. È stato notato che l'aumento di articoli diminuiva il rapporto tra *word type* e *word token*, ma aumentava il numero di parole originali utilizzate una sola volta. Questo è un effetto abbastanza comune nel linguaggio giornalistico sia perché i temi sono sempre diversi, sia perché gli articoli di giornale sono infarciti di luoghi, nomi, prestiti o neologismi che non si ripetono (cfr. Šarac 2019).

Inoltre, analizzando meglio gli *hapax* con una prospettiva qualitativa, si è notato che alcuni di essi erano costituiti da parole o singole lettere russe traslitterate in alfabeto latino, oppure forme diverse dello stesso verbo (es. "dedicata", "dedicati", "dedicò" nel *corpus* DeepL). Nonostante il numero di *hapax* elevato, soprattutto nel *corpus* di articoli tradotti da essere umani, si è deciso di continuare comunque le analisi, e di tenere in considerazione questa problematica se ci fossero presentati dei risultati ambigui in fase di analisi dei *corpora*.

Dopo aver analizzato la struttura generale dei singoli *corpora*, è stata condotta un'ulteriore analisi focalizzata sul vocabolario e la struttura interna dei singoli *corpora*. Facendo dei calcoli più approfonditi sui *word type* di un *corpus*, è possibile ottenere molte informazioni sul vocabolario. Una di queste sono le fasce o zone di frequenza, ovvero "una suddivisione delle parole in classi sulla base della frequenza con cui compaiono in un *corpus*" (Tuzzi 2003, p. 77; Bolasco 1999, p. 202). Le classi utilizzate in letteratura sono tre: *high* (alte), *medium* (medie), *low* (basse). Le parole ad alta frequenza sono quelle più utilizzate e occupano le prime posizioni del vocabolario del *corpus*. Per individuarle si parte dal primo *word token* con frequenza massima fino a trovare la prima coppia consecutiva di *word token* in cui la frequenza si ripete. Come afferma A. Tuzzi (2003, p. 77), è stato notato empiricamente che le "prime parole del vocabolario hanno tutte frequenza diversa con ampi salti tra posizioni consecutive […] fino ad arrivare al punto in cui si ha la prima differenza nulla, cioè una frequenza che si ripete uguale per due termini consecutivi diversi". Successivamente, si trovano le parole a bassa frequenza d'uso partendo dagli *hapax* e salendo la lista di *word type* fino ad arrivare al primo salto di frequenza tra due parole. Partendo dagli *hapax*, infatti, si trovano molti *word type*

con frequenza 1, 2, 3, 4 e così via, e questa successione continua sequenzialmente fino ad arrivare a un punto in cui l'ordine numerico si spezza e si trova una lacuna nella successione (cfr. Tuzzi 2003, p. 80). Le parole a media frequenza sono i *word type* contenuti tra quelli ad alta e bassa frequenza. Dopo aver individuato le parole ad alta, media e bassa frequenza, è interessante calcolare i tassi di copertura, ovvero il *corpus coverage rate* (ccr) e il *vocabulary coverage rate* (vcr). Il primo indica la percentuale di copertura del *corpus* di determinate forme grafiche con una frequenza superiore ad una certa soglia γ e si ottiene dal "rapporto tra il numero di *word token* che appartengono alle classi di frequenza superiore a γ e la dimensione N del *corpus*" (Tuzzi 2003, p. 81):

$$ccr = \frac{\sum_{m=\gamma}^{f_{max}} mV_m}{N}$$

Fig. 3 - *Formula per il calcolo del ccr* (da Tuzzi 2003, p. 81)

Il secondo, analogamente al ccr, indica la percentuale di copertura del vocabolario di determinati *word type* con una frequenza superiore ad una certa soglia γ. Il vcr è dato, quindi, dal "rapporto tra il numero di *word type* rappresentanti delle classi di frequenza superiore a γ e la dimensione V(N) del vocabolario" (Tuzzi 2003, p. 81):

$$vcr = \frac{\sum_{m=\gamma}^{f_{max}} V_m}{V}$$

Fig. 4 - *Formula per calcolare il vcr* (da Tuzzi 2003, p. 81)

Per ottenere questi calcoli, sono stati collezionati gli elenchi di *word type* presenti nei singoli *corpora* utilizzando una funzione di *AntConc*. Il programma permette di salvare in formato .txt una lista completa di *word type*, in cui ogni *word type* è associato ad un *rank* (posizione di un *word type* nella lista, dove *rank 1* indica la parola più utilizzata e a scalare quelle meno utilizzate) e un'occorrenza, ovvero quante volte un certo *word type* viene ripetuto all'interno di un *corpus*. Questi dati

sono stati importati su un foglio di calcolo .xlsx, in cui è stato possibile eseguire i calcoli, visibili nella tabella Excel riportata di seguito:

Rank	Word Types	Occorrenze	Frequenza relativa	Rate%1000	Vcr%100	Frequenza cumulata	Ccr%100	Zona
1	di	2029	0,0383	38,3228	0,011	2029	3,8323	high
2	e	1303	0,0246	24,6104	0,023	3332	6,2933	high
3	il	1128	0,0213	21,3051	0,034	4460	8,4238	high
4	la	1118	0,0211	21,1163	0,045	5578	10,5355	high
5	è	1101	0,0208	20,7952	0,056	6679	12,6150	high
6	un	916	0,0173	17,3010	0,068	7595	14,3451	high
7	che	881	0,0166	16,6399	0,079	8476	16,0091	high
8	in	742	0,0140	14,0145	0,090	9218	17,4105	high
9	non	690	0,0130	13,0324	0,102	9908	18,7138	high
10	a	682	0,0129	12,8813	0,113	10590	20,0019	high
11	per	675	0,0127	12,7491	0,124	11265	21,2768	high
12	i	602	0,0114	11,3703	0,136	11867	22,4138	high
13	l	548	0,0104	10,3504	0,147	12415	23,4489	high
14	ha	543	0,0103	10,2559	0,158	12958	24,4745	high
15	del	523	0,0099	9,8782	0,169	13481	25,4623	high

Tab. 2 - *DeepL corpus description*

dove per "frequenza relativa" si intende il rapporto tra l'occorrenza di una parola e la somma di tutte le occorrenze; "rate%1000" indica la frequenza relativa moltiplicata per 1000; il "Vcr%100" ottenuto dal rapporto tra il *rank* di una parola e il *rank* dell'ultima parola moltiplicato per 100; la "frequenza cumulata" ottenuta dalla somma della frequenza di una parola con le frequenze delle parole precedenti; il "Ccr%100" dato dal rapporto tra la frequenza cumulata di una parola e la somma delle frequenze cumulate di tutti i *word type* moltiplicato per 100; "zona", infine, indica la zona o fascia di frequenza di ogni singolo *word type*. La Tabella 2 rappresenta i calcoli effettuati sui primi 15 *word type* presenti nel *corpus* contenente le traduzioni automatiche di DeepL. I calcoli presenti nella tabella sono stati effettuati per ogni *word type* del *corpus* DeepL (in totale: 8856 *word types*), inoltre gli stessi calcoli sono stati effettuati negli altri tre *corpora* utilizzando le medesime procedure. I risultati principali possono essere riassunti nei seguenti schemi:

- ***Corpus* Articoli Russi**

Numero di *word type* con frequenza alta (*high*): 27 (occorrenza ≥ 105) e ricoprono il:

- 0,20% del vocabolario (vcr);
- 22,40% del *corpus* (ccr).

Numero di *word type* con frequenza alta e media (*high* + *medium*): 63 (occorrenza ≥ 54) e ricoprono:

- 0,46% del vocabolario (vcr);
- 28,62% del *corpus* (ccr).

Numero di *word type* con frequenza bassa (*low*): 13.773 (occorrenza ≤ 50) e ricoprono:

- 99,54% del vocabolario (vcr);
- 71,25% del *corpus* (ccr).

Tra i *word type* a bassa frequenza (*low*) si trovano 9.379 *hapax*, che rappresentano:

- 67,78% del vocabolario (vcr);
- 22,64% del *corpus* (ccr).

Per ottenere una copertura maggiore del 70% (*corpus coverage rate*) si ha bisogno di *word type* con le seguenti frequenze o occorrenze:

- ≥ 2 ricopre il 77,35% del *corpus* (4.457 *word type*);

Purtroppo, all'interno di questo *corpus* per ottenere una copertura maggiore del 77,35% è necessario includere gli *hapax*, infatti per raggiungere una copertura totale del corpus esclusi gli *hapax* si ha bisogno di *word type* con frequenza o occorrenza:

- ≥ 2 per ricoprire il 77,35% del *corpus* (4.457 *word types*).

- ***Corpus* Traduzioni Umane**

Numero di *word type* con frequenza alta (*high*): 22 (occorrenza ≥ 326) e ricoprono il:

- 0,23% del vocabolario (vcr);
- 29,79% del *corpus* (ccr).

Numero di *word type* con frequenza alta e media (*high* + *medium*): 124 (occorrenza ≥ 51) e ricoprono:

- 1,29% del vocabolario (vcr);
- 48,46% del *corpus* (ccr).

Numero di *word type* con frequenza bassa (*low*): 9.503 (occorrenza ≤ 49) e ricoprono:

- 98,70% del vocabolario (vcr);
- 51,45% del *corpus* (ccr).

Tra i *word type* a bassa frequenza (*low*) si trovano 5.416 *hapax*, che rappresentano:

- 56,25% del vocabolario (vcr);
- 9,58% del *corpus* (ccr).

Per ottenere una copertura maggiore del 70% (*corpus coverage rate*) si ha bisogno di *word type* con le seguenti frequenze o occorrenze:

- ≥ 8 ricopre il 71,68% del *corpus* (939 *word type*);
- ≥ 4 ricopre 81,34% del *corpus* (2.018 *word type*);
- ≥ 3 ricopre l'85,28% del *corpus* (2.760 *word type*).

Per ottenere una copertura totale del *corpus* esclusi gli *hapax* si ha bisogno di *word type* con frequenza o occorrenza:

- ≥ 2 per ricoprire il 90,42% del *corpus* (4.211 *word types*).

- ***Corpus* DeepL**

Numero di *word type* con frequenza alta (*high*): 26 (occorrenza ≥ 234) e ricoprono il:

- 0,29% del vocabolario (vcr);
- 32,81% del *corpus* (ccr).

Numero di *word type* con frequenza alta e media (*high* + *medium*): 76 (occorrenza ≥ 72) e ricoprono:

- 0,86% del vocabolario (vcr);
- 44,54% del *corpus* (ccr).

Numero di *word type* con frequenza bassa (*low*): 8780 (occorrenza ≤ 70) e ricoprono:

- 99,13% del vocabolario (vcr);
- 55,33% del *corpus* (ccr).

Tra i *word type* a bassa frequenza (*low*) si trovano 4852 *hapax*, che rappresentano:

- 54,78% del vocabolario (vcr);
- 9,16% del *corpus* (ccr).

Per ottenere una copertura maggiore del 70% (*corpus coverage rate*) si ha bisogno di *word type* con le seguenti frequenze o occorrenze:

- ≥ 9 ricopre il 70,13% del *corpus* (757 *word type*);
- ≥ 4 ricopre l'81,55% del *corpus* (1.877 *word type*);
- ≥ 3 ricopre l'85,30% del *corpus* (2.539 *word type*).

Per ottenere una copertura totale del *corpus* esclusi gli *hapax* si ha bisogno di *word type* con frequenza o occorrenza:

- ≥ 2 per ricoprire il 90,84% del *corpus* (4.004 *word types*).

- ***Corpus* Google Translate**

Numero di *word type* con frequenza alta (*high*): 26 (occorrenza ≥ 250) e ricoprono il:

- 0,29% del vocabolario (vcr);
- 32,36% del *corpus* (ccr).

Numero di *word type* con frequenza alta e media (*high* + *medium*): 125 (occorrenza ≥ 46) e ricoprono:

- 1,04% del vocabolario (vcr);
- 49,39% del *corpus* (ccr).

Numero di *word type* con frequenza bassa (*low*): 8794 (occorrenza ≤ 44) e ricoprono:

- 98,59% del vocabolario (vcr);
- 50,53% del *corpus* (ccr).

Tra i *word type* a bassa frequenza (*low*) si trovano 4898 *hapax*, che rappresentano:

- 54,91% del vocabolario (vcr);
- 9,18% del *corpus* (ccr).

Per ottenere una copertura maggiore del 70% (*corpus coverage rate*) si ha bisogno di *word type* con le seguenti frequenze o occorrenze:

- ≥ 9 ricopre il 70,31% del *corpus* (787 *word type*);
- ≥ 4 ricopre l'81,69% del *corpus* (1.924 *word type*);
- ≥ 3 ricopre l'85,51% del *corpus* (2.605 *word type*).

Per ottenere una copertura totale del *corpus* esclusi gli *hapax* si ha bisogno di *word type* con frequenza o occorrenza:

- ≥ 2 per ricoprire il 90,82% del *corpus* (4.021 *word types*).

Come si poteva prevedere, data la grande quantità di *hapax*, ci sono in tutti i *corpora* un'alta percentuale di *low frequency words* che occupano più del 50% di tutti i *corpora*. Per quanto riguarda i *corpora* dei traduttori automatici si trovano delle percentuali simili nella copertura delle parole ad alta frequenza (*high frequency words*), che occupano poco meno del 33% degli interi *corpora*, mentre si trovano delle differenze nella copertura delle parole a media e alta frequenza:

DeepL raggiunge il 45% mentre Google Translate quasi il 50% di copertura. Le traduzioni umane, invece, nonostante le parole ad alta frequenza non riescano a coprire più del 30% del *corpus*, le parole ad alta e media frequenza riescono a coprire quasi il 49% del *corpus*, mostrando risultati più simili al traduttore automatico Google Translate. Sorprendentemente la percentuale di copertura più alta dei *low frequency words* si trova nel traduttore automatico DeepL (55,33%), che presenta invece il minor numero di *hapax*.

Un dato positivo che accomuna tutti i *corpora* in italiano è che gli *hapax*, pur essendo molto elevati, non occupano molto spazio nei *corpora* e si raggiunge una copertura del 90% senza di essi. Questo risultato non si ripete nel *corpus* in russo, che invece presenta un'altissima copertura di *hapax*, tanto che senza di essi si riesce a coprire solo il 77,35% dell'intero *corpus*. Da una prospettiva qualitativa, osservando più attentamente gli *hapax* russi riconosciuti da *AntConc*, si nota che non solo molte forme della stessa parola sono indicate separatamente in quanto si tratta di forme diverse dal punto di vista della logica di riconoscimento (es. *abchazskimi*, *abchazskogo*, *abchazskom*[26] sono declinazioni diverse dello stesso aggettivo), ma anche molte parole scritte in alfabeto latino sono presenti tra gli *hapax* (es. "act", "action", "amnesty"). Questo significa che negli articoli giornalistici russi non tutte le parole di origine straniera vengono traslitterate in alfabeto cirillico, ma alcune rimangono in alfabeto latino (es. "bank", "capital") e sono comunque poco frequenti. Inoltre, bisogna specificare che gli articoli raccolti nei *corpora* "Articoli Russi" e "Traduzioni Umane" hanno molti autori diversi e non sempre trattano gli stessi temi; quindi, può essere che la grande vastità di vocaboli diversi sia dovuta alla varietà di autori e temi (gli articoli di giornale restano molto vari sebbene sia stata fatta una selezione degli argomenti di ambito politico-sociale). Analogamente le traduzioni automatiche, traducendo articoli con autori e temi diversi, riportano caratteristiche simili agli altri due *corpora*.

[26] Russo: "абхазскими", "абхазского", "абхазском"

3. Metodi e strumenti di analisi

Dopo aver terminato la fase di preparazione dei *corpora* e aver analizzato le caratteristiche generali di ogni *corpus*, si passa all'analisi effettiva dei contenuti e delle principali caratteristiche di interesse. In questa fase, illustrata successivamente nel capitolo 4, si utilizzeranno diversi metodi quantitativi di analisi dei testi, realizzabili attraverso l'uso di *software* differenti. All'interno di questo capitolo verranno elencati i metodi e gli strumenti, che saranno poi utilizzati durante l'analisi delle differenze tra traduzione umana e automatica. I riferimenti bibliografici dei singoli *software* sono inclusi nei riferimenti bibliografici, elencati al capitolo 7.

3.1. Analisi delle traduzioni in *AntConc*: uso delle funzioni *Concordance plot* e *File View*

Il primo studio che verrà effettuato è a livello lessicale e si utilizzerà il *software* gratuito *AntConc*, sviluppato da Laurence (2019), per realizzare le analisi lessicali. Questo programma è una tipologia di *concordance program*[27] che permette di avere accesso ai dati testuali di un *corpus* e analizzare determinati fenomeni linguistici tramite una serie di funzioni disponibili nel programma (cfr. Römer & Wulff 2010, p.103). La prima funzione utilizzata per raggiungere gli obiettivi della presente tesi, è stato il calcolo di *word type* e *word token*, descritto nel paragrafo 2.4.2 e verrà utilizzata anche in fase di analisi, per controllare se determinati *word type* siano stati effettivamente tradotti in modo univoco in italiano. In fase di analisi, verranno scelte una serie di acronimi, abbreviazioni, prestiti stranieri, forestierismi, termini e traslitterazioni dall'alfabeto cirillico a quello latino e viceversa, di cui si confronterà:

1. la capacità dei *software* di traduzione automatica di tradurre correttamente questi termini, considerando come variante corretta la traduzione umana (cfr. Li et al. 2014);

[27] trad. programma di concordanza

2. successivamente, verrà controllato se il numero delle occorrenze di una serie di termini sia lo stesso sia nei testi originali che nelle traduzioni umane o automatiche, per verificare se la loro traduzione sia effettivamente univoca e non siano stati impiegati dei sinonimi.

La scelta dei *word type* da analizzare verrà effettuata controllando la *word list* proposta da *AntConc* per ogni *corpus* e cercando termini, acronimi, abbreviazioni, prestiti stranieri, forestierismi e traslitterazioni che potrebbero essere tradotti in modo equivoco dai *software* di traduzione automatica. Per agevolare questa ricerca, verrà fatto uso di *stoplist*, ovvero una lista di parole che vengono molto utilizzate in una lingua e spesso coincidono con articoli, congiunzioni, preposizioni, ecc. Dato il loro ampio utilizzo, queste parole figurano quasi sempre tra le parole più utilizzate in un *corpus*, ma in realtà non forniscono alcuna informazione rilevante sui contenuti lessicali e distintivi di un testo (cfr. Dolamic & Savoy 2009, p. 200). Per questo motivo si può decidere di escludere queste parole (dette *stopword*) nella produzione di una *word list*, inserendo una *stoplist* all'interno di *AntConc*. Il problema degli *stopword* è che non esiste una metodologia chiara e univoca per la definizione di queste parole (cfr. Dolamic & Savoy 2009, p. 200). Se in generale le preposizioni, i pronomi personali o gli aggettivi possessivi sono molto utilizzati in una lingua, non sempre è proficuo escluderli da un corpus. In un testo di propaganda politica, ad esempio, parole come "noi", "con", "nostro" possono essere utili per capire lo stile di un determinato politico ed è meglio non escluderle. Di conseguenza è importante controllare le *stopword* della *stoplist* che si deciderà utilizzare per togliere o aggiungere determinate parole in base alle domande di ricerca alle quali si desidera rispondere. All'interno di questa tesi, per identificare le parole più utilizzate (*most frequent words*) si utilizzeranno due *stoplist*, una in italiano e l'altra in russo. La lista di *stopword* italiane è stata presa dal programma online di analisi testuale *Voyant-tools* e non sono state apposte modifiche. Per quanto riguarda la *stoplist* in russo, non è stato possibile scaricarla dallo stesso sito, perché il *software* non è in grado di leggere documenti in lingua russa. È stato, quindi, necessario consultare un'altra fonte e si è deciso di utilizzare la *stoplist* proposta dal sito *Ranks.nl*.

La scelta dei termini da analizzare verrà effettuata sulla base sia dei primi *most frequent words* presenti nel *corpus* "Articoli Russi" sia di alcune traduzioni scorrette di acronimi o nomi propri che sono state individuate durante la collezione dei testi. Ad esempio, si è notato che la parola russa "телеграм канал", che in italiano si riferisce al *social network* Telegram, è tradotto in DeepL con "canale telegrafico" e in Google Translate con "canale telegramma". Questa problematica verrà approfondita analizzando altre parole che presentano traduzioni simili.

Durante l'analisi, si esaminerà la traduzione delle parole scelte attraverso due funzioni di *AntConc*: *Concordance plot* e *File View*. La prima permette di visualizzare velocemente in quali testi è presente una determinata parola e in quale punto dei testi occorre (Römer & Wulff 2010, p. 110). Una rappresentazione della funzione è disponibile qui sotto:

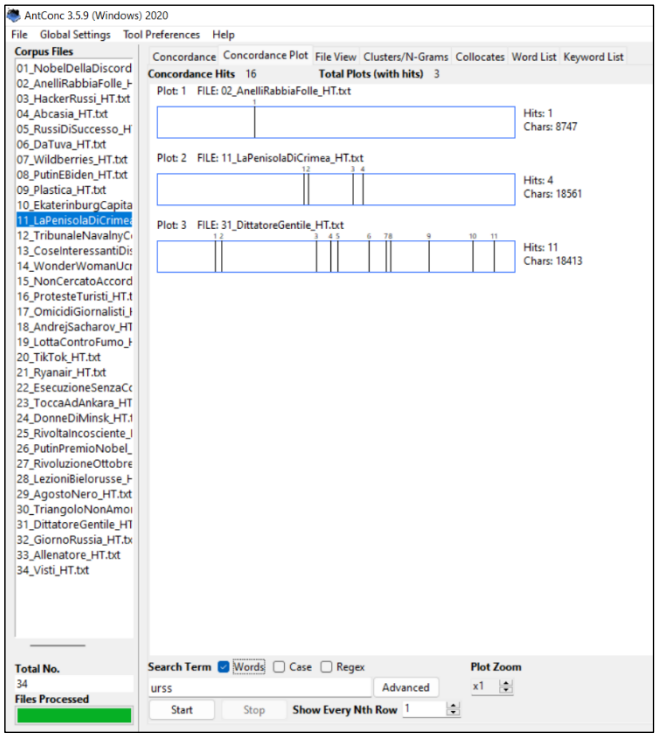

Fig. 5 - *Schermata della funzione Concordance Plot in AntConc*

dove "FILE" indica il documento in cui si trova la parola che si sta analizzando, le strisce verticali rappresentano la posizione dell'occorrenza di tale parola in un testo e "Hits" ne indica l'occorrenza. La funzione *File View*, invece, sarà utilizzata per vedere in modo più specifico dove si trova una determinata parola all'interno del testo (cfr. Römer & Wulff 2010, p. 107-108). Questa funzione servirà a vedere in

modo più specifico la traduzione di una determinata parola, nel caso in cui non fosse possibile trovarla attraverso il *Concordance plot*. Può accadere, infatti, che un traduttore non traduca in modo letterale determinate parole e l'unico modo veloce per capire com'è stato tradotto un *word-type*, è effettuare un controllo incrociato tra la posizione della parola nel testo originale e la sua posizione nel testo tradotto. In questo modo, sarà possibile individuare la traduzione della parola presa in esame. Una rappresentazione della funzione *File View* è disponibile qui sotto:

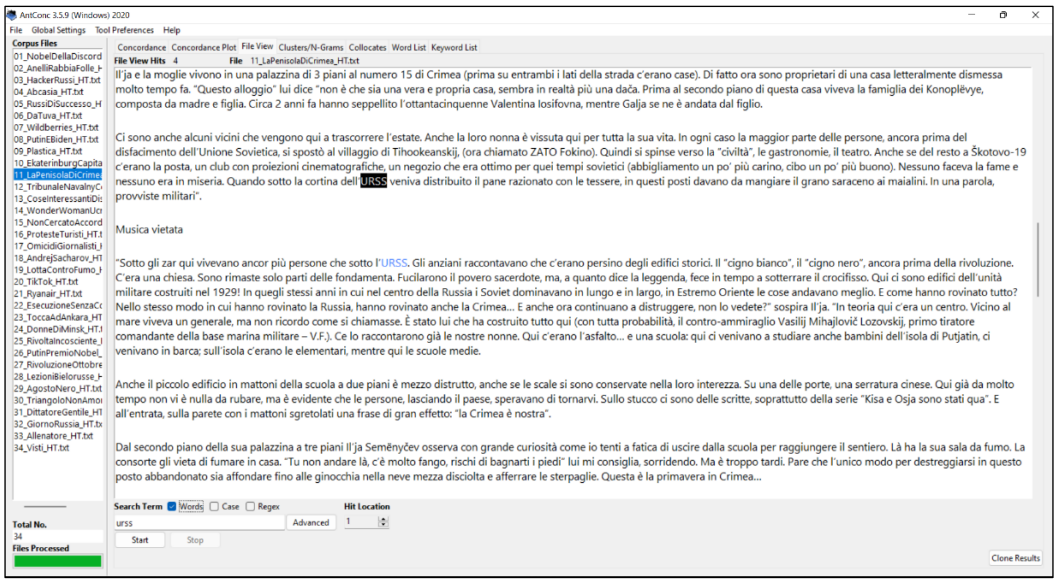

Fig. 6 - *Schermata della funzione* File View *in* AntConc

dove il *word-type* che si sta analizzando è visibile in un altro colore rispetto al testo, in modo da facilitarne l'identificazione durante lo scorrimento del testo.
Successivamente, durante la fase di controllo della corretta traduzione di termini, che ammettono solo una traduzione possibile, si porrà maggiore attenzione non solo alla traduzione ma anche al numero delle occorrenze.

3.2. Livello di *semplificazione* delle traduzioni: confronto tra il TTR e la lunghezza media di una frase

Nella seconda fase di analisi si è deciso di esaminare il grado di *semplificazione* delle traduzioni italiane, utilizzando come parametri il *Type Token Ratio* e la lunghezza media delle frasi. Questi sono due indicatori che vengono utilizzati per analizzare l'universale traduttivo della *semplificazione* negli studi di traduttese (cfr. cap. 1.2). In questo contesto, con il termine *semplificazione* si intende la tendenza

del traduttore a semplificare la lingua durante la produzione del testo di arrivo (cfr. Ondelli & Viale 2010, pp. 3-5).

Inoltre, il TTR e la lunghezza media della frase, sono alcuni dei criteri utilizzati da Kunilovskaya et al. (2018) per comparare le traduzioni dall'inglese al russo realizzate da traduttori professionisti e studenti di traduzione. Prima di tutto, i ricercatori hanno calcolato la lunghezza media di un ogni testo dei tre *corpora* sotto esame, ovvero un *corpus* con i testi originali inglesi, uno con i testi tradotti da professionisti e l'ultimo con i testi tradotti da studenti di traduzione. Al termine delle analisi si è scoperto che la lunghezza media delle frasi tradotte da professionisti era maggiore rispetto a quella degli studenti, confermando il fatto che uno studente è più influenzato dalla struttura sintattica del testo di partenza rispetto al traduttore professionista, che adatta maggiormente la sintassi alla lingua di arrivo (cfr. Kunilovskaya et al. 2018, p. 38-39). In secondo luogo, per verificare la varietà lessicale delle traduzioni, Kunilovskaya et al. (2018) hanno calcolato il *Type Token Ratio* di un ogni testo dei tre *corpora* presi in esame. Successivamente è stata effettuata una media dei TTR per ogni *corpus* e si è scoperto che il TTR delle traduzioni degli studenti era minore rispetto alla percentuale di *Type Token Ratio* presente nel *corpus* dei traduttori professionisti. In questo modo, i ricercatori hanno potuto concludere che i testi tradotti dagli studenti erano più semplici e meno complessi da un punto vista lessicale rispetto alle traduzioni dei professionisti (Kunilovskaya et al. 2018, p. 40).

Nella presente tesi, verranno effettuate le stesse misurazioni per confrontare il livello di *semplificazione* della traduzione automatica da quella umana, e trovare eventuali differenze tra i due *software* di traduzione automatica. I programmi di *machine translation* tendono a tradurre più letteralmente rispetto ad un essere umano (cfr. Ibanez 2021); quindi, ci si aspetta che la traduzione automatica presenti valori minori sia nella lunghezza media della frase sia nel TTR, analogamente al *corpus* di studenti analizzato da Kunilovskaya et al. (2018). In questo modo sarà possibile studiare l'universale traduttivo della *semplificazione* e verificare se effettivamente le traduzioni automatiche sono più semplici da un punto di vista lessicale e sintattico, rispetto alle traduzioni umane.

Le operazioni sopraindicate verranno effettuate con il programma *Voyant-tools* sviluppato da Sinclair, Stéfan & Rockwell 2016. Il *software* permette di calcolare automaticamente il *Type Token Ratio* e la lunghezza media delle frasi per ogni testo di un *corpus*. Il *software*, purtroppo, non legge testi in lingua russa; quindi, non sarà possibile effettuare il confronto tra le traduzioni e i testi originali, presente nello studio di Kunilovskaya et al. (2018). Per questo motivo, i calcoli verranno effettuati solo nei tre *corpora* di traduzioni italiane. In ogni caso, la verifica di questi parametri risulta affidabile solo nei casi in cui i testi presentano una lunghezza simile e sono scritti nella stessa lingua, perché il TTR e la lunghezza media delle frasi sono dei parametri che dipendono dal numero *n* di *word-type* e *word-token* presenti in un *corpus* e ovviamente variano da lingua a lingua. Un approfondimento dei motivi per cui questa analisi rimane valida è visibile al capitolo 4.2.

Infine, per verificare l'attendibilità dei calcoli di *Voyant-tools* sono state effettuate delle prove, calcolando manualmente il *Type Token Ratio* e la lunghezza media delle frasi di alcuni testi. I risultati erano gli stessi, quindi verranno considerati attendibili tutti i calcoli del programma.

3.3. Differenze tra traduzione umana e automatica attraverso la classificazione dei *topic*

Nella fase successiva si effettuerà un'analisi del contenuto, in cui si verificherà se i *topic* (argomenti), utilizzati dei testi russi, vengono rispettati nelle traduzioni umane e automatiche e riconosciuti in maniera coerente dai sistemi di analisi automatica. Verranno considerate le traduzioni più fedeli ai testi originali, quelle che mostreranno i contenuti più simili a quelli originali.

Come già accennato nel capitolo 2.1, per *analisi del contenuto* si intende "un processo di acquisizione, sintesi e restituzione delle informazioni presenti in una comunicazione" (Tuzzi 2003, p.17), che nella presente tesi verrà effettuato in modo automatico attraverso l'uso del metodo Reinert nel *software Iramuteq*. Prima di parlare nello specifico di tale metodo, è importante capire come si conduce un'analisi del contenuto. Essa si suddivide in 5 diversi step (cfr. Sbalchiero 2018):

1. Lettura iniziale dei testi che compongono il *corpus*;
2. Identificare temi e informazioni specifiche all'interno dei testi;

3. Etichettare i temi trovati per creare delle categorie di contenuto;
4. Unire le categorie che sembrano simili per evitare una sovrapposizione (*overlap*) e ridondanza (*redundancy*) degli stessi temi;
5. Interpretare e rappresentare le categorie trovate per restituire il contenuto della propria ricerca in modo comprensibile per gli altri.

É possibile eseguire questi passaggi senza l'ausilio di alcun *software*, ma ovviamente il procedimento diventa molto più lungo e quasi impossibile da eseguire nel momento in cui si analizzano grandi quantità di testi. Riprendendo la distinzione di Giuliano e La Rocca (2008), esistono sia *software* di "analisi dei dati qualitativi assistita dai computer", sia *software* di "analisi statistica dei dati testuali". Nel caso specifico dell'analisi del contenuto, è possibile eseguire i 5 step sopracitati in modo più immediato attraverso l'uso di programmi di analisi dei dati qualitativi assistita dai computer. In questo caso, il ricercatore riesce a velocizzare il processo di analisi grazie a programmi come *Nvivo* o *AtlasT* che permettono di "facilitare la lettura e interrogazione dei documenti […] per trarne sistematicamente delle risposte sulla base di domande a priori, oppure essere di aiuto per la costruzione di ipotesi e teorie che emergono dalla esplorazione diretta delle fonti stesse" (Giuliano & La Rocca 2008, p. 8). Ne consegue che l'uso di questi *software* velocizza il lavoro del ricercatore, che ad esempio non deve "perdere tempo" a leggere per intero i documenti, ma lo studioso ha ancora il compito di trovare autonomamente le categorie tematiche che si trovano nei testi.

Con l'invenzione di algoritmi, come il metodo Reinert, si è resa ancora più immediata l'analisi del contenuto, dato che in questo caso è il computer stesso a individuare automaticamente le categorie tematiche. Il metodo che verrà utilizzato all'interno della presente tesi è il già citato metodo Reinert (*Reinert's method*) implementato nel *software Iramuteq*. Lo scopo del *Reinert's method* è l'analisi di un *corpus* a partire dalle co-occorrenze delle parole come appaiono in *chunk* di testo, e identificare, così, le classi semantiche. Con il termine classe semantica o *semantic class* ci si riferisce a un gruppo di parole associate tra di loro, che formano una manifestazione concreta e osservabile di *topoi* o temi convenzionali (*conventional themes*) (cfr. Sbalchiero 2018, p. 202). All'interno di *Iramuteq*, per permettere al *software* di individuare i *topic*, è necessario dividere il *corpus* testuale

in segmenti di testo, chiamati ECU, ovvero *Elementary Context Unit*[28]. Di default il programma imposta il limite di una ECU a 40 parole, ma questo limite può essere cambiato dal ricercatore in base alla tipologia di testi che deve analizzare. Successivamente, per trovare i *topic*, il programma crea una matrice che incrocia gli ECU e le parole (*word*), in cui le celle della matrice (*matrix*) indicano la presenza (es. 1) o l'assenza (es. 0) di una determinata parola all'interno dell'ECU. Un esempio di matrice è il seguente:

	Word 1	Word 2	Word 3	Word 4	Word 5
ECU 6	1	0	1	0	0
ECU 1	1	0	1	0	1
ECU 3	0	1	0	1	1
ECU 5	1	1	0	1	1
ECU 4	0	1	0	1	1
ECU 2	0	1	0	1	1

Tab. 3 - *Esempio di matrice parola x ECU* (da Sbalchiero 2018, p. 204)

I risultati di questo *matrix* vengono riassunti in un *cluster* gerarchico discendente (v. Fig. 7). La procedura di *clustering*, che verrà spiegata in modo più approfondito nel paragrafo 3.5, identifica gerarchicamente i gruppi di parole (*cluster*) che meglio rappresentano una classe semantica a partire dalla distanza x^2 tra le classi (cfr. Reinert 1983 in Sbalchiero 2018, p. 203).

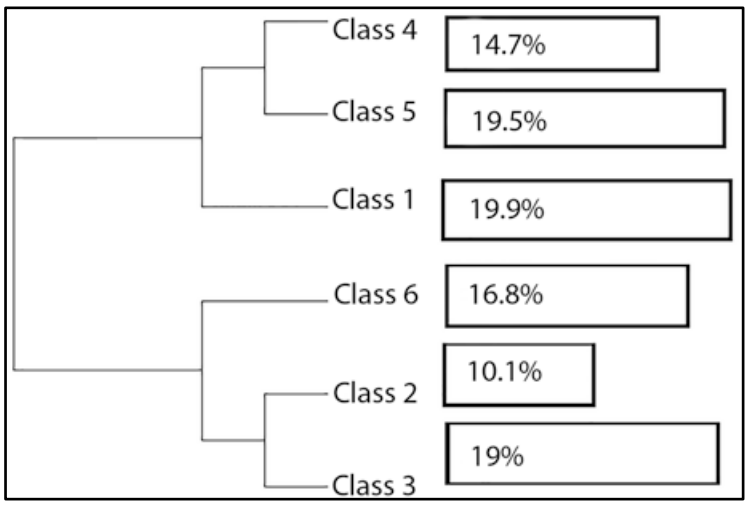

Fig. 7 - *Esempio di divisione in cluster nel metodo Reinert* (da Sbalchiero 2018, p. 204)

[28] Trad. Unità di Contesto Elementare

Infine, l'elenco delle parole più significative, che meglio rappresentano una classe semantica, viene individuato dal *software* associando la distanza x^2 tra parole e classi e produce graficamente un dendrogramma simile a quello in Figura 8.

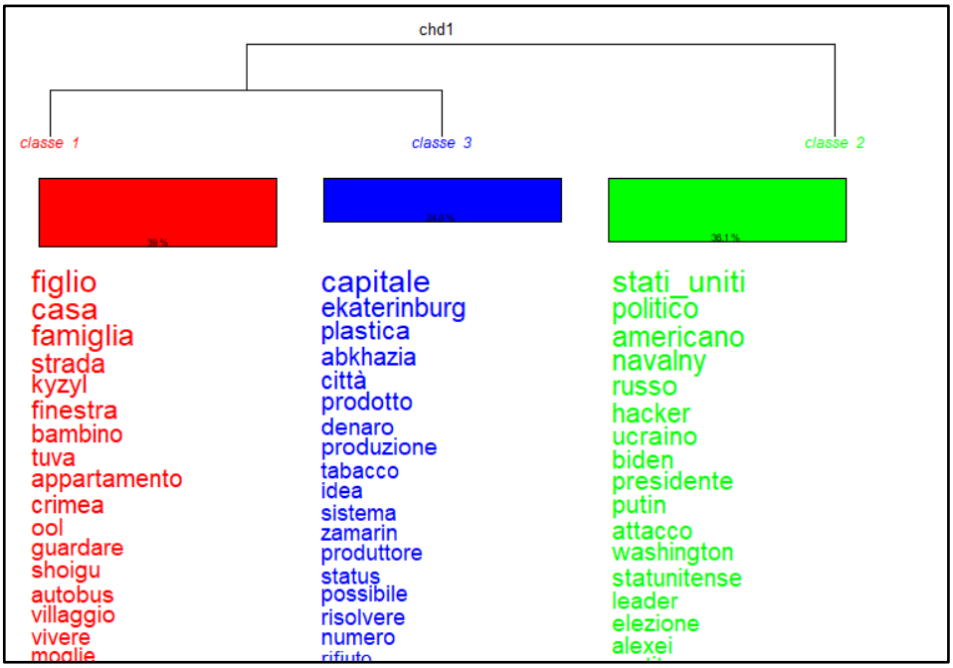

Fig. 8 - *Esempio di dendrogramma creato in Iramuteq con il metodo Reinert*

Dopo aver individuato le classi di semantiche, il ricercatore ha il compito di analizzare le classi e capire che tipo di argomento esprimono. Ad esempio, analizzando la Figura 8 si può dire che la "classe 1" sia composta da parole che fanno riferimento al tema della famiglia; la "classe 2" si riferisce al tema della città e della metropoli; infine, la "classe 3" raggruppa parole che hanno a che fare con la politica.

3.4. Differenze tra traduzione umana e automatica con la suddivisione in *cluster*

La quarta analisi che verrà realizzata ha lo scopo di trovare le differenze tra traduzione umana e automatica attraverso la *cluster analysis*. Con il termine *text clustering* si intende una tipologia specifica di classificazione non supervisionata di documenti, facenti parte di un *corpus* di testi in formato elettronico. Lo scopo del *text clustering* è quello di raggruppare testi simili in uno stesso *cluster* (gruppo) e di separare testi dissimili in *cluster* diversi (cfr. Tuzzi 2010, p. 81). Diversamente

dal metodo Reinert e dagli algoritmi di *machine learning*, il *text clustering* non è un algoritmo specifico, ma un compito (*task*) che può essere svolto da diversi algoritmi.

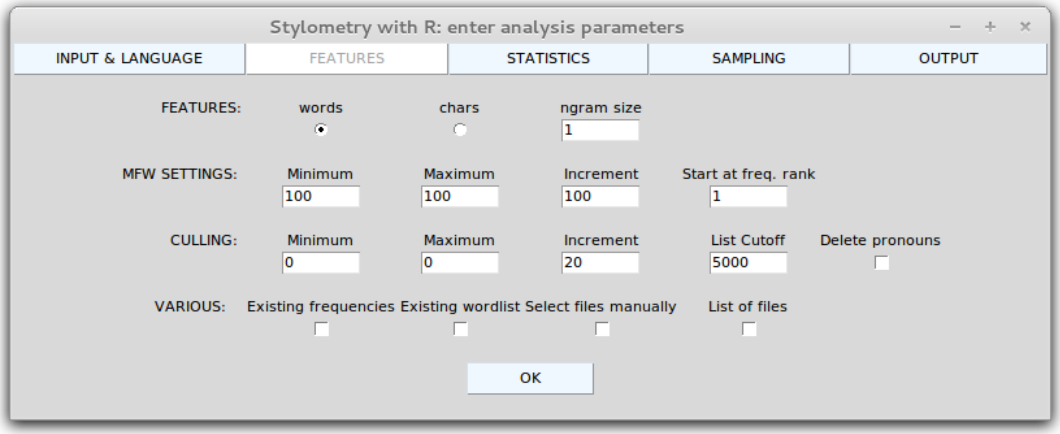

Fig. 9 - *Interfaccia grafica di stylo* (da Eder et al. 2016, p. 114)

Nella presente tesi si utilizzerà un algoritmo di classificazione gerarchica agglomerativa con metodo del legame completo che si basa sulla *Burrow's Classic Delta* disponibile del pacchetto *stylo* scritto nel linguaggio R, utile per effettuare analisi stilometriche (Eder et al. 2016, p. 107). Per stilometria (*stylometry*) in questo ambito si intende un metodo computazionale che si occupa dello studio quantitativo dello stile di scrittura di testi e dei metadati (*meta-data*) dei testi (es. la data, il genere testuale, il gender, la paternità del testo). Ne consegue che la ricerca stilometrica può essere utile, ad esempio, per dedurre la data di composizione dei testi in base agli aspetti stilistici (Eder et al. 2016, p. 107-106). Una delle maggiori applicazioni di *stylometry* si trova oggi negli studi di *authorship*, ovvero paternità di un testo. Nell'ambito degli studi letterari, ad esempio, si è scoperto che un *software* è in grado di unire i testi scritti da uno stesso autore in un unico *cluster* perché ne riconosce automaticamente lo stile. La grande svolta di questo risultato è stato notare che lo stile di un autore può essere riconosciuto solo dall'analisi dei bigrammi, trigrammi o delle prime *n* parole più frequenti presenti nel testo. Data la possibilità di condure ricerche in ambiti molto diversi, i creatori di *stylo* (Maciej Eder, Jan Rybicki, Mike Kestemont, Steffen Pielstroem) hanno creato un'interfaccia che permettesse anche a ricercatori con poca o nessuna conoscenza di programmazione di condurre ricerche di *stylometry*.

Come si può notare dalla Figura 9, è possibile decidere se effettuare l'analisi contando le parole più frequenti o gli *n*-grammi. Inoltre, è possibile scegliere il tipo di algoritmo e rappresentazione grafica che si desidera creare. Nella presente tesi tutte le analisi si baseranno sulle prime 200 MFW (*most frequent* words) e verrà utilizzato il *Burrow's Classic Delta*, che produrrà graficamente un dendrogramma simile a quello in Figura 10.

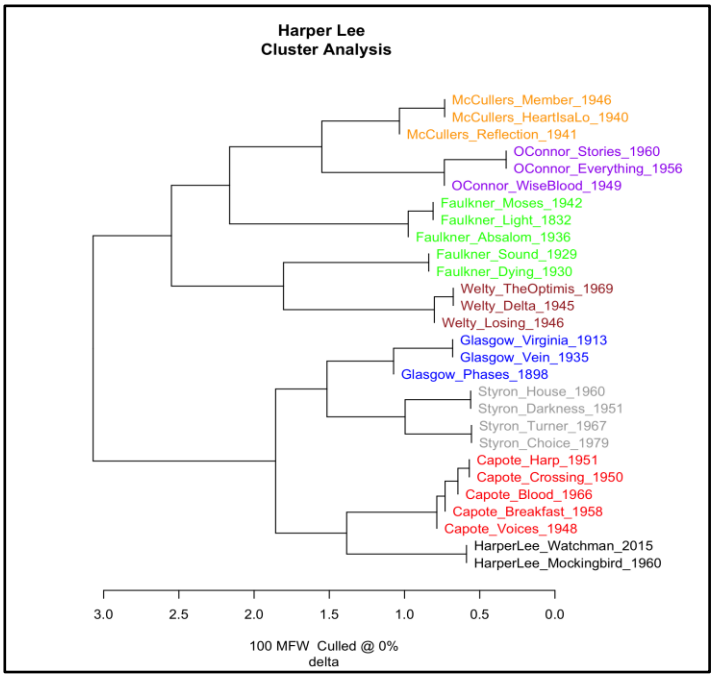

Fig. 10 - *Esempio di dendrogramma* (da Ebert 2016, p. 115)

Il dendrogramma, prodotto attraverso *stylo*, permetterà di visualizzare la divisione in *cluster* dei diversi testi che compongono il *corpus* e i loro rapporti gerarchici. Come si può vedere in Figura 10, alcuni testi (es. "HaperLee_Whatchman_2015"; "HarperLee_Mockingbird_1960") formano un unico *cluster*, ma anche le opere scritte da Harper Lee e Capote si uniscono in un unico *cluster*. Questo significa, che i testi che sono vicini graficamente presentano una minor distanza e quindi sono più simili, rispetto a quelli lontani graficamente che hanno una distanza maggiore. Ad esempio, Harper Lee e Capote sembrano molto simili stilisticamente, mentre Capote e McCullers sembrano non avere molte similitudini. Da un punto di vista statistico, la distanza tra coppie di *cluster* in Figura 9 è ottenuta dalla distanza massima tra tutte le coppie di elementi di due *cluster* e

le coppie di un *cluster* con una distanza minima vengono unite (cfr. Tuzzi 2010, p. 89).

Nella presente tesi verrà fatto uso di *stylo* per produrre dendrogrammi simili a quello visibile in Figura 10. Lo scopo è quello di verificare la coerenza delle traduzioni rispetto ai testi di partenza, confrontando la divisione in *cluster* dei testi originali russi con quella delle traduzioni umane e automatiche. A questo proposito verranno creati quattro dendrogrammi diversi, ovvero uno per ogni *corpus*. Vedendo le similitudini e differenze tra la divisione in *cluster* dei testi originali e delle traduzioni, verrà notato quali traduzioni sono più fedeli allo stile dei testi originali e quali no. Verranno considerate come traduzioni più fedeli quelle che presentano una divisione in *cluster* il più simile possibile al dendrogramma di articoli originali russi.

Inizialmente era stato deciso di effettuare la *cluster analysis*, per vedere se i traduttori umani fossero influenzati dallo stile dei giornalisti russi che traducevano. Purtroppo, data la vastità di autori diversi sia per le traduzioni che per gli articoli in russo, non sarà possibile effettuare questa distinzione perché il materiale collezionato su un singolo autore non è sufficiente per avere abbastanza informazioni sul suo stile. Si è pensato allora di non abbandonare la *cluster analysis*, ma di cambiare la domanda di ricerca come sopra esposto. Si ritiene che sia possibile distinguere la traduzione umana da quella automatica utilizzando *stylo* grazie agli studi già condotti in materia da Cembrzyńska et al. (2021) e Żak et al. (2021). Entrambe le ricerche, descritte nel capitolo 2.3, si occupavano del confronto tra le traduzioni umane e automatiche di testi letterari inglesi attraverso l'uso di *stylo*. Dato che gli studi sopranominati hanno provato l'attendibilità del pacchetto *stylo* anche in ricerche di traduttologia, si è deciso di utilizzarlo per lo stesso scopo anche nella presente tesi.

3.5. Riconoscimento automatico di una traduzione umana attraverso il *machine learning*

Nella quinta ed ultima fase di analisi si verificherà la capacità di un *software* di distinguere un testo tradotto da un essere umano, rispetto a un testo tradotto da un programma di traduzione automatica. A questo scopo si farà uso del *machine*

learning, nello specifico un algoritmo che permette a un computer di riconoscere in modo automatico l'autore di un testo (*authorship attribution*). Prima di spiegare nello specifico in cosa consiste l'*authorship attribution*, è necessario capire cos'è e come funziona il *machine learning*.

Il *machine learning*, precedentemente nominato nel capitolo 1, è lo studio scientifico di algoritmi e modelli statistici che permettono a un computer di compiere un determinato compito (*task*) senza essere esplicitamente programmato per farlo (cfr. Mahesh 2019, p. 381). Da un punto di vista pratico si può dire che nei metodi che si basano sul *machine learning* un essere umano programma un algoritmo, contente le istruzioni del *task* che deve compiere il computer. Successivamente, il computer elabora autonomamente i dati e trova delle soluzioni al problema posto inizialmente. Dato che il programmatore non ha creato un algoritmo che dice passo per passo al computer cosa fare, ma ha fornito solo degli *input* iniziali, il computer è in grado di elaborare i dati con un certo grado di autonomia. Per questo motivo, nel definire il *machine learning* si parla di algoritmi "non esplicitamente programmati" per compiere un *task* specifico.

Esistono diversi metodi usati nel *machine learning*, da cui derivano algoritmi con diversi gradi di autonomia, che hanno bisogno di maggiore o minore assistenza esterna (cfr. Mahesh 2019, p. 381). Tra questi ricordiamo il *neural network* spiegato al capitolo 1.1. Nella presente tesi, si farà uso del *supervised learning*, ovvero un *task*, che consiste nella creazione di un algoritmo che mappa una serie di dati in entrata, detti *input*, in dati in uscita, detti *output*, basandosi su esempi di coppie *input-output* (Mahesh 2019, p. 381). Un algoritmo di *supervised machine learning* ha bisogno di assistenza esterna e funziona nel seguente modo (cfr. Mahesh 2019, p. 381; Gatti &Tuzzi 2020):

1. Il *dataset* (collezione di dati) iniziale è diviso in *training set* e *test set*;
2. Il *training set* viene utilizzato dall'algoritmo per creare un modello interpretativo per rispondere a certi parametri impostati inizialmente (es. far memorizzare ad un algoritmo una serie di testi scritti da un determinato autore);

3. Successivamente, il *test set* è utilizzato per verificare se il modello costruito precedentemente è in grado di rispondere ad una domanda di ricerca specifica (es. chi è l'autore di un determinato testo);
4. Infine, molti algoritmi di *machine learning* fanno uso di un *validation test*, che estrae casualmente dei campioni di dati dal *training set* e stima l'accuratezza dell'algoritmo nel classificare i dati.

Il metodo sopracitato verrà utilizzato in sede di analisi per permettere a un computer di distinguere tra traduzioni umane e automatiche. Studi di questo tipo sono già stati realizzati negli ultimi anni e hanno dimostrato come il *machine learning* sia un metodo affidabile che permette a un computer di distinguere in modo automatico le traduzioni umane da quelle automatiche. In letteratura si possono trovare vari studi che hanno affrontato questo tema e in questa sede ne verranno citati due.

Nel 2015 i ricercatori Li, Wang e Zhai hanno verificato la capacità di un algoritmo di distinguere a livello frasale le *human translation* dalle *machine translation* utilizzando la settima versione del corpus "Europarl"[29], dal quale hanno ricavato quattro coppie linguistiche (francese-inglese, tedesco-inglese, italiano-inglese, danese-inglese). La tipologia di algoritmo che hanno utilizzato è il *supervised machine learning*, che è stato in grado di distinguere con successo le traduzioni. Nel 2021 i ricercatori Fu e Nederhof, precedentemente citati al capitolo 1.3.2., hanno nuovamente verificato la capacità di un computer di distinguere la *machine translation* dalla *human translation*, utilizzando il BERT model. In questo caso la distinzione è stata analizzata a livello testuale e non frasale e si è notato che l'algoritmo è riuscito comunque con successo a distinguere la traduzione umana da quella automatica.

All'interno della presente tesi verrà utilizzato il modello *Author's Multilevel N-Gram Profiles*, presente nello studio di Mikros e Perifanos (2013 e 2015), e due algoritmi che utilizzano il metodo *supervised machine learning*, ovvero il *Support Vector Machine* e *Random Forest*. Il modello e i due algoritmi sopracitati sono stati

[29] http://www.statmt.org/europarl/

utilizzati nello studio condotto da Gatti e Tuzzi (2020), che aveva lo scopo di "studiare i 71 messaggi di fine anno pronunciati dagli 11 Presidenti della Repubblica Italiana per stabilire se è possibile rilevare una frattura netta all'interno del corpus, con l'ausilio di metodi di *supervised machine learning* applicati ai testi" (Gatti & Tuzzi 2020, p. 1). Anche se la ricerca non si occupa di traduzione, verrà comunque presa come punto di riferimento durante l'analisi, perché il modello e gli algoritmi utilizzati saranno gli stessi e verranno spiegati brevemente qui di seguito.

- *Author's Multilevel N-Gram Profiles* (AMNP)

AMNP (Mikros, Perifanos 2013 e 2015) è un modello utile per preparare il testo di un *corpus* per analisi successive da effettuare con algoritmi di classificazione, perché permette non solo di analizzare e dividere il *corpus* in un numero specifico di *chunk* (porzioni di testo), ma anche di impostare molti parametri per rendere il *corpus* leggibile ad altri algoritmi (cfr. Gatti & Tuzzi 2020, p. 5). Nella presente analisi il *corpus* verrà diviso inizialmente in *chunk* di 200 parole e verranno analizzate le caratteristiche linguistiche dei testi, individuando i *bigram* (sequenze di 2 caratteri consecutivi), *trigram* (sequenze di 3 caratteri consecutivi), *word* e *word bigrams* (coppia di parole consecutive) più frequenti. Successivamente, si produrranno delle tabelle con le frequenze relative (*relative frequency*) di *bigram*, *trigram*, *word* e *word bigram*, che verranno riassunte in un un *matrix* con potenzialmente i 200 *bigram* più frequenti, i 200 *trigram* più frequenti, le 200 parole più frequenti e i 200 *word bigram* più frequenti (cfr. Gatti & Tuzzi 2020, p. 5).

- *Support Vector Machine*

L'algoritmo *Support Vector Machine* (SVM) analizza dati usati per la classificazione e analisi di regressione (cfr. Mahesh 2019, p. 282). Nelle fasi di *trainig set* e *testing set*, l'algoritmo SVM divide il *dataset* in due parti, trovando l'*hyperplane* ideale con la distanza massima tra i punti di ogni classe. In questo modo l'algoritmo inizia a distinguere i dati in due classi, che nella presente tesi sono la traduzione automatica e umana. Per creare un SVM con un *polynomial kernel* (*scale*, *degree*, *c*), l'algoritmo creerà 3^3 combinazioni per un totale di 27, per trovare le differenze tra le due classi con la maggior accuratezza possibile. Dopodiché,

verrà effettuato un *validation test*, che consta in un controllo incrociato effettuato 5 volte (*5-fold cross-validation procedure*), per proiettare i risultati finali della classificazione (cfr. Gatti & Tuzzi 2020, p. 6).

- *Random Forest*

Il secondo algoritmo di classificazione che verrà utilizzato è il *Random Forest* (RF). RF, così come SVM, sono due algoritmi di *supervised machine learning*, ma, diversamente da SVM, *Random Forest* lavora con i *decision tree*, ovvero una tipologia di *machine learning* che divide i dati in modo continuo seguendo un certo parametro impostato dal programmatore. RF crea molteplici *decision tree*, che vengono poi uniti per dare una soluzione accurata e stabile (cfr. Gatti & Tuzzi 2020, p. 6).

Verranno utilizzati entrambi gli algoritmi per ottenere una suddivisione più accurata possibile tra traduzione umana e automatica. In questa fase di analisi, non verranno utilizzati i testi originali russi, ma solo testi provenienti dai *corpora* "Traduzioni Umane", "DeepL" e "GoogleTranslate". Per creare il *training set*, verranno selezionati i *corpora* di traduzione nella loro totalità, ovvero 34 testi per *corpus*. Come *sample*, che il *software* dovrà riconoscere come traduzione umana o automatica, verrà utilizzato un articolo non presente nei *corpora*, tradotto da un essere umano, da DeepL e da Google Translate. Infine, saranno utilizzati gli stessi algoritmi per vedere se sono in grado di distinguere le traduzioni automatiche di DeepL, da quelle di Google Translate.

Il *software* che verrà utilizzato per eseguire i calcoli sopracitati è R, un *software* gratuito per realizzare calcoli statistici realizzato da *R Core Team*. Può essere scaricato dal sito ufficiale ed è eseguibile su un'ampia gamma di piattaforme UNIX, Windows e MacOS. Alcune funzioni di R, però, funzionano meglio sul sistema operativo Windows, rispetto al MacOS.

4. Elaborazione dei dati

4.1. Verifica della corretta traduzione di acronimi, abbreviazioni, traslitterazioni di nomi propri, prestiti stranieri e termini

La prima fase di analisi ha lo scopo di confrontare le traduzioni di acronimi, abbreviazioni, traslitterazioni di nomi propri, prestiti stranieri e termini, considerando come traduzione corretta quella umana. Il *software* utilizzato è *AntConc*, in particolare è stata utilizzata la "Concordance plot", che permette di vedere in quali articoli sono utilizzate determinate parole; e il "File View", che evidenzia le *word type* oggetto di analisi all'interno del testo dell'articolo in cui compaiono (rappresentazioni visive delle due funzioni sono disponibili al cap. 3.1, Figg. 5-6). Sono state scelte delle liste di acronimi, termini e prestiti di lingue straniere con diverso grado di occorrenza e sono stati analizzati solo i *word type* più significativi e rappresentativi.

La prima serie di *word type* che è stata studiata è costituita dagli acronimi. La lingua russa fa un uso molto più ampio di acronimi rispetto alla lingua italiana e spesso la loro traduzione rappresenta una sfida anche per i traduttori professionisti. Nella traduzione di un acronimo, infatti, è necessario trovare non solo l'equivalente esatto italiano/russo, ma anche capire come viene utilizzato quel termine nell'uso comune della lingua (cfr. Torresin 2022, p. 57). Per questo motivo si è deciso di mettere a confronto la traduzione di una serie di acronimi dalle frequenze diverse, per verificare se un traduttore automatico sia in grado di riconoscere anche gli acronimi meno usati.

In un primo momento si è deciso di effettuare un ulteriore confronto tra le traduzioni umane e le traduzioni letterali fornite dai dizionari russo-italiano (Kovalev e https://ru.glosbe.com/ru/it). È stato notato che le traduzioni letterali degli acronimi e quelle dei professionisti non sempre coincidono, probabilmente perché il traduttore professionista non deve tradurre semplicemente alla lettera, ma deve impiegare termini che vengono comunemente utilizzati nella lingua di arrivo (cfr. Torresin 2022, p. 57).

Acronimo russo	Traduzione letterale	Traduzione umana (TU)
США	USA	USA; Stati Uniti
СССР	URSS	URSS; Unione Sovietica
РФ	Federazione Russa	Federazione Russa
СМИ	media	media
КНР	Repubblica popolare cinese	Cina
МВД	Ministero degli affari interni della Federazione Russa	Ministero degli affari interni della Federazione Russa; Ministero degli interni; Ministero degli affari interni
НАТО	NATO	NATO
НКР	Repubblica del Nagorno Karabakh	Nagorno-Karabach; Nagorno-Karabakh
ЕС	Unione Europea	Unione Europea
МОК	Comitato olimpico internazionale; CIO	Comitato olimpico internazionale; CIO
СИЗО	casa circondariale	centro di detenzione; SIZO
МИД	ministero degli affari esteri	ministero degli affari esteri; ministero degli esteri; ministro degli esteri
НТИ	NTI	NTI
ООН	ONU	ONU; Nazioni Unite
АП	Amministrazione presidenziale	Amministrazione presidenziale
ТАСС	TASS	TASS
Соцсети	social network	social
ЕСПЧ	Corte Europea dei Diritti dell'Uomo	Corte Europea dei Diritti dell'Uomo

Tab. 4 - *Confronto tra Traduzione letterale e Traduzione Umana*

Come si può notare nella Tabella 4 in molti casi sono stati utilizzati dei sinonimi, per evitare ripetizioni. Ad esempio, i traduttori professionisti hanno deciso di utilizzare la parola “Unione Sovietica” per tradurre “CCCP”, anche se la traduzione letterale dell’acronimo russo è “URSS”. Un esempio analogo lo si trova con la traduzione di “МИД”, che letteralmente si traduce con “Ministero degli affari esteri”, ma i traduttori hanno preferito utilizzare anche dei sinonimi, come “ministero degli esteri” e in un caso “ministro degli esteri” perché ci si riferiva a una persona e non all’istituzione in sé.

Successivamente, si è proseguito con il confronto tra le traduzioni umane e quelle automatiche di DeepL e Google Translate (v. Tabella 5). In tutte le tabelle

riportate di seguito, il numero scritto tra parentesi accanto ad ogni parola indica il numero di occorrenze delle parole o gruppo di parole nei diversi *corpora*.

Acronimo russo	Traduzione umana	Traduzione DeepL	Traduzione Google Translate
США (78)	USA (17); Stati Uniti (57)	USA (8); Stati Uniti (64)	Stati Uniti (62); USA (6)
СССР (20)	URSS (16); Unione Sovietica (8)	URSS (18); sovietico (42); Unione Sovietica (7)	URSS (20)
РФ (18)	Federazione Russa (21)	Federazione Russa (9); russo (423)	Federazione Russa (20); russo (431)
СМИ (13)	media (27)	Media (40)	media (30)
КНР (7)	Cina (22)	Cina (19); RPC (4); cinese (11)	Cina (19); RPC (3)
МВД (7)	Ministero degli affari interni della Federazione Russa (1); Ministero degli interni (1); Ministero degli affari interni (4)	ministro degli interni (1); ministero dell'Interno (6); ministro dell'interno (1)	Ministero degli affari interni (2); Ministero dell'Interno (4); ministro dell'Interno (1)
НАТО (7)	NATO (7)	NATO (7)	Nato/NATO (7)
НКР (7)	Nagorno-Karabach (11); Nagorno-Karabakh (3)	NKR (7)	NKR (7)
ЕС (6)	Unione Europea (4)	UE (8)	UE (7)
МОК (5)	Comitato olimpico internazionale (1); CIO (4)	CIO (5)	Cio (5)
СИЗО (5)	centro di detenzione (3); SIZO (3)	centro di detenzione preventiva (1); detenzione preventiva (1); custodia cautelare (2); carcerario (1)	centro di detenzione preventiva (4)
МИД (4)	ministero degli affari esteri (1); ministero degli esteri (2); ministro degli esteri (4)	ministero degli esteri (4)	ministero degli Esteri (3); ministro degli Esteri (2)
НТИ (4)	NTI (4)	NTI (4)	NTI (4)
ООН (4)	ONU (3); Nazioni Unite (1)	ONU (3); Nazioni Unite (1)	ONU (2); Nazioni Unite (2)
АП (3)	Amministrazione presidenziale (9)	amministrazione presidenziale (7); AP (2)	Amministrazione presidenziale (6)
ТАСС (3)	TASS (3)	TASS (3)	TASS (3)
Соцсети (3)	Social (18); social network (6)	social media (11)	social network (19)
ЕСПЧ (3)	Corte Europea dei Diritti dell'Uomo (1)	CEDU (1)	CEDU (1)

Tab. 5 - *Confronto tra traduzioni umane e automatiche di acronimi russi*

Dalla Tabella 5 si può notare che la traduzione automatica di tutti gli acronimi è corretta. Ad esempio, "МОК", che significa "Comitato Olimpico Internazionale" ed è possibile utilizzare l'acronimo CIO, è stato correttamente tradotto sia da DeepL che da Google Translate che hanno entrambi utilizzato l'acronimo corretto, senza utilizzare una traslitterazione dell'acronimo originale. Un ulteriore esempio è "ЕСПЧ", che è stato tradotto correttamente da entrambi i *software* di traduzione automatica in "CEDU", acronimo corretto per "Corte Europea dei Diritti dell'Uomo". Sembra, quindi, che i due programmi siano in grado di riconoscere e distinguere gli acronimi russi senza problemi. Inoltre, confrontando Tabella 4 e Tabella 5, si può notare che anche i traduttori automatici non hanno sempre tradotto letteralmente alcuni acronimi. Ad esempio, l'acronimo "США" è stato tradotto spesso con "Stati Uniti" e non "USA" sia dai traduttori umani sia da quelli automatici. Un altro esempio è la traduzione di "КНР". L'acronimo corrispondente in italiano è "RPC", ma è stato tradotto con: "Cina" dai traduttori umani; "Cina", "RPC", "cinese" da DeepL; "Cina", "RPC" da Google Translate. La differenza nel numero di occorrenze tra l'originale russo e le traduzioni è giustificata dal fatto che "Cina" in russo si dice anche "Китай" e nelle traduzioni la parola "Cina" è stata utilizzata sia per tradurre "КНР" sia "Китай".

L'unico caso di traslitterazione che si è riscontrato in entrambi i programmi di *machine translation* è nella traduzione dell'acronimo "НКР", che è stato tradotto come "NKR" sia da DeepL che da Google Translate, mentre il traduttore umano ha tradotto in modo esplicito con "Nagorno-Karabach" o "Nagorno-Karabakh". In questo caso, può essere che i *software* di traduzione automatica non abbiano riconosciuto il significato di questo acronimo e lo abbiano semplicemente traslitterato. Sulla base degli acronimi selezionati, questo è l'unico che non sembra essere riconosciuto dai *software*. Un discorso a parte deve essere fatto per la parola "СМИ", che in italiano significa "media" nell'accezione di "mezzo di comunicazione di massa". Come si può notare nelle traduzioni italiane le occorrenze del *word-type* "media" sono significativamente più alte rispetto al corrispondente russo, perché "media" in italiano ha anche altri significati. Di conseguenza, la grande differenza di frequenza tra la parola originale e le traduzioni è giustificata dalla polisemia della parola italiana "media". Infine, si è notato che

nella traduzione degli acronimi, il numero di occorrenze non sempre coincide con l'originale russo, perché spesso vengono utilizzati dei sinonimi o altre parti del discorso. Ad esempio, l'acronimo "МИД" nelle seguenti frasi:

Originale	Traduzione Umana
Так, госсекретарь США Энтони Блинкен пообщался с главой **МИД** Украины Дмитрием Кулебой, пообещав активную военную и экономическую помощь Киеву.	Dunque, il Segretario di Stato americano Anthony Blinken ha parlato con il **ministro degli Esteri** ucraino Dmitro Kuleba, promettendo assistenza militare ed economica attiva a Kiev.

Traduzione DeepL	Traduzione Google Translate
Per esempio, il segretario di stato americano Anthony Blinken ha parlato al **ministro degli esteri** ucraino Dmytro Kuleba, promettendo un'assistenza militare ed economica attiva a Kiev.	Così, il segretario di Stato americano Anthony Blinken ha parlato con il **ministro degli Esteri** ucraino Dmitry Kuleba, promettendo assistenza militare ed economica attiva a Kiev.

In nessuna delle traduzioni esposte sopra, l'acronimo "МИД" è stato tradotto letteralmente, perché in italiano, in questo caso, si usa il nome "ministro degli esteri" e non "ministero degli esteri".

Successivamente, sono state analizzate le traduzioni di due abbreviazioni che sono state trovate nel corpus di articoli originali russi, ovvero "др" e "РЕД".

Abbreviazione	Traduzione Umana	Traduzione DeepL	Traduzione Google Translate
др (4)	ecc. (10)	ecc. (15)	ecc. (14)
РЕД (4)	ndr (4)	ndr (3)	ndr (4)

Tab. 6 - *Confronto di traduzioni: abbreviazioni*

Anche in questo caso si può notare che i programmi di traduzioni automatica non hanno avuto problemi a tradurre correttamente le abbreviazioni, perché le traduzioni proposte da DeepL e Google Translate sono le stesse dei traduttori umani. Dalla Tabella 6 è possibile notare la differenza di occorrenze tra l'abbreviazione "др" e le sue traduzioni. Questo squilibrio è dovuto dal fatto che l'abbreviazione "ecc." in italiano è la traduzione sia di "др" sia di "и так далее",

per questo motivo si presentano tali differenze di frequenza. Per quanto riguarda la parola "РЕД", al contrario, le frequenze sono uguali sia nel *corpus* di traduzione umana che in quello di Google Translate e solo il *corpus* di DeepL presenta un numero occorrenze inferiore. Dopo aver effettuato un'analisi più approfondita, si è notato che il *software* DeepL ha omesso la frase in cui era presente il quarto "ndr" e ha tradotto erroneamente l'abbreviazione con l'acronimo I.T., come si può notare nelle frasi riportate sotto.

Traduzione Umana	Traduzione DeepL
(guerra georgiano-abcasa 1992-93, **ndr**)	(nella guerra abkhaza-georgiana del 1992-93 - **I.T.**)

Questo sembra essere l'unico errore di DeepL nella corretta traduzione delle abbreviazioni.

In un secondo momento, sono state messe a confronto le trascrizioni di nomi propri dall'alfabeto cirillico all'alfabeto latino. I risultati sono riassunti nella Tabella 7.

Nomi propri	Traduzione Umana	Traduzione DeepL	Traduzione Google Translate
навальн* (77)	Naval'nyj (57); Naval'nyi (2); Naval'ny(10); Naval'nij (1)	Navalny (76)	Navalny (76)
абхази* (71)	Abcasia (44); abcas_agg (31); Abkhazia (2)	Abkhazia (49); abkhaz_agg (22)	Abkhazia (53); abkhaz_agg (18)
путин* (64)	Putin (64)	Putin (64)	Putin (63)
крым* (39)	Crimea (37); crimean_agg (3)	Crimea (33); crimean_agg (2)	Crimea (37); della Crimea (7)
сахаров* (18)	Sacharov (19)	Sacharov (10); Sakharov (8)	Sacharov (9); Sakharov (99
тув* (33)	Tuva (29); tuvan_agg (9)	Tuva (21); tuvan_agg (5)	Tuva (22); tuvan_agg (9)
Кыргызстан* (1)	Kyrgyzstan (1)	Kyrgyzstan (1)	Kirghizistan (9)

Tab. 7 - *Confronto traduzioni: Traslitterazione*

Gli asterischi presenti al termine delle parole russe hanno permesso di controllare l'occorrenza della singola parola russa al di là del modo in cui è stata declinata all'interno del testo. Ad esempio, per capire quante volte fosse stata

utilizzata la parola “Навальный” è stato necessario aggiungere un asterisco per poter includere tutte le occorrenze delle forme declinate della parola “Навальный” (nominativo singolare), come “Навального” (genitivo singolare). Inoltre, alcune traduzioni sono state scritte sotto forma di lemmi (es. “tuvan_agg”) perché una parola russa poteva essere resa in italiano con un aggettivo o un nome che poteva avere declinazioni diverse (es. “tuvano”; “tuvani”). Per vedere l’occorrenza generale del nome o aggettivo in questione, è stato usato l’asterisco al termine della parola ed è stato riassunto sotto forma di lemma nella tabella. Come si può notare dagli esempi esposti in Tabella 7, entrambi i traduttori automatici traslitterano in modo simile al traduttore professionista, ma non riconoscono il segno debole “ь” nella lingua russa, che in traslitterazione scientifica dovrebbe essere reso da un apostrofo “ʼ” (cfr. Torresin 2022, p. 172). Come si può notare nella traslitterazione del nome “Навальный”, entrambi i traduttori automatici traducono con “Navalny” e omettono il segno debole. Inoltre, è possibile notare che i traduttori automatici non traslitterano mai le lettere russe “ый” con “yj”, come previsto dalle regole di traslitterazione scientifica. È necessario, infine, osservare che i traduttori umani non seguono sempre la convenzione italiana di traslitterazione scientifica data dalla norma ISO-09 (cfr. Torresin 2022, p. 33) perché nel linguaggio dei media viene spesso utilizzata la traslitterazione commerciale, che è ha delle regole leggermente diverse dalla norma ISO-09 (cfr. Torresin 2022, pp. 172-173). Ad esempio, il nome Навальн**ый** viene traslitterato rispettando le regole del codice ISO-09 57 volte (v. Naval'nyj), ma nelle restanti 13 occorrenze si trovano le seguenti traslitterazioni: Naval'nyi; Naval'ny; Naval'nij. Di conseguenza, si può affermare che la traslitterazione dei traduttori automatici è corretta e segue maggiormente le regole della traslitterazione commerciale. In un caso si è notato che sia DeepL che Google Translate, per traslitterare la lettera “x”, hanno utilizzato sia la traslitterazione scientifica “ch” sia quella commerciale “kh”. Il nome in questione è “Сахаров”: tradotto da DeepL 10 volte con “Sacharov” e 8 volte con “Sakharov”; tradotto da Google Translate 9 volte con “Sacharov” e 9 volte con “Sakharov”. Inoltre, dopo un’attenta analisi delle tabelle di traslitterazione della lettera “x”, si è notato che in un caso i traduttori automatici hanno traslitterato meglio dei traduttori umani. Il nome in questione è “Аб**х**азия” che è stato traslitterato correttamente sia da DeepL

che da Google Translate con "Ab**kh**azia", mentre i traduttori umani hanno traslitterato solo due volte in questo modo e in tutti gli altri hanno scritto "Ab**c**asia". Nonostante la traslitterazione automatica in questo caso sia stata più precisa, è importante far notare che nella lingua italiana è presente il nome "Abcasia", quindi i traduttori umani non hanno sbagliato la traduzione.

Il quarto confronto di traduzioni a livello lessicale riguarda l'analisi delle traduzioni in italiano di prestiti stranieri e forestierismi nei testi russi. I primi sono prestiti non adattati di un'altra lingua mantenuti in caratteri latini (es. "ryanair"), i secondi sono parole di origine straniera (cfr. Fusco 2016, p. 53; Torresin 2022, p. 54).

Prestiti stranieri o forestierismi	**Traduzione Umana**	**Traduzione DeepL**	**Traduzione Google Translate**
Wildberries (10)	Wildberries (10)	Wildberries (9); bacche selvatiche (1)	wildberries (7); frutti di bosco (3)
телеграм* (6)	telegram (6)	telegram (2); canale telegrafico (1); telegraf_nom (3)	telegram (1); telegramm_nom (5)
вотсапп* (1)	whatsapp (1)	waps (1)	whatsapp (1)
ютуб* (3)	Youtube (3)	youtube (3)	Youtube (3)
WADA (13)	WADA (14)	WADA (13)	WADA (13)
bloomberg (9)	bloomberg (9)	bloomberg (8)	bloomberg (9)
amazon (5)	amazon (5)	amazon (5)	amazon (5)
ryanair (5)	ryanair (5)	ryanair (5)	ryanair (5)
forbes (4)	forbes (4)	forbes (4)	forbes (4)
blackmatter (3)	blackmatter (3)	blackmatter (4)	blackmatter (3)
darkside (3)	darkside (2); dark side (1)	darkside (3)	darkside (3)
middle class (2)	middle class (2)	classe media (2)	classe media (2)
moonlight (3)	moonlight (3)	moonlight (2); chiaro di luna (1)	moonlight (3)
vicious (3)	vicious (3)	vicious (3)	vicious (3)
тикток* (6)	tiktok (6)	tiktok* (5); tris (1)	tiktok* (6)
IRC (1)	IRC (1)	IRC (1)	IRC (1)
экс (5)	ex (32)	ex (32)	ex (32)
бум (3)	boom (4)	boom (6)	boom (3)
инстаграмно-карамельного (1)	instagram-caramello (1)	istagram-caramello (1)	instagram-caramello (1)

Tab. 8 - *Traduzioni a confronto: prestiti stranieri e forestierismi*

Come si può notare nella Tabella 8 alcune parole presenti nei testi originali russi sono state traslitterate in russo, mentre altre sono state lasciate in alfabeto latino. Questo ha creato confusione nei traduttori automatici, che non sempre hanno tradotto correttamente le parole di origine straniera. Ad esempio, "Wildberries", che è il nome proprio di un'azienda online e deve rimanere invariato nella traduzione, è stato tradotto una volta con "bacche selvatiche" da DeepL e tre volte con "frutti di bosco" da Google Translate. La parola in totale è presente 10 volte nel testo originale ed è importante far notare che i traduttori automatici non hanno tradotto sempre in modo scorretto, ma nella maggior parte dei casi riconoscevano che "Wildberries" è il nome proprio di un'azienda russa e quindi non va tradotto (cfr. Torresin 2022, p. 56). Interessante, invece, è la traduzione dei *social network* "телеграм" e "вотсапп", che sono stati tradotti spesso erroneamente da entrambi i programmi di *machine translation*. DeepL ha tradotto "телеграм" una volta con "canale telegrafico", tre volte con "telegrafo" o "telegrafi" e solo due volte ha tradotto correttamente con "telegram". Anche Google Translate non ha sempre riconosciuto il nome del *social network*, che ha tradotto cinque volte con "telegramma" o "telegrammi" e solo una volta con "telegram". Per quanto riguarda "вотсапп", invece, Google Translate l'ha tradotto correttamente, mentre DeepL ha reso il termine con "waps". Gli altri nomi stranieri sono stati tradotti correttamente, eccetto per la traduzione dell'azienda "moonlight", che è stata tradotta una volta da DeepL con "chiaro di luna". Sembrerebbe che i traduttori automatici riescano più facilmente a riconoscere parole di origine straniera se scritte in alfabeto latino, mentre quelle traslitterate in russo sembrano più difficili da comprendere. Ad esempio, тикток è stato tradotto una volta con la parola "tris" da DeepL.

Infine, sono stati analizzati una serie di termini la cui traduzione in italiano dovrebbe essere univoca e non presentare alcun sinonimo, quindi il numero di occorrenze dei termini dovrebbe essere lo stesso sia nei testi originali russi sia nelle traduzioni. I risultati sono riassunti nella tabella a pagina seguente.

Termini	Traduzione Umana	Traduzione DeepL	Traduzione Google Translate
рубл* (64)	rubli (45); rublo (27)	rubli (42); rublo (27)	rubli (39); rublo (27)
хакер* (39)	hacker (43)	hacker (36); hacking (4)	hacker (38); hacking (2)
коронавирус* (18)	coronavirus (16)	coronavirus (18)	coronavirus (18)
кремл* (26)	cremlino (26)	cremlino (26)	cremlino (26)
нобел* (47)	nobel (47)	nobel (47)	nobel (47)
доллар* (13)	dollar_nom (17)	dollar_nom (18)	dollar_nom (15)

Tab. 9 - *Traduzioni a confronto: termini*

Dalla Tabella 9 si può vedere che i termini "Cremlino" e "Nobel" hanno lo stesso numero di occorrenze in tutti i testi e questo conferma l'impossibilità di rendere queste parole in altro modo nelle traduzioni. Per quanto riguarda il termine "коронавирус", invece, il numero di occorrenze è lo stesso nei testi originali e nei traduttori automatici, ma è diverso nelle traduzioni umane. Si è notato infatti che in due punti del *corpus* i traduttori umani hanno preferito utilizzare la parola "pandemia" e non la traduzione letterale "coronavirus", come si può vedere nelle frasi riportate sotto.

Testo originale	Traduzione umana
во время **коронавирус**ных ограничений	durante il periodo di restrizioni dovute alla **pandemia**
во время **коронавирус**а	durante la **pandemia**

Per quanto riguarda la traduzione delle valute "рубль" e "доллар" si è notato che in russo le parole "rublo" e "dollaro" vengono omesse in frasi in cui in italiano vengono esplicitate. Per questo motivo le occorrenze di "rublo" e "dollaro" sono maggiori nei testi in italiano.

Infine, è stato osservato che la parola russa "хакер" non viene sempre tradotta letteralmente in italiano e anche i traduttori automatici utilizzano parole derivate, come "hacking". Si è comunque notato che le occorrenze della parola "hacker" nelle traduzioni umane è maggiore rispetto alle occorrenze originali, a conferma del fatto che i traduttori umani di articoli giornalistici tendono a adattare il testo per il lettore finale modificando maggiormente le strutture del testo di partenza e rendendo più esplicite determinate forme (cfr. Torresin 2022, pp. 54-57).

4.2. Grado di *semplificazione* delle traduzioni

La seconda analisi effettuata ha lo scopo di verificare il livello di *semplificazione* delle traduzioni italiane attraverso il calcolo del *Type Token Ratio* (TTR) e la lunghezza media di una frase per ogni testo contenuto nei *corpora* in italiano. Si ricorda che con il termine *semplificazione*, si fa riferimento a un universale traduttivo che indica la tendenza del traduttore a semplificare la lingua durante la produzione del testo di arrivo. I parametri che vengono considerati maggiormente per verificare questo universale sono la lunghezza media delle frasi e la ricchezza lessicale, calcolata in base al rapporto tra *word tokens* (N) e *word types* (V). In caso di semplificazione, entrambi i calcoli daranno valori inferiori nelle traduzioni rispetto ai testi originali (cfr. Ondelli & Viale 2010, pp. 3-5). Inoltre, si ricorda che in questa analisi è stato preso come esempio la ricerca effettuata da Kunilovskaya et al. (2018) (v. cap. 3.2). Per effettuare questi calcoli è stato fatto uso di *Voyant.tools*, che misura automaticamente il TTR e la lunghezza media di una frase in un testo. Questi calcoli verranno effettuati solo nei *corpora* in lingua italiana, perché il programma *Voyant.tools* non legge la lingua russa e non sarà possibile un eventuale confronto con i testi originali. In ogni caso, la verifica di questi parametri risulta affidabile solo nei casi in cui i testi presentano una lunghezza simile e sono scritti nella stessa lingua. Il TTR, ad esempio, è un valore che dipende dal numero *n* di *word-type* e *word-token* presenti in un *corpus* e, come mostrato nel capitolo 2.4.2, la percentuale di TTR del *corpus* "Originali Russi" era nettamente superiore rispetto a quello degli altri *corpora* in italiano, perché le caratteristiche morfosintattiche della lingua russa sono diverse da quelle della lingua italiana (cfr. Torresin 2022, pp. 31-33). Per poter comparare due lingue diverse bisognerebbe lavorare sui *corpora* iniziali fino ad ottenere porzioni di testo di lunghezza simile. Nello studio di Kunilovskaya et al. (2018), ad esempio, i *corpora* sono stati campionati fino ad ottenere porzioni di testo comparabili di 100 testi casuali. Nella presente tesi, non è stato effettuato questo campionamento perché si è deciso di comparare solo i testi in lingua italiana, i quali, essendo traduzioni degli stessi testi originali, hanno lunghezze simili. Per questo motivo, non è stato necessario apportare modifiche ai *corpora*.

In un primo momento, è stata calcolata la percentuale di TTR in ogni traduzione prima del *corpus* di "Traduzioni Umane", poi nei due *corpora* di traduzione automatica. Successivamente i risultati della ricerca sono stati sintetizzati nel grafico sottostante.

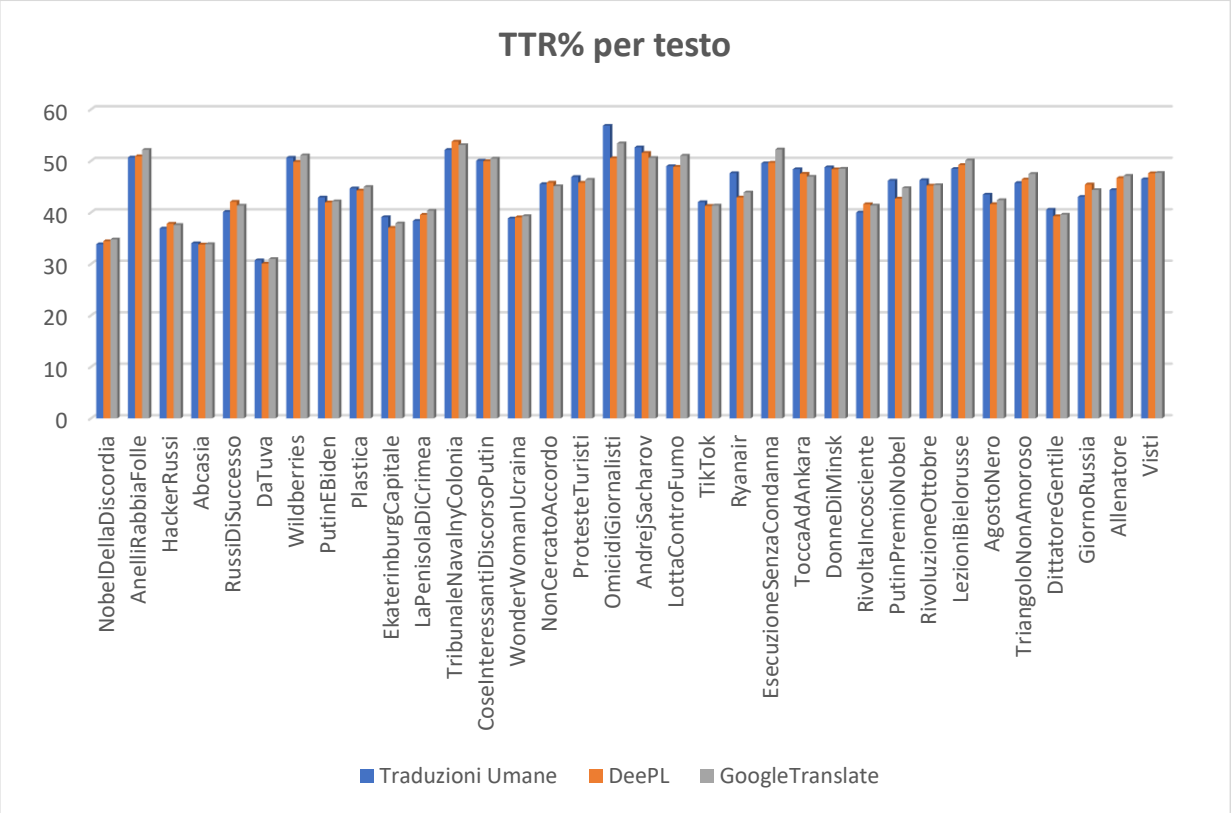

Fig. 11 - *Grafico: confronto tra i TTR delle singole traduzioni dei* corpora *"Traduzioni Umane", "DeepL" e "Google Translate"*

Contrariamente da quanto ci si aspettasse, i valori di TTR delle traduzioni umane non sono sempre maggiori rispetto alle traduzioni automatiche di DeepL e Google Translate. Tranne il TTR della traduzione "Omicidi Giornalisti", in cui la traduzione umana sembra presentare un rapporto V/N nettamente maggiore rispetto alle traduzioni automatiche, negli altri casi il TTR delle traduzioni umane sembra pari o minore rispetto alle traduzioni automatiche. Molto interessante è il fatto che in molti testi il rapporto tra *word type* e *word token* è maggiore nelle traduzioni automatiche effettuate da DeepL o Google Translate rispetto a quelle realizzate da traduttori umani. Sembrerebbe, quindi, che in questo caso il TTR non possa essere considerato un termine di paragone sufficiente per determinare la semplicità

lessicale delle singole traduzioni, dal momento che i valori di TTR appaiono pressoché simili in tutte le traduzioni.

Successivamente, è stata calcolata la media del numero di parole presenti nelle singole frasi delle traduzioni presenti nei *corpora* in lingua italiana. I risultati sono sintetizzati nel grafico sottostante.

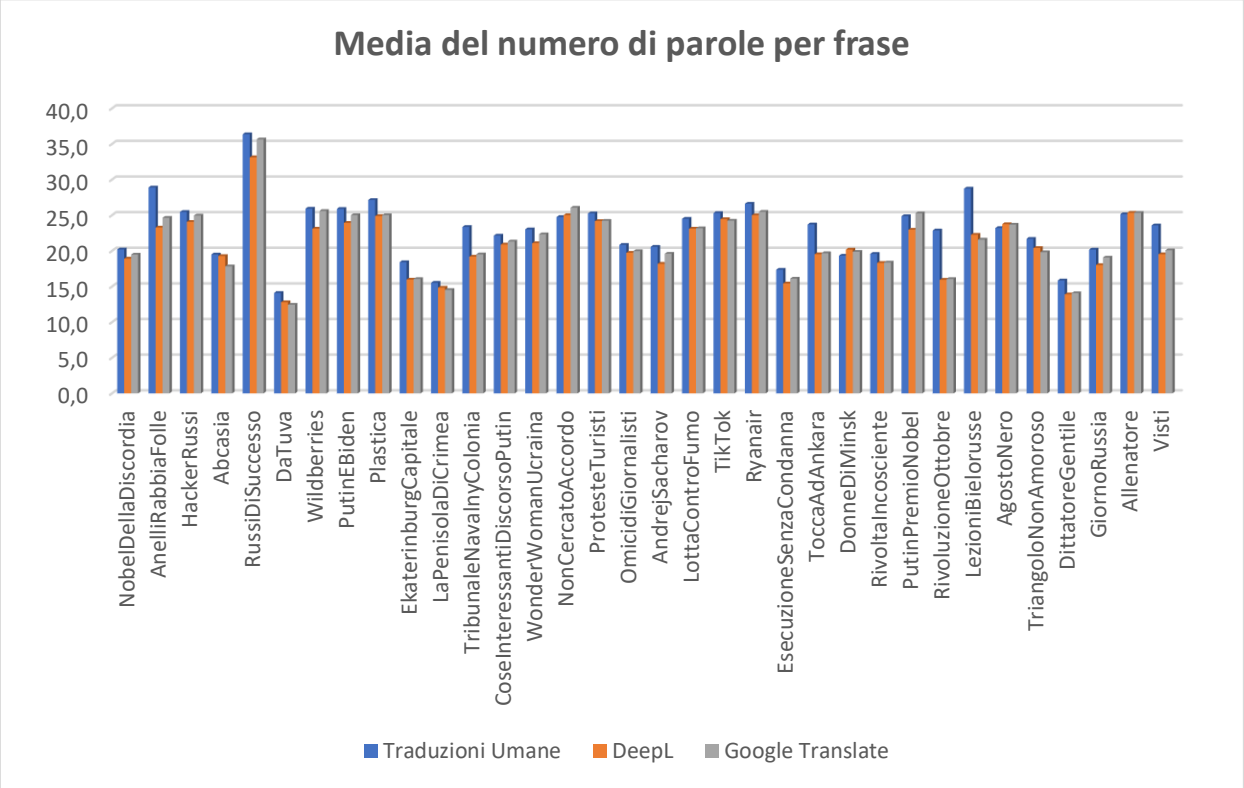

Fig. 12 - *Grafico: media del numero di parole presenti in una frase nelle traduzioni umane e automatiche*

Diversamente dal grafico in Figura 11, nel grafico in Figura 12 è nettamente più visibile la complessità sintattica nella maggior parte delle traduzioni umane. Si può osservare che in molte traduzioni il numero di parole medio per frase è maggiore nelle traduzioni umane, rispetto alle traduzioni automatiche. Di conseguenza, tenendo sempre in considerazione i parametri della *semplificazione* (Ondelli & Viale 2010, pp. 3-5), si può concludere che le frasi presenti nelle traduzioni umane siano più complesse e lunghe rispetto a quelle automatiche. È importante notare che questo fenomeno non avviene in tutti i testi. In alcune traduzioni, come "Abcasia" e "Donne di Minsk", il numero di parole per frase delle traduzioni umane è molto simile o inferiore rispetto alle traduzioni di DeepL e Google Translate; quindi, non sempre il traduttore umano presenta strutture sintattiche più complesse rispetto al

traduttore automatico. È comunque possibile concludere che la maggior parte delle traduzioni analizzate in questa ricerca presentano una maggior varietà sintattica nei testi prodotti da essere umani rispetto a quelli dei traduttori automatici.

4.3. Rispetto dei *topic* all'interno delle traduzioni

Pur avendo selezionato articoli che trattano notizie di area politico-sociale, la terza analisi che è stata effettuata aveva lo scopo di confrontare più in profondità i *topic* (argomenti) presenti nelle traduzioni umane e verificare la loro coerenza con i *topic* presenti nelle traduzioni russe. Per effettuare quest'analisi è stato utilizzato *Iramuteq*, che riconosce in modo automatico gli argomenti presenti in un *corpus* attraverso il metodo di Reinert (cfr. paragrafo 3.3). Purtroppo, il *software* non riconosce in modo adeguato la lingua russa né con la codifica di caratteri utf-8, né con le codifiche cp866 e koi8_r specifiche per la lingua russa. Dopo numerosi tentativi di far leggere correttamente il *corpus* in russo, si è deciso di non includere il *corpus* con i testi in lingua russa all'interno di questa analisi ed effettuare il confronto dei *topic* solo con i *corpora* in lingua italiana, che vengono letti senza problemi dal programma Iramuteq. Verranno presi come punto di riferimento i *topic* presenti nelle traduzioni umane.

Durante l'analisi sono stati utilizzati diversi valori di "dimensioni di numeri di *cluster* finali in fase 1" per ottenere dei dendrogrammi delle stesse dimensioni con un totale di 6 classi. I valori che sono stati utilizzati sono: 12 per il corpus di traduzioni umane, 17 per il *corpus* di traduzioni automatiche di DeepL e 13 per il *corpus* di traduzioni automatiche di Google Translate. Dato che il numero di forme (*word type*) riconosciute da *Iramuteq* è diverso per ogni *corpus* ("Traduzioni Umane": 9740; "DeepL": 8948; "Google Translate": 9003), è stato necessario utilizzare un valore diverso per il "numero massimo di forme analizzate": 10000 per le traduzioni umane; 9000 per le traduzioni di DeepL; 9500 per le traduzioni di Google Translate. Infine, le analisi di tutti i *corpora* sono state effettuate senza lemmatizzazione, quindi le indagini sono state realizzate sui *word type* e non sui lemmi.

Le analisi hanno prodotto i seguenti dendrogrammi:

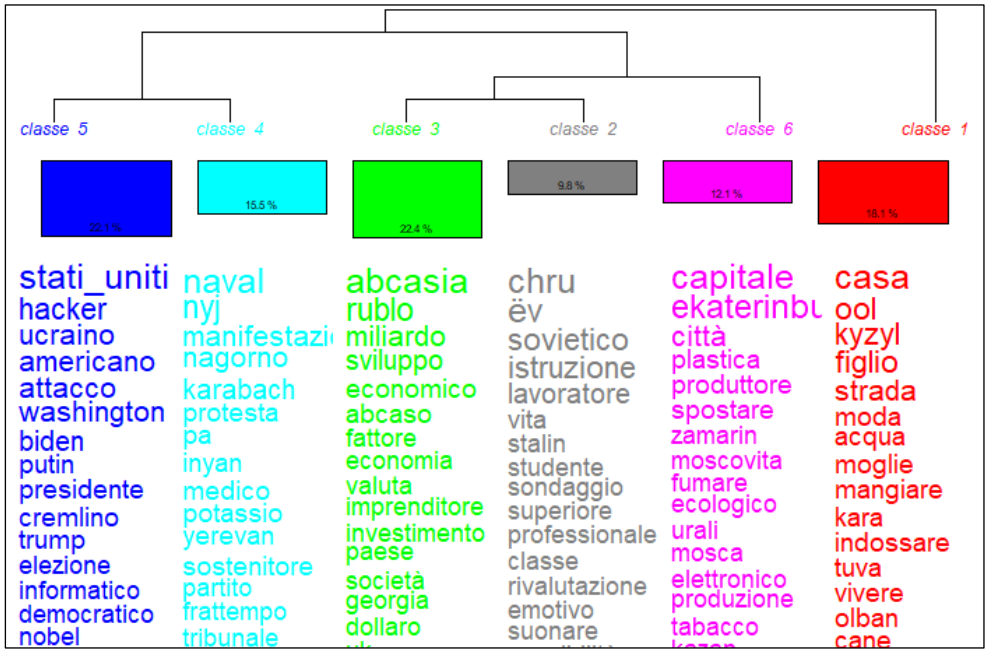

Fig. 13 - *Dendrogramma dei topic utilizzati nelle traduzioni umane. Metodo Reinert.*

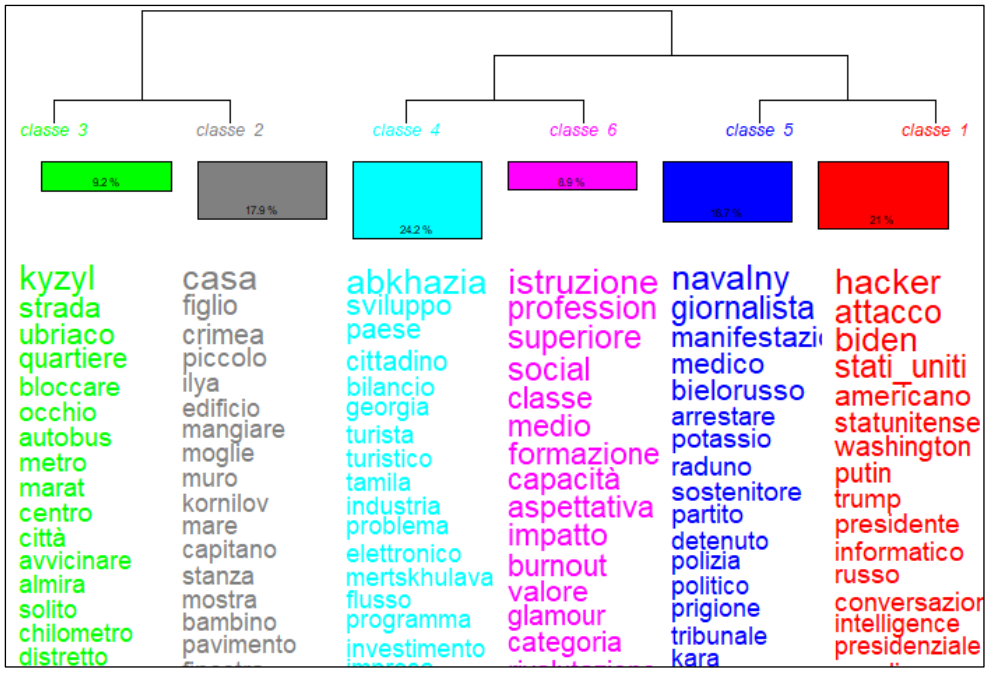

Fig. 14 - *Dendrogramma dei topic utilizzati nelle traduzioni di DeepL. Metodo Reinert.*

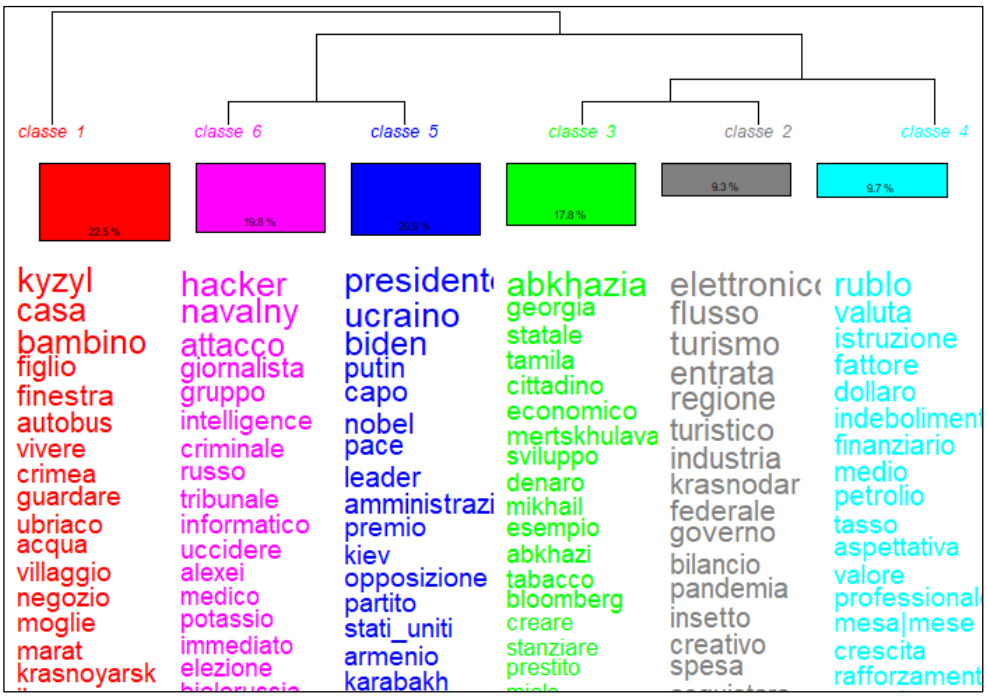

Fig. 15 - *Dendrogramma dei topic utilizzati nelle traduzioni di Google Translate. Metodo Reinert.*

Dai dendrogrammi illustrati sopra è possibile notare che il *cluster* formato da "classe 5" e "classe 4" delle traduzioni umane è molto simile sia al *cluster* "classe 5" e "classe 1" delle traduzioni di DeepL sia al *cluster* formato da "classe 6" e "classe 5" delle traduzioni di Google Translate. Come illustrato in Figura 16 (a pagina seguente), è possibile notare che le parole appartenenti a questo *cluster* riguardano il tema della politica estera. Sono presenti, infatti, le parole "cremlino"; "stati_uniti"; "americano"; i nomi dei presidenti "Putin", "Biden", "Trump". Inoltre, è possibile osservare che i *cluster* delle traduzioni umane e delle traduzioni automatiche di DeepL sono molto più simili rispetto al *cluster* di Google Translate, dove, ad esempio, le parole "hacker" e "attacco" fanno parte di una classe diversa rispetto a quelle degli altri *corpora* (cfr. Fig. 16 a pagina seguente).

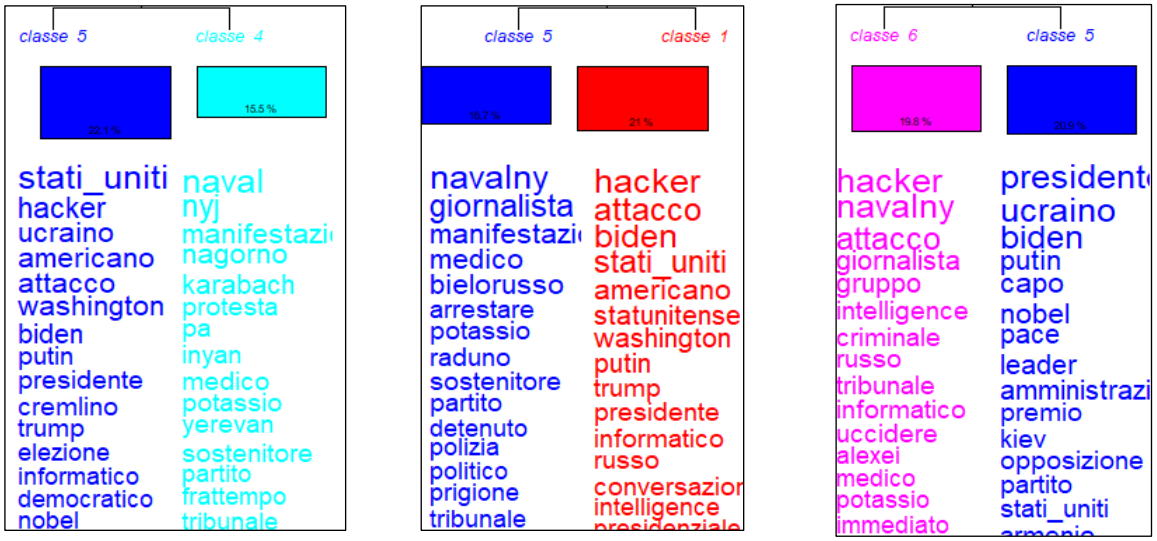

Fig. 16 - *Cluster "Politica". A sinistra il cluster delle traduzioni umane, al centro il cluster delle traduzioni di DeepL e a destra il cluster delle traduzioni di Google Translate.*

Inoltre, si osserva che, nonostante le singole classi presentino percentuali di copertura diverse, la somma delle percentuali di copertura delle classi all'interno del *cluster* chiamato "Politica" sono simili: 37,7% per le traduzioni umane; 39,7% per le traduzioni di DeepL; 40,7% per le traduzioni di Google Translate.

La stessa similitudine tra i vari *corpora* non si trova nel secondo *cluster* individuato nella "classe 3" e "classe 2" delle traduzioni umane, "classe 4" e "classe 6" delle traduzioni di DeepL, "classe 3" e "classe 2" delle traduzioni di Google Translate (v. Fig. 17).

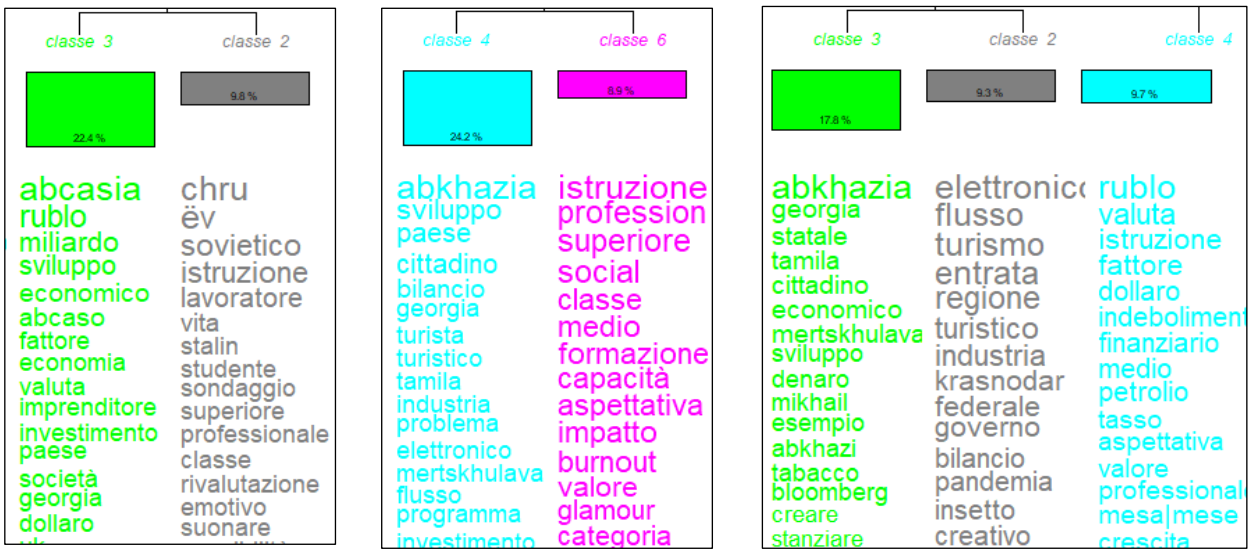

Fig. 17 - *Cluster "Società e formazione". A sinistra il cluster delle traduzioni umane, al centro il cluster delle traduzioni di DeepL e a destra il cluster delle traduzioni di Google Translate*

In Figura 17 a pagina 85 è possibile osservare che le parole “abcasia”, “sviluppo” si trovano in tutti i *corpora* (“classe 3” per le traduzioni umane e di Google Translate; “classe 4” per DeepL) e sempre nella stessa classe di tutti i *corpora* si trovano parole inerenti alla società e all’economia come “economia” (in “classe 3” delle traduzioni umane), “bilancio” e “industria” (in “classe 4” delle traduzioni di DeepL), “economico”, “cittadino” e “stanziare” (in “classe 3” delle traduzioni di Google Translate). Se queste prime classi contengono lo stesso tema in tutti i *corpora*, le classi ad esse collegate in un *cluster* (“classe 2” per traduzioni umane e Google Translate; “classe 6” per DeepL) non contengono gli stessi temi. Si nota, infatti, che solo tra i *corpora* “Traduzioni umane” e “DeepL” c’è una similitudine ed entrambe contengono parole inerenti alla formazione (“istruzione”, “professionale”, “superiore”). Nel *corpus* “Google Translate”, invece, la classe, che forma un *cluster* con il tema società, presenta parole completamente diverse e inerenti ad un tema diverso (v. “elettronico”, “flusso”, “industria”). Alcuni *word type* simili alle classi 2 e 6 degli altri due *corpora* si trovano in una classe separata (v. “classe 4” di Google Translate), che comunque contiene parole inerenti anche al tema dell’economia (v. “valuta”, “rublo”, “finanziario”).

Sembra esserci nuovamente una maggiore similitudine tra il *corpus* “Traduzioni umane” e “DeepL” anche nell’ultimo *cluster* creato da *Iramuteq* (v. Fig. 18).

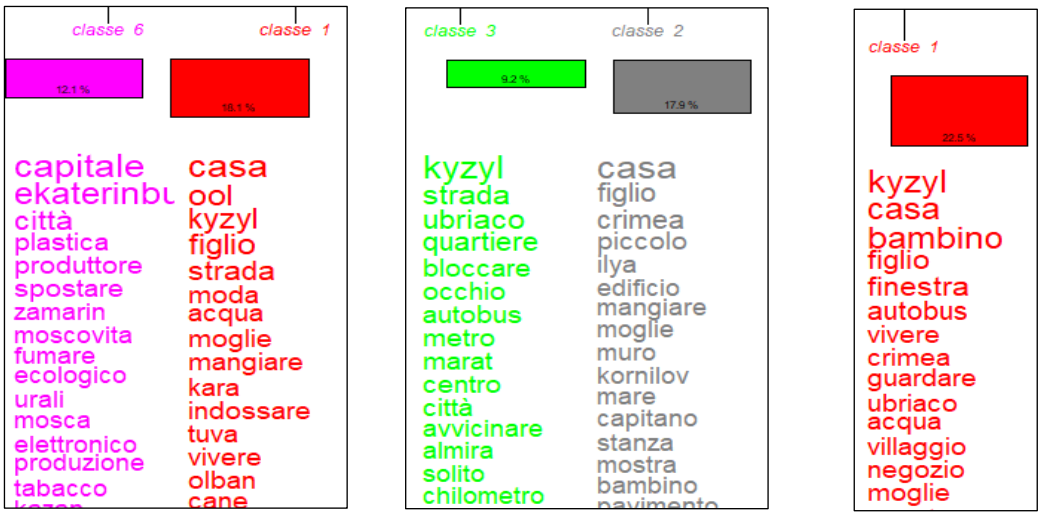

Fig. 18 - *Cluster “Casa e città”. A sinistra il cluster delle traduzioni umane, al centro il cluster delle traduzioni di DeepL e a destra il cluster delle traduzioni di Google Translate*

Come si può notare in Fig. 18 nella pagina precedente, la “classe 1” delle traduzioni umane, la “classe 2” delle traduzioni di DeepL e la “classe 1” delle traduzioni di Google Translate presentano tutte un lessico inerente alla casa e alla famiglia (v. “casa”, “figlio”, “moglie” presenti in tutti i *corpora*). Queste classi formano un *cluster* con la “classe 6” delle traduzioni umane e la “classe 3” delle traduzioni di DeepL, che sono formate da parole inerenti al tema della città. Al contrario delle classi precedenti, qui è evidente che i *word type* presenti in “classe 6” e “classe 3” non sono gli stessi, ma è possibile notare che le parole “capitale”, “mosca”, “ekaterinburg” in “classe 6” e le parole “strada”, “quartiere”, “autobus”, “metro” in classe 3, facciano tutte parte di un macro-tema inerente alla città. La stessa similitudine non si trova nel *corpus* “Google Translate”, in cui la classe inerente al tema casa e famiglia si trova isolata e non è presente una classe che parli del tema città (v. Fig. 15, p. 84).

In conclusione, si può affermare che esistono molte più similitudini in termini di contenuti tra la traduzione automatica di DeepL e la traduzione umana, rispetto alla traduzione automatica di Google Translate e la *human translation*. Infatti, come scritto al capitolo 3.3, si ricorda che i *cluster* si formano sulla base della co-occorrenza di parole nelle stesse porzioni di testo (*chunk*), di conseguenza se due *corpora* producono classi simili significa che tendono ad utilizzare le stesse parole negli stessi contesti. Si può quindi supporre che il *software* DeepL riesca a produrre traduzioni più simili all’essere umano, rispetto al programma Google Translate, perché utilizza le stesse parole dei traduttori umani negli stessi contesti.

4.4. Similitudini e differenze delle traduzioni rispetto ai testi originali russi

La quarta analisi che è stata effettuata ha lo scopo di trovare le differenze tra traduzione automatica e umana attraverso una classificazione automatica dei testi (*cluster analysis*) descritta nel paragrafo 3.4. Inoltre, si verificherà se la stessa divisione in *cluster* presente nei testi originali in russo verrà rispettata nelle traduzioni umane e automatiche. Per effettuare quest'analisi è stato utilizzato il pacchetto *stylo* presente nel *software* R. Si ricorda, inoltre, che in questo caso i *cluster* non verranno suddivisi in base agli argomenti, come in *Iramuteq*, bensì in base allo stile grammaticale presente in un testo (cfr. Eder et al. 2016; Tuzzi 2010). Ad esempio, articoli con i *word-type* simili più frequenti verranno riuniti all'interno di un *cluster*, mentre quelli con *word-type* dissimili saranno separati. Nella presente tesi si è deciso di utilizzare solo i *most frequent word-type* presenti nei *corpora* e non sono state eseguite analisi con bigrammi o trigrammi.

La prima analisi realizzata in *stylo* aveva lo scopo di verificare l'attendibilità del programma nel distinguere lo stile delle traduzioni umane da quelle automatiche. Si è deciso di utilizzare i 200 *most frequent word* presenti in tutti i *corpora* in italiano, che coprono il 54% del vocabolario dei *corpora* "Traduzioni Umane" e "Google Translate" e il 55% del vocabolario del *corpus* "DeepL". I risultati sono visibili in Fig. 19.

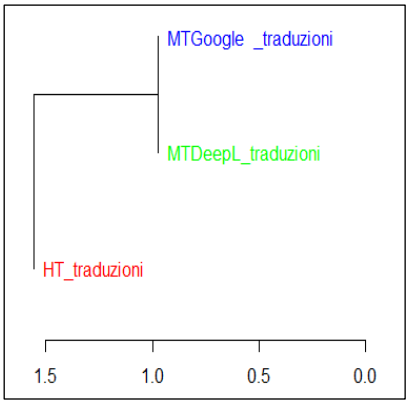

Fig. 19 - *Distinzione tra lo stile delle traduzioni umane da quelle automatiche con 200 MFW e distanza Delta. Dendrogramma risultante da un'analisi dei cluster gerarchica.*

Come si può notare dall'immagine sopraesposta (Fig. 19, p. 88), il pacchetto *stylo* distingue senza problemi le traduzioni umane ("HT_traduzioni") da quelle automatiche ("MTGoogleT_traduzioni" e "MTDeepL_traduzioni") con il solo utilizzo dei 200 MFW presenti nei *corpora*, che coprono circa il 55% del vocabolario di tutti i *corpora*. Si ricorda che i dendrogrammi prodotti da *stylo* sono la rappresentazione grafica di un processo di classificazione basato sulla misura *Burrow's Classic Delta*. Questa misurazione permette di calcolare la distanza intertestuale tra i testi e successivamente *stylo* associa i testi con una distanza minore e allontana quelli con una distanza maggiore, ovvero realizza un'analisi dei *cluster* gerarchica. Minore è la distanza tra i testi, maggiore è la loro somiglianza; mentre maggiore è la distanza, minore è la somiglianza tra i testi. I testi che si assomigliano di più vengono rappresentati graficamente vicini e formano un primo *cluster*, successivamente gruppi di testi simili vengono rappresentati in un *cluster* e via dicendo fino alla rappresentazione dei testi più dissimili. In Figura 19 i testi analizzati sono formati da tutte le traduzioni presenti nei diversi *corpora* in italiano e *stylo* ha individuato una distanza minore tra "MTGoogle_traduzioni" e "MTDeepL_traduzioni" rispetto a "HT_traduzioni", che presenta una distanza intertestuale maggiore.

Dopo aver verificato l'attendibilità di *stylo* nel distinguere traduzioni umane da quelle automatiche, si è inserito il *corpus* in russo e verificato la sua divisione in *cluster* considerando i testi singolarmente. In questo caso si è deciso di aumentare il numero di *most frequent word* per ottenere la più ampia copertura di vocabolario possibile e permettere al programma *stylo* di capire non solo lo stile grammaticale degli articoli, ma di intuire anche i *topic* presenti nei singoli testi. Ci si aspetta, infatti, che un aumento del numero di parole analizzate da stylo, aiuti il programma a trovare meglio gli articoli che si assomigliano, da quelli che si diversificano. Nella prima analisi sono stati utilizzati i 2478 *most frequent word* presenti nel *corpus* "Originali Russi", che hanno una frequenza ≥ 3 e coprono il 67,8% del vocabolario. Dopo aver effettuato una prima divisione in *cluster* degli articoli in russo, è stato notato che tutti gli articoli presentavano colori differenti, perché i codici numerici presenti prima del simbolo *underscore* ("_") erano tutti diversi. Il programma *stylo*, infatti, rappresenta con gli stessi colori solo i testi rinominati allo stesso modo prima

del carattere *underscore*. Per agevolare la ricerca di similitudini e differenze tra i testi originali russi e le traduzioni in italiano, si è deciso di rinominare gli articoli fornendo un codice numerico uguale ai testi che si trovavano nello stesso *cluster* (es. "10_Абказия"[30] e "10_ОтТувы"[31]). Inoltre, questa rinomina ha facilitato anche i confronti successivi tra i diversi *corpora*. Sono stati effettuati diversi tentativi di rinomina, fino ad arrivare al risultato finale visibile in Figura 20.

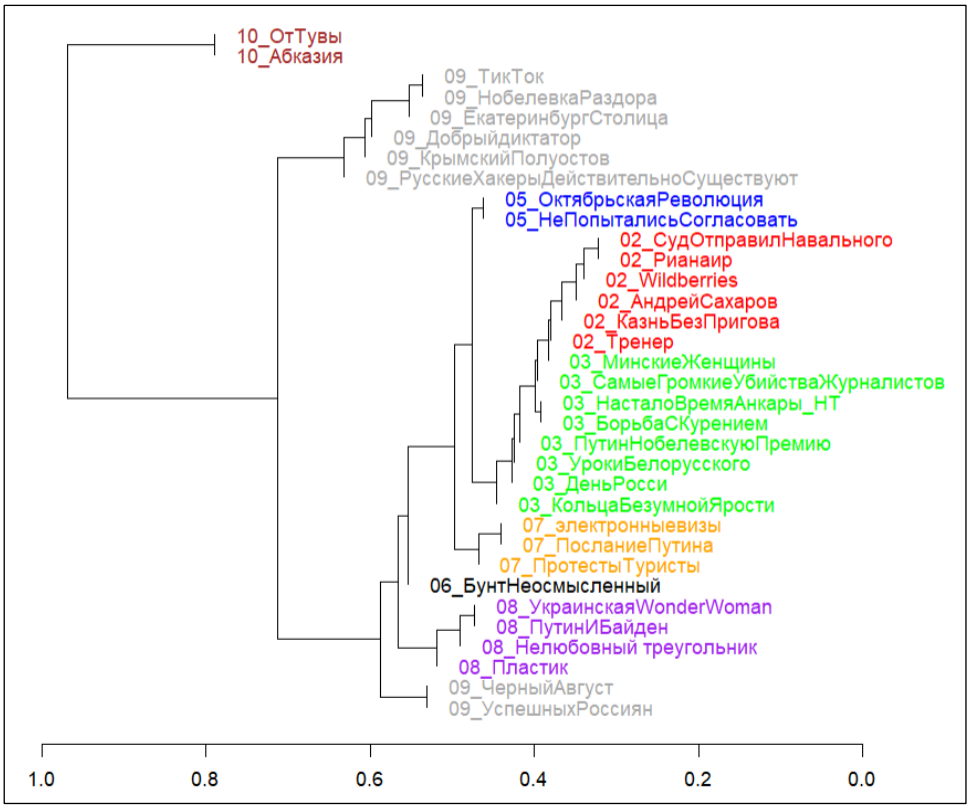

Fig. 20 - *Divisione in cluster degli articoli originali in russo con 4457 MFW e distanza Delta. Dendrogramma risultante da un'analisi dei* cluster *gerarchica.*

Come è possibile notare in Figura 20, per ottenere una copertura del vocabolario del 77,35%, si è deciso di considerare anche tutti i *word-type* con una frequenza ≥ 2. Inizialmente erano stati presi in considerazione solo i *word-type* con una frequenza ≥ 3, ma ricoprivano solo il 67,8% del vocabolario del *corpus*; quindi, si

[30] trasl. Abkazija

[31] trasl. Ot Tuvy

è preferito aumentare il numero di *most frequent word* analizzati per ottenere la maggior copertura di vocabolario possibile escludendo gli *hapax*.

Dopo aver ottenuto il dendrogramma in Figura 20, si è passati all'analisi delle traduzioni in italiano, i cui testi sono stati rinominati con gli stessi codici visibili in Figura 20. Durante l'analisi delle traduzioni, si è selezionato un numero *n* di MFW che presentasse una copertura di vocabolario simile a quella utilizzata nei testi originali russi (77,35%), per ottenere una suddivisione in *cluster* delle traduzioni umane il più attendibile possibile. Nel *corpus* di traduzioni umane ("HT_traduzioni") sono state prese in considerazione le parole con una frequenza ≥ 5 con un tasso di copertura del vocabolario pari al 78,22%, che hanno prodotto il dendrogramma in Figura 21.

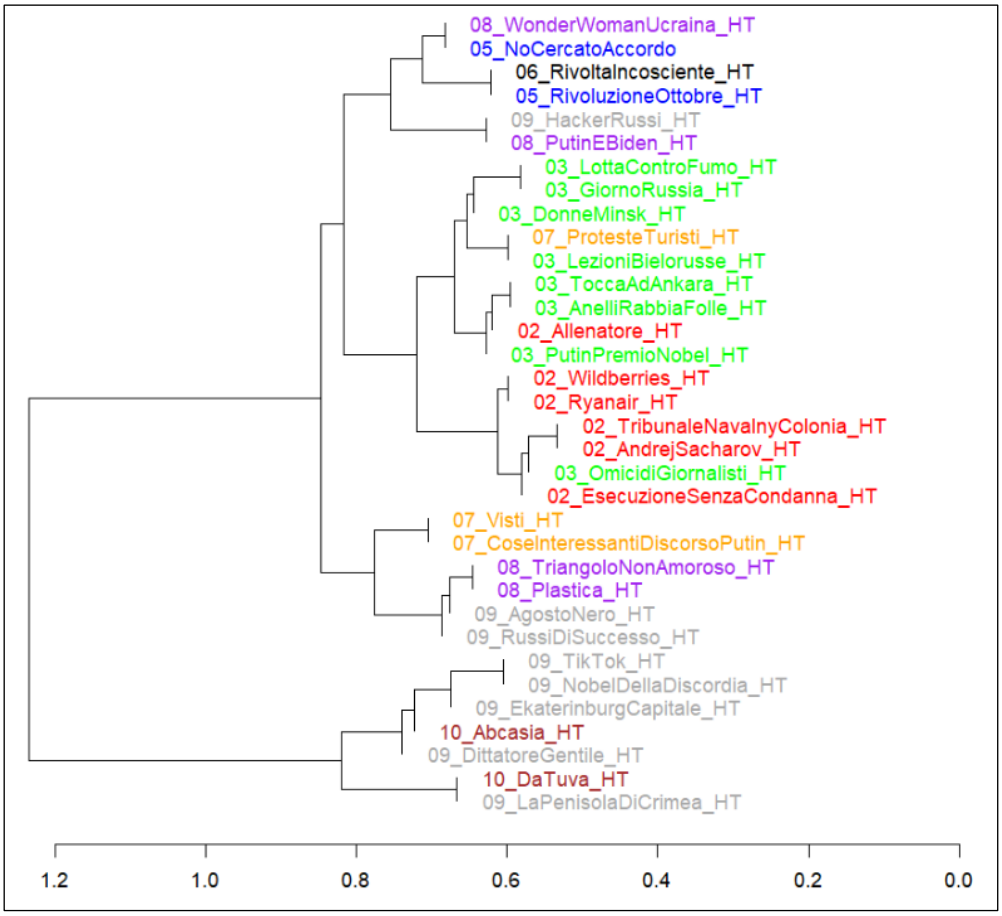

Fig. 21 - *Divisione in cluster delle traduzioni umane con 1577 MFW e distanza Delta. Dendrogramma risultante da un'analisi dei* cluster *gerarchica.*

Nel *corpus* di traduzioni automatiche di DeepL ("MTDeepL_traduzioni") sono state prese in considerazione le parole con una frequenza ≥ 5 con un tasso di copertura del vocabolario pari al 78,58%, che ha prodotto il dendrogramma in Figura 22.

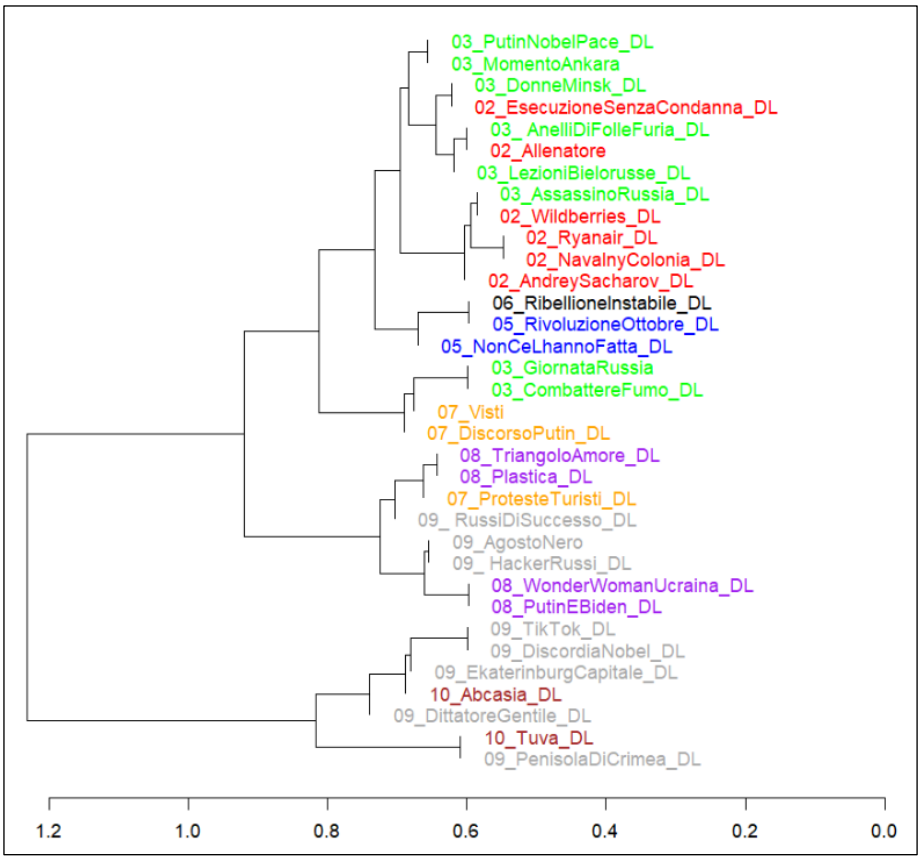

Fig. 22 - *Divisione in cluster delle traduzioni automatiche di DeepL con 1484 MFW MFW e distanza Delta. Dendrogramma risultante da un'analisi dei* cluster *gerarchica.*

Nel *corpus* di traduzioni automatiche di Google Translate ("MTGoogleT_traduzioni") sono state prese in considerazione le parole con una frequenza ≥ 5 con un tasso di copertura del vocabolario pari al 78,51%, che ha prodotto il dendrogramma in Figura 23.

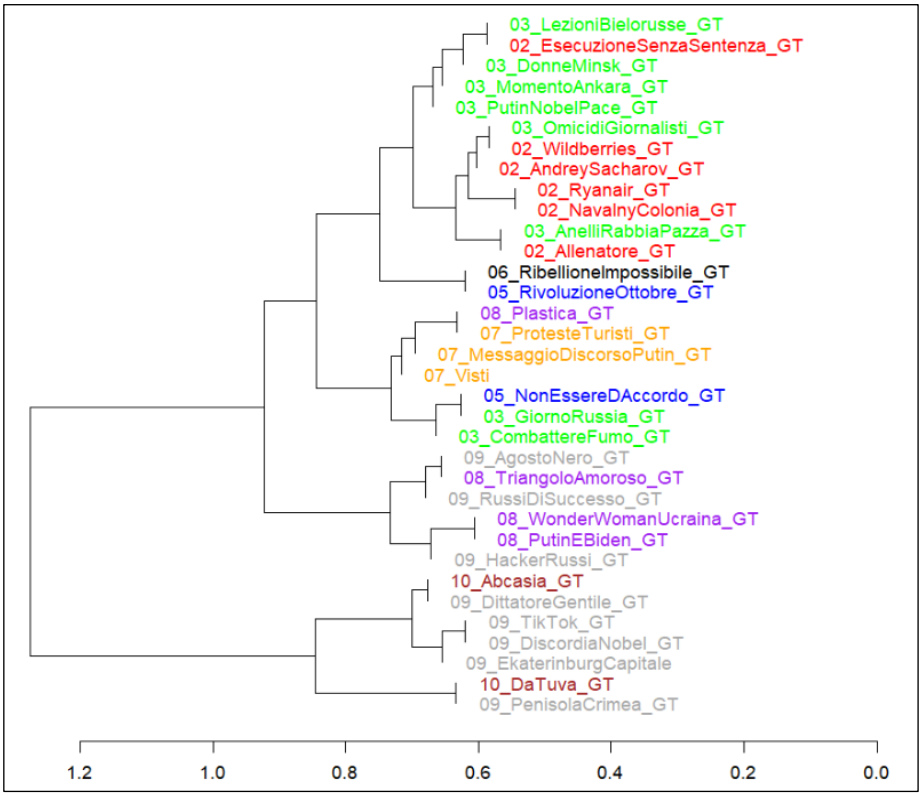

Fig. 23 - *Divisione in cluster delle traduzioni automatiche di Google Translate con 1501 MFW MFW e distanza Delta. Dendrogramma risultante da un'analisi dei* cluster *gerarchica.*

La prima caratteristica che si nota è che la distribuzione in *cluster* dei testi originali russi non è pienamente rispettata nelle traduzioni né automatiche né umane. Analizzando più attentamente i singoli *cluster* si nota che alcune traduzioni presentano una suddivisione più simile all'originale, mentre altre meno.

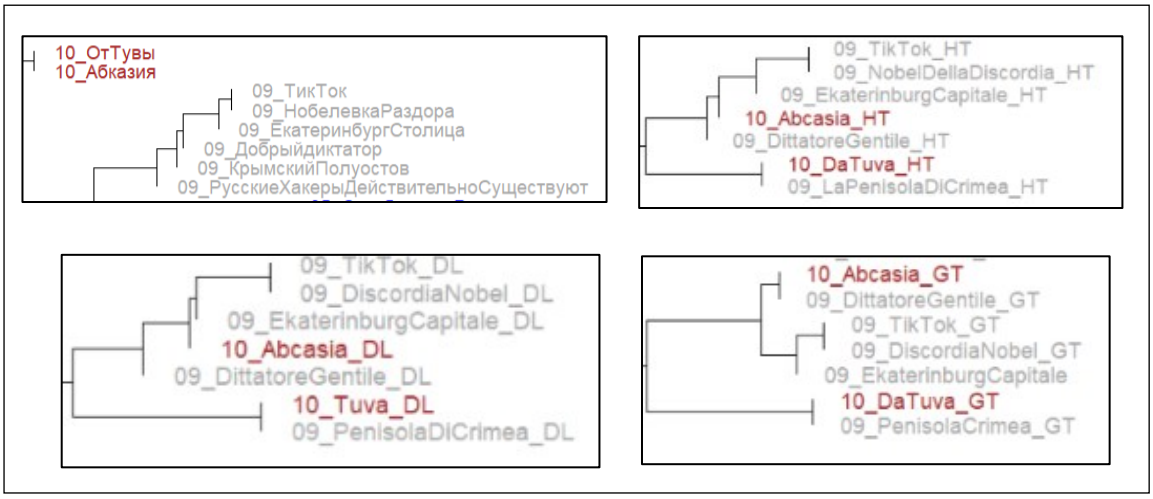

Fig. 24 - *Differenze tra* cluster *testi originali e traduzioni. In alto a sx* corpus *di testi originali russi; in alto a dx* corpus *di traduzioni umane; in basso a sx* corpus *di traduzioni di DeepL; in basso a dx* corpus *di traduzioni di Google Translate.*

Come si può notare dalla Figura 24 in alto, il *cluster* formato da "10_Абказия"[32] e "10_ОтТувы"[33] non viene rispettato in nessuna delle traduzioni, perché viene incorporato nei *cluster* contenenti gli articoli con codice "09_". Si nota, però, una somiglianza tra il *cluster* di traduzioni umane e quello di traduzioni automatiche di DeepL. Si osserva, infatti, che non solo sono traduzioni degli stessi articoli, ma si presentano anche nello stesso ordine; a differenza di Google Translate che presenta un *cluster* con gli stessi articoli delle traduzioni umane e di DeepL, ma l'ordine è differente. Ad esempio, Google Translate racchiude nello stesso *cluster* "10_Abcasia" e "09_Dittatore Gentile", mentre, negli altri due *corpora* di traduzioni, i due articoli non formano lo stesso *cluster*. Inoltre, confrontando nuovamente i testi originali russi con le traduzioni, è possibile notare che tutti gli articoli con codice "09_" presenti nel *cluster* del *corpus* "Originali Russi" sono presenti anche negli altri dendrogrammi, eccetto per l'articolo "09_РусскиеХакерыДействительноСуществуют"[34], che non è presente in nessun *cluster* di traduzioni illustrato sopra.

[32] trasl. Abkazija

[33] trasl. Ot Tuvy

[34] trasl. Russkie Hakery Dejstvitel'no Sušestvujut

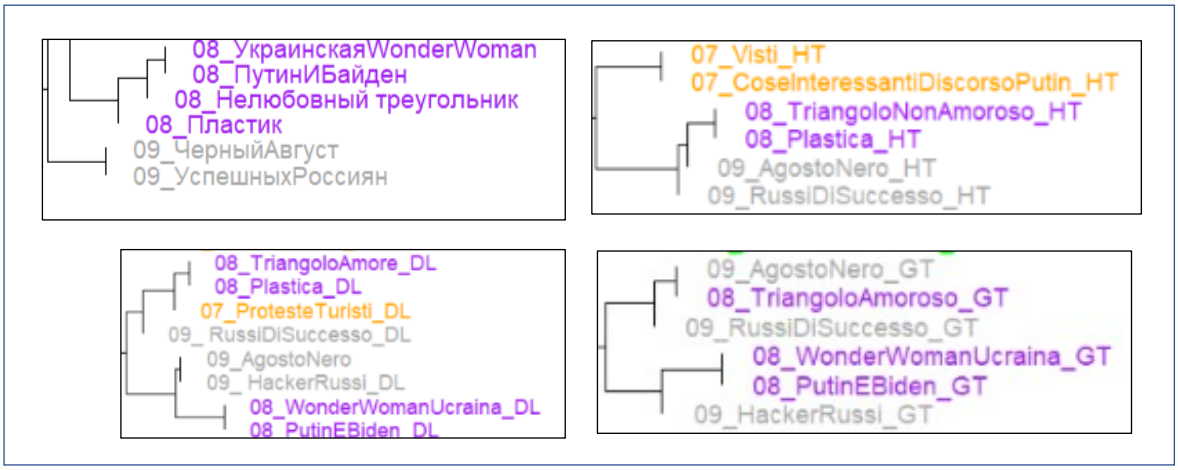

Fig. 25 - *Differenze tra cluster testi originali e traduzioni. In alto a sx corpus di testi originali russi; in alto a dx corpus di traduzioni umane; in basso a sx corpus di traduzioni di DeepL; in basso a dx corpus di traduzioni di Google Translate.*

Successivamente, è stato osservato che gli ultimi due articoli russi con codice "09_" ("09_ЧерныйАвгуст"[35] e "09_УспешныхРоссиян"[36]) formano un *cluster* separato dagli altri articoli con lo stesso codice e questa divisione è stata rispettata nei *corpora* "Traduzioni Umane" e "DeepL", ma non si trova nel *corpus* "Google Translate" (v. Fig. 25). Inoltre, gli articoli originali russi con codice "08_" formano un *cluster* a parte, ma questa suddivisione non viene rispettata in nessun *corpus* di traduzioni. Il *corpus* che contiene i quattro articoli nello stesso macro-cluster è "DeepL", mentre le traduzioni umane riuniscono in unico *cluster* solo "08_TriangoloNonAmoroso" e "08_Plastica" e non gli altri articoli.

[35] trasl. Černyi Avgust

[36] trasl. Uspešnyi Rossijan

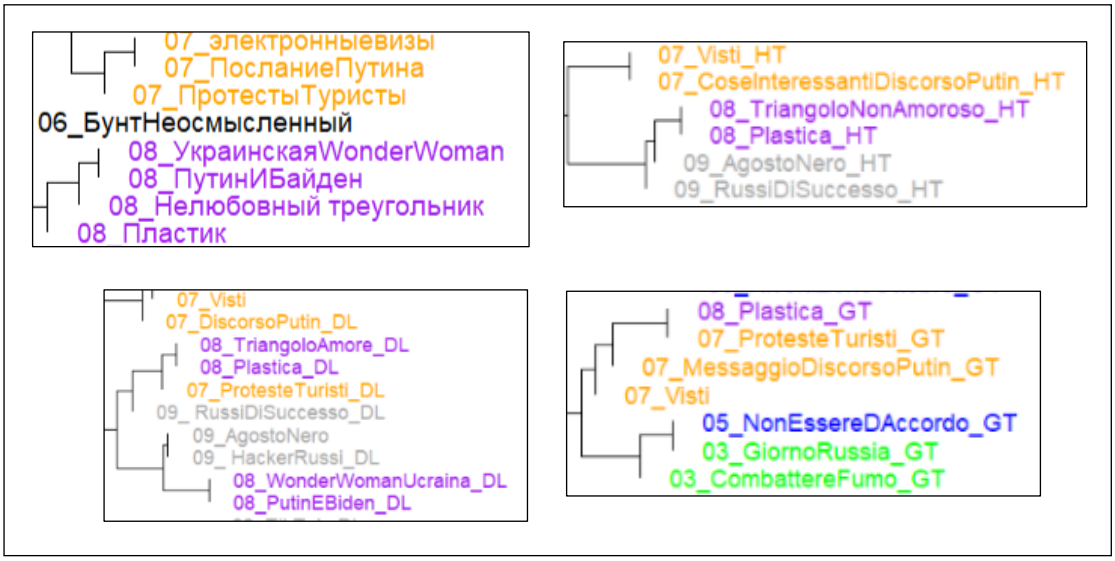

Fig. 26 - *Differenze tra cluster testi originali e traduzioni. In alto a sx corpus di testi originali russi; in alto a dx corpus di traduzioni umane; in basso a sx corpus di traduzioni di DeepL; in basso a dx corpus di traduzioni di Google Translate.*

In seguito, dalla Figura 26, si può notare che nuovamente la traduzione automatica ha rispettato maggiormente la suddivisione in *cluster* dei testi originali russi. Il *cluster* formato dagli articoli con codice "07_", infatti, si ritrova solo nel *corpus* "Google Translate". Il dendrogramma delle traduzioni umane presenta nello stesso *cluster* solo gli articoli "07_Visti" e "07_CoseInteressantiDiscorsoPutin" e una divisione simile si può trovare nel *corpus* "DeepL" con le traduzioni degli stessi articoli. Inoltre, è possibile notare un'ulteriore similitudine tra la traduzione automatica di DeepL e quella umana. Entrambi i dendrogrammi, infatti, presentano nello stesso *cluster* le traduzioni degli articoli "08_TriangoloAmore" e "08_Plastica".

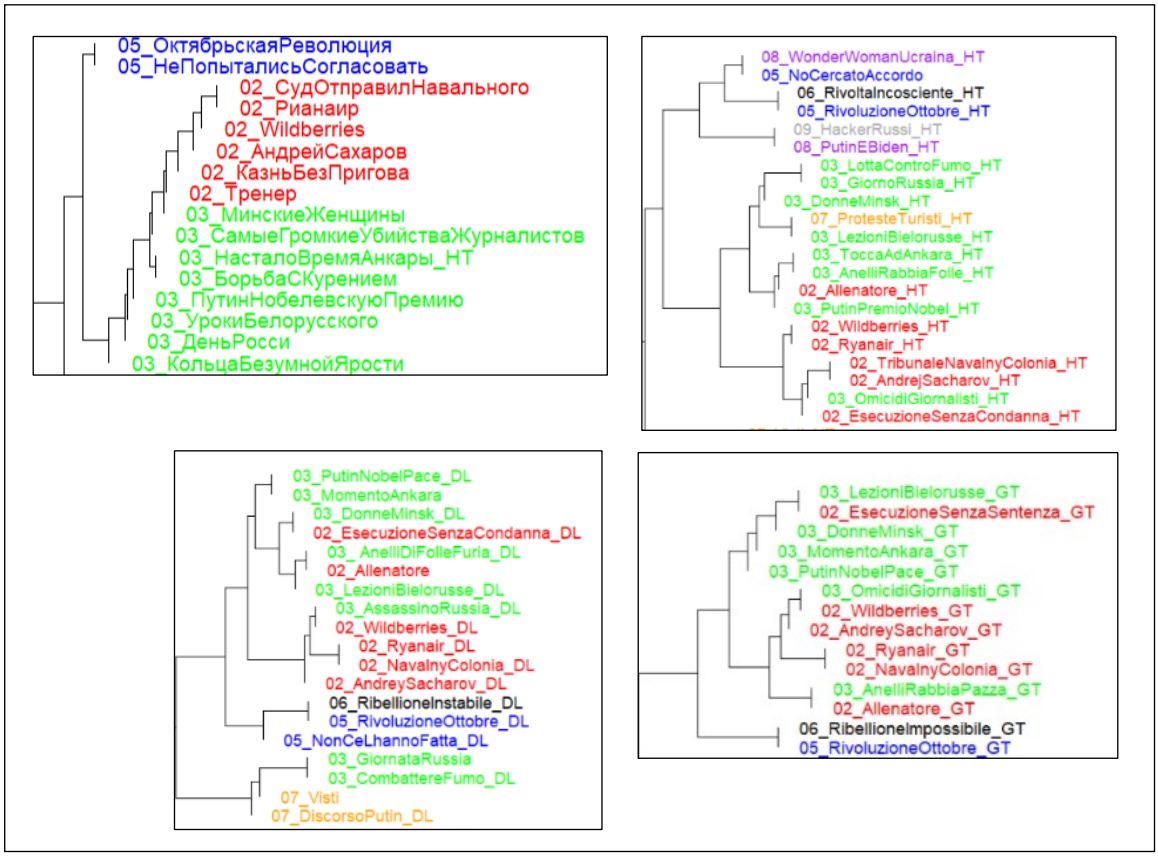

Fig. 27 - *Differenze tra cluster testi originali e traduzioni. In alto a sx corpus di testi originali russi; in alto a dx corpus di traduzioni umane; in basso a sx corpus di traduzioni di DeepL; in basso a dx corpus di traduzioni di Google Translate.*

Dalla Figura 27 è possibile osservare che la suddivisione in *cluster* degli articoli con codice "03_" non è stata completamente rispettata in nessun *corpus* di traduzioni, ma il dendrogramma che raggruppa in maniera più fedele questi articoli all'interno di un unico macro-*cluster* è quello delle traduzioni umane. Un discorso simile si può fare anche con gli articoli aventi codice "05_" che si presentano all'interno dello stesso macro-*cluster* sia nel *corpus* "Traduzioni Umane" sia in "DeepL", sottolineando nuovamente le similitudini tra le traduzioni umane e quelle automatiche di DeepL. Infine, per quanto riguarda gli articoli con codice "02_", sembra che la rappresentazione più fedele all'originale sia la suddivisione in *cluster* delle traduzioni umane, che risulta comunque molto simile a quella delle traduzioni di DeepL. In conclusione, si può affermare che non sempre gli stili dei traduttori automatici sono i più simili all'originale, ma a volte le traduzioni umane somigliano maggiormente ai testi originali rispetto alle traduzioni automatiche. Inoltre, si è notata una forte somiglianza tra le traduzioni umane e le traduzioni automatiche di DeepL.

4.5. Riconoscimento automatico delle HT e MT

La quinta e ultima analisi ha lo scopo di verificare la capacità di un algoritmo di *machine learning* di distinguere le traduzioni umane da quelle automatiche. In questa ricerca sono stati usati gli algoritmi *Support Vector Machine* (SVM) e *Random Forest* (RF), i cui funzionamenti sono descritti nel paragrafo 3.5. Solitamente per allenare correttamente un algoritmo di *machine learning* si ha bisogno di collezionare un *corpus* di grandi dimensioni, che contenga milioni di dati testuali al suo interno (cfr. Wu et al., 2016; Popel et al. 2020). È stato provato, però, che gli algoritmi SVM e RF riescono a distinguere facilmente lo stile comunicativo dei politici (per esempio se si mettono a confronto discorsi di Obama e Trump) anche solo prendendo in considerazione un numero limitato dei loro discorsi politici. Per questo motivo, nonostante le dimensioni dei *corpora* "Traduzioni Umane", "DeepL" e "Google Translate" siano ridotte, si è comunque deciso di provare a realizzare questo studio.

Sono state eseguite 9 analisi in totale, in cui veniva richiesto agli algoritmi di riconoscere l'autore di una traduzione e capire se si trattasse di un essere umano, di DeepL o di Google Translate. Come esempio di traduzione è stata utilizzata una traduzione extra non presente nei *corpora* in italiano, le cui fonti russe e italiane sono visibili al paragrafo 6.3. Tutte le traduzioni sono state precedentemente rinominate nei seguenti modi: "HT_*titolo dell'articolo*" per le traduzioni umane; "MTDL_*titolo dell'articolo*" per le traduzioni di DeepL; "MTGT_*titolo dell'articolo*" per le traduzioni di Google Translate. In tutte le indagini sono state inserite le traduzioni, di cui bisognava capire l'autore, con le seguenti nomine: "Unknown_HT" per la traduzione umana, "Unknown_DL" per la traduzione di DeepL; "Unknown_GT" per la traduzione di Google Translate. Attraverso la modifica dei nomi dei diversi *file* da analizzare, gli algoritmi sono stati in grado di distinguere senza problemi le traduzioni umane, da quelle di DeepL e Google Translate e individuare il testo di cui dovevano riconoscere l'autore grazie alla *keyword* "Unknown". È importante far notare che la porzione di testo presente dopo l'*underscore* ("_") è ignorata dagli algoritmi; quindi, la presenza dell'acronimo HT in "Unknown_HT" non ha influenzato nella capacità di SVM e RF di riconoscere o meno l'autore della traduzione.

Nelle prime due indagini gli algoritmi sono stati allenati prima con i *corpora* "Traduzioni Umane" e "DeepL", poi con i *corpora* "Traduzioni Umane" e "Google Translate". In entrambe le analisi è stata inserita la traduzione "Unknown_HT", il cui autore umano doveva essere riconosciuto dagli algoritmi.

SVM: Unknown_HT		
Previsione	HT	MTDeepL
HT	41,6	14,2
MTDeepL	12,6	35,4
Accuratezza	0,74	
Previsione finale		
	HT	MTDeepL
	0	6

RF: Unknown_HT		
Previsione	HT	MTDeepL
HT	33,6	27,6
MTDeepL	20,6	22,0
Accuratezza	0,54	
Previsione finale		
	HT	MTDeepL
	6	0

Tab. 10 - *Riconoscimento traduzione umana dalla traduzione di DeepL. A sinistra i risultati dell'algoritmo SVM, a destra i risultati dell'algoritmo RF.*

SVM: Unknown_HT		
Previsione	HT	MTGT
HT	42,8	13,0
MTGT	11,4	36,6
Accuratezza	0,76	
Previsione finale		
	HT	MTGT
	0	6

RF: Unknown_HT		
Previsione	HT	MTGT
HT	32,0	29,2
MTGT	22,2	20,4
Accuratezza	0,5	
Previsione finale		
	HT	MTGT
	6	0

Tab. 11 - *Riconoscimento traduzione umana dalla traduzione di Google Translate. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

In entrambe le analisi è possibile notare che il livello di accuratezza, ovvero la capacità dell'algoritmo di distinguere tra i due gruppi di testi utilizzati nel *training*, non è buono in nessun algoritmo. L'accuratezza di SVM (0,74 e 0,76) è maggiore rispetto a RF (0,54 e 0,50), ma l'algoritmo SVM pare non distinguere la traduzione umana né dalla traduzione di DeepL né da quella di Google Translate, perché un'accuratezza del 74% o 76% indica che l'algoritmo ha ancora difficoltà a riconoscere le differenze tra i *corpora* inseriti nel *training*. Al contrario RF sembrerebbe riuscire a distinguere la traduzione umana dalle traduzioni di DeepL e Google Translate, ma il livello di accuratezza registrato è del 50% e quindi non è sufficiente per affermare che l'algoritmo sia in grado di distinguere la traduzione umana da quella automatica. Un'accuratezza del 50%, infatti, indica che l'algoritmo

non riesce a riconoscere con sicurezza l'autore del testo e ha una probabilità del 50% di identificare l'autore corretto o quello scorretto; per questo motivo, la sua risposta non è affidabile.

Successivamente, è stata inserita una traduzione realizzata da DeepL, chiamata "Unknown_DL", che i due algoritmi dovevano distinguere prima dalle traduzioni umane (v. Tabella 12), poi dalle traduzioni di Google Translate (v. Tabella 13).

SVM: Unknwon_DL		
Previsione	HT	MTDeepL
HT	41,8	14,2
MTDeepL	12,4	35,4
Accuratezza	0,74	
Previsione finale		
	HT	MTDeepL
	0	6

RF: Unknown_DL		
Previsione	HT	MTDeepL
HT	34,0	27,2
MTDeepL	20,2	22,4
Accuratezza	0,54	
Previsione finale		
	HT	MTDeepL
	5	1

Tab. 12 - *Riconoscimento traduzione di DeepL dalla traduzione umana. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

SVM: Unknwon_DL		
Previsione	MTDeepL	MTGT
HT	43,6	6,6
MTDeepL	6,0	43,0
Accuratezza	0,87	
Previsione finale		
	MTDeepL	MTDeepL
	3	3

RF: Unknown_DL		
Previsione	MTDeepL	MTGT
HT	14,0	36,0
MTDeepL	35,6	13,6
Accuratezza	0,28	
Previsione finale		
	MTDeepL	MTGT
	1	5

Tab. 13 - *Riconoscimento traduzione di DeepL dalla traduzione di Google Translate. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

Anche nelle seconde analisi si è notato che l'algoritmo SVM (0,74 e 0,87) presenta un'accuratezza maggiore rispetto all'algoritmo RF, che non riesce nemmeno a raggiungere un'accuratezza del 50% nel confronto tra le traduzioni di DeepL e quelle di Google Translate. Inoltre, si può notare che l'algoritmo SVM è in grado di distinguere le traduzioni automatiche di DeepL, da quelle umane, mentre RF confonde le traduzioni umane con quelle di DeepL. Nel confronto tra traduzioni automatiche, si è notato che SVM, nonostante la capacità di apprendere le differenze tra i due *corpora* superi l'80%, di fronte al testo "Unknown_" non riesce a distinguere le traduzioni automatiche di DeepL da quelle di Google Translate, mostrando che esiste un 50% di probabilità che l'autore della traduzione possa

essere DeepL o Google Translate. RF, invece, confonde completamente la traduzione di DeepL con quelle di Google Translate, riconoscendo Google Translate come autore della traduzione.

Infine, è stata inserita la traduzione di Google Translate, nominata "Unknown_GT", ed è stata messa a confronto prima con le traduzioni umane e poi con le traduzioni automatiche di DeepL. Come si può notare dalle Tabelle 14 e 15, gli algoritmi hanno fornito nuovamente previsioni diverse.

SVM: Unknwon_GT		
Previsione	HT	MTGT
HT	43,2	13,2
MTDeepL	11,0	36,4
Accuratezza	0,77	
Previsione finale		
	HT	MTGT
	0	6

RF: Unknwon_GT		
Previsione	HT	MTGT
HT	32,2	30,2
MTDeepL	22,0	19,4
Accuratezza	0,5	
Previsione finale		
	HT	MTGT
	5	1

Tab. 14 - *Riconoscimento traduzione di Google Translate dalla traduzione umana. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

SVM: Unknwon_GT		
Previsione	MTDeepL	MTGT
HT	43,6	6,6
MTDeepL	6,0	43,0
Accuratezza	0,87	
Previsione finale		
	MTDeepL	MTGT
	4	2

RF: Unkwnon_GT		
Previsione	MTDeepL	MTGT
HT	15,2	35,6
MTDeepL	34,4	14,0
Accuratezza	0,3	
Previsione finale		
	MtDeepL	MTGT
	2	4

Tab. 15 - *Riconoscimento traduzione di Google Translate dalla traduzione di DeepL. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

SVM sembra che riesca a distinguere la traduzione di Google Translate dalle traduzioni umane. Al contrario, pare che SVM non riesca a distinguere la traduzione di Google Translate da quella di DeepL. L'algoritmo RF, invece, sembra che riconosca meglio la traduzione di Google Translate da quella di DeepL, rispetto alla traduzione di Google Translate da quella umana. Le percentuali di accuratezza di RF, però, sono troppo basse per affermare con certezza che l'algoritmo riesca a distinguere l'autore di una traduzione rispetto ad un altro. Di conseguenza questo risultato di RF non può essere considerato valido.

Infine, è stata effettuata un'analisi generale utilizzando tutti i *corpora* ("HT_"; "MTDeepL_"; "MTGT_") e inserendo prima la traduzione umana ("Unknown_HT"), poi la traduzione di DeepL ("Unknown_DL") e la traduzione di Google Translate ("Unkown_GT"). I risultati sono visibili nelle Tabelle 16, 17 e 18.

SVM: Unknwon_HT			
Previsione	HT	MTDeepL	MTGT
HT	36,4	14,6	11,2
MTDeepL	11,0	31,2	4,8
MTGT	6,8	3,8	33,6
Accuratezza	0,7		

Previsione finale		
HT	MTDeepL	MTGT
0	3	3

RF: Unknwon_HT			
Previsione	HT	MTDeepL	MTGT
HT	33,0	14,8	14,2
MTDeepL	10,0	7,2	27,6
MTGT	11,2	27,6	7,8
Accuratezza	0,31		

Previsione finale		
HT	MTDeepL	MTGT
6	0	0

Tab. 16 - *Riconoscimento traduzione di DeepL dalle traduzioni di Google Translate e umana. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

SVM: Unkwnon_DL			
Previsione	HT	MTDeepL	MTGT
HT	37,0	14,2	11,4
MTDeepL	10,4	31,4	4,8
MTGT	6,8	4,0	33,4
Accuratezza	0,67		

Previsione finale		
HT	MTDeepL	MTGT
0	3	3

RF: Unkwnon_DL			
Previsione	HT	MTDeepL	MTGT
HT	33,6	14,4	16,2
MTDeepL	10,0	8,6	26,4
MTGT	10,6	26,6	7,0
Accuratezza	0,32		

Previsione finale		
HT	MTDeepL	MTGT
5	0	1

Tab. 17 - *Riconoscimento traduzione umana dalla traduzione di DeepL e Google Translate. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

SVM: Unkwnon_GT			
Previsione	HT	MTDeepL	MTGT
HT	37,0	14,2	11,4
MTDeepL	10,4	31,4	4,8
MTGT	6,8	4,0	33,4
Accuratezza	0,67		

Previsione finale		
HT	MTDeepL	MTGT
0	4	2

RF: Unknwon_GT			
Previsione	HT	MTDeepL	MTGT
HT	33,4	15,0	13,4
MTDeepL	10,0	7,2	29,6
MTGT	10,8	27,4	6,6
Accuratezza	0,30		

Previsione finale		
HT	MTDeepL	MTGT
5	0	1

Tab. 18 - *Riconoscimento traduzione di Google Translate dalle traduzioni di DeepL e umana. A sinistra i risultati dell'algoritmo SVM, a destra i risutati dell'algortimo RF.*

Si può notare che i risultati delle precedenti analisi sono confermati anche nell'analisi di tutti i *corpora* in contemporanea. Si osserva infatti che SVM non riconosce la traduzione umana e la confonde con le traduzioni automatiche, al contrario RF sembra riconosce la traduzione umana, ma nuovamente la percentuale di accuratezza del 31% è troppo bassa per accertare questo dato (v. Tabella 16). Successivamente si nota come SVM non distingua completamente la traduzione di DeepL da quella di Google Translate e nemmeno la traduzione di Google Translate da quella di DeepL, ma differenzia interamente la traduzione automatica da quella umana (v. Tabella 17 e Tabella 18 a sinistra). Al contrario RF confonde la traduzione automatica con quella umana sia inserendo la traduzione automatica di DeepL che quella di Google Translate (v. Tabella 17 e Tabella 18 a destra).

Si può concludere che, con una divisione dei testi in *chunk* da 200 parole e un'analisi dei 200 *bigram*, *trigram*, *word* e *word bigram* più frequenti, nessuno dei due algoritmi lavora bene con il piccolo corpus a disposizione. *Random Forest* presenta sempre una percentuale di accuratezza troppo bassa ($\leq 54\%$) e, per questo motivo, i suoi risultati non possono essere considerati affidabili. Anche nei casi in cui RF è riuscito a riconoscere l'autore corretto della traduzione (cfr. Tabella 10), può essere che questo risultato sia casuale e non dovuto a un vero riconoscimento dell'autore da parte dell'algoritmo. Un discorso analogo vale anche per *Support Vector Machine*. In questo caso l'algoritmo ha presentato sempre un'accuratezza $\geq$ 74%, che è diminuita fino al 67% solo nelle ultime tre analisi in cui l'algoritmo è stato allenato con tutti i *corpora* di traduzione.

Nonostante le percentuali di accuratezza siano basse ma quasi sempre accettabili, si è notato che l'algoritmo SVM fatica a riconoscere correttamente i vari autori delle traduzioni. Probabilmente questa mancanza è dovuta alle dimensioni limitate dei *corpora* presi in esame. Tutti i *corpora* in italiano, infatti, contengono 34 traduzioni e ogni testo tradotto ha dimensioni molto variabili. Il testo più lungo ("Da Tuva") contiene 6614 *word-token* nel *corpus* "Traduzioni Umane", 6377 nel *corpus* "DeepL" e 6178 nel *corpus* "Google Translate"; quello più breve ("Navalny Colonia"), invece, contiene 514 *word-token* nel *corpus* "Traduzioni Umane", 480 nel *corpus* "DeepL" e 488 nel *corpus* "Google Translate". Questa grande differenza nella lunghezza dei testi ha delle conseguenze nella fase di *training* degli algoritmi

SVM e RF. Come spiegato in precedenza, ogni testo presente nei diversi *corpora* è stato diviso in *chunk* da 200 parole, creando una disparità di dimensioni tra i vari testi, che può aver impedito il corretto allenamento degli algoritmi nella fase di *training set*.

In conclusione, i risultati di questa analisi sono stati considerati non accettabili e, in eventuali studi futuri, si consiglia di collezionare un numero maggiore di testi dalle dimensioni simili per ottenere risultati migliori.

5. Discussione dei risultati e considerazioni finali

5.1. Ipotesi e obiettivi iniziali

Il presente elaborato si è concentrato sullo studio delle differenze tra traduzione umana e automatica attraverso l'utilizzo di metodi di analisi quantitativa dei testi. Uno degli scopi della tesi, infatti, era capire se fosse possibile analizzare le caratteristiche della traduzione umana e automatica senza uno studio qualitativo dei testi, ma solo quantitativo. A questo proposito è stato deciso di prendere come riferimento la *corpus linguistics*, ovvero una disciplina che studia i contenuti di un *corpus* (Ondelli 2018, pp. 134-135). In linea di principio, con il termine *corpus* si fa riferimento, a una collezione di testi "coerente con gli scopi perseguiti dalla ricerca" (Tuzzi, 2003, p. 29). Tuttavia, è necessario ricordare che, come scrive Ondelli (2018), non tutte le collezioni di testi possono essere considerate un *corpus*, inteso come oggetto di analisi della *corpus linguistics*. In questo caso, i testi vengono raccolti in formato elettronico, in modo che possano essere elaborati e analizzati automaticamente da *software*, e, in fase di raccolta dei testi, vengono seguiti una serie di fattori (socio)linguistici, come le cinque variazioni della lingua di Berruto (variazione diacronica, diatopica, diafasica, diastratica e diamesica) (Barbera et al. 2007, p. 70; Ondelli 2018, p. 134; Berruto 1987, pp. 19-279).

Nella presente tesi si è deciso di raccogliere quattro differenti *corpora*:

1. Il primo *corpus*, chiamato "Originali russi", raccoglie 34 articoli di giornale scritti in russo e pubblicati in diverse testate giornalistiche online. Le fonti di questi articoli sono disponibili al capitolo 6.1.
2. Il secondo *corpus*, chiamato "Traduzioni Umane", è costituito da 34 traduzioni in italiano degli articoli russi presenti nel *corpus* "Originali russi". Le traduzioni sono state effettuate da diversi traduttori umani e sono pubblicate nel sito *Russia In Translation*. Le fonti di queste traduzioni sono disponibili al capitolo 6.2.
3. Il terzo *corpus*, chiamato "DeepL", è formato da 34 traduzioni in italiano degli articoli originali in russo. Le traduzioni sono state effettuate dal traduttore automatico DeepL.

4. Il quarto *corpus*, chiamato "Google Translate", è costituito da 34 traduzioni in italiano del *corpus* "Originali russi" e le traduzioni sono state realizzate dal *software* di traduzione automatica Google Translate.

Oltre a verificare l'attendibilità dell'uso di metodi di analisi quantitativa dei testi in studi di traduttologia, l'obiettivo di questa tesi era capire le differenze tra traduzione umana e automatica e analizzare quanto la *machine translation* si sia avvicinata alla traduzione umana. La ricerca è stata limitata allo studio della traduzione di un unico genere testuale, ovvero l'articolo di giornale pubblicato online. Si ritiene necessario ribadire che tutti gli articoli russi raccolti nel *corpus* trattano temi inerenti alla politica e alla società, ma non comprendono alcun articolo inerente all'attuale guerra russo-ucraina. Tutti gli articoli collezionati, infatti, sono stati pubblicati tra il 30/03/2020 e il 13/12/2021 e sono stati esclusi articoli pubblicati nel 2022 per evitare la collezione di testi propagandistici.

Per raggiungere gli obiettivi della presente tesi sono state formulate le seguenti ipotesi:

1. Un traduttore automatico non traduce sempre correttamente abbreviazioni, acronimi, nomi stranieri e non traslittera in modo corretto nomi propri dall'alfabeto cirillico a quello latino e viceversa, perché la traduzione automatica non è precisa nel tradurre costruzioni lessicali, semantiche e pragmatiche complesse (Li et al. 2014, p. 190). Inoltre, nella traduzione di termini che prevedono almeno in teoria un'unica traduzione possibile, il traduttore automatico tenderà a tradurre sempre in modo univoco, mentre il traduttore umano utilizzerà dei sinonimi in alcuni casi. Un traduttore automatico, infatti, tende a produrre una traduzione letterale senza tenere in considerazione il contesto culturale o le particolarità del pubblico a cui si rivolge (Ibanez 2021). Al contrario, il traduttore umano può ricorrere a diverse tecniche di traduzione per rendere più comprensibili determinati termini ai lettori della traduzione, utilizzando delle traduzioni descrittive o sinonimi comprensibili ai lettori (cfr. Torresin 2022, pp. 54-55);
2. Inoltre, è stato ipotizzato che un traduttore automatico produce testi lessicalmente e sintatticamente più semplici rispetto a un traduttore umano. Per valutare l'universale traduttivo della *semplificazione* sono stati

confrontanti i valori di *Type Token Ratio* e lunghezza media delle frasi per ogni testo presente nei *corpora* in italiano. (cfr. Ondelli & Viale 2010, pp-3-5; Kunilovskaya et al. 2018, pp. 6-12).

3. In un confronto di *topic* presenti nelle traduzioni e nei testi originali in russo, la traduzione automatica presenta *topic* più simili ai testi originali rispetto ai traduttori umani. La traduzione automatica, infatti, traducendo più letteralmente rispetto al traduttore umano, può rispettare maggiormente la struttura di *topic* dei testi originali, rispetto a un traduttore umano che può decidere di allontanarsi dall'originale per rendere il testo di arrivo più comprensibile al lettore finale (cfr. Ibanez 2021; Torresin 2022, pp. 54-55);
4. Successivamente, confrontando gli stili di scrittura degli autori originali e dei traduttori, è stato ipotizzato che i traduttori umani siano meno influenzati dallo stile di scrittura degli autori originali dei testi, perché non traducono sempre letteralmente come i traduttori automatici. Al contrario, un traduttore automatico produce traduzioni con uno stile più simile all'autore originale del testo in russo. È importante specificare che, in un'analisi quantitativa dei dati testuali, per "stile" di un autore si intende l'insieme di bigrammi, trigrammi o delle prime *n* parole più frequenti all'interno di un testo (Eder et al. 2016, pp. 107-106);
5. A seguire, è stato ipotizzato che algoritmi di *machine learning* adibiti al riconoscimento dell'autore di un testo, siano in grado di distinguere le traduzioni umane da quelle automatiche e le traduzioni automatiche di DeepL da quelle di Google Translate. Sulla base di studi precedenti, infatti, è stato provato che il metodo del *machine learning* sia in grado di distinguere una traduzione umana da una automatica (cfr. Li et al. 2015; Fu et al. 2021).
6. Infine, è stato presupposto che in generale la traduzione automatica di DeepL sia più precisa rispetto a quella di Google Translate. Nonostante entrambi utilizzino la *neural machine translation* (v. cap. 1.1), solo in DeepL vengono effettuati regolarmente dei test in cui traduttori professionisti giudicano la resa migliore tra una serie di traduzioni automatiche proposte (cfr. https://www.deepl.com/en/whydeepl).

Per confermare o confutare le ipotesi sopra esposte, sono state eseguite una serie di indagini, applicando metodi e strumenti di analisi quantitativa dei testi.

5.2. Discussione dei risultati trovati in *AntConc*

La prima analisi ha riguardato lo studio e il confronto delle traduzioni di acronimi, abbreviazioni, nomi stranieri, termini e traslitterazioni, prendendo sempre come punto di riferimento i testi originali scritti in russo. Oltre alle traduzioni, sono state studiate anche le occorrenze delle parole prese in esame e si è verificato se l'occorrenza di una parola russa era la stessa anche nelle traduzioni. I risultati di queste analisi sono riassunti nelle Tabelle 4, 5, 6, 7, 8, 9 presenti nel capitolo 4.1. Il *software* utilizzato per effettuare queste indagini è *AntConc* e in particolare sono state utilizzate le funzioni "Concordance Plot" e "File View", i cui funzionamenti sono visibili in Fig. 5 e Fig. 6 al capitolo 3.1.

Attraverso un confronto dei risultati raccolti, si è notato che nella traduzione degli acronimi russi, i traduttori umani non hanno sempre rispettato la traduzione letterale presente nei dizionari online e cartacei. Ad esempio, l'acronimo russo "США", che letteralmente in italiano si traduce con "USA", non viene sempre tradotto in questo modo nelle traduzioni umane, perché la lingua italiana non ricorre ampiamente all'uso di acronimi come nella lingua russa. Per questo motivo, l'acronimo "США" è stato tradotto spesso con "Stati Uniti" e non "USA". Lo stesso fenomeno è stato riscontrato anche nei traduttori automatici, che non hanno sempre tradotto letteralmente gli acronimi, ma hanno fatto uso di sinonimi. Ad esempio, la parola "США", con occorrenza 78 nel *corpus* "Originali russi", è stata tradotta da DeepL 8 volte con "USA" e 64 volte con "Stati Uniti"; Google Translate ha tradotto 6 volte "USA" e 62 volte "Stati Uniti"; i traduttori umani, invece, hanno tradotto 17 volte con "USA" e 57 volte con "Stati Uniti". In generale si è notato che sia DeepL che Google Translate sono in grado di riconoscere e tradurre senza problemi gli acronimi russi in italiano, senza ricorrere a traslitterazioni perché l'acronimo non è stato compreso. L'unico caso in cui sembra sia avvenuta una traslitterazione è nella traduzione di "НКР" (occorrenza 7), tradotto con "Nagorno-Karabach" (occorrenza 11) o "Nagorno-Karabakh" (occorrenza 3) dai traduttori umani, ma con "NKR" (occorrenza 7) da entrambi i *software* di traduzione automatica. In questo caso specifico si può supporre che i traduttori automatici non abbiano compreso il

significato dell'acronimo e, invece di tradurre, hanno traslitterato il nome in caratteri latini. Si può, quindi, supporre che i traduttori automatici riescono generalmente a riconoscere e tradurre correttamente gli acronimi russi più usati (es. "США", "СМИ"), ma possono ricorrere a semplici traslitterazioni quando gli acronimi russi sono più specifici e meno frequenti (es. "НКР"). Tuttavia, è necessario far notare che "НКР" è l'unico acronimo che è stato traslitterato e non tradotto dai *software* di traduzione automatica: Di conseguenza, per confermare questa tesi è necessario uno studio più approfondito delle traduzioni di acronimi meno frequenti nell'uso comune della lingua russa.

Successivamente, per quanto riguarda le due abbreviazioni analizzate ("др" con occorrenza 4 e "РЕД" con occorrenza 4), si è notato che i traduttori automatici non hanno avuto problemi a tradurre correttamente le due abbreviazioni, perché entrambi hanno tradotto rispettivamente con "ecc." e "ndr". Solo in un caso il traduttore DeepL non è stato in grado di tradurre correttamente "РЕД" con "ndr", ma ha utilizzato l'abbreviazione "I.T.". Lo stesso errore non si trova nelle traduzioni di Google Translate, che, al contrario, ha sempre tradotto le due abbreviazioni correttamente ed è stato più preciso di DeepL.

A seguire, l'analisi delle traslitterazioni ha permesso di far notare caratteristiche interessanti non solo nei traduttori automatici, ma anche in quelli umani. Dalla Tabella 7 a pagina 74, è possibile notare che i traduttori umani del sito *Russia in Translation* non hanno utilizzato le stesse regole di traslitterazione. Ad esempio, secondo la norma ISO-09 utilizzata in Italia (cfr. Torresin 2022, p. 33), la traslitterazione corretta di "Навальный" è "Naval'nyj" (occorrenza 57 nel *corpus* "Traduzioni Umane"). Questa forma, però, non viene sempre utilizzata dai traduttori umani, che in alcuni casi hanno traslitterato il nome proprio "Навальный" con "Naval'nyi" (occorrenza 2), "Naval'ny" (occorrenza 10) e "Naval'nij" (occorrenza 1). Tali traslitterazioni non sono scorrette, perché nei testi pubblicati sul web si utilizza la traslitterazione commerciale, che presenta regole leggermente diverse dalla traslitterazione scientifica prevista dalla norma ISO-09 (cfr. Torresin 2022, pp. 34; 56; 172-173). Prendendo quindi come variante corretta sia la traslitterazione scientifica che commerciale, si può affermare che i *software* di traduzione automatici tendano a traslitterare seguendo maggiormente le regole

della traduzione commerciale rispetto a quella scientifica. Infine, si è notato che il segno debole "ь" non viene apparentemente mai traslitterato con un apostrofo dai *software* di traduzione automatica, come previsto dalla norma ISO-09, ma viene omesso, come concesso dalla traslitterazione commerciale (es. "Nava**l**ny" e non "Nava**l'**ny").

Successivamente, sono state analizzate le traduzioni dei prestiti stranieri e forestierismi (v. Tab. 8, p. 76). È stato notato che generalmente entrambi i traduttori automatici presi in esame traducono in modo corretto sia i forestierismi che i prestiti stranieri, trovando l'equivalente corretto in lingua italiana. Questo, però, non accade in tutte le traduzioni. I forestierismi "телеграм" e "вотсапп", che indicano i due *social network* "Telegram" e "Whatsapp", sono stati tradotti erroneamente da DeepL con "canale telegrafico" o "telegrafo" e "waps". Google Translate, al contrario, ha tradotto in modo corretto "вотсапп", ma ha confuso la parola "телеграм" con "telegramma". Altri forestierismi, come "тикток" o "ютуб", sono stati tradotti in modo corretto dai traduttori automatici con "TikTok" e "Youtube", eccetto per DeepL che ha tradotto una volta "тикток" con "tris". Per quanto concerne i prestiti stranieri, il nome proprio dell'azienda "Wildberries" non è sempre stato mantenuto invariato, ma è stato tradotto una volta con "bacche selvatiche" da DeepL e tre volte con "frutti di bosco" da Google Translate. Un fenomeno analogo si trova anche nella traduzione del nome proprio dell'azienda "moonlight", che è stato tradotto una volta con "chiaro di luna" da DeepL, mentre Google Translate non presenta questo errore.

Infine, sono state analizzate le traduzioni di termini, che almeno in teoria dovrebbero presentare una traduzione univoca. Si è notato che, contrariamente a quanto si ipotizzava all'inizio, i traduttori umani hanno quasi sempre tradotto in modo univoco i termini scelti (v. Tab. 9 p. 78), presentando quasi sempre le stesse occorrenze dei termini originali russi. In alcuni casi dei termini sono stati resi con dei sinonimi per esigenze della lingua italiana. Ad esempio, in due casi il traduttore umano ha preferito rendere la parola "коронавирус" con "pandemia" e non "coronavirus", perché in italiano aveva una resa migliore.

In conclusione, si può affermare che in generale i traduttori automatici non presentano problemi né nella traduzione di acronimi né di abbreviazioni. Le

traslitterazioni di DeepL e Google Translate rimangono corrette perché seguono la traslitterazione commerciale presente nel web e non diversificano di molto dalla traslitterazione umana. Le uniche tipologie di parole che hanno creato di fatto dei problemi ai traduttori automatici sono i prestiti stranieri e i forestierismi, che in alcuni casi vengono tradotti erroneamente rendendo difficile la loro comprensibilità. Bisogna ricordare che sono stati trovati maggiori errori nelle traduzioni di DeepL, rispetto a quelle di Google Translate, che sembra più preciso di DeepL nella traduzione a livello lessicale. Infine, è stata trovata conferma del fatto che il traduttore umano tenda ad utilizzare dei sinonimi anche di termini, che dovrebbero avere una traduzione univoca. È necessario, però, precisare che questa tendenza è stata ritrovata in pochi casi e che tendenzialmente il traduttore umano ha tradotto in modo univoco questi termini.

5.3. Discussione dei risultati in *Voyant-tools*: verifica del grado di *semplificazione* delle traduzioni italiane

Durante questa analisi sono stati utilizzati solo i *corpora* in lingua italiana, perché il *software Voyant-tools* non legge la lingua russa. In ogni caso, una comparazione corretta del *Type Token Ratio* e della lunghezza media della frase all'interno di un testo funziona solo con testi scritti nella stessa lingua con lunghezza simile (cfr. cap. 4.2). All'interno di questa tesi sono state analizzate traduzioni in italiano degli stessi testi di partenza; quindi, le traduzioni hanno tutte circa la stessa lunghezza. I testi dei diversi *corpora* sono stati analizzati separatamente all'interno di *Voyant-tools*, che ha permesso di calcolare in modo automatico il TTR e la lunghezza media delle frasi in ogni singola traduzione. I risultati sono stati successivamente riassunti in Fig. 11 pagina 80 e Fig. 12 a pagina 81. Al termine dell'analisi, non sono state osservate particolari differenze nei TTR delle varie traduzioni, ma si è riscontrata una lunghezza media della frase maggiore nelle traduzioni umane, rispetto a quelle automatiche. Attraverso questi due parametri, prendendo come riferimento gli studi di Ondelli e Viale (2019) e di Kunilovskaya et al. (2018), si può affermare che, le traduzioni automatiche non sono necessariamente più semplici delle traduzioni umane, perché i valori dei TTR dei vari testi sono pressoché gli stessi in tutti i testi. Al contrario, il calcolo della

lunghezza media delle frasi in un testo ha permesso di evidenziare la maggiore complessità sintattica delle traduzioni umane rispetto a quelle automatiche.

Infine, è necessario far notare che il calcolo del TTR e della lunghezza media della frase sono solo alcuni dei parametri utili nel verificare la ricchezza lessicale di un *corpus* di traduzioni. Un altro parametro da poter utilizzare è, ad esempio, la densità lessicale, ovvero il rapporto tra parole piene e vuote all'interno di un testo, che potrebbe risultare minore nelle traduzioni automatiche e maggiore nelle traduzioni umane (cfr. Ondelli & Viale 2010, pp. 3-5).

In conclusione, potrebbe essere interessante analizzare in modo più approfondito tutti i principali universali traduttivi (cfr. cap. 1.2) e verificare se effettivamente le traduzioni umane sono più complesse di quelle automatiche.

5.4. Discussione dei risultati in *Iramuteq*: analisi dei *topic*

L'obiettivo iniziale dell'analisi dei *topic* era quello di confrontare la suddivisione dei *topic* del *corpus* "Originali russi" con le suddivisioni di *topic* nei *corpora* di traduzioni, per verificare quali *corpus* di traduzioni presentavano la suddivisione in *topic* più simile all'originale. Purtroppo, non è stato possibile realizzare questo confronto perché con il *software Iramuteq* non siamo stati in grado di leggere correttamente la lingua russa con nessun tipo di *encoding*. Si è deciso di svolgere ugualmente questa analisi e confrontare solo i *corpora* in italiano, una lingua che *Iramuteq* legge senza problemi. Si è scelto, allora, di prendere come punto di riferimento la suddivisione in *topic* delle traduzioni umane e verificare le similitudini dei traduttori automatici con le traduzioni umane. Sono stati utilizzati valori diversi di "dimensioni di numeri di *cluster* finali in fase 1" per ottenere dendrogrammi delle stesse dimensioni con un totale di 6 classi. Dopo un'attenta analisi dei *topic* presenti nei tre dendrogrammi (v. cap. 4.3 pp. 83-84), si è notato che i *corpora* "Traduzioni Umane" e "DeepL" condividono una suddivisione dei *topic* molto più simile rispetto a quella del *corpus* "Google Translate". Ad esempio, i *corpora* "Traduzioni umane" e "DeepL" uniscono nello stesso *cluster* le classi a tema società e formazione, mentre il *corpus* "Google Translate" presenta una suddivisione in *cluster* differente (v. Fig. 17, p. 85). Un ulteriore similitudine tra le traduzioni umane e le traduzioni di DeepL è stata trovata nel *cluster* che unisce la

classe di temi inerenti alla casa e alla famiglia e la classe con *word-type* inerenti al tema della città. La stessa similitudine non si trova nel corpus "Google Translate", in cui la classe inerente al tema casa e famiglia si trova isolata e non è presente una classe che parli del tema città (v. Fig. 15 p. 84; Fig. 18 p. 86). Pare, quindi, che le traduzioni di DeepL presentino contenuti più simili alle traduzioni umane, rispetto a quelle Google Translate.

In conclusione, si può affermare che l'analisi dei *topic* in *Iramuteq* è stata comunque utile ai fini di questa tesi perché ha permesso di trovare molte similitudini tra le traduzioni umane e quelle automatiche di DeepL. Anche se non è stato possibile realizzare il confronto della suddivisione in *topic* con i testi originali, si può comunque confermare la riuscita di questa analisi che ha messo in evidenza la capacità del *software* DeepL di produrre traduzioni più simili all'essere umano, rispetto al programma Google Translate. All'interno di possibili ricerche future sarebbe interessante poter implementare il *software Iramuteq* anche per la lingua russa e approfondire il confronto dei *topic* della presente tesi, includendo il *corpus* "Originali Russi".

5.5. Discussione dei risultati in *stylo*: *cluster analysis*

Lo scopo della *cluster analysis* all'interno della presente tesi era quello di trovare le differenze tra traduzione automatica e umana, attraverso l'analisi dello stile di scrittura dei diversi autori dei testi. Si ricorda che per "stile di scrittura" qui si intende una misura basata sui *most frequent words* presenti nei vari *corpora* che serve a stabilire somiglianze e differenze tra testi. Non sono state eseguite analisi prendendo in considerazione i bigrammi e trigrammi più frequenti. Le similitudini e differenze di stile sono state analizzate attraverso l'osservazione dei diversi dendrogrammi prodotti da *stylo*.

In primo luogo, è stata verificata la capacità del pacchetto *stylo* di distinguere la traduzione umana da quella automatica, prendendo in considerazione i primi 200 *most frequent words* presenti in tutti i *corpora* di traduzione. Come si può notare dalla Figura 19 a pagina 88, il *software* è riuscito a distinguere senza problemi i diversi tipi di traduzione, inserendo nello stesso *cluster* le traduzioni di DeepL e Google Translate e separando il *corpus* "Traduzioni Umane". In questa prima

analisi non è stato inserito il *corpus* in lingua russa perché la *cluster analysis* si basa su misurazioni di distanze intertestuali e ovviamente un testo scritto in una lingua diversa è molto distante dai testi scritti nella stessa lingua, perché le strutture linguistiche sono differenti. Per questo motivo non è stato ritenuto pertinente aggiungere il *corpus* "Originali Russi" in questa prima *cluster analysis*.

Successivamente, il primo dendrogramma prodotto è stato quello dei testi originali russi, in cui sono stati presi in considerazione tutti i *word-type* con una frequenza ≥ 2 per ottenere una copertura del vocabolario del 77,35%. Si è deciso di prendere in esame un ampio numero di *most frequent word* per ottenere la più ampia copertura di vocabolario possibile e permettere al programma *stylo* di capire non solo lo stile grammaticale degli articoli, ma di intuire anche i *topic* presenti nei testi e quindi trovare meglio i testi più simili da inserire nello stesso cluster. Dopo aver ottenuto la suddivisone in *cluster* del *corpus* "Originali russi", si è passati all'analisi dei *corpora* in italiano in cui sono stati presi in esame un numero *n* di MFW, che desse un tasso di copertura del vocabolario simile a quello dei testi originali russi. In tutti i *corpora* di traduzione sono stati presi in considerazione tutti i *word-type* con una frequenza ≥ 5 per ottenere un tasso di copertura del vocabolario pari al 78,22% per il *corpus* "Traduzioni Umane", 78,58% per il *corpus* "DeepL" e 78,51% per il *corpus* "Google Translate".

Dopo una prima osservazione dei diversi dendrogrammi prodotti da *stylo* (v. Fig. 20 p. 90; Fig. 21 p. 91; Fig. 22 p. 92; Fig. 23 p. 93) si è subito notato che la suddivisione in *cluster* di tutti i *corpora* di traduzione è diversa da quella del *corpus* "Originali russi". Successivamente sono stati confrontati separatamente gruppi di *cluster* e sono state osservate in modo più specifico similitudini e differenze tra i vari *corpora*. È stato notato che non sempre la traduzione automatica di DeepL o Google Translate mostrava la suddivisione in cluster più simile all'originale russo. In alcuni casi la traduzione umana rispettava maggiormente la suddivisione in *cluster* rispetto ai programmi di traduzione automatica. Questo fenomeno si può trovare nei *cluster* contenenti gli articoli "09_ЧерныйАвгуст"[37] e

[37] trasl. Černyi Avgust

“09_УспешныхРоссиян”[38] (v. Fig. 25, p. 95) e gli articoli con i codici “03_” e “02_” (v. Fig. 27 p. 97). In questi casi il *corpus* “Traduzioni Umane” mostra la suddivisione in un unico *cluster* più fedele all’originale rispetto agli altri *software* di traduzione. Una rappresentazione simile degli articoli con i codici “03_” e “02_” si trova anche nel *corpus* “DeepL”, sottolineando in questo modo la similitudine tra le traduzioni umana e quella di DeepL. Ulteriori analogie tra i *corpora* “Traduzioni umane” e “DeepL” sono state riscontrate anche nella suddivisione in *cluster* degli articoli con codici “09_”, “07_” e “08_” (v. Fig. 24 p. 94; Fig. 26 p. 96), ma le stesse similitudini non sono state trovate tra i *corpora* “Traduzioni umane” e “Google Translate”. Infine, è stato notato che nei seguenti casi i *corpora* di traduzione automatica avevano la suddivisione più simile al *corpus* “Originali russi”: il *cluster* contenente gli articoli con codice “08” nel *corpus* “DeepL” e il *cluster* contenente gli articoli con codice “07” nel *corpus* “Google Translate” (v. Fig. 25 p. 95; Fig. 26 p. 96).

In conclusione, si può affermare che non sempre gli stili grammaticali delle traduzioni automatiche sono più simili ai testi originali delle traduzioni umane. Inoltre, le similitudini tra la traduzione automatica di DeepL e le traduzioni umane sono state confermate anche dalla *cluster analysis* di *stylo*, confermando che il fatto che il *software* di traduzione DeepL produce traduzioni “più umane” di Google Translate.

5.6. Discussione dei risultati del metodo di classificazione basato sul *machine learning*: distinzione tra traduzione umana e automatica

L’ultima analisi che è stata effettuata aveva lo scopo di capire se un algoritmo di *machine learning* fosse in grado di distinguere una traduzione umana da una automatica e una traduzione automatica di DeepL da una di Google Translate. Gli algoritmi di *machine learning* utilizzati sono stati il *Support Vector Machine* (SVM) e il *Random Forest* (RF). Inoltre, per preparare adeguatamente i vari *corpora* alle

[38] trasl. Uspešnyi Rossijan

analisi di *machine learning*, è stato utilizzato l'algoritmo *Author's Multilevel N-Gram Profiles* (AMNP), che ha permesso di dividere i *corpora* in *chunk* da 200 parole e individuare i 200 *bigram*, *trigram*, *word* e *word bigram* più frequenti e svolgere ulteriori analisi descritte al paragrafo 3.5. Sono state effettuate 9 analisi in totale, in cui è stato richiesto agli algoritmi di riconoscere l'autore di una traduzione e capire se si trattasse di un essere umano, di DeepL o di Google Translate. Per poter allenare gli algoritmi a riconoscere la scrittura di un essere umano, da quella di DeepL e Google Translate, sono stati utilizzati tutti gli articoli contenuti nei vari *corpora* ed è stato scelto un articolo fuori dai *corpora* da far riconoscere agli algoritmi SVM e RF.

Le prime sei analisi sono state condotte prendendo separatamente coppie di *corpora* in italiano: ad esempio, prima è stata eseguita un'analisi considerando i *corpora* "Traduzioni umane" e "DeepL", poi i *corpora* "Traduzioni Umane" e "Google Translate". Solo nelle ultime tre analisi sono stati analizzati tutti i *corpora* insieme per far riconoscere agli algoritmi prima la traduzione umana, poi quella di DeepL e infine quella prodotta da Google Translate. Si è deciso di suddividere in questo modo l'analisi per vedere se gli algoritmi avrebbero riconosciuto con una maggiore accuratezza l'autore della traduzione quando avrebbero dovuto scegliere tra due variabili invece di tre.

Al termine delle analisi, purtroppo, è stato notato che né l'algoritmo SVM né RF sono riusciti a lavorare bene. Ad esempio, nel riconoscimento della traduzione umana tra i *corpora* "Traduzioni Umane" e "DeepL" (v. Tab. 10 p. 99), SVM ha registrato un'accuratezza del 74%, ma nonostante questo non è riuscito a riconoscere la traduzione umana. RF ha sempre registrato un'accuratezza troppo bassa per eseere utilizzabile. L'algoritmo SVM, invece, ha sempre avuto un'accuratezza maggiore del 74% in tutte le prime sei analisi, che può essere considerata accettabile in certi contesti ma, in fase di riconoscimento, non è comunque stato in grado di identificare sempre in modo corretto l'autore della traduzione.

Inoltre, anche l'accuratezza di SVM è diventata troppo bassa per essere utilizzabile quando gli algoritmi sono stati allenati con tutti i tre *corpora* in italiano (v. Tabb. 16-17-18 p. 102).

Dati i risultati caotici ottenuti e i bassi valori di accuratezza, non è stato possibile stabilire quale algoritmo funzioni meglio o se SVM e RF siano in grado di distinguere la traduzione umana da quella automatica. È probabile che la divisione in chunk di 200 *parole* e gli altri parametri presenti nell'algoritmo AMNP non siano adeguati ad analizzare i *corpora* in italiano di questi testi. Inoltre, è plausibile ritenere che la lunghezza limitata dei singoli testi e la presenza di soli 34 articoli per *corpus* non sia sufficiente per allenare algoritmi di *machine learning* (cfr. cap. 4.5). Per questi motivi i risultati della presente analisi non sono stati considerati utilizzabili per la ricerca. Infine, è necessario puntualizzare che il fallimento di questi esperimenti non dimostra che una macchina non sia in grado di distinguere lo stile di una traduzione umana da una automatica. Come si è potuto notare nella precedente analisi, *stylo* è riuscito a distinguere correttamente lo stile delle traduzioni umane da quelle automatiche. In questo caso, *stylo* ha analizzato i *corpora* nel loro insieme e non articolo per articolo, come negli algoritmi di *machine learning*. Inoltre, gli studi di Li et al. (2015) e Fu et al. (2021) hanno già provato che il *machine learning* è in grado di distinguere la *human translation* dalla *machine translation*, ed è quindi probabile che con l'utilizzo di *corpora* più ampi si possano ottenere risultati migliori sia da SVM che RF.

Dopo aver terminato la rassegna dei risultati delle singole analisi, si può affermare che i metodi quantitativi di analisi dei testi sono utili anche negli studi di traduttologia e nel confronto tra *machine translation* e *human translation*. Al termine delle indagini condotte, è stato notato come al giorno d'oggi la traduzione automatica sia simile alla traduzione umana. Di fronte a questo fatto, un traduttore può utilizzare la precisione di un *software* di traduzione automatica per poter tradurre più velocemente dei testi e non vedere la *machine translation* come un concorrente. Forse, nei prossimi anni, la traduzione automatica neurale raggiungerà i livelli di un essere umano nella traduzione di alcune tipologie testuali (cfr. Popel et al. 2020), ma si ritiene che ci sarà sempre bisogno del controllo finale di un essere umano per evitare di pubblicare testi ambigui o non scorrevoli nella lettura. Al contrario, se un utente ha bisogno di una traduzione veloce di un testo in lingua

straniera, allora la traduzione automatica è sicuramente la soluzione più veloce e meno dispendiosa (cfr. Li et al. 2014).

Inoltre, questa tesi ha dimostrato le differenze tra due *software* di traduzione automatiche diversi. DeepL e Google Translate non realizzano traduzioni della stessa qualità, ma con l'analisi del contenuto e la *cluster* analysis è stato dimostrato che le traduzioni di DeepL si avvicinano maggiormente alla traduzione umana. Tuttavia, nell'analisi in *AntConc*, si è notato che in alcuni casi Google Translate traduce in modo più preciso determinati termini (es. forestierismi) rispetto a DeepL. Sembra, quindi, che il *software* DeepL sia molto accurato da un punto di vista stilistico, ma lessicalmente necessita di qualche miglioramento (v. "waps" invece di "WhatsApp").

Si precisa che tutte le considerazioni si limitano alla coppia linguistica russo-italiano, ed è possibile che in altri casi (es. inglese-italiano), i *software* analizzati abbaino una resa migliore o peggiore. In vista di eventuali ricerche future si consiglia la collezione di *corpora* con altre coppie linguistiche e l'analisi di altre tipologie testuali.

5.7. L'analisi automatica dei dati testuali in possibili ricerche future sul *machine translation*

L'analisi presentata in questa tesi è poco comune tra le ricerche che si occupano di analisi delle differenze tra traduzione umana e automatica. In letteratura, infatti, i maggiori studi riguardanti il confronto tra *human translation* e *machine translation* si basano soprattutto su un sistema di valutazione effettuato da *raters* scelti in precedenza (cfr. Wu et al., 2016; Läubli et al. 2018; Popel et al. 2020; Takakusagi et al. 2021), a volte integrato da una modifica dell'algoritmo del *software* di traduzione automatica per migliorarne la *performance* (cfr. Wu et al., 2016; Popel et al. 2020). Spesso questi studi vengono presentati in occasione del WMT (Workshop on Machine Translation), una conferenza tenuta annualmente in cui si discute di traduzione automatica e ricerche in tale ambito, con lo scopo di migliorare la *machine translation*, studiandone i progressi e i limiti. Il WMT organizza anche competizioni, chiamate *shared tasks*, in cui i promotori dell'evento forniscono ai concorrenti una serie di *dataset* e istruzioni da seguire. I partecipanti

hanno il compito di risolvere una serie di *task* di traduzione attraverso l'uso di un *software* di traduzione e inviarne i risultati agli organizzatori del WMT. Successivamente le traduzioni proposte dai partecipanti vengono giudicate da esseri umani prima dell'inizio della conferenza. Infine, i vari concorrenti presentano i risultati dei propri lavori durante il *Workshop on Machine Translation* e vengono annunciati i vincitori della competizione (cfr. https://machinetranslate.org/wmt). In base alle ricerche effettuate sul *Workshop on Machine Translation* e ai maggiori studi sul *machine translation* legati ai *dataset* forniti dal WMT (cfr. Wu et al., 2016; Läubli, et al., 2018; Hassan et al., 2018; Popel et al., 2020; Fu & Nederhof, 2021) non sono state trovate ricerche sullo studio delle differenze tra traduzione umana e automatica attraverso metodi di analisi quantitativa dei testi. In base ai risultati ottenuti in studi precedenti (cfr. Lee 2019; Cembrzyńska et al. 2021; Żak et al. 2021) e nella presente tesi, è stato provato che l'uso di metodi di analisi quantitativa dei testi è utile a capire le differenze tra *machine* e *human translation*. L'uso di questi metodi può fornire un punto di vista diverso all'interno di studi futuri su altre coppie linguistiche o altre tipologie testuali. Si ritiene, infatti, che studi di questo tipo possano essere utili agli sviluppatori di *software* di traduzione automatica per capirne i limiti in modo più immediato e senza l'utilizzo di *rater* umani. Inoltre, questi studi possono essere utili per i traduttori professionisti per capire quali programmi di traduzione automatica sono più attendibili e per quali tipologie di testo possono essere utilizzati come aiuto nella traduzione.

In conclusione, si considerano utili maggiori studi sulla *machine translation* e *human translation* attraverso metodi di analisi quantitativa dei testi. In base ai risultati ottenuti in questa ricerca, si consiglia l'uso di *corpora* dall'estensione maggiore per ottenere risultati ancora più affidabili.

6. Lista degli articoli inseriti nel *corpus*

6.1. Articoli originali scritti in russo utilizzati nel *corpus*

Козлов, В. (13/12/2021), *Протесты, туристы, устойчивость и сообщества — четыре слова для понимания 2021 года*. Главная: Мнения. In: Эксперт. https://expertsouth.ru/comments/protesty-turisty-ustoychivost-i-soobshchestva-chetyre-slova-dlya-ponimaniya-2021-goda/ (ultima consultazione: 26/08/2022)

Пискунова, Н. (08/12/2021), *И Тик, И Ток, и все не так*. In: Газета.Ру. https://www.gazeta.ru/comments/column/n_piskunova/14286367.shtml (ultima consultazione: 26/08/2022)

Котляр, М. (24/10/2021), *Борьба с курением в неположенном месте. Нарушителей найдут и оштрафуют*. Общество. In: Газета.Ру. https://www.gazeta.ru/social/2021/10/24/14127967.shtml (ultima consultazione: 26/08/2022)

Autore sconosciuto (08/10/2021), *«Нобелевка раздора»: как на присуждение премии Муратову отреагировали в России*. Новости. In: The Insider. https://theins.ru/news/245577 (ultima consultazione: 26/08/2022)

Зарубина, А. (08/04/2021), *Андрей Сахаров — выдающийся гуманист 20-го века*. In: Дилетант. https://diletant.media/articles/45309691/ (ultima consultazione: 26/08/2022)

Пащенко, Д. (07/10/2021), *Самые громкие убийства журналистов в современной России*. In: Дилетант. https://diletant.media/articles/37700782/ (ultima consultazione: 26/08/2022)

Галеев, А. (20/08/2021), *«Русские хакеры» действительно существуют. Откуда они появились, как стали звездами и почему их боится Запад?*. Интернет и СМИ: Киберпреступность. In: Lenta.ru. https://lenta.ru/articles/2021/08/20/russians_are_coming/ (ultima consultazione: 26/08/2022)

Серебряный, И. (29/06/2021), *Социально успешных россиян накрыла волна кризиса ожиданий*. Здоровье. In: Эксперт. https://expert.ru/2021/07/29/sotsialno-uspeshnykh-rossiyan-nakryla-volna-krizisa-ozhidaniy/ (ultima consultazione: 26/08/2022)

Успенский, А. (24/06/2021), *Кольца безумной ярости*. Комментарий: Спорт. In: Новая Газета. https://novayagazeta.ru/articles/2021/07/24/koltsa-bezumnoi-iarosti (ultima consultazione: 26/08/2022)

Серебряный, И. (28/05/2021), *В России запретят пластик. Но не весь*, Право: Экология. In: Эксперт. https://expert.ru/2021/05/28/v-rossii-zapretyat-plastik-no-ne-ves/ (ultima consultazione: 26/08/2022)

Autore sconosciuto (24/05/2021), *Вынужденную посадку рейса Ryanair в Минске обсудят на саммите ЕС. Авиакомпании отказываются летать над страной*. Новости. In: Медуза. https://meduza.io/feature/2021/05/24/vynuzhdennuyu-posadku-reysa-ryanair-v-minske-obsudyat-na-sammite-es-aviakompanii-otkazyvayutsya-letat-nad-belorussiey (ultima consultazione: 26/08/2022)

Васильчук, Т. & Докшин, В. (22/05/2021), *От Тувы до самых до окраин*. Репортажи: Общество. In: Новая Газета. https://novayagazeta.ru/articles/2021/05/22/ot-tuvy-do-samykh-do-okrain (ultima consultazione: 26/08/2022)

Логинова, К. (24/04/2021), *Украинская Wonder Woman: Киев отправил нового посла в США*. Мир. In: Известия. https://iz.ru/1156095/kseniia-loginova/ukrainskaia-wonder-woman-kiev-otpravil-novogo-posla-v-ssha (ultima consultazione: 26/08/2022)

Autore sconosciuto (21/04/2021), *Самое интересное из послания Владимира Путина*. Политика. In: Ведомости. https://www.vedomosti.ru/politics/articles/2021/04/21/866993-samoe-interesnoe-iz-poslaniya-vladimira-putina (ultima consultazione: 26/08/2022)

Акимов, И. (19/04/2021), *«Даже «для галочки» не попытались согласовать»*. Общество. In: Газета.Ру. https://www.gazeta.ru/social/2021/04/19/13564808.shtml (ultima consultazione: 26/08/2022)

Латынина, Ю. (21/04/2021), *«Это казнь без приговора»*. Комментарий: Политика. In: Новая Газета.

https://novayagazeta.ru/articles/2021/04/18/eto-kazn-bez-prigovora (ultima consultazione: 26/08/2022)

Тумакова, И. (31/03/2021), *Принеси, подай, дай денег — не мешай*. Репортажи: Общество. In: Новая Газета. https://novayagazeta.ru/articles/2021/03/31/prinesi-podai-dai-deneg-ne-meshai (ultima consultazione: 26/08/2022)

Матвеева, А. (15/03/2021), *Что если столица переедет в Екатеринбург? Отвечают екатеринбуржцы*. Город: Люди. In: Москвич. https://moskvichmag.ru/gorod/chto-esli-stolitsa-pereedet-v-ekaterinburg-otvechayut-ekaterinburzhtsy/ (ultima consultazione: 26/08/2022)

Федоренко, В. (14/03/2021), *Крымский полуостров. В Приморском крае есть свой Крым. Сегодня этот некогда цветущий поселок мертв, но не совсем*. Репортажи: Общество. In: Новая Газета. https://novayagazeta.ru/articles/2021/03/14/krymskii-poluostov (ultima consultazione: 26/08/2022)

Максимушкина, Е. (24/02/2021), *Wildberries вышла на рынки Франции, Италии и Испании*. Мнения: Темы недели. In: Ведомости. https://www.vedomosti.ru/opinion/articles/2021/02/24/859165-wildberries-vishla (ultima consultazione: 26/08/2022)

Autore sconosciuto (02/02/2021), *Суд отправил Навального в колонию. Главное*. Новости. In: Медуза. https://meduza.io/feature/2021/02/02/sud-otpravil-navalnogo-v-koloniyu-glavnoe (ultima consultazione: 26/08/2022)

Панов, А. (28/01/2021), *Путин и Байден поговорили.* Комментарий: Политика. In: Новая Газета. https://novayagazeta.ru/articles/2021/01/28/88921-putin-i-bayden-pogovorili (ultima consultazione: 26/08/2022)

Эггерт, К. (18/11/2020), *Комментарий: Время Москвы на Южном Кавказе вышло. Настало время Анкары*. Мнения. In: DW. https://learngerman.dw.com/ru/kommentarij-vremja-moskvy-na-juzhnom-kavkaze-vyshlo-nastalo-vremja-ankary/a-55652196 (ultima consultazione: 26/08/2022)

Забродин, А. (12/11/2020), *Бунт неосмысленный: как Ереван переживает окончание войны в Карабахе*. In: Известия. https://iz.ru/1085944/aleksei-zabrodin/bunt-neosmyslennyi-kak-erevan-perezhivaet-okonchanie-voiny-v-karabakhe (ultima consultazione: 26/08/2022)

Кармазин, И., Байкова, Т., Байназаров, Э. (06/10/2020), *Октябрьская революция: экс-президента Киргизии вынесли из СИЗО*. In Известия. https://iz.ru/1069960/igor-karmazin-tatiana-baikova-elnar-bainazarov/oktiabrskaia-revoliutciia-eks-prezidenta-kirgizii-vynesli-iz-sizo (ultima consultazione: 26/08/2022)

Autore sconosciuto. (24/09/2020), *Путин выдвинут на Нобелевскую премию мира*. In: РИА Новости. https://ria.ru/20200924/premiya-1577710461.html (ultima consultazione: 26/08/2022)

Сивцова, А. (09/09/2020), *История одной фотографии: минские женщины, которых собираются задерживать силовики*. Истории. In: Медуза. https://meduza.io/feature/2020/09/09/istoriya-odnoy-fotografii-zhenschiny-kotoryh-sobirayutsya-zaderzhivat (ultima consultazione: 26/08/2022)

Трегубова, Е. (31/08/2020), *«Черный» август позади? Что будет с рублем, долларом и евро в сентябре*. In: Аргументы и Факты https://aif.ru/money/economy/chernyy_avgust_pozadi_chto_budet_s_rublem_dollarom_i_evro_v_sentyabre (ultima consultazione: 26/08/2022)

Мельничук, С. (21/08/2020), *Нелюбовный треугольник. Где место России в холодной войне Китая и США*. In: РИАНовости. https://ria.ru/20200821/rossiyakitay-1576057823.html (ultima consultazione: 26/08/2022)

Вишневский, Б. (17/08/2020), *Уроки белорусского. О чем думают россияне, глядя на события в Минске*. Колонка: Политика. In: Новая Газета. https://novayagazeta.ru/articles/2020/08/17/86712-uroki-belorusskogo (ultima consultazione: 26/08/2022)

Орлов, Д. (19/07/2020), *«Добрый» диктатор. Хрущёв искренне пытался улучшить жизнь народа. Но получилось… как всегда?*. In: Дилетант. https://diletant.media/articles/45290845/ (ultima consultazione: 26/08/2022)

Закурдаев, М. (22/06/2020), *Не сейчас же: на электронные визы призвали ввести мораторий.* Туризм. In: Правда.Ру. https://www.pravda.ru/travel/1505829-opasnost_elektronnih_viz/ (ultima consultazione: 26/08/2022)

Перцев, А. (07/06/2020), *Кремль собирается праздновать День России всю неделю.* Истории. In: Медуза. https://meduza.io/feature/2020/06/07/administratsiya-prezidenta-gotovitsya-prazdnovat-den-rossii-vsyu-nedelyu-budut-pirogi-stikery-s-medvedem-v-vatnike-i-dazhe-chellendzh-v-tiktoke-s-bezrukovym (ultima consultazione: 26/08/2022)

Кузнецов, А. (30/03/2020), *"Нас бьют, потому что мы не защищаемся". Тренер российских велогонщиков об иске к WADA*. Интервью. In: ТАСС. https://tass.ru/interviews/8108993?utm_source=google.com&utm_medium=organic&utm_campaign=google.com&utm_referrer=google.com (ultima consultazione: 26/08/2022)

6.2. Traduzioni umane in italiano utilizzate nel *corpus*

Groppi, E. (30/01/2022), *Proteste, turisti, sostenibilità e comunità: quattro parole per comprendere il 2021*. Politica: Economia. In: Russia in Translation. http://russiaintranslation.com/2022/01/30/proteste-turisti-sostenibilita-e-comunita-quattro-parole-per-comprendere-il-2021/ (ultima consultazione: 26/08/2022)

Groppi, E. (16/01/2022), *E Tik, e Tok, e qualcosa non va*. Società: Tecnologia. In: Russia in Translation. http://russiaintranslation.com/2022/01/16/e-tik-e-tok-e-qualcosa-non-va/ (ultima consultazione: 26/08/2022)

Gentile, C. (29/11/2021), *I più clamorosi omicidi di giornalisti nella Russia contemporanea*. Società. In: Russia in Translation. http://russiaintranslation.com/2021/11/29/i-piu-clamorosi-omicidi-di-giornalisti-nella-russia-contemporanea/ (ultima consultazione: 26/08/2022)

Groppi, E. (29/11/2021), *La lotta contro il fumo in luoghi non adibiti. I trasgressori vengono rintracciati e multati*. Società: Salute. In: Russia in

Translation. http://russiaintranslation.com/2021/11/29/la-lotta-contro-il-fumo-in-luoghi-non-adibiti-russia/ (ultima consultazione: 26/08/2022)

Angelone, E. (05/11/2021), *Abcasia – quale (in)dipendenza?*. Società: Economia. In Russia in Translation. http://russiaintranslation.com/2021/11/05/abcasia-quale-indipendenza/ (ultima consultazione: 26/08/2022)

Groppi, E. (05/11/2021), *I russi di successo nella società sono travolti da una crisi delle aspettative*. Società: Economia. In: Russia in Translation. http://russiaintranslation.com/2021/11/05/i-russi-di-successo-nella-societa-sono-travolti-da-una-crisi-delle-aspettative/ (ultima consultazione: 26/08/2022)

Oberto, M. (19/10/2021), *«Il Nobel della discordia»: come hanno reagito in Russia all'assegnazione del premio a Muratov.* Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/10/19/il-nobel-della-discordia/ (ultima consultazione: 26/08/2022)

Ticozzi, P. (06/10/2021), *Gli anelli di una rabbia folle*. Società. In: Russia in Translation. http://russiaintranslation.com/2021/10/06/gli-anelli-di-una-rabbia-folle/ (ultima consultazione: 26/08/2022)

Ghitti, B. (06/10/2021), *Gli "hacker russi" esistono davvero*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/10/06/gli-hacker-russi-esistono-davvero/ (ultima consultazione: 26/08/2022)

Gentile, C. (10/07/2021), *Il dittatore "buono"*. Politica: Storia. In: Russia in Translation. http://russiaintranslation.com/2021/07/10/il-dittatore-buono/ (ultima consultazione: 26/08/2022)

Angelone, E. (29/06/2021), *Da Tuva alla periferia*. Politica: Società. In: Russia in Translation. http://russiaintranslation.com/2021/06/29/da-tuva-alla-periferia/ (ultima consultazione: 26/08/2022)

Groppi, E. (20/06/2021), *In Russia vieteranno la plastica. Ma non tutta*. Società: Natura. In: Russia in Translation. http://russiaintranslation.com/2021/06/20/in-russia-vieteranno-la-plastica-ma-non-tutta/ (ultima consultazione: 26/08/2022)

Autore sconosciuto (20/06/2021), *La penisola di Crimea*. Società: Economia. In: Russia in Translation. http://russiaintranslation.com/2021/06/20/la-penisola-di-crimea/ (ultima consultazione: 26/08/2022)

Cori, G. (24/05/2021), *Arresto di Roman Protasevič: cosa sappiamo*. Politica. In: Russia in Translation. http://russiaintranslation.com/2021/05/24/arresto-di-roman-protasevic-cosa-sappiamo/ (ultima consultazione: 26/08/2022)

Gentile, C. (15/05/2021), *Andrej Sacharov, eminente umanista del ventesimo secolo*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/05/15/andrej-sacharov-eminente-umanista-del-ventesimo-secolo/ (ultima consultazione: 26/08/2022)

Cannarella, L. (09/05/2021), *Le cose più interessanti del discorso di Vladimir Putin*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/05/09/le-cose-piu-interessanti-del-discorso-di-vladimir-putin/ (ultima consultazione: 26/08/2022)

Vaccaro, E. (03/05/2021), *La Wonder Woman ucraina: Kiev ha inviato un nuovo ambasciatore negli Stati Uniti*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/05/03/la-wonder-woman-ucraina-kiev-ha-inviato-un-nuovo-ambasciatore-negli-stati-uniti/ (ultima consultazione: 26/08/2022)

Vaccaro, E. (21/04/2021), *«Non hanno cercato di trovare un accordo neanche "per mostra"»*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/04/21/non-hanno-cercato-di-trovare-un-accordo-neanche-per-mostra/ (ultima consultazione: 26/08/2022)

Gentile, C. (21/04/2021), *"Questa è un'esecuzione senza condanna"*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/04/21/questa-e-unesecuzione-senza-condanna/ (ultima consultazione: 26/08/2022)

Savelli, E. (05/04/2021), *WildBerries entra nei mercati di Francia, Italia e Spagna*. Società: Economia. In: Russia in Translation. http://russiaintranslation.com/2021/04/05/wildberries-entra-nei-mercati-di-francia-italia-e-spagna/ (ultima consultazione: 26/08/2022)

Conti, G. (05/04/2021), *E se la capitale venisse spostata a Ekaterinburg? Ecco le opinioni di chi vive negli Urali*. Politica: Società. In Russia in Translation. http://russiaintranslation.com/2021/04/05/e-se-la-capitale-venisse-spostata-a-ekaterinburg-ecco-le-opinioni-di-chi-vive-negli-urali/ (ultima consultazione: 26/08/2022)

Savelli, E. (16/02/2021), *Putin e Biden si sono parlati*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2021/02/16/putin-e-biden-si-sono-parlati/ (ultima consultazione: 26/08/2022)

Cannarella, L. (16/02/2021), *Il tribunale manda Naval'ny in colonia. L'essenziale*. Politica. In: Russia in Translation. http://russiaintranslation.com/2021/02/16/il-tribunale-manda-navalny-in-colonia/ (ultima consultazione: 26/08/2022)

Vaccaro, E. (21/11/2020), *Opinione: il tempo di Mosca nel Caucaso meridionale è scaduto. Ora tocca ad Ankara*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2020/11/21/opinione-il-tempo-di-moscanel-caucaso-e-finito-ora-tocca-ad-ankara/ (ultima consultazione: 26/08/2022)

Lazzaro, B. (21/11/2020), *Storia di una fotografia: le donne di Minsk che stanno per essere arrestate dalle forze dell'ordine*. Politica: Paesi Ex Urss. In: Russia in Translation. http://russiaintranslation.com/2020/11/21/storia-di-una-fotografia-le-donne-di-minsk-che-stanno-per-essere-arrestate-dalle-forze-dellordine/ (ultima consultazione: 26/08/2022)

Spigno, A. (13/11/2020), *Una rivolta incosciente: come Yerevan sta attraversando la fine della guerra nel Karabach*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2020/11/13/una-rivolta-incosciente-come-yerevan-sta-attraversando-la-fine-della-guerra-nel-karabach/ (ultima consultazione: 26/08/2022)

Giannotti, F. (07/10/2020), *Putin candidato al premio Nobel per la pace*. Politica. In: Russia in Translation. http://russiaintranslation.com/2020/10/07/putin-

candidato-al-premio-nobel-per-la-pace/ (ultima consultazione[39]: 30/06/2022)

Vaccaro, E. (07/10/2020), *La rivoluzione d'Ottobre: liberato l'ex presidente del Kirghizistan*. Politica: Paesi Ex Urss. In: Russia in Translation. http://russiaintranslation.com/2020/10/07/la-rivoluzione-dottobre-liberato-lex-presidente-del-kirghizistan/ (ultima consultazione: 26/08/2022)

Giannotti, F. (26/09/2020), *Non è ancora il momento: richiesta la moratoria sui visti elettronici*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2020/09/26/non-e-ancora-il-momento-richiesta-la-moratoria-sui-visti-elettronici/ (ultima consultazione: 26/08/2022)

Savelli, E. (24/09/2020), *L' Agosto "nero" è alle spalle? Cosa succederà al Rublo, al Dollaro e l'Euro a Settembre*. Società: Economia. In: Russia in Translation. http://russiaintranslation.com/2020/09/24/l-agosto-nero-e-alle-spalle-cosa-succedera-al-rublo-al-dollaro-e-leuro-a-settembre/ (ultima consultazione: 26/08/2022)

Lazzari, A. (19/09/2020), *Triangolo non amoroso: il ruolo della Russia nella guerra fredda tra USA e Cina*. Politica: Relazioni Internazionali. In: Russia in Translation. http://russiaintranslation.com/2020/09/19/triangolo-non-amoroso-il-ruolo-della-russia-nella-guerra-fredda-tra-usa-e-cina/ (ultima consultazione: 26/08/2022)

Ticozzi, P. (13/09/2020), *Lezioni bielorusse. Cosa pensano i russi degli avvenimenti di Minsk*. Politica. In: Russia in Translation. http://russiaintranslation.com/2020/09/13/lezioni-bielorusse/ (ultima consultazione: 26/08/2022)

Lazzaro, B. (21/06/2020), *Il Cremlino si appresta a festeggiare il Giorno della Russia per tutta la settimana.* Società. In: Russia in Translation. http://russiaintranslation.com/2020/06/21/il-cremlino-si-appresta-a-

[39] In data 30/06/2022 è stato scoperto che questo articolo è stato cancellato, probabilmente a seguito delle vicende che coinvolgono Russia e Ucraina. In data 09/03/2022 l'articolo era ancora visibile nel link soprariportato e fa parte del *corpus* di articoli analizzati.

festeggiare-il-giorno-della-russia-per-tutta-la-settimana-ci-saranno-pirogi-adesivi-dellorso-in-giubba-militare-e-anche-una-tiktok-challenge-con-bezrukov/ (ultima consultazione: 26/08/2022)

Lazzari, A. (01/05/2020), *"Ci colpiscono perché non ci difendiamo". L'allenatore dei ciclisti russi sulla causa contro la WADA*. Politica. In: Russia in Translation. http://russiaintranslation.com/2020/05/01/ci-colpiscono-perche-non-ci-difendiamo-lallenatore-dei-ciclisti-russi-sulla-causa-contro-la-wada/ (ultima consultazione: 26/08/2022)

6.3. Articoli extra usati nei test di riconoscimento di traduzione umane e automatiche

Autore sconosciuto (14/10/2015 e 16/06/2022), *Где отдыхал Советский Союз*. In: Дилетант. https://diletant.media/articles/26251007/ (ultima consultazione: 26/08/2022)

Gentile, C. (19/10/2021), *Dove riposava il popolo sovietico*. Società. In: Russia in Translation. http://russiaintranslation.com/2021/10/19/dove-riposava-il-popolo-sovietico/ (ultima consultazione: 26/08/2022)

6.4. Informazioni sulle singole testate giornalistiche contenenti gli articoli in russo: fonti

The Insider: https://theins.ru/about; https://theins.ru/en/about (ultima consultazione: 17/06/2022)

Novaja Gazeta: https://novayagazeta.ru/articles/2022/03/28/my-priostanavlivaem-rabotu; https://novayagazeta.ru/ (ultima consultazione: 17/06/2022)

Lenta.ru: https://lenta.ru/info/ (ultima consultazione: 17/06/2022)

Expert: https://expert.ru/about/ (ultima consultazione: 21/06/2022)

Meduza: https://meduza.io/en/pages/about (ultima consultazione: 21/06/2022)

RIA Novosti: https://ria.ru/ (ultima consultazione: 21/06/2022)

Gazeta.ru: https://www.gazeta.ru/about/?updated (ultima consultazione: 21/06/2022)

Izvestija: https://iz.ru/o-kompanii (ultima consultazione: 21/06/2022)

Vedomosti: https://www.vedomosti.ru/info/mission (ultima consultazione: 21/06/2022)

Moskvič: https://moskvichmag.ru/redaktsiya/ (ultima consultazione: 21/06/2022)

Diletant: https://diletant.media/ (ultima consultazione: 21/06/2022)

Argumenti i Facti: https://aif.ru/ (ultima consultazione: 21/06/2022)

Deutsche Welle (DW): https://www.dw.com/en/about-dw/s-30688 (ultima consultazione: 21/06/2022)

TASS: https://tass.ru/tass-today (ultima consultazione: 29/06/2022)

Pravda.ru: https://www.pravda.ru/missioncoverageprioritiespolicy.html; https://www.pravda.ru/about.html (ultima consultazione: 29/06/2022)

7. Riferimenti bibliografici

Ali, S. M., Hussein K. S. (2014), *The Comparative Power of Type/Token and Hapax legomena/Type Ratios: A* Corpus-*based Study of Authorial*, in: "Advances in Language and Literary Studies. Vol. 5 No. 3", Australian International Academic Centre, Footscray, pp. 112-119.

Anthony, Laurence (2019). AntConc (Version 3.5.8) [Computer Software], Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015), *Neural machine translation by jointly learning to align and translate*. Paper presented at 3rd International Conference on Learning Representations, ICLR 2015, San Diego, United States.

Baker, P., Hardie, A., & McEnery, T. (2006), *A Glossary of Corpus Linguistics*. In: "A Glossary of Corpus Linguistics", Edinburgh University, Edimburgo, pp. 7-174.

Barbera, M., Corino, E., & Onesti, C. (2007). *Cosa è un* corpus*? Per una definizione più rigorosa di corpus, token, markup*, in: M. Barbera, E. Corino, & C. Onesti (Eds.), *Corpora e linguistica in rete*, Guerra, Perugia, pp. 25-88.

Berruto, G. (1987), *Sociolinguistica dell'italiano contemporaneo*, La Nuova Italia Scientifica, Roma.

Bertazzoli R. (2015), *La traduzione: teorie e metodi*. Carocci editore Bussole, Roma.

Bolasco, S. (1999), *Analisi multidimensionale dei dati*, Carocci Editore, Roma.

Cembrzyńska, A., Rybicki, J., & Świątek, J. (2021), *Trurl's Electronic Bard and DeepL: a stylometric analysis of the works by Stanisław Lem in human and machine translation* (master thesis), Jagiellonian University, Cracovia.

Cevese, C., Dobrovolskaja, J., Magnanini, E. (2000), *Grammatica russa. Morfologia: teoria ed esercizi*, Hoepli, Milano.

Chang, S.L. (2019), *Stylometric Comparative Analysis of Style in Human vs. Machine Literary Translations*, in: "The Journal of Translation Studies, 20(2)", The Korean Association for Translation Studies, pp. 111-130.

Costa-jussa M.R., Farrús M., Marino J., & Fonollosa J. A. R. (2012), *Study and comparison of rule-based and statistical catalan-spanish machine translation systems*, in: "Computing and Informatics. 31 (2)", pp. 245-270.

De Mauro, T. (1995), *Quantità-qualità: un binomio indispensabile per comprendere il linguaggio*, in: Bolasco S., Cipriani R., *Ricerca qualitativa e computer: teorie metodi e applicazioni*, Franco Angeli, Milano, pp. 21-30.

De Mauro, T., Chiari, I. (2005, a cura di), *Parole e Numeri. Analisi quantitative dei fatti di lingua*, Aracne, Roma.

Dmitrieva, A. & Tiedemann, J. (2021), *Creating an Aligned Russian Text Simplification Dataset from Language Learner Data*, in: B. Babych [et al.] (eds.), *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, Kiev, pp. 73-79.

Dolamic, L. & Savoy, J. (2010), *When Stopword Lists Make the Difference*. In: "Journal of the American Society for Information Science and Technology. 61", pp. 200-203.

Eder, M., Rybicki, J., Kestemont, M. (2016), *Stylometry with R: a package for Computational Text Analysis*, in: "The R Journal 8 (1)", pp. 107-121.

Fu, Y. & Mark-Jan Nederhof, M. (2021), *Automatic Classification of Human Translation and Machine Translation: A Study from the Perspective of Lexical* Diversity, in: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, Association for Computational Linguistics, online, pp. 91–99.

Fusco, F. (2016). *Che cos'e l'interlinguistica*. Carocci, Roma.

Gatti, F.M.T. & Tuzzi, A. (2020), *Can a troubled political era be detected by machine learning methods? An application on the End of Year speeches of the Italian Presidents*, in: "JADT 2020: 15th Journées internationales d'Analyse statistique des Données Textuelles".

Gellerstam, M. (1986), *Translationese in Swedish novels translated from English*, in: L. Wollin & H. Lindquist (Eds.) *Translation Studies in Scandinavia*: *Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, CWK Gleerup, Lund, pp. 88-95.

Giuliano, L. & La Rocca, G. (2008), *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*, Edizioni Universitarie di Lettere Economia Diritto, Milano.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W.D., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., & Zhou, M. (2018), *Achieving Human Parity on Automatic Chinese to English News Translation*. In: *ArXiv, abs/1803.05567*.

Hutchins, J. (2001), *Machine Translation and Human Translation: In Competition or in Complementation?* in: "International Journal of Translation, 13 (1-2)", pp. 5-20.

Kelih, E. (2010), *The type-token relationship in Slavic parallel texts*, in: "Glottometrics, 20", pp. 1-11.

Kovalev, V. (2014), *Il Kovalev. Dizionario russo-italiano, italiano-russo*, Zanichelli, Bologna.

Kunilovskaya, M., Morgoun, N., & Pariy, A. (2018), *Learner vs. professional translations into Russian: Lexical profiles*, in: D. Behr, & M. Sha (Eds.) *Translation & Interpreting. 10* (1), pp. 33-52.

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", Association for Computational Linguistics, Bruxelles, pp. 4791-4796.

Li, H., Graesser, A.C., & Cai, Z. (2014), *Comparison of Google Translation with Human Translation*, in: "FLAIRS Conference", pp. 190 – 195.

Li, Y., Wang, R., & Zhao, H. (2015), *A Machine Learning Method to Distinguish Machine Translation from Human Translation*, in: H. Zhao (Ed.), *29th Pacific Asia Conference on Language Information and Computation. Proceedings of PACLIC 2015: Poster papers, Shangai, 30 October – 1 November 2015. 29*, pp. 354-360.

Lopez, A. (2008), *Statistical machine translation*, in: *ACM Comput. Surv. 40, 3, Article 8 (August 2008)*,

Mahesh, B. (2019), Machine Learning Algorithms - A Review. In *International Journal of Science and Research. Volume 9 Issue 1*, Association for Computing Machinery, New York, pp. 381- 386.

Mikros, Georgios K.; Perifanos, Kostas (2013), *Authorship attribution in Greek tweets using multilevel author's n-gram profiles*, in: Hovy E.; Markman V.; Martell C. H. and Uthus D. (eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013*, Stanford, California. Palo Alto, California, pp. 17-23.

Mikros, Georgios K.; Perifanos, Kostas (2015), *Gender Identification in Modern Greek Tweets*, in: Tuzzi A.; Benešová M. & Macutek J. (Eds.), *Recent Contributions to Quantitative Linguistics (Vol. 70)*, De Gruyter, Berlino, pp. 75-88.

Ondelli, S., & Vitale, M. (2010), *L'assetto dell'italiano delle traduzioni in un corpus giornalistico. Aspetti qualitativi e quantitativi*, in: "Rivista internazionale di tecnica della traduzione = International Journal of Translation, 12", pp. 1-62.

Ondelli S. (2018), *Treat Texts as Data but Remember They Are Made of Words: Compiling and Preprocessing Corpora*. In: Tuzzi A. (ed), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, Springer Nature Switzerland, pp. 133-150.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020), *Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals*, in: "Nature Communications, 11(1)".

R Core Team (2020), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ratinaud, P. (2009), Iramuteq: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. http://www.iramuteq.org/

Reinert, M. (1983), *Une methode de classification descendante hierarchique: Application a l'analyse lexicale par context*, in: *Les Cahiers de l'Analyse des Données, 8(2)*, pp. 187-198.

Römer, U., & Wulff, S. (2010), *Applying corpus methods to written academic texts: Explorations of MICUSP*. In: "Journal of Writing Research, 2(2)", pp. 99–127.

Rubino, R., Lapshinova-Koltunski, E., & van Genabith, J. (2016), *Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification*, in: "Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", Association for Computational Linguistics, San Diego, pp. 960-970.

Šarac, A. (2019), *Il linguaggio giornalistico: la frequenza d'uso dei prestiti e neologismi* (master's thesis), Università di Spalato, Spalato.

Sbalchiero, S. (2018), *Topic Detection: A Statistical Model and a Quali-Quantitative Method*, in: Tuzzi A. (ed.), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, Springer Nature Switzerland, pp. 189-210.

Sinclair, Stéfan and Geoffrey Rockwell (2016), Voyant Tools. Web. http://voyant-tools.org/.

Takakusagi, Y., Oike, T., Shirai, K., Sato, H., Kano, K., Shima, S., Tsuchida, K., Mizoguchi, N., Serizawa, I., Yoshida, D., Kamada, T., & Katoh, H. (2021), *Validation of the Reliability of Machine Translation for a Medical Article from Japanese to English Using DeepL Translator*, "Cureus, 13(9): e17778", online.

Torresin, L. (2022), *Tradurre dal russo: teoria e pratica per studenti italofoni*, Hoepli, Milano.

Tuzzi, A. (2003), *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*, Carocci, Roma.

Tuzzi, A. (2010), *What to put in the bag? Comparing and contrasting procedures for text clustering*. In: "Statistica Applicata. Italian Journal of Applied Statistics. 22(1)", pp. 77-94.

Wang, X., Lu, Z., Tu, Z., Li, H., Xiong, D., & Zhang, M. (2017), *Neural machine translation advised by statistical machine translation*, in: “Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)”. AAAI Press, pp. 3330–3336.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J.R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., & Dean, J. (2016), *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, in: “ArXiv, abs/1609.08144”.

Żak, A., Kurtyka, A., & Rybicki, J. (2021). *Stylometry of Selected Works of J. R. R. Tolkien, Christopher Tolkien, C. S. Lewis and Charles Williams, and Their Polish Translations* (master thesis), Jagiellonian University, Cracovia.

Мифтахова Р. Г. (2015), *Основные факторы улучшения машинного перевода*, Вестник Башкирского университета, 20 (1), с. 188–192.

8. Siti internet consultati

Berruto, G. (2010). *Italiano standard*. Enciclopedia dell'italiano. Treccani. https://www.treccani.it/enciclopedia/italiano-standard_(Enciclopedia-dell%27Italiano) (ultima consultazione: 26/08/2022)

Cran.r-project: https://cran.r-project.org/web/packages/stylo/index.html (ultima consultazione: 11/07/2022)

DeepL: http://www.deepl.com/; https://www.deepl.com/pro?cta=header-pro; https://www.deepl.com/whydeepl/ (ultima consultazione: 15/08/2022)

Faloppa, F. (2010). *Lemma, tipi di*. Enciclopedia dell'italiano. Treccani. https://www.treccani.it/enciclopedia/tipi-di-lemma (ultima consultazione: 26/08/2022)

Glosbe: https://ru.glosbe.com/ru/it (ultima consultazione: 03/08/2022)

Google Translate: http://www.googletranslate.com/ (ultima consultazione: 21/06/2022)

Ibanez, F. (2020). Cosa è assolutamente necessario sapere sulla traduzione neurale. TraductaItaly. https://www.traducta.it/notizie/traduzione-automatica-neurale (ultima consultazione: 26/08/2022)

Ibanez, F. (2021). I vantaggi e gli inconvenienti della traduzione umana/automatica. TraductaItaly. https://www.traducta.it/notizie/vantaggi-inconvenienti-traduzione-umana-automatica (ultima consultazione: 26/08/2022)

Machine Translate: https://machinetranslate.org/wmt (ultima consultazione: 15/08/2022)

Proz: http://www.proz.com/ (ultima consultazione: 3/05/2022)

Ranks.nl: http://www.ranks.nl/stopwords/russian (ultima consultazione: 08/08/2022)

R-project: https://www.r-project.org/ (ultima consultazione: 08/07/2022)

Treccani: https://www.treccani.it/vocabolario/stringa2/ (ultima consultazione: 2/05/2022); http://www.treccani.it/ (ultima consultazione: 01/07/2022)

TreeTagger: https://cental.uclouvain.be/treetagger/ (ultima consultazione: 13/06/2022)

Voyant-tools: https://voyant-tools.org/ (ultima consultation: 16/08/2022)

Резюме дипломной работы

Темой настоящей дипломной работы является наблюдение разницы между человеческими и машинными переводами. **Целью** исследования является сравнение и анализ оригинальных текстов, написанных на русском языке, с соответствующими переводами на итальянский язык, выполненными переводчиком и двумя программами машинного перевода, т. е. DeepL и Google Translate. Следует напомнить, что объектом количественного анализа текста является корпус текстов. Термин «корпус» в основном означает совокупность текстов в соответствии с определенными принципами, преследуемым исследованием (ср. Мифтахова 2015, стр. 188; Tuzzi, 2003, стр. 29).

Дипломная работа состоит из пяти глав. Первая глава называется «Контекст» и знакомит с темами машинного перевода (см. 1.1) и концепцией «translationese» (см. 1.2). В данной работе также был описан общий обзор основных современных исследований, в которых изучаются различия между машинным и человеческим переводом (см. 1.3). Вторая глава под названием «Анализ текстовых данных: работа с корпусами текстов». В этой главе объясняется, что означает «Анализ текстовых данных» (см. 2.1 и 2.2), и описываются основные исследования в области переводоведения, в которых использовался данный метод (см. 2.3). Кроме того, в этой главе было описано содержание анализируемых корпусов как с качественной (см. 2.4.1), так и с количественной точки зрения (см. 2.4.2). Затем в третьей главе описаны методы и программы, использованные в ходе анализа корпусов (см. 3.1, 3.2, 3.3, 3.4 и 3.5), а результаты анализа представлены в следующей главе (см. 4.1, 4.2, 4.3, 4.4 и 4.5). В последней главе подведены итоги дипломной работы и проанализированы результаты анализа (см. 5.1, 5.2, 5.3, 5.5 и 5.6), а также возможные будущие исследования (см. 5.7).

В дальнейшем в этом резюме будут изложены основные цели исследования и использованные методы. Кроме того, будут описаны основные результаты различных анализов и возможные будущие исследования.

Для достижения цели дипломной работы было необходимо решить следующие задачи:

1. Проанализировать переводы итальянских аббревиатур, акронимов, слов иностранного происхождения, слов, предусмотренных теоретически только одним правильным переводом, а также проанализировать правильные транслитерации русских слов в латинских буквах. Анализ проводился с помощью программы количественного анализа текстов *AntConc*. Предполагалось, что машинный перевод не всегда переводит правильно аббревиатуры и акронимы, поскольку он не является точным при переводе сложных лексических, семантических и прагматических конструкций (Li и др. 2014, стр. 190). Например, может быть, что слово «Телеграм» неправильно переводится машинными переводчиками, которые могут ошибочно переводить как «telegramma» на итальянский. Кроме того, предполагалось, что иногда перевод выполнен человеком содержит большее количество синонимов, чем машинные переводчики, которые переводят более буквально.
2. Вычислить TTR (*Type Token Ratio*, на русском Коэффициент Лексического Разнообразия, объясняется далее в пункте 2) и среднюю длину предложения каждого перевода в корпусах текстов, чтобы проверить какой корпус лексически проще. Вычисления сделаны посредством онлайн программы *Voyant-tools*. К сожалению, данная программа не читает кириллический алфавит, поэтому было невозможным выполнить вышеуказанные расчеты в корпусе русских текстов, а только в корпусе текстов на итальянском языке. В любом случае это не составит проблемы, поскольку проверка этих параметров возможна только в тех случаях, когда тексты имеют похожую длину и написаны на одном языке. Предполагалось, что переводы выполнены машинными переводчиками имеют большее лексическую простоту чем человеческие, поскольку они используют меньше синонимов чем человеческие переводчики. Важно отметить, что *лексическая*

простота является переводческой универсалией и значит низкие уровни TTR и среднюю длину предложения.

3. Проанализировать темы разных корпусов текстов посредством программы Iramuteq. Целью исследования является проверить корректность тем корпусов переводов в сравнении с корпусом оригинальных русских текстов. К сожалению, не было возможно использовать корпус на русском языке в программе *Iramuteq*, потому что он не считывает правильно русский язык ни с кодом utf-8, ни со специфическими кодами кодирования для русского языка, т. е. cp866 и koi8_r. Поэтому было решено, в любом случае, провести анализ только с итальянскими корпусами, рассматривая корпус текстов, выполненных переводчиками как правильный вариант. Предполагалось, что корпусом с переводами DeepL эквивалент перевода будет более подобен человеческому, чем корпусом с переводами Google Переводчиком, поскольку сайт DeepL регулярно тестируется профессиональными переводчиками, которые оценивают наилучший результат из ряда предложенных машинных переводов (ср. www.deepl.com).
4. Сравнить оригинальные статьи на русском языке со сравнительными переводами на итальянском языке посредством кластерного анализа, проведенного с помощью программы R с помощью пакета *stylo*. Цель анализа состоит в том, чтобы увидеть различия и сходства между кластерами русских текстов, сделанных как с помощью машинного перевода, так и посредством переводов, сделанных профессиональным переводчиком используя при этом дендрограммы. Предполагалось, что кластеры машинных переводов, сделанных с помощью DeepL и Google Translate, имеют больше шансов быть похожими на кластеры оригинальных русских текстов. Данная гипотеза обосновывается тем, что программы машинного перевода создают переводы, более буквальные и поэтому более похожи на стиль оригинальных текстов (ср. Ibanez 2021).

5. Проверить способность двух алгоритмов машинного обучения распознавать человеческий перевод от машинного. Анализ приводился посредством программы R и тремя алгоритмами: *Authorship Attribution*, *Sector Vector Machine* и *Random Forest*. Первый алгоритм полезен для подготовки корпусов текстов к анализу, потому что может разделить тексты на части и способствовать анализу. Другие алгоритмами, т. е. *Sector Vector Machine* (SVM) и *Random Forest* (RF), являются настоящими алгоритмами машинного обучения и должны понимать автора перевода среди корпусов текстов, выполненных как переводчиком, так и машинных переводов. В ходе анализа использовались только корпусы текстов на итальянском языке. Предполагалось, что два алгоритма могут отличать переводы, поскольку предыдущие исследования продемонстрировали способность машинного обучения отличать человеческий перевод от машинного (ср. Li и др. 2015; Fu и др. 2021).

Решение поставленных задач было осуществлено за счет применения следующих методов анализа текстовых данных:

1. Анализ «появлений» и переводов некоторых слов в текстах, найденных в различных корпусах. В этом случае термин «появление» значит сколько раз слово повторяется в корпусе текстовых данных. Прежде всего, русские слова выбирались из списков *word-type*, т. е. отдельные графические формы корпуса (ср. Tuzzi 2003, стр. 72). Для облегчения поиска в русском и итальянском корпусах использовались *stop-word*, т. е. список слов, которые широко используются в языке часто совпадают с артиклями, союзами, предлогами и т. д.. Учитывая их широкое употребление, эти слова почти всегда входят в число наиболее часто употребляемых слов в корпусе текстов, но на самом деле они не дают никакой релевантной информации о лексическом и отличительном содержании текста (ср. Dolamic & Savoy 2009, стр. 200). Поэтому данные слова были исключены, чтобы получить список слов, содержащий только наиболее значимые *word-type*. В корпусе на

русском языке использовался список *stop-word* из сайта *Ranks.nl*, в то время как, во всех корпусах на итальянском языке использовался список *stop-word* из сайта *Voyant.tools*. После выбора слов для анализа, переводы различных корпусов на итальянском языке были проанализированы и сравнены с русскими оригиналами. Данное сравнения было проведено с помощью программы *AntConc*, в частности, функции «*Concordance Plot*» и «*File View*», которые позволили лучше идентифицировать перевод и контекст его использования в разных корпусах на итальянском языке.

2. Вычисление TTR и средней длины предложения каждого перевода в корпусах текстов на итальянском языке. Английский акроним TTR значит *Type Token Ratio* и рассчитывается следующим образом:

$$\frac{V}{N} * 100$$

В данном случае N значит сумма всех *word-token* (т. е. графических форм) текста или корпуса, наоборот V показывается сумма всех *word-type* присутствуют в тексте или в корпусе текстов. Кроме того, средняя длина предложения в тексте была рассчитана автоматически программой Voyant-tool. Данные расчеты используются для понимания лексической простоты текста и, чем ниже их результат, тем больше лексической простоты присутствует в тексте (ср. Kunilovskaja и др. 2018; Ondelli & Viale 2010).

3. Контент-анализ корпусов текстов на итальянском языке. Контент-анализ относится к процессу получения, синтеза и возвращения информации, присутствующей в сообщении (ср. Tuzzi 2003, стр. 17), и примером может быть устное изложение сюжета фильма в его основных моментах. Для проведения автоматического контент-анализа можно использовать программу Iramuteq, которая использует метод *Reinert* для автоматического определения тем корпусов текстов. Метод *Reinert* состоит из алгоритма, который анализирует корпус в соответствии с совпадениями слов, которые появляются во фрагментах текста, и таким образом определяет семантические классы. Термин семантический класс относится к группе слов, связанных друг с

другом, образующих тему (ср. Sbalchiero 2018, стр. 202). Чтобы определить темы в *Iramuteq*, необходимо разделить корпус на сегменты текста, называемые ECU, т. е. *Elementary Context Unit* (элементарная единица контекста) и создать матрицу, в которой пересекаются ECU и слова (*word*), где ячейки матрицы (*matrix*) указывают на наличие (например, 1) или отсутствие (например, 0) определенного слова в ECU (см. таблицу 3 стр. 60). Результаты этой матрицы представлены графически в виде дендрограммы, состоящей из нескольких кластеров. С помощью кластерного анализа были проанализированы различия между человеческим переводом, переводом Deepl и переводом Google Translate.

4. Кластерный анализ корпусов текстов на русском и на итальянском языке. Цель кластерного анализа - сгруппировать похожие тексты в один и тот же кластер (группу) и разделить несхожие тексты на разные кластеры (ср. Tuzzi 2010, стр. 81). Важно, что кластерным анализом является не конкретный алгоритм, а задача, которую могут выполнять разные алгоритмы. В данном исследовании использовался алгоритм *Borrow's Classic Delta*, доступный из пакета *stylo*, написанного на языке программирования R, который полезен для проведения стилометрического анализа (ср. Eder и др. 2016, стр. 107). Термин «стилометрия» относится к вычислительному методу, связанному с количественным изучением стиля написания текстов (например жанр текста, дата издания, авторство). На сегодняшний день одним из основных применений стилометрии является исследование авторства, в котором было установлено, что можно определить автора текста, проанализировав первые наиболее часто встречающийся слова или биграммы (ср. Eder и др. 2016, стр. 107–108). Кроме того, кластерный анализ также успешно используется для сравнения человеческого и машинного перевода (ср. Cembrzyńska и др. 2021; Żak и др. 2021). По этой причине он был использован в данном исследовании. Сначала была создана дендрограмма с текстовым корпусом на русском языке, а затем с другими корпусами на итальянском языке. Затем отдельные

кластеры были проанализированы и выявлены сходства и различия между оригинальными русскими текстами и итальянскими переводами.

5. Машинное обучение для распознавания человеческого перевода из машинного перевода. Машинное обучение — это научное изучение алгоритмов и статистических моделей, которые позволяют компьютеру выполнять определенную задачу без явного программирования (ср. Mahesh 2019, стр. 381). Машинное обучение использует различные методы, в результате чего получаются алгоритмы с разной степенью автономности, требующие большей или меньшей помощи человека (ср. Mahesh 2019, стр. 381). В данной работе использовались *supervised learning* (т. е. контролируемое обучение), в частности алгоритмы *Support Vector Machine* (SVM) и *Random Forest* (RF) (ср. Gatti & Tuzzi 2020). Перед использованием алгоритмов машинного обучения использовался алгоритм *Authorship Attribution*, который позволяет подготовить корпуса текстов на итальянском языке для SVM и RF анализа (ср. Gatti & Tuzzi 2020, стр. 5; Mikros & Perifanos 2013 and 2015). С помощью алгоритма *Authorship Attribution* три корпуса текстов на итальянском языке были разделены на части от 200 слов и были проанализированы следующие лингвистические характеристики текстов: наиболее частые *bigram* (последовательности из 2 символов подряд), *trigram* (последовательности из 3 символов подряд), *word-type*, *word bigram* (пара последовательных слов). После подготовки текстовых корпусов можно было начать анализ с помощью SVM и RF, которые должны были распознать автора перевода среди следующих: переводчик-человек, переводчик DeepL и переводчик Google Translate. На этом этапе был включен дополнительный перевод, которого не было ни в одной из текстовых корпусов, и тексты из различных корпусов были использованы для обучения алгоритмов машинного обучения, чтобы отличить переводчика-человека, переводчика DeepL и переводчика Google Translate.

В данном проекте были проанализированы переводы статей, изданных в разных русских газетах[40], которые можно прочесть онлайн. Полный список источников статей на русском языке показывается в главе н. 6.1. Все переводы, сделанные переводчиками на итальянском языке, были опубликованы на одном и том же сайте, под названием russiaintranslation.com. Авторы переводов являются разными и их источники представлены в главе н. 6.2.

Критерием отбора текстов для анализа являются следующие:

1. В первую очередь были собраны только те статьи, опубликованные в русских газетах в период с 30 марта 2020 года по 13 декабря 2021 года. Это значит, что в корпусе текстов для анализа нет статей о войне в Украине, а также нет статей на тему текущей российской политической пропаганды.
2. Выбор текстов для анализа проводился на итальянском сайте russiaintranslation.com, где были выбраны 34 перевода на итальянский язык с русских статей о теме политики и общества. Было решено начать отбор текстов, сделанных переводчиками, чтобы быть уверенным в существовании и достоверности переводов. Кроме того, в конце каждого перевода на сайте russiaintranslation.com указывалась ссылка на оригинальный текст на русском языке. Таким образом, можно было проверить достоверность источников.
3. Благодаря ссылкам на оригинальные тексты можно было собрать 34 статьи, опубликованные онлайн в разных русских газетах.
4. После окончания сбора статей на русском языке, все тексты были переведены на итальянский язык с помощью DeepL Переводчика и Google Переводчика. Все данные машинные переводы были собраны в двух разных корпусах текстов.

[40] Примечание: только статья «Комментарий: Время Москвы на Южном Кавказе вышло. Настало время Анкары» была опубликована в немецкой газете DW, но автор статьи — русский и написал на русском языке.

Необходимо отметить, что все четыре корпуса текстов состоят из 34 текстов с расширением .txt.

На первом этапе анализа поставлена цель сравнить переводы акронимов, аббревиатур, слов иностранного происхождения, транслитерации несколько имен, терминов, принимая человеческий перевод за правильный, следующие результаты предоставляются в конце данного анализа. Важно отметить, что, обе программы машинного перевода могут правильно переводить большинство акронимов. Например, русский акроним «МОК» был переведен соответствующим итальянским акронимом «CIO» как с программы DeepL, так и с Google Translate. Этот же перевод найден и в корпусе «Человеческие переводы» и означает, что машинный перевод смог перевести «МОК» аббревиатурой, реально используемой в итальянском языке. Другим примером является «ЕСПЧ», который был правильно переведен с помощью обеих программ машинного перевода на «CEDU». Кроме того, было отмечено, что русские аббревиатуры не всегда переводились буквально человеческими переводчиками, поскольку в итальянском языке употребляется не так много аббревиатур, как в русском языке. Поэтому, аббревиатуры как «США» или «КНР», не всегда были переведены на итальянский как «USA» и «RPC», но человеческие переводчики также употребили слова «Stati Uniti» и «Cina». Очень интересным является тот факт, что такая же тенденция была замечена и в машинных переводчиках, которые почти всегда переводили «Stati Uniti» и «Cina» вместо «США» и «КНР». Только в одном случае оказалось, что машинные переводчики не распознали значение аббревиатуры. Речь идет об аббревиатуре «НКР», которая была переведена человеческими переводчиками словами «Nagorno-Karabach» или «Nagorno-Karabakh» на итальянский язык. И наоборот, вместо того чтобы правильно перевести акроним «NKR» оба машинных переводчика DeepL и Google Translate транслитерировали русскую аббревиатуру. Однако это единственный случай из 18, в котором машинные переводчики, вероятно, использовали транслитерацию. По этим причинам можно сказать, что DeepL и Google Translate без проблем правильно переводят наиболее часто

употребляемые русские аббревиатуры на итальянский язык, даже используя синонимы (см. «Stati Uniti» вместо «USA», стр. 71). Кроме того, существует гипотеза, что обе программы машинного перевода имеют проблемы с переводом менее часто встречающихся аббревиатур, но для подтверждения этой гипотезы необходимо провести специальное исследование на эту тему.

В дальнейшем были проанализированы две русские аббревиатуры, т. е. «др» и «РЕД». Машинные переводчики без проблем перевели данные две аббревиатуры с правильными версиями, т. е. «ecc.» и «ndr.». Только в одном случае DeepL неправильно перевел слово «РЕД» с «I.T.». Кажется, что в переводе аббревиатур Google Translate был более точен, чем DeepL.

Затем анализ транслитераций показал интересные особенности не только у машинных, но и у человеческих переводчиков. Прежде чем обсуждать результаты этого исследования, важно напомнить, какие правила транслитерации используются в Италии. Первым правилом научной транслитерации кириллицы является стандарт ISO 9, который был введен в 1929 году славистом Ettore Lo Gatto (ср. Torresin 2022, ст. 33). Однако, как пишет Torresin (2022), в текстах, опубликованных на сайтах, обычно используется коммерческая транслитерация вместо научной. Между ними существуют некоторые различия. Например, согласно научной транслитерации фамилию «Навальный» следует переводить так «Naval'nyj», а согласно коммерческой транслитерации – так «Navalny». В корпусе «Человеческие переводы» было отмечено, что переводчики не соблюдают единого правила транслитерации, но используют как научные, так и коммерческие нормы транслитерации. Например, фамилия «Навальный» была переведена 57 раз как «Naval'nyj», 2 раза как «Naval'nyi», 10 раза как «Naval'ny» и 1 раз как «Naval'nij». С другой стороны и DeepL, и Google Translate, всегда переводили фамилию «Навальный» так «Navalny». Данное означает, что машинные переводчики не ошибаются в транслитерации латинского алфавита и чаще соблюдают правила коммерческой транслитерации.

Следующее сравнение переводов на лексическом уровне касается анализа итальянских переводов двух категорий слов: заимствованное слова

иностранного происхождения (напр. «Ютуб») и иностранные слова, написанные латинскими буквами (напр. «Wildberries»). В конце анализа было отмечено, что машинные переводчики в большинстве случаев переводили правильно. Но в нескольких случаях перевод был совершенно неправильным. Машинный переводчик DeepL перевел правильно слово «телеграм» только два раза и потом перевел это слово как «canale telegrafico» (1 раз) или «telegrafo» (3 раза). Потом DeepL перевел «вотсапп» как «waps». С другой стороны, машинный переводчик Google Translate перевел «вотсапп» правильно, но перепутал слово «телеграм» с «telegramma» 5 раз и правильно перевел данное слово только 1 раз. Наоборот, «тикток» и «ютуб», были правильно переведены машинными переводчиками как «TikTok» и «Youtube», за исключением DeepL, который однажды перевел «тикток» как «tris». Похожие ошибки были обнаружены и в переводе иностранных слов, написанные латинскими буквами. Например, название компании «Wildberries», которое должно было остаться неизменным, было переведено один раз как «bacche selvatiche» в DeepL и три раза как «frutti di bosco» в Google Translate. Еще один пример можно найти в переводе названия компании «moonlight», которое DeepL однажды перевел как «chiaro di luna», тогда как в Google Translate этой ошибки не было обнаружено. Важно отметить, что существуют только эти две ошибки в переводе иностранных слов, поскольку слова как «amazon» или «ryanair» были переведены без ошибок. Таким образом, похоже, что машинные переводчики лучше переводят иностранные слова, написанные латинским алфавитом, а не слова иностранного происхождения. По итогу, были проанализированы несколько терминов, которые теоретически должны переводиться только одним способом. Было отмечено, что, вопреки первоначальным ожиданиям, человеческие переводчики почти всегда однозначно переводили выбранные термины. Только в редких случаях человеческий переводчик использовал синонимы, поскольку при использовании итальянского языка это было разрешено. Например, человеческий переводчик перевел два раза термин «коронавирус» как «pandemia», поскольку это оказался лучший перевод на итальянском языке (см. стр. 78).

В заключение следует отметить следующее, что на лексическом уровне различий между машинным и человеческим переводом не так много.

Целью второго этапа анализа была поставлена задача сравнить уровни *Type Token Ratio* и средней длины предложения каждого текста для расчета уровня простоты переводов в корпусах текстов «Человеческий Перевод», «DeepL» и «Google Translate». Отмечалось, что различия в расчетах TTR не слишком велики среди переводов и иногда несколько машинных переводов показывали больший уровень TTR чем человеческие переводы (см. рисунок 11, стр. 81). Было отмечено, что машинный перевод отличается средней длинной предложения от человеческого тем, что в машинном переводе используются короче предложения. Это означает, что машинный переводчик использует более упрощенные предложения чем человеческие переводчики (см. рисунок 12, стр. 81), тогда как было замечено, что машинные переводы не обязательно менее лексически сложны, чем человеческие. Так же было отмечено, что человеческие переводы более сложны синтаксически.

Третий этап анализа касается проверки корректности предметов корпусов переводов в сравнении с корпусом оригинальных русских текстов. К сожалению, с программой *Iramuteq* возникли проблемы. Тем не менее, было решено провести данное исследование, используя только корпусы на итальянском языке, который программа читает без проблем. В ходе анализа в *Iramuteq* были введены различные значения в «*dimensioni di numeri di cluster finali in fase 1»,* чтобы получить дендрограммы одинакового размера, с 6 классами для каждого корпуса на итальянском языке (см. рисунки 13-14-15, стр. 83–84). При сравнении дендрограмм было решено взять за точку отсчета корпус человеческих переводов. После внимательного анализа отдельных кластеров, представленных на рисунках 16, 17, 18 на стр. 85–86, обнаружилось, что между человеческим и машинными переводами DeepL существует много аналогий. Темы корпуса «DeepL» всегда были очень похожи на темы корпуса «Человеческие Переводы», фактически разделение на тематические кластеры было почти таким же. В отличие от этого, разделение Google Translate на тематические классы часто были менее

похожими на человеческий перевод, чем DeepL. Хотя сравнить русские тексты с переводами не было возможным, данное исследование, тем не менее, было полезно для выявления аналогий между DeepL и человеческим переводом.

Четвертый этап анализа касается сравнения оригинальных статей на русском языке с переводами на итальянском языке посредством кластерного анализа. Цель анализа состоит в том, чтобы найти разницы между человеческими и машинными переводами, с помощью исследования стиля написания разных авторов текстов. В данном исследовании термин «стиль» означает грамматический стиль, исходя из самых частых *word-type* в корпусах. В первую очередь тестировалась способность пакета *stylo* отличать человеческий перевод от машинного. В этом анализе учитывались первые 200 наиболее часто встречающихся слов *most frequent word* (самые частые *word-type*) в корпусах на итальянском языке. Оказалось, что *stylo* смог отличить разницу между стилем человеческого и машинного перевода, поскольку корпус «Человеческие Переводы» находился в отдельном кластере, а корпусы «DeepL» и «Google Translate» находились в одном и том же кластере (см. рисунок 19, стр. 88). Далее были созданы четыре дендрограммы, по одной для каждого корпуса. Для того, чтобы создать как можно более точное кластерное разделение, для каждого корпуса были выбраны *most frequent word*, занимающие 77/78% лексикона каждого корпуса. Первая дендрограмма была построена для корпуса «Русские Оригиналы», в которой были рассмотрены все *word-type* с частотой ≥ 2, для получения словарного запаса 77,35%. Во всех итальянских корпусах рассматривались все *word-type* с частотой ≥ 5, для получения словарного запаса 78,22% для корпуса «Человеческие Переводы», 78,58% для корпуса «DeepL» и 78,51% для корпуса «Google Translate». После первого просмотра различных дендрограмм, сделанных *stylo* (см. рисунок 20 стр. 90; рисунок 21 стр. 91; рисунок 22 стр. 92; рисунок 23 стр. 93), первое, что вы могли заметить, это то, что кластерное разделение всех корпусов на итальянском языке отличается от кластерного разделения корпуса «Русские Оригиналы». После анализа групп кластеров отдельно, были отмечены более

специфические сходства и различия между разными корпусами. Было отмечено, что в некоторых случаях человеческий перевод больше напоминает кластеризацию корпуса «Русские Оригиналы», чем программы машинного перевода (см. статьи «09_ЧерныйАвгуст» и «09_УспешныхРоссиян» на рисунке 25, стр. 95; статьи с кодами «03_» и «02_» на рисунке 27, стр. 97). Кроме того, более тщательное изучение статей с кодами «03_» и «02_» показывает, что кластеры в корпусе «Русские Оригиналы» также представлены в корпусе «DeepL», что указывает на сходство между человеческим переводом и переводом DeepL. Дальнейшее сходство между корпусами «Human Translations» и «DeepL» было также найдено в кластерной группировке статей с кодами «09_», «07_» и «08_» (см. рисунок 24 стр. 94; см. рисунок 26 стр. 96), но такое же сходство не было найдено между корпусами «Человеческие Переводы» и «Google Translate». Наконец, было отмечено, что в следующих случаях корпусы машинного перевода показали наибольшее сходство с корпусом «Русские оригиналы»: кластер, в котором содержатся статьи с кодом «08_» в корпусе «DeepL» и кластер, в котором содержатся статьи с кодом «07_» в корпусе «Google Translate» (см. рисунок 25 стр. 95; рисунок 26 стр. 96). В заключение можно сказать, что в некоторых случаях человеческий перевод более близок к стилю оригинальных текстов, чем машинный. Более того, сходство между машинным переводом DeepL и человеческим переводом было отмечено и кластерным анализом *stylo*, который подтвердил тот факт, что программа перевода DeepL производит «более человеческие» переводы, чем Google Translate.

На последнем этапе анализа поставлена цель проверить способность двух алгоритмов машинного обучения (*Sector Vector Machine* и *Random Forest*) распознавать человеческий перевод от машинного. Всего было проведено девять анализов, в которых алгоритмы должны были узнать автора перевода и определить, кто является автором перевода, человеческий переводчик, программа DeepL или программа Google Translate. Для обучения алгоритмов определения человеческого почерка от почерка DeepL и Google Translate были использованы все статьи, содержащиеся в разных корпусах,

тексты, в которых алгоритмы должны были узнать автора, не были найдены в корпусах и поэтому они специально были отобраны для этого анализа. В конце анализа, к сожалению, было отмечено, что ни SVM, ни RF не дали приемлемых результатов. Например, при распознавании человеческого перевода между корпусами «Человеческие Переводы» и «DeepL» (см. табл. 10 стр. 99) SVM показала точность 74%, но все же не смогла узнать человеческий перевод. В отличие от этого, RF показал низкую точность 54%, но смог распознать автора перевода как человека. В других случаях (см. таб. 13 стр. 100; таб. 15 стр. 101) точность RF падала до 28% и 30%. Хотя в некоторых случаях RF распознал правильного автора текстов, такие низкие цифры точности делают результаты неприемлемыми. Процент точности менее 50% означает, что алгоритм плохо распознает различия между авторами и принимает случайные решения. Из-за неясных или несоответствующих результатов на вопросы исследования, связанные с данным анализом, получить нужные ответы не удалось. Безрезультатность этого анализа, вероятно, связана с небольшим размером анализируемых корпусов. Действительно, чтобы правильно обучить алгоритм машинного обучения, необходимо собрать большой корпус, содержащий миллионы текстовых данных (ср. Wu и др. 2016; Popel и др. 2020). Несмотря на сказанное, было решено все равно провести этот анализ, поскольку было доказано, что алгоритмы SVM и RF могут различать коммуникативный стиль политиков Обамы и Трампа, принимая во внимание только пять их политических речей. Однако результат оказался неудовлетворительным, вероятно, потому что стилистические различия между человеческим и машинным переводами меньше, чем между двумя политиками с абсолютно разными способами выражения. Провал этого анализа не означает, что данные два алгоритма машинного обучения не способны отличить человеческие переводы от машинных. На самом деле, считается, что с большим количеством текстовых данных алгоритмы SVM и RF могли бы работать лучше (ср. Li и др. 2015; Fu и др. 2021). В пользу этой гипотезы говорит и тот факт, что *stylo*, анализируя корпусы в целом, а не текст за текстом, смог без проблем отличить человеческий перевод от машинного.

В завершение этого исследования можно сказать, что количественные методы анализа текста также полезны в исследованиях перевода и в сравнении между машинным и человеческим переводом в целом. Более того, доказано, что в данный момент машинный перевод очень похож на человеческий. Переводчик (человек) может сам решить, считать ли машинные переводчики конкурентами или использовать точность программы машинного перевода, чтобы переводить тексты быстрее. Некоторые исследователи считают, что машинный перевод достигнет уровня человека при переводе некоторых типов текстов (ср. Popel и др. 2020). Тем не менее, считается, что всегда будет необходимо проводить окончательную проверку человеком, чтобы избежать публикации текстов, которые не являются на 100 процентов точными. С другой стороны, если пользователю требуется быстрый перевод текста на иностранном языке, то машинный перевод является самым быстрым и недорогим решением (ср. Li и др. 2014). Кроме того, это исследование показало, что не все машинные переводчики одинаковы, однако некоторые из них более точны и «человечны», чем другие. По результатам анализа в *Iramuteq* и *stylo*, становится ясно, что переводы DeepL более похожи на человеческий перевод, чем Google Translate. Однако анализ в AntConc показал, что Google Translate более точен в переводе некоторых слов, чем DeepL (например, слов иностранного происхождения, таких как «telegram»). Поэтому, кажется, что DeepL очень точен со стилистической и синтаксической точки зрения, но в плане лексики он всё ещё нуждается в доработке.

Анализ, проведенный в данном исследовании, является необычным среди исследований, рассматривающих различия между человеческим и машинным переводом. В литературе наиболее важные исследования, связанные со сравнением человеческого и машинного перевода, основаны на системе подсчета баллов, проводимой предварительно выбранными оценщиками (ср. Wu и др., 2016; Läubli и др. 2018; Popel и др. 2020; Takakusagi и др. 2021), а иногда и на модификации алгоритма программного обеспечения машинного перевода для улучшения его работы (ср. Wu и др. 2016; Popel и др. 2020). Эти

исследования часто представляются на WMT (*Workshop on Machine Translation*) - ежегодной конференции, на которой обсуждаются пути улучшения машинного перевода. На основании результатов, полученных в предыдущих работах (ср. Lee 2019; Cembrzyńska и др 2021; Żak и др 2021) и в данном исследовании было доказано, что использование методов количественного анализа текста полезно для понимания различий между машинным и человеческим переводом. По этой причине рекомендуется, чтобы возможные будущие исследования также включали методы количественного анализа текста при изучении различий между человеческим и машинным переводом.