UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Studio dei fenomeni linguistici correlati al coinvolgimento suscitato da un testo

Candidato: *Anikeeva Evgeniia*

Relatore: *Felice Dell'Orletta*

Correlatori: *Alessandro Lenci*

Dominique Brunato

Anno Accademico 2020-2021

Indice

1. Introduzione

La mia tesi prende spunto dall'attività di tirocinio che ho svolto presso il laboratorio ItaliaNLP Lab dell'Istituto di Linguistica Computazionale "Antonio Zampolli", un istituto del CNR di Pisa che svolge l'attività di ricerca, di valorizzazione e trasferimento tecnologico e di formazione in settori scientifici strategici della Linguistica Computazionale. La tesi si propone di indagare se esiste una relazione tra il grado di interesse che un testo suscita in un lettore e la forma linguistica del testo stesso, focalizzandosi in particolare su caratteristiche linguistiche legate allo stile del testo, piuttosto che al suo contenuto. La metodologia adottata si è basata su tre fasi principali: nella prima fase, è stato raccolto un corpus di testi rappresentativi del genere testuale dei blog di viaggio e omogenei rispetto alla tematica. I testi del corpus sono stati fatti valutare da un campione di lettori, selezionati tramite crowdsourcing, rispetto al grado di interesse suscitato dal testo stesso dopo la lettura. Nella fase successiva, i testi valutati sono stati annotati linguisticamente in modo automatico e analizzati tramite tecnologie e strumenti di *Natural Language Processing* (NLP) deputati al monitoraggio linguistico del testo. Questo processo ha permesso di estrarre, per ciascun testo, il suo profilo linguistico, ovvero una ricca descrizione basata su un ampio insieme di caratteristiche linguistiche identificative di fenomeni legati prevalentemente alla struttura di base, morfo-sintattica e sintattica.

Nella terza fase, ci siamo occupati dello studio dei fattori linguistici maggiormente implicati nel coinvolgimento suscitato dalla lettura di testo. A questo scopo, è stato usato un test di correlazione statistica volto ad identificare le caratteristiche linguistiche che risultano maggiormente associate ai testi giudicati mediamente più interessanti dai lettori.

Data questa articolazione generale del lavoro, la tesi si compone dei seguenti capitoli: il capitolo 2 si propone di presentare il corpus che fa da sfondo allo studio. In particolare, nel paragrafo 2.1 verranno presentati i criteri che hanno guidato la selezione dei testi e l'approccio adottato per definire, in maniera operativa, cosa rende un testo coinvolgente per un lettore. Il paragrafo 2.2 descrive come sono stati costruiti i test da somministrare agli utenti e una analisi del campione di annotatori reclutato rispetto ad alcune variabili di sfondo (età, genere, livello di istruzione). Infine, il paragrafo 2.3 riporta una prima descrizione dei dati ottenuti

dall'annotazione, ovvero la distribuzione media dei punteggi ottenuti da ciascun testo e la relativa deviazione standard, con alcuni esempi di testi giudicati tra i più e i meno interessanti.

Nel capitolo 3 verranno introdotti i fondamenti della metodologia di monitoraggio linguistico del testo e i prerequisiti su cui si basa. Sarà illustrato lo strumento con cui questa metodologia è stata implementata in questo studio, chiamato Profiling-UD, che si basa sul framework di annotazione linguistica sviluppato nell'ambito del progetto delle *Universal Dependencies* e verrà data una breve descrizione della tipologia di caratteristiche che lo strumento permette di estrarre. Nei paragrafi 3.2 si riporteranno alcune statistiche relative alla distribuzione media di queste caratteristiche nel corpus di testi analizzati.

Infine, il capitolo 4 sarà dedicato allo studio dei fenomeni linguistici che caratterizzano i testi giudicati più interessanti dai lettori. A questo scopo si discuteranno i risultati delle correlazioni tra le caratteristiche linguistiche estratte da ciascun testo e il giudizio medio di interesse espresso dagli annotatori, selezionando quelle significative in base a un test statistico non parametrico di correlazione.

2. Descrizione del corpus

In questa sezione verrà descritta la tipologia di testi che sono stati raccolti per la costruzione del corpus. Verrà anche spiegato come sono stati ottenuti i giudizi di interesse per ciascuno di questi testi, quindi verranno illustrati tutti i passaggi per la creazione del questionario; e infine, come sono stati estratti e analizzati i risultati di tali indagini.

2.1 Tipologia dei testi

Il primo passo necessario per la realizzazione di questo progetto è stato quello di selezionare i testi da analizzare.
È stato preso in considerazione un unico genere testuale, quello dei blog di viaggio, e un argomento comune, ovvero blog relativi alla descrizione di viaggi in Russia. Questa scelta è stata motivata dal tentativo di limitare quanto più possibile la variabilità nella percezione dell'interesse di un testo dovuto al *topic* trattato e quindi permettere di concentrare le analisi sui fenomeni legati prevalentemente allo stile di scrittura. I testi del corpus sono 120 sono stati raccolti dal web all'interno di siti di blog italiano. In quanto segue, si presentano alcuni esempi di testi.
Il primo esempio riguarda un blog di viaggio "Mi prendo e mi porto via" dal quale a sua volta è stato scelto un post con il seguente titolo "Viaggio in Russia fra Mosca e San Pietroburgo":

Figura 1. Schermata della pagina web di un blog di viaggio "Viaggio in Russia fra Mosca e San Pietroburgo".

Questo blog racconta le esperienze di viaggio di una coppia, che aspira a visitare posti insoliti e interessanti.
Ecco un estratto del testo dedicato del loro viaggio fra Mosca e San Pietroburgo:

> *"Non vi dirò nulla del Cremlino di Mosca, né del Palazzo d'Inverno di San Pietroburgo, niente di Serghiev Posad e nemmeno di San Basilio. Sono così ben restaurate da sembrare finte e parzialmente lo sono anche, quindi se l'effetto è sicuramente abbagliante, l'idea è sempre quella di guardare una copia della cappella sistina … è bella sì, ma quanto può essere emozionante?"*[1]

[1] Il blog di viaggio "Miprendoemiportovia" https://www.miprendoemiportovia.it/2013/06/05/viaggio-in-russia-fra-mosca-e-san-pietroburgo

Al riguardo del secondo esempio di testo che è stato preso di un altro blog di viaggio che si chiama “Viaggi di Gusto”:

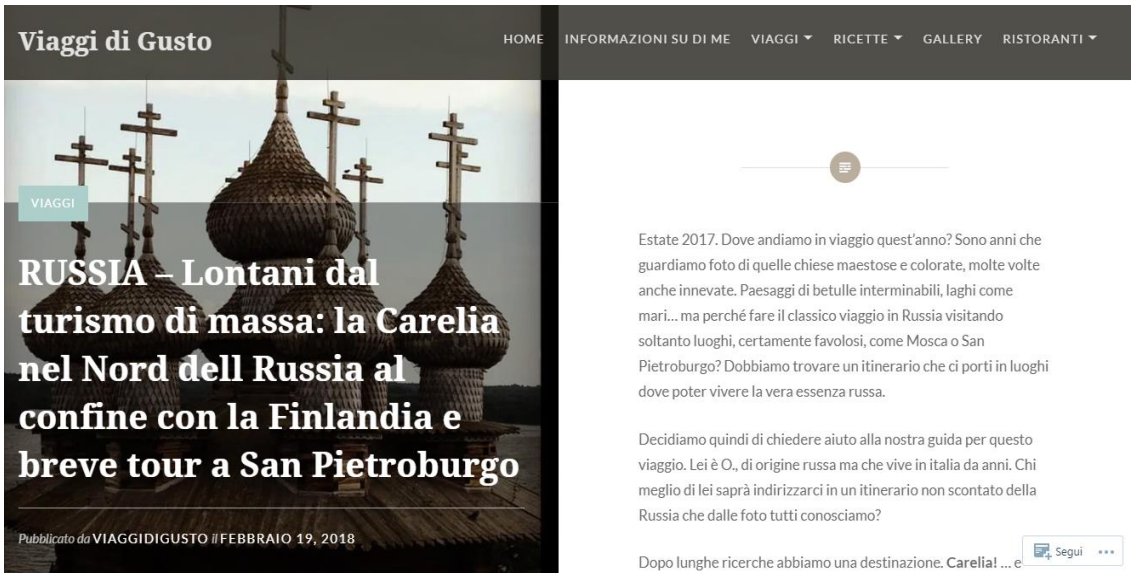

Figura 2. Schermata della pagina web di un blog di viaggio “RUSSIA – Lontani dal turismo di massa…”

Questa parte di blog racconta una storia di viaggio in Carelia situata a Nord-Ovest della Russia vicino San Pietroburgo e ha i suoi confini con la Finlandia.

> *“Estate 2017. Dove andiamo in viaggio quest’anno? Sono anni che guardiamo foto di quelle chiese maestose e colorate, molte volte anche innevate. Paesaggi di betulle interminabili, laghi come mari… ma perché fare il classico viaggio in Russia visitando soltanto luoghi, certamente favolosi, come Mosca o San Pietroburgo? Dobbiamo trovare un itinerario che ci porti in luoghi dove poter vivere la vera essenza russa.”*[2]

Tutti i testi sono stati scaricati manualmente dai siti web dei vari blog di viaggio e salvati in un formato .txt. Ai fini dello studio, è stata selezionata solo l’introduzione corrispondente alle prime *3-4 righe* del testo per un totale di *circa 60 parole*. Questa scelta è stata motivata dall’approccio seguito nella creazione dei sondaggi per cercare di formalizzare un aspetto molto soggettivo quale il grado di interesse veicolato da un testo. Si è pensato infatti di sottoporre alla valutazione umana non

[2] Il blog di viaggio “Viaggi di Gusto” https://viaggidigusto.wordpress.com/2018/02/19/russia/

tutto il testo bensì solo l'introduzione, chiedendo al lettore quanto sarebbe stato interessato a proseguire nella lettura dell'intero testo.

Nei paragrafi seguenti saranno presentati i passaggi successivi per la creazione dei questionari di valutazione e risultati delle annotazioni dei lettori, con alcuni esempi di stesti giudicati "più interessanti" dai lettori, ovvero quelli che hanno preso un voto altro.

2.2 Creazione di un questionario per l'annotazione

È stato creato un unico questionario per permettere ad un campione abbastanza vario dei lettori di esprimere il proprio giudizio riguardo l'interesse suscitato da ogni porzione di testo. Inizialmente sono stati selezionati dal *web* 60 testi dei blog di viaggio e aggiunti nel questionario *QuestBase*[3]. Dopo aver raccolto tutti i giudizi assegnati ad ogni di 60 testi sono stati selezionati e messi altri 60 testi in questionario, fino ad arrivare ad un totale di 120 testi per una questione di completezza. A partire dalla raccolta dei 120 testi e della selezione delle loro introduzioni, in questo paragrafo vengono descritti i passaggi relativi alla creazione dei questionari di valutazione.

Per lo svolgimento di questo lavoro è stata utilizzata la piattaforma *QuestBase*.

Si tratta di un'applicazione web che offre gli strumenti necessari per creare e gestire dei test, dei questionari, delle verifiche, dei sondaggi e dei quiz; utilizzabile direttamente on-line e eventualmente, i risultati sono stampabili su carta.

Inizialmente il questionario presenta un messaggio iniziale che spiega o l'obiettivo di questo progetto e le istruzioni per il suo svolgimento, quindi il tempo totale richiesto e la necessità di essere madrelingua italiani.

[3] La piattaforma QuestBase https://www.questbase.com/

Nella figura seguente (Figura 3) viene mostrato il messaggio iniziale di questionario:

Ciao a tutti o come dicono in Russia - privet!

Il sondaggio a cui stai per partecipare richiede circa 35 minuti per essere completato.
Prima di proseguire e, quindi, di dare il consenso alla partecipazione, ti spieghiamo brevemente In che cosa consiste.

In questo sondaggio ti mostreremo 120 incipit di testi tratti da blog di viaggio che raccontano un'esperienza di viaggio nello spazio post-sovietico. Ti chiediamo di leggere ciascun incipit e valutare quanto ti piacerebbe proseguire nella lettura del testo. Esprimi il tuo giudizio sulla seguente scala:

1 - per niente, 2 - poco, 3 - abbastanza, 4 - molto, 5 - moltissimo

Nell'assegnare il punteggio, tieni presente che non esiste una risposta giusta o sbagliata: quello che conta è semplicemente quello che tu pensi!

La tua partecipazione al sondaggio è completamente libera. Se in qualsiasi momento dovessi cambiare idea e volessi interrompere il test, sarai libero / a di farlo.

Un'ultima cosa: prima di iniziare il sondaggio, ti chiediamo di darci alcune informazioni anagrafiche su di te, che ci serviranno solo a fini statistici. I dati rimarranno completamente anonimi e in nessun modo le tue risposte verranno associate alla tua persona.

Se hai dubbi, curiosità o chiarimenti puoi mandarmi una mail all'indirizzo:

e.anikeeva@studenti.unipi.it

Grazie e buona lettura!

Figura 3. Messaggio iniziale con le istruzioni per la compilazione del questionario relativoall'interesse suscitato da blog di viaggio.

Al fine di controllare la varietà del campione delle persone intervistate, sono state inoltre richieste come mostrato in Figura 4 le seguenti informazioni personali:

- L'età,
- Il sesso,
- Livello di istruzione.

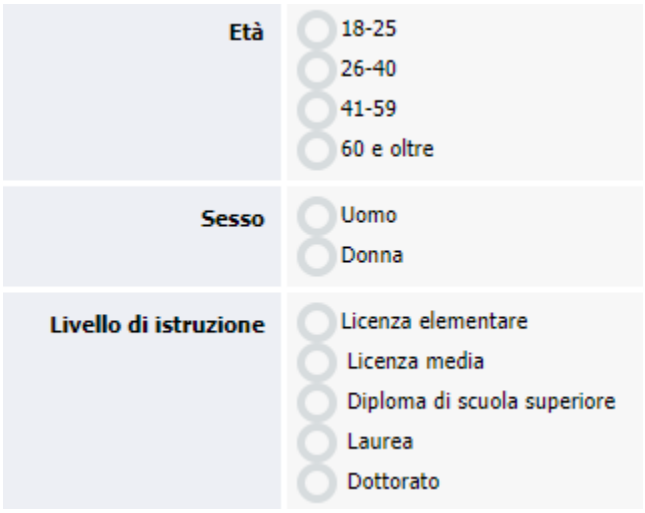

Figura 4. Schermata della pagina del questionario relativa all'inserimento delle informazioni personali.

Successivamente sono stati riportati, dieci per pagina, tutti i testi in ordine casuale e per ognuno di essi è stato chiesto al lettore di leggere con attenzione il testo e di valutare quanto fosse interessato a continuare a leggere quel testo. Ad ogni lettore è stato chiesto di valutare il livello del suo interesse nei confronti di ogni testo sulla base della seguente scala:

per niente – poco – abbastanza – molto – moltissimo.

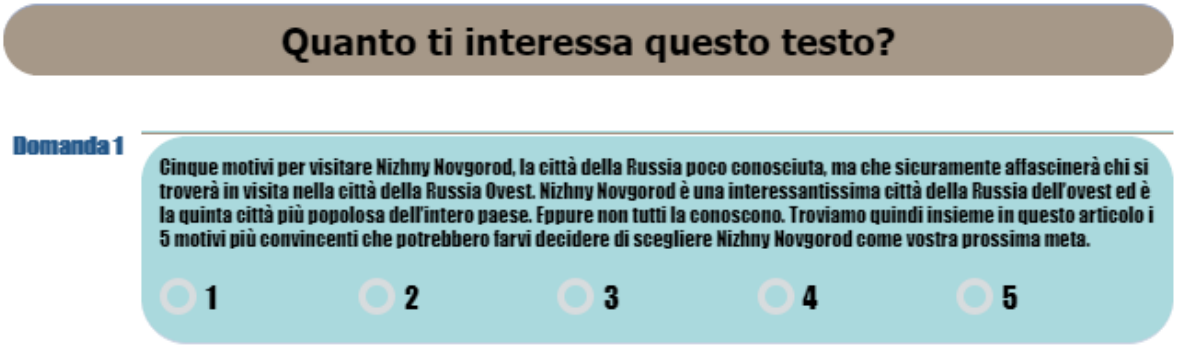

Figura 5. Schermata della pagina del questionario relativa all'inserimento dei giudizi di interesse nei confronti di un'introduzione di un blog di viaggio.

Il questionario è stato condiviso sulla mia pagina personale di *Facebook*, *Instagram* e altri *social network* chiedendo agli utenti di dare un contributo alla mia ricerca

rispondendo alle domande di questionario, e possibilmente di ricondividerlo a loro volta il seguente post.

Figura 6. Schermata della mia pagina su Facebook con la richiesta di contribuire.

2.3 Raccolta delle risposte

Tramite una funzione offerta dalla piattaforma *Questbase*[4] è stato possibile esportare i risultati del sondaggio e salvarli in un file con la estensione .csv. Questo file contiene tutte le informazioni personali richieste al lettore, il suo indirizzo IP, il tempo totale impiegato per la compilazione e i voti che ha espresso per ogni introduzione del testo. Tutti i giudizi, riportati nella scala descritta precedentemente, sono stati convertiti in un valore numerico compreso tra 1 e 5 (dove 1 = per niente e 5 = moltissimo). Alla fine è stato possibile eliminare tutti i tentativi incompleti (quando gli utenti non hanno dato un voto ad un testo) o quelli che avevano una durata inferiore ad 1 ora o 1,5 ora (il tempo minimo necessario per svolgere completamente e efficacemente il questionario con 120 domande).

[4] Pagina di Questionario su QuestBase: https://my.questbase.com/take.aspx?pin=2566-9954-0778

Di seguito vengono illustrati alcuni grafici (Figura 7, 8, 9) che descrivono la variabilità del campione rispetto alle variabili di sfondo raccolte nella fase iniziale.

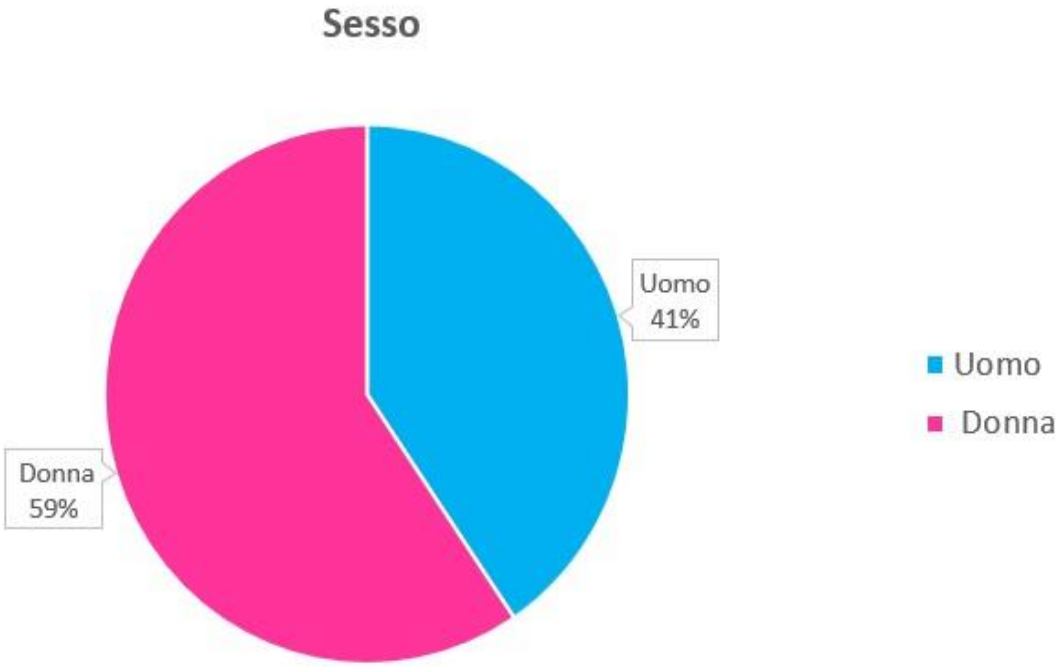

Figura 7. Grafico di percentuale e numero di risposte valide suddivise per il sesso degli utenti.

All'interno del questionario è stata registrata una maggioranza delle donne: sono 59% dei tentativi di valutazione dei blog che sono stati effettuati dalle donne e 41% valutazione effettuata dagli uomini.

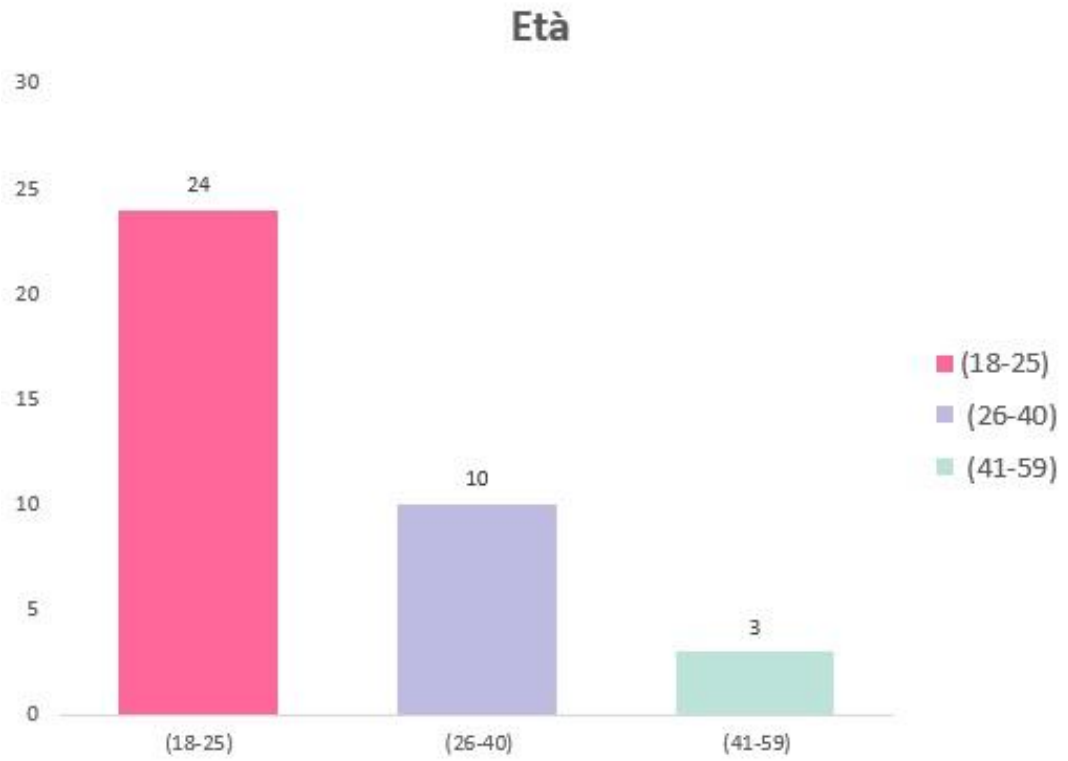

Figura 8. Grafico di numero delle risposte valide suddivise per l'età dei lettori.

Per l'età degli annotatori esiste una prevalenza, all'interno questionario, di una determinata classe d'età degli utenti di 24 anni di età. Da notare che l'età degli utenti che hanno compilato il questionario varia dai 18 ai 25 anni. Ciò può essere dovuto al fatto che la maggior parte degli utenti intervistati sono studenti universitari. Mentre un altro caso è quello di utenti che possedono al più di 60 anni che non hanno partecipato al questionario. Di conseguenza, come è possibile notare dal precedente grafico, non risulta alcuna risposta valida da poter essere considerata all'interno delle mie analisi.

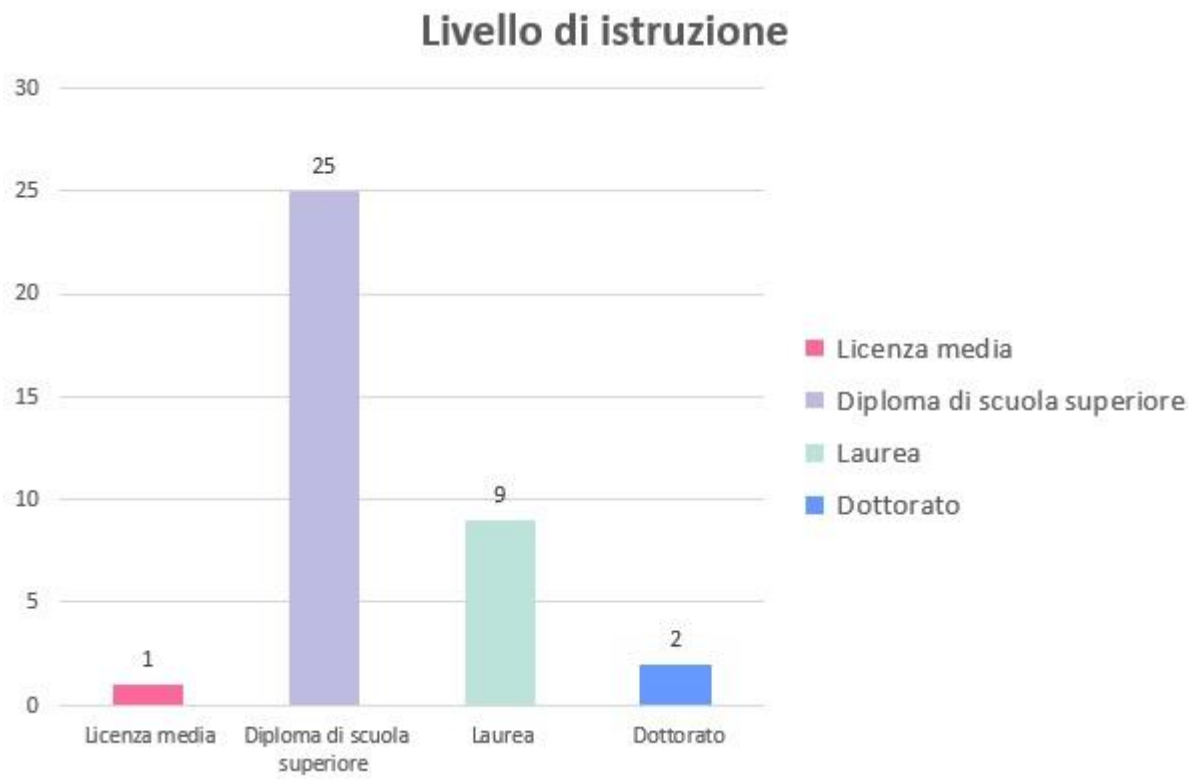

Figura 9. Grafico di numero delle risposte valide suddivise per il grado d'istruzione dei lettori.

Nella Figura 9, possiamo notare che la distribuzione del grado di istruzione degli annotatori è abbastanza varia: la maggioranza delle recensioni data dai lettori possiede un diploma di scuola superiore con una risultante del 25%, invece il l'1% viene rappresentato da chi possiede una licenza media, il 2% è composto da chi ha un dottorato. Infine, il 9% dei lettori è rappresentato da chi possiede una laurea. In conclusione, delle analisi tecniche dei grafici, si registra una predominanza delle lettrici donne, i parametri relativi all'età e al grado di istruzione variano in modo sostanziale da caso a caso.

2.3.1 Calcolo di media e deviazione standard sui giudizi raccolti con il questionario

Per ciascuno dei 120 testi è stata calcolata la "*media aritmetica*" di tutti i giudizi espressi dai lettori per ottenere una prima stima approssimativa del coinvolgimento dei lettori sull'intero campione. In primis è stata calcolata la media del primo gruppo di 60 testi e poi gli altri 60 sullo stesso questionario.
Quindi è stato creato un documento con estensione "*.csv*" contenente:

- Il nome del file .txt che contiene il testo (15_motivi_per_visitare_la_Russia);
- La media aritmetica dei giudizi espressi dagli utenti per ogni testo;
- La deviazione standard.

TESTI	MEDIA	DEVIAZIONE STANDARD
5_motivi_per_andare_in_vacanza_a_Nizhny_Novgorod	2.857	0.557
7_cose_da_sapere_prima_andare_in_Russia	3.679	0.186
10_Luoghi_da_visitare_in_Russia	3.214	0.186
15_motivi_per_visitare_la_Russia	3.964	0.742
Attraverso_la_Russia_Europea	3.214	0.928
Carelia_Kizhi_e_larcipelago_delle_isole_Solovki	3.036	0.186
Cimiteri_comunisti_in_Russia_Mosca_e_San_Pietroburgo	3.393	0.742
Cosa_vedere_a_Mosca	3.429	0.371
Da_Mosca_a_Pechino_in_treno	3.714	0.186
Frammenti_di_San_Pietroburgo	4	0.928

Tabella 1. Tabella con i testi e risultati dei calcoli della media aritmetica e deviazione standard.

Dopo aver raccolto tutti i giudizi assegnati ad ogni testo sono state in totale circa 60 persone di quelle intervistate in una porzione dei 120 testi. La seguente tabella (*Tabella 2*) riporta il valore massimo e quello minimo delle medie dei voti assegnati ai primi 60 testi. Dopo aver fatto la pulizia dei voti non assegnati, il primo gruppo di 60 testi aveva 43 annotatori in totale. In seguito ne sono rimasti 28 veri e propri che hanno dato i voti a tutti i testi in modo da eliminare tutti i tentativi non validi, incompleti o caratterizzati da un tempo di svolgimento inferiore. Tutti i voti assegnati (da 1 a 5) per ogni gruppo dei testi sono state messe in file .csv e sulla base dei criteri di esclusione delle risposte:

1) Non assegnate (un campo vuoto);

2) Grazie alla piattaforma *QuestBase* è stato possibile vedere il tempo dedicato alla compilazione del questionario, quindi sono stati esclusi dalla tabella gli annotatori che hanno messo troppo tempo per compilare il questionario (più di 1 ora e mezzo);

3) Le annotazioni che sono state caratterizzate come estremamente basse o estremamente elevate rispetto alla media dei voti assegnati.

	Blog di viaggio
Media più alta	4
Media più bassa	2,786

Tabella 2. Massimo e minimo delle medie dei voti assegnati al primo gruppo di 60 testi.

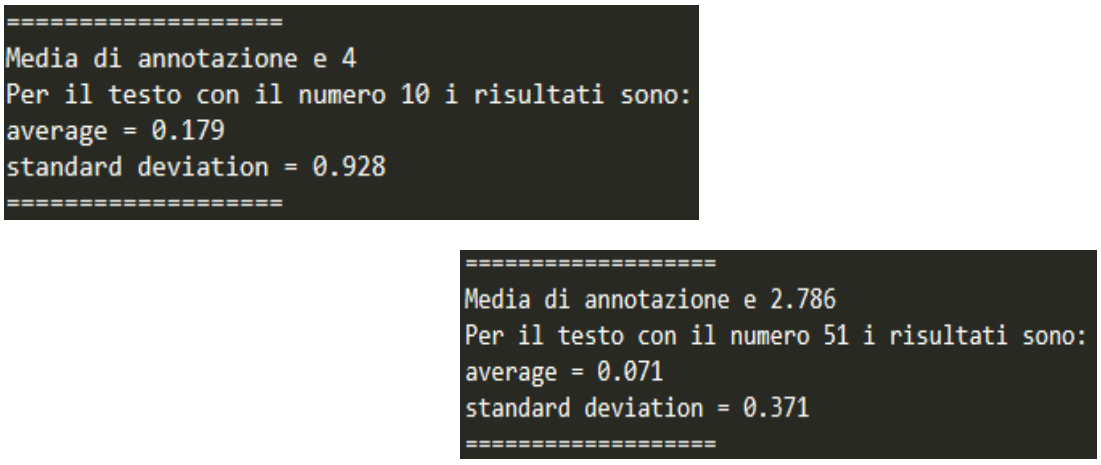

Figura 10. Schermata dei risultati del calcolo per il primo gruppo di 60 testi.

La seguente tabella *Tabella 3* riporta il valore massimo e quello minimo delle medie dei voti assegnati del secondo gruppo di 60 testi. Dopo aver fatto la pulizia dei voti non assegnati, questo gruppo di 60 testi aveva 15 annotatori: di questi ne sono rimasti 13 veri e propri.

	Blog di viaggio
Media più alta	4
Media più bassa	2.923

Tabella 3. Massimo e minimo delle medie dei voti assegnati al secondo gruppo di 60 testi.

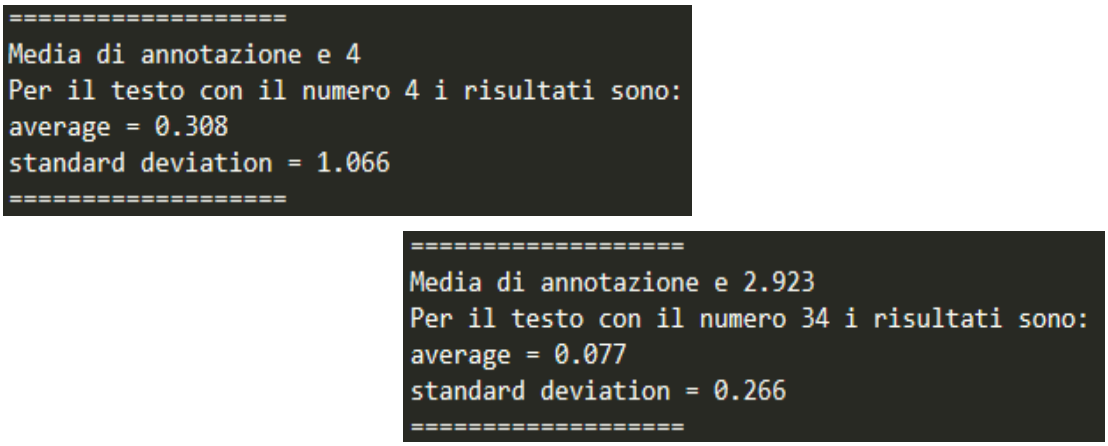

Figura 11. Schermata dei risultati del calcolo per il primo gruppo di 60 testi.

La *deviazione standard* permette di comprendere la dispersione dei dati attorno alla media presenti nel campione. Una deviazione prossima allo 0 è indicativa di un insieme di dati che non mostrano una variazione significativa; se invece il valore è vicino alla media indicherà una maggiore dispersione dei singoli dati.

Lo scarto quadratico medio (deviazione standard) di un carattere rilevato su una popolazione di N unità statistiche si definisce esplicitamente come:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

Formula 1. Lo scarto quadratico.

Dopo aver fatto questa analisi si può vedere quali testi del corpus sono stati valutati *maggiormente interessanti* (ovvero quelli che ottengono un punteggio medio più alto). Nella sezione dei primi 60 testi sono contenuti due testi che hanno ottenuto un valore medio pari a 4, il valore più alto registrato nel corpus. In particolare, si tratta dei testi 10 e 29, che sono stati valutati come più interessanti.
Qui di seguito si riporta il testo 10:

> *"Scriveva, seduta al tavolino di quella vecchia caffetteria, sulla riva del canale Gribaedova, ricavata dal pian terreno di un antico palazzo nobiliare, che, sicuramente, aveva vissuto tempi migliori.. dalla vetrata opaca della porta di ingresso riusciva a scorgere i candidi ed incredibilmente grandi fiocchi di neve che scivolavano lenti attraverso le splendide guglie della Chiesa del Salvatore fin verso la strada ghiacciata."*[5]

Una scrittura così ben dettagliata si trova di solito nel genere narrativo, non in un blog. Pertanto, molto probabilmente, i lettori sono stati incuriositi della particolare descrizione in terza persona di San Pietroburgo. La deviazione standard dei voti assegnati a questo frammento di testo corrisponde a (0.928).
La seconda porzione di testo che è stata valutata come la più interessante è il numero 29:

> *"La sensazione che ti avvolge all'arrivare a Sochi è quella di essere a casa. Le stradine strette di ciottoli delimitate da alberi in fiore, il caldo un po' umido dell'estate in arrivo, il passo rilassato della gente in strada, le vecchine che trasportano le buste della spesa chiacchierando tra loro, gli ampi sorrisi. Mentre ti inerpichi per i sentieri in salita e il cane della casa accanto ti viene incontro per farsi accarezzare non puoi far altro che respirare questo senso di pace, insieme al profumo verde delle piante che divorano il paesaggio."*[6]

Si tratta di una descrizione dettagliata della città di Sochi, che viene descritta come una "seconda casa" in un paese straniero, destando l'attenzione del lettore. La deviazione standard dei voti assegnati a questo testo è uguale al testo precedente,

[5] Tratto da: https://www.simonasacri.com/emozioni/frammenti-san-pietroburgo.php

[6] Tratto da: https://www.scorcidimondo.it/russia-benvenuti-sochi-trote-fiumi-ghiacciati-foreste-mowgli/

ovvero (0.928). Questo valore dimostra come i giudizi di gradimento non varino in modo significativo ma si avvicinino tutti al valore medio.
Il testo valutato meno positivamente è il numero 51 dei primi 60 testi:

> *"Enrico, 43 anni e viaggiatore racconta a PimpMyTrip.it in questa intervista la sua transiberiana da Mosca a Pechino. La transiberiana è, tra tutti, il viaggio per eccellenza. Ho sempre pensato che questo fosse un viaggio di quelli di cui davvero raccontare e per questo avevo una gran voglia di scrivere qualcosa a riguardo, finché un giorno non mi si presenta l'occasione grazie proprio ad Enrico che quando gliel'ho chiesto ha acconsentito a lasciarmi questa intervista sul suo viaggio dei viaggi."*[7]

Questo testo riporta l'intervista di un viaggiatore di nome Enrico sulla sua avventura durante un viaggio nella transiberiana. Il testo non è coinvolgente ed è scritto in modo semplice. La media di questo testo è minore rispetto agli altri due: (2.786), quindi quasi la metà. Inoltre la deviazione standard ha un valore molto basso: (0.371). Quando non c'è dispersione dei dati, la deviazione standard è pari a 0. Si illustra un confronto sui testi appena riportati. Si notano subito alcune differenze specifiche: nei testi valutati come i più interessanti del corpus sono riportate opinioni personali sui luoghi, che suscitano nel lettore interesse e curiosità. Un testo meno dettagliato e poco descritto suscita poco interesse.
In un altro gruppo di 60 testi sono quattro frammenti che hanno ottenuto una media alta, ovvero 4.
Tra tutti i testi presentati, 64, 66, 77, 78 e 95 sono stati valutati come più interessanti. Qui di seguito riporto il testo 64:

> *"Le persone sensibili facciano attenzione, questa tappa ricorda un momento triste della storia con dettagli inquietanti: ad Auschwitz ti aspetta una visita davvero impressionante. Il campo di concentramento di Auschwitz è tristemente celebre per tutte le atrocità che vi hanno avuto luogo, dall'inizio degli anni 40' fino al 1944, poco prima della fine della seconda guerra mondiale."* [8]

[7] Tratto da: hhttps://www.pimpmytrip.it/transiberiana-mosca-pechino-di-enrico/
[8] Tratto da: https://www.evaneos.it/polonia/viaggio/destinazioni/3800-auschwitz/

In questo frammento di blog, l'autore descrive il campo di concentramento di *Auschwitz*. La triste storia legata a questo luogo è probabilmente ciò che ha spinto gli annotatori ad assegnare un alto giudizio di gradimento e interesse a questo testo. Tuttavia, la deviazione standard è più elevata rispetto a quella osservata nel gruppo precedente, ovvero (1.066).
La seconda porzione di testo 66 che ha la media più alta è la seguente:

> *"La Belovezhskaya Pushcha (foresta di Białowieża in polacco) si trova al confine tra Polonia e Bielorussia ed è una delle foreste più antiche d'Europa. L'età media degli alberi di questo parco protetto è di 80 anni, ma molti esemplari hanno anche più di 300 anni. Per il suo valore ecologico, la foresta è ufficialmente riconosciuta anche come World Heritage UNESCO dal 1992."*[9]

Qui un viaggiatore racconta di un bel luogo in Bielorussia che si chiama "*Belovezhskaya Pushcha*": è una foresta di *World Heritage UNESCO* che cattura subito l'attenzione e l'interesse del lettore. A questo testo è stato assegnato il voto massimo.
La terza porzione di testo 77 che ha la media più alta è la seguente:

> *"Natura selvaggia, bellissime spiagge, edifici in legno: in Lettonia troverete questo e molto altro ancora. Infatti, sono innumerevoli le cose da vedere in Lettonia questo splendido Paese degli Stati Baltici offre moltissime attrazioni turistiche e vanta dei paesaggi spettacolari, spesso protetti dai confini dei molti parchi nazionali."*[10]

È un testo che descrive la bellezza della Lettonia in modo molto semplice e coinvolgente. Ha una deviazione standard più alta degli altri testi: (1.332).
La quarta porzione di testo 78 che ha la media più alta è la seguente:

[9] Tratto da: https://www.ilpasseggero.eu/belovezhskaya-pushcha-bielorussia/
[10] Tratto da: https://www.nomavic.it/lettonia-cosa-vedere/

"Tagi che? Probabilmente tra gli 'stan countries il Tagikistan è il meno conosciuto (non che gli altri lo siano eh..), e ammetto che fino a 6 mesi fa io stessa ne ignoravo l'esistenza. Quando poi ho iniziato a studiare la Via della Seta mi sono imbattuta in questo isolato paese dell'Asia Centrale e ho deciso, al contrario, che sarebbe stato il focus del mio intero viaggio."[11]

Il testo numero 78 ha media e deviazione standard uguali a primi due: (4) e (1.066). Queste sono le esperienze del viaggio di Nizhny Tagil in Russia raccontate in prima persona, dando consigli su un paese dell'Asia Centrale.
L'ultima porzione di testo 95 con la media più alta è la seguente:

"Certamente la più bella destinazione ucraina. Leopoli ha un centro assolutamente splendido. A Leopoli c'è qualcosa di decisamente centro-europeo, nelle stradine ciottolate del centro, nei suoi bellissimi edifici barocchi dipinti di tutti i colori, nelle sue belle chiese e nei grandi viali alberati. Come Praga o Cracovia, ma prima della liberalizzazione, prima che queste città venissero invase dall'industria del turismo. "[12]

Il viaggiatore sta raccontando nel suo blog un'esperienza turistica a Leopoli, in Ucraina. È rilevante la descrizione della somiglianza di Leopoli con città europee come Cracovia o Praga. Qui i lettori sono più interessati. La deviazione standard e la media sono simili ai testi precedenti, ovvero (4) e (1.066).

[11] Tratto da: https://www.beborghi.com/tagikistan-pamir-viaggio-asia-centrale/

[12] Tratto da: https://www.evaneos.it/ucraina/viaggio/destinazioni/5678-leopoli/

3. Metodologia di monitoraggio linguistico dei testi

In questo capitolo vengono spiegati i fondamenti della metodologia di monitoraggio linguistico del testo applicata ai testi contenuti nel corpus. Sono illustrate le fasi di annotazione linguistica automatica, prerequisito per la ricostruzione del profilo linguistico dei testi che è stato eseguito tramite *Profiling-UD*[13], un tool sviluppato dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) dall'ItaliaNLP Lab nel 2020. Il paragrafo 3.1 fornisce una panoramica generale sui fondamenti della metodologia di monitoraggio linguistico dei testi, la sua definizione e spiegazione. Nel paragrafo seguente viene spiegato il processo dell'annotazione linguistica automatica, quindi la distinzione tra i tipi di annotazione, e le fasi principali; infine, sarà introdotto lo strumento *Profiling-UD.*

3.1 Fondamenti di metodologia

Da un punto di vista generale, la definizione di monitoraggio e la comprensione dei suoi risultati fanno intendere che questo concetto è utilizzato più spesso quando si tratta dello svolgimento di una serie di azioni specifiche o alcune attività miste svolte sulla base sia di principi teorici che pratici. Per condurre un'indagine di monitoraggio, vengono utilizzati mezzi cognitivi tradizionali quali l'osservazione, la ricerca, l'analisi comparativa, la verifica e infine, il controllo. Come illustrato da *Montemagni (2013),* in ambito linguistico-computazionale, si fa riferimento al monitoraggio linguistico del testo come ad un processo di analisi che si basa sul uso di tecnologie per l'annotazione linguistica automatica che consentono di monitorare un ampio spettro di parametri che riguardano i diversi livelli di descrizione linguistica. Il monitoraggio linguistico si fonda su un'annotazione linguistica multi-livello e, attraverso regole che agiscono sull'output di questa annotazione permette di identificare specifici costrutti sintattici, morfosintattici e informazioni relative alle strutture semantiche. L'annotazione linguistica automatica, che è il prerequisito per ricostruire il profilo linguistico del testo, è un processo incrementale che si compone di una serie di moduli di complessità crescente, quali:

[13] Profiling-UD http://linguistic-profiling.italianlp.it

tokenizzazione, analisi morfo-sintattica, lemmatizzazione, parsing in termini di relazioni di dipendenza sintattica. Il monitoraggio linguistico si applica in una varietà di casi applicativi in cui è necessario studiare lo *stile del testo*, ad esempio per classificarne il genere testuale o il livello di leggibilità in base ad una determinata categoria di lettori. Nel mio studio la metodologia di monitoraggio linguistico è stata implementata attraverso il tool di *Profiling-UD*.

3.1.1 L'Annotazione linguistica automatica

Esistono due tipi di annotazione:

1. annotazione automatica, tramite un programma automatizzato;
2. annotazione manuale, eseguita da linguisti esperti.

L'analisi della struttura linguistica di un testo avviene solitamente con un grado di complessità crescente. Per ciascuna fase di annotazione il risultato dell'analisi precedente viene utilizzato come dato di input per il successivo. Le principali fasi di annotazione sono le seguenti:

1. *Sentence splitting*: divisione del testo in frasi;
2. *Tokenizzazione*: divisione del testo in *tokens* - unità minima del testo digitale;
3. *Lemmatizzazione* e *analisi morfo-sintattica*: assegnazione ad ogni *token* (unità minima) del testo l'informazione che riguarda la categoria grammaticale della parola in un contesto specifico, e poi l'associazione di ogni parola del testo al suo lemma;
4. *Parsing*: l'analisi di una stringa nei simboli, nel linguaggio naturale, nei linguaggi informatici o nelle strutture di dati, conforme alle regole di una grammatica formale. Gli approcci al *parsing* sintattico si distinguono in due grandi tipologie: *parsing* a costituenti e *parsing* a dipendenze. In questo studio, la struttura sintattica è stata formalizzata in base alla rappresentazione sintattica a dipendenze.

Come anticipato al inizio del paragrafo, a partire dal testo automaticamente annotato, è stato possibile applicare le tecnologie di monitoraggio per ricostruire il

profilo linguistico del testo. Sia l'annotazione che il monitoraggio sono stati eseguiti tramite lo strumento *Profiling-UD*, descritto nel prossimo paragrafo.

Profiling-UD è un'applicazione *web-based* ispirata alla metodologia inizialmente presentata da *Montemagni (2013)*, sviluppata per effettuare la profilazione linguistica di un testo, o di un'ampia raccolta di testi, per più lingue. Consente l'estrazione di oltre 130 *features* su diversi livelli di descrizione linguistica, ed è stato specificamente concepito per essere multilingue poiché si basa sul framework delle dipendenze universali (*UD*) [14]. Si tratta di uno sforzo comunitario aperto con oltre 300 contributori che producono quasi 200 *treebanks* in oltre 100 lingue.

L'interfaccia *Profiling-UD* offre all'utente la possibilità di aggiungere il testo nell'area proposta oppure caricare un file di testo *.txt* o lo *.zip* (collezione dei testi in una cartella).

Lo strumento implementa un processo strutturato in due fasi:

1. annotazione linguistica,
2. profilazione linguistica.

Come accennato in precedenza, l'annotazione dei testi viene eseguita dalla catena di analisi *UDPipe*[15], che utilizza i modelli addestrati sulle *treeban*k *UD* disponibili, versione 2.5, per la lingua di input. Per ogni testo caricato, il risultato dell'annotazione è un file nel formato *"CoNLLU–tab–separated"*. I testi annotati automaticamente vengono utilizzati come input per il passaggio successivo, eseguito dal componente di profilazione linguistica che definisce le regole per estrarre e quantificare le proprietà formali.

[14] Lo strumento è accessibile a questo link: http://www.italianlp.it/demo/profiling-UD/

[15] Dipendenze universali (UD) http://lindat.mff.cuni.cz/services/udpipe/

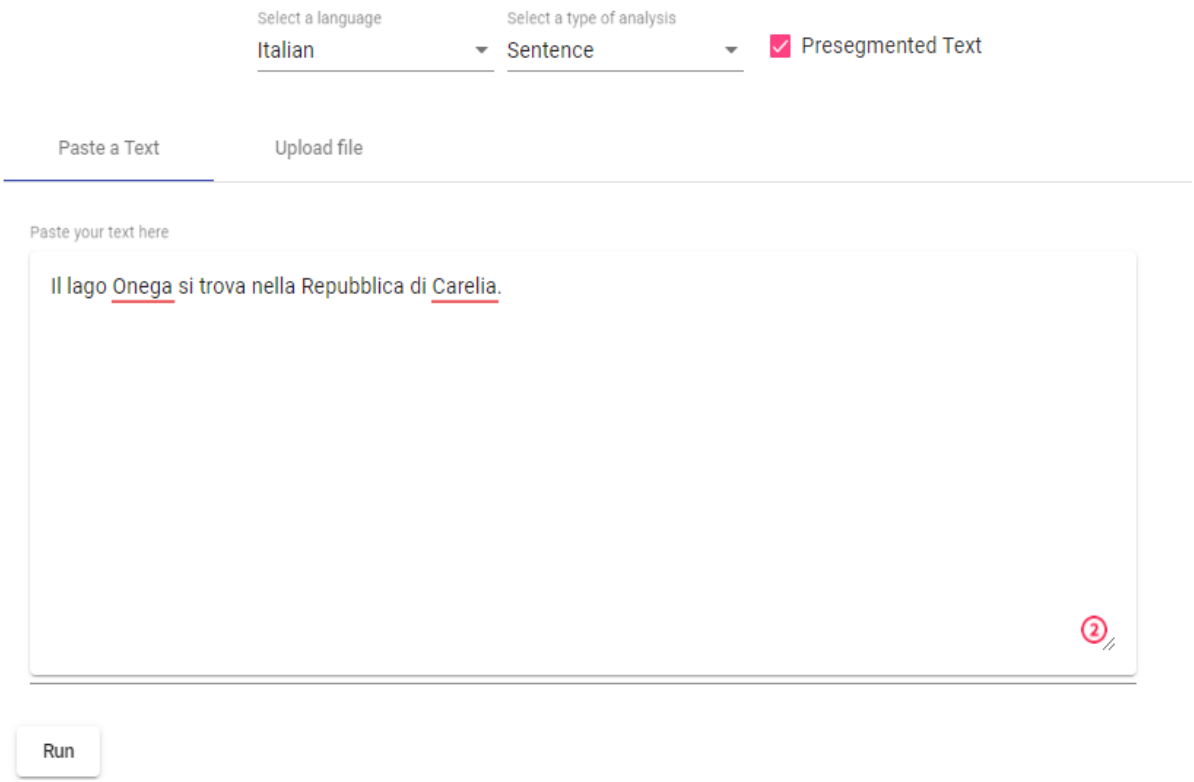

Figura 12. Profiling-UD tool per annotazione linguistica automatica.

Prima viene scelta la lingua con quale bisogna lavorare: nel mio caso ho scelto la lingua italiana. In seguito viene selezionato un documento o una frase da inserire. Avviato il tool, viene eseguito il processo di annotazione linguistica con l'input indicato come un testo semplice o caricato direttamente un file come archivio compresso di file con estensione *.txt* e codifica *UTF-8*.

I 120 testi selezionati sono stati prima inseriti in una cartella, poi compressa e caricata sul *Profiling-UD.*

L'output per il testo selezionato si compone di tre file :

Figura 13. I file di output prodotti da Profiling-UD rispetto al corpus analizzato.

1. Annotazione linguistica: file scaricato come un file di testo con estensione *.txt*:

```
# newdoc
# newpar
# sent_id = 1
# text = Il lago Onega si trova nella Repubblica di Carelia.
1       Il      il      DET     RD      Definite=Def|Gender=Masc|Number=Sing|PronType=Art       2       det     _       _
2       lago    lago    NOUN    S       Gender=Masc|Number=Sing 5       nsubj   _       _
3       Onega   Onega   PROPN   SP      _       2       nmod    _       _
4       si      si      PRON    PC      Clitic=Yes|Person=3|PronType=Prs        5       expl    _       _
5       trova   trovare VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       _
6-7     nella   _       _       _       _       _       _       _
6       in      in      ADP     E       _       8       case    _       _
7       la      il      DET     RD      Definite=Def|Gender=Fem|Number=Sing|PronType=Art        8       det     _       _
8       Repubblica      Repubblica      PROPN   SP      _       5       obl     _       _
9       di      di      ADP     E       _       10      case    _       _
10      Carelia Carelia PROPN   SP      _       8       nmod    _       SpaceAfter=No
11      .       .       PUNCT   FS      _       5       punct   _       _

# sent_id = 2
# text = La sua superficie occupa oltre 9.000 km².
1       La      il      DET     RD      Definite=Def|Gender=Fem|Number=Sing|PronType=Art        3       det     _       _
2       sua     suo     DET     AP      Gender=Fem|Number=Sing|Poss=Yes|PronType=Prs    3       det:poss        _       _
3       superficie      superficie      NOUN    S       Gender=Fem|Number=Sing  4       nsubj   _       _
4       occupa  occupare        VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       _
5       oltre   oltre   ADV     B       _       6       advmod  _       _
6       9.000   9.000   NUM     N       NumType=Card    7       nummod  _       _
7       km²     km²     PROPN   SP      _       4       obj     _       SpaceAfter=No
8       .       .       PUNCT   FS      _       4       punct   _       _

# sent_id = 3
# text = Le sue rive sono molto frequentate dagli escursionisti e la temperatura dell'acqua può arrivare fino a 24°C nel mese di agosto.
1       Le      il      DET     RD      Definite=Def|Gender=Fem|Number=Plur|PronType=Art        3       det     _       _
2       sue     suo     DET     AP      Gender=Fem|Number=Plur|Poss=Yes|PronType=Prs    3       det:poss        _       _
3       rive    riva    NOUN    S       Gender=Fem|Number=Plur  6       nsubj:pass      _       _
4       sono    essere  AUX     V       Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   6       aux:pass        _       _
```

Figura 14. Esempio di una rappresentazione dell'annotazione di una frase presa da questionario.

Si illustra un esempio di annotazione di una frase presa da un blog di viaggio:

"*Il lago Onega si trova nella Repubblica di Carelia.*"

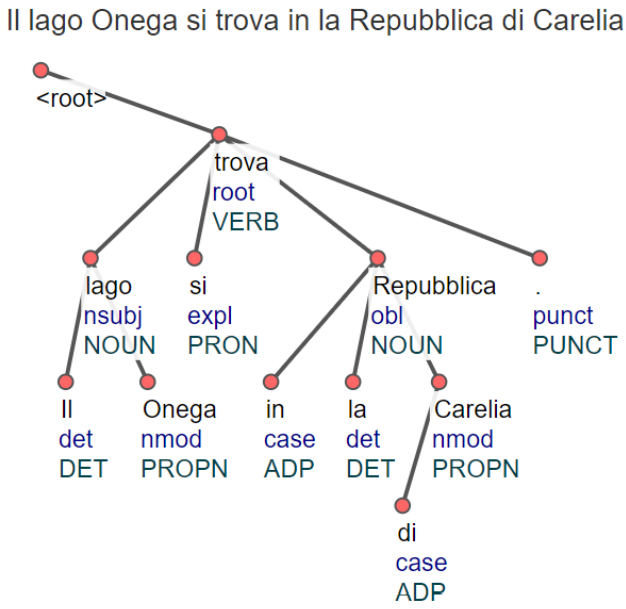

Figura 15. Esempio rappresentazione grafica dell'albero sintattico.

Ad ogni riga della tabella corrisponde un *token*, mentre le colonne indicano le proprietà del *token* a diversi livelli di analisi.
La Figura 15 mostra un esempio di visualizzazione della annotazione linguistica di una frase del testo. L'albero sintattico è composto da nodi e rami. In particolare, ha una struttura che parte dalla radice (*root*) e termina con le foglie.

2. il file scaricato è una cartella Excel con estensione *.xlsx*:

Filename	n_sentences	n_tokens	tokens_per_sent	char_per_tok
input_text_1.conllu	1	11	11.0	4.1

Figura 16: Esempio di un estratto dell'output del monitoraggio linguistico scaricato da Profiling-UD.

Nel file *.xlsx* sono presenti le caratteristiche di base:

- Il numero di frasi (*n_sentences*);
- Il numero totale dei token nel testo (*n_tokens*);
- La lunghezza media della frase in tokens (*tokens_per_sent*);
- La lunghezza media delle parole in caratteri (*char_per_tok*).

3. Legenda di un profilo linguistico: il file scaricato è un file di testo con estensione *.txt*:

```
>>> Raw Text Properties:

[n_sentences]: total number of sentences
[n_tokens]: total number of tokens
[tokens_per_sent]: average length of sentences in a document, calculated in terms of the number of words per sentence
[char_per_tok]: average number of characters per word (excluded punctuation)

>>> Lexical Variety:

[ttr_lemma_chunks_100]: Type/Token Ratio (TTR) calculated with respect to the lemmata in first 100 tokens of a document. It ranges between 1
[ttr_lemma_chunks_200]: Type/Token Ratio (TTR) calculated with respect to the lemmata in first 200 tokens of a document. It ranges between 1
[ttr_form_chunks_100]: Type/Token Ratio (TTR) calculated with respect to the word forms in first 100 tokens of a document. It ranges between
[ttr_form_chunks_200]: Type/Token Ratio (TTR) calculated with respect to the word forms in first 200 tokens of a document. It ranges between


>>> Morphosyntactic information:

[upos_dist_*]: distribution of the 17 core part-of-speech categories defined in the Universal POS tags, as detailed at the following link: ht

[lexical_density]: the value corresponds to the ratio between content words (nouns, proper nouns, verbs, adjectives, adverbs) over the total

Inflectional morphology:

[verbs_tense_dist_*]: distribution of verbs according to their tense: https://universaldependencies.org/u/feat/Tense.html
[verbs_mood_dist_*]: distribution of verbs according to their moods: https://universaldependencies.org/u/feat/Mood.html
[verbs_form_dist_*]: distribution of verbs according to their forms: https://universaldependencies.org/u/feat/VerbForm.html
[verbs_gender_dist_*]: distribution of verbs according to the gender of participle forms, for the languages that have this features: https://
[verbs_num_pers_dist_*]:  distribution of verbs according to their number and person: https://universaldependencies.org/u/feat/Person.html

>>> Syntactic features:

Verbal Predicate Structure:
```

Figura 17. Esempio di file di un profilo linguistico con la legenda scaricato da Profiling-UD.

Questo esempio (Figura 17) mostra un insieme delle caratteristiche linguistiche monitorate da *Profiling-UD*, come illustrato nell'articolo di *Brunato (2020)*, in cui tutte le caratteristiche linguistiche estratte sono raggruppati in 7 macro-categorie di fenomeni linguistici, riportati nella Figura 18 e 19:

1. Raw Text Properties	2. Lexical Variety	3. Morpho–syntactic information	4. Verbal Predicate Structure
Lunghezza del documento	Type/Token Ratio (TTR)	Distribuzione delle categorie grammaticali	Distribuzione delle teste verbali
Lunghezza della frase		Densità lessicale	Distribuzione delle radici verbali
Lunghezza della parola		Morfologia flessiva	Varianza dei verbi

Figura 18. Tipologia di fenomeni linguistici monitorati nel testo tramite Profiling-UD (prima parte di tabella).

5. Global and Local Parsed Tree Structures	6. Syntactic Relations	7. Subordination phenomena
Profondità media dell'albero sintattico	Distribuzione dei rapporti di dipendenza	Distribuzione delle proposizioni subordinate e principali
Lunghezza media delle clausole		Ordine relativo dei subordinati rispetto al capo verbale
Lunghezza dei collegamenti di dipendenza		Profondità media delle clausole subordinate incorporate
Profondità media delle catene		
Fenomeni di ordine delle parole		

Figura 19. Tipologia di fenomeni linguistici monitorati nel testo tramite Profiling-UD (seconda parte di tabella).

1. Il primo gruppo rappresenta le caratteristiche di base del testo come la lunghezza del documento, della frase e della parola;
2. Il secondo gruppo rappresenta la varietà lessicale. La *Type Token Ratio* mette a rapporto il numero totale di parole uniche (parole tipo) con quello di parole complessive *(token*);
3. Il terzo gruppo rappresenta le caratteristiche morfosintattiche estratte dall'analisi morfosintattica, quali la distribuzione di tutte le categorie grammaticali;
4. Il quarto gruppo include proprietà legate alla struttura dei predicati verbali che sono solitamente rappresentati da una forma verbale finita, o "lessicale".
5. Il quinto gruppo rappresenta le proprietà derivanti dalla struttura ad albero, sia globali che locali: la profondità dell'albero sintattico, la lunghezza media delle catene di dipendenza, fenomeni legati all'ordine sintattico delle parole (es. soggetto, oggetto) etc..
6. Il sesto gruppo a sua volta include le proprietà quali la distribuzione delle relazioni sintattiche definite dallo schema *UD*.

7. L'ultimo gruppo, settimo, descrive proprietà legate al fenomeno di subordinazione: ad esempio, la distribuzione delle proposizioni subordinate e principali e la lunghezza media delle catene di subordinazione.

3.1.2 Monitoraggio linguistico usando lo strumento Profiling-UD

Infine, viene qui mostrato un estratto i risultati della distribuzione di queste caratteristiche rispetto ai testi del corpus analizzato per eseguire il monitoraggio linguistico, sono stati creati due file *.zip* rispettivamente con il primo gruppo di 60 testi e secondo di 60 testi. Tramite il tool *Profiling-UD* sono stati scaricati due file con estensione *.xlsx* che contengono tutti i dati relativi al profilo linguistico e all'annotazione linguistica per le due parti di corpus:

Filename	n_sentences	n_tokens	tokens_per_sent	char_per_tok
testi/Un_viaggio_in_Siberia.conllu	3	99	33.0	4.823529411764706
testi/Da_Mosca_a_Pechino_in_treno.conllu	2	71	35.5	4.298507462686567
testi/Scopri_Mosca_e_San_Pietroburgo.conllu	7	108	15.428571428571429	4.872340425531915
testi/Scopri_Il_lago_Onega.conllu	4	80	20.0	4.068493150684931
testi/Attraverso_la_Russia_Europea.conllu	2	80	40.0	4.257575757575758
testi/Russia_viaggio_a_Mosca.conllu	6	120	20.0	4.4495412844036695
testi/Cimiteri_comunisti_in_Russia_Mosca_e_	3	114	38.0	4.660194174757281
testi/La_città_Astrakhan.conllu	3	113	37.666666666666664	4.204081632653061
testi/La_città_Salekhard_Siberia.conllu	3	102	34.0	4.7032967032967035
testi/Transiberiana_da_Mosca_a_Pechino.conl	3	91	30.333333333333332	4.928571428571429

Figura 20. File Excel scaricato da Profiling-UD che riguarda primo gruppo di 60 testi (oltre 100 features linguistiche).

Filename	n_sentences	n_tokens	tokens_per_sent	char_per_tok
altri 60 testi/10_Luoghi_da_visitare_in_Lettonia.conllu	4	100	25.0	4.609195402298851
altri 60 testi/Come_organizzare_un_viaggio_in_Georgia.conllu	4	88	22.0	4.6075949367088604
altri 60 testi/VIAGGIO_IN_ESTONIA_ESCURSIONE_AL_LAHEMA/	3	93	31.0	4.174418604651163
altri 60 testi/Armenia.conllu	4	117	29.25	4.4672897196261685
altri 60 testi/Crimea_Ucraina.conllu	5	112	22.4	4.989583333333333
altri 60 testi/Viaggio_zaino_in_spalla_in_Armenia.conllu	2	98	49.0	5.235955056179775
altri 60 testi/Vilnius_10_cose_da_fare_e_vedere_nella_Capital	2	89	44.5	5.012987012987013
altri 60 testi/Estonia.conllu	2	73	36.5	4.621212121212121
altri 60 testi/COSA_FARE_E_VEDERE_A_RIGA_LETTONIA.conllu	2	62	31.0	4.702127659574468
altri 60 testi/Turkmenistan_in_bici.conllu	2	83	41.5	5.108108108108108

Figura 21. File Excel scaricato da Profiling-UD che riguarda secondo gruppo di 60 testi (oltre 100 features linguistiche).

La prima colonna "*Filename*" contiene 60 porzioni di testo con un identificatore univoco. Ogni porzione di testo è scaricata in formato *CONLL-U.*
Invece le altre colonne contengono più di 100 caratteristiche linguistiche (*features*) del corpus presentato. Per esempio, le features di base che coprono prime 4 colonne della tabella sono: (*n_sentences, n_tokens, tokens_per_sent, char_per_tok)* dei quali parlerò nel terminata caratteristica linguistica.

3.2 Analisi delle distribuzioni riscontrate nei testi

Come mostrato nelle Figure 20 e 21 l'output del monitoraggio linguistico prodotto da *Profiling-UD* corrisponde ad un vettore numerico, in cui ogni valore rappresenta una determinata caratteristica linguistica. Per ogni testo sono state estratte più di 100 caratteristiche linguistiche, suddividendole in quattro categorie:

- L'Analisi delle caratteristiche linguistiche di base;
- L'Analisi delle caratteristiche lessicali;
- L'Analisi delle caratteristiche linguistiche morfo-sintattiche;
- L'Analisi delle caratteristiche linguistiche sintattiche.

In quanto segue si riportano i valori della media e della deviazione standard nel corpus per una selezione di queste caratteristiche.

3.2.1 Analisi delle caratteristiche linguistiche di base

La prima fase della catena di analisi linguistica è segmentazione in frasi e *tokenizzazione* ovvero segmentazione del testo in parole ortografiche o tokens (unità minime).
Per ogni feature di base (*n_sentences, n_tokens, tokens_per_sent, char_per_tok),* è stata calcolata la *media aritmetica* dei valori e la *deviazione standard* rispetto ai valori relativi ai testi del corpus. Questi risultati vengono riportati nella tabella seguente *Tabella 5.*

Features di base	Primo gruppo di 60 testi		Secondo gruppo di 60 testi	
	Media aritmetica	**Deviazione standard**	**Media aritmetica**	**Deviazione standard**
Totale frasi	3.2	1.1	3.4	1.3
Totale tokens	87.4	15.3	90.4	14.8
Lunghezza media della frase	28.7	8.8	29.8	13.3
Lunghezza media della parola	4.1	0.5	4.08	0.2

Tabella 5. Media aritmetica e deviazione standard delle caratteristiche di base calcolate sulla prima e seconda sezione dei testi.

Essa rappresenta le caratteristiche di base del testo come la lunghezza del documento, della frase e della parola. Come era indicato prima, ogni corpus contiene solo l'introduzione per sondare il coinvolgimento del lettore, non tutto il testo del blog. Quindi le caratteristiche sono state estratte da piccole parti dei testi, e per questo motivo il valore del numero totale di frasi nel testo non sono molto alti: nel corpus dei primi 60 testi il numero medio di frasi per frammento è di 3.293, mentre nel secondo gruppo è di 3.483.

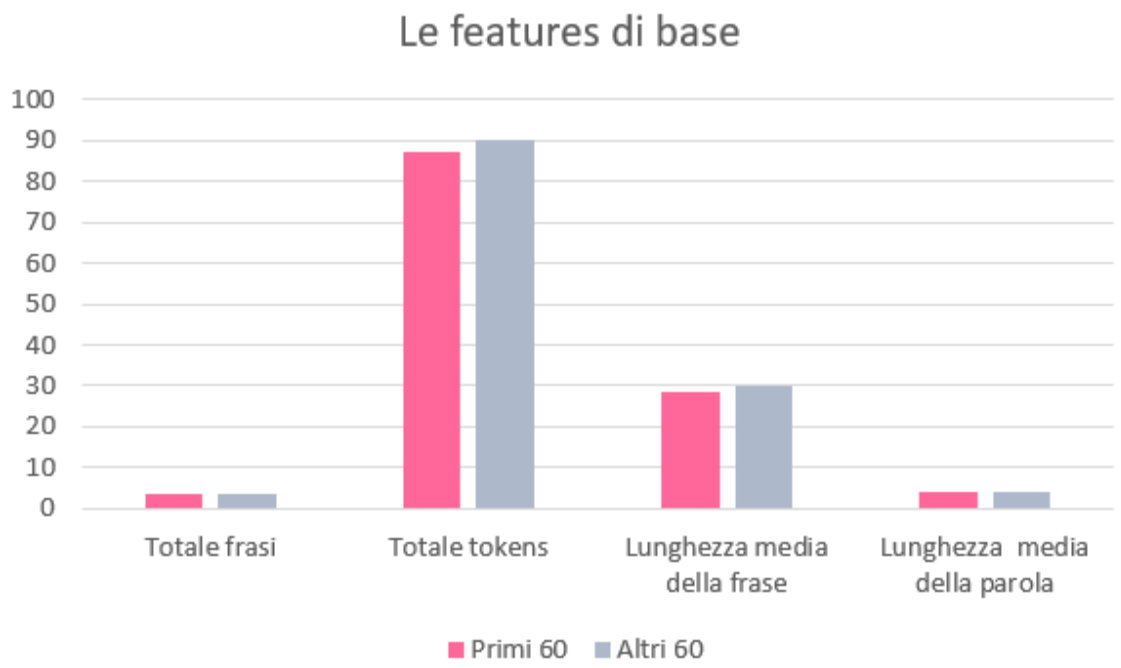

Figura 22. Grafico di media aritmetica delle caratteristiche di base calcolate sui testi dei due sezioni del corpus.

Nella Figura 22 per quanto riguarda la lunghezza media delle frasi dei testi, i due gruppi sono piuttosto omogenei: il secondo gruppo mostra una lunghezza media solo leggermente maggiore (29.896) rispetto al primo gruppo (28.414). Questa similarità fra i due gruppi è probabilmente dovuta al fatto che i testi sono stati selezionati tenendo conto di criteri legati alla lunghezza dei testi: il corpus infatti contiene solo testi contenenti circa 60 parole (tokens) e 400 caratteri.

3.2.2 Analisi delle caratteristiche lessicali

Il prossimo passo è il monitoraggio linguistico del testo, che richiede l'analisi della *lemmatizzazione* e *dell'annotazione morfo-sintattica*. Tale analisi permette di acquisire tratti legati al profilo lessicale del testo. La caratteristica lessicale monitorata, presentata nel file scaricato da *Profiling-UD* è la *Type/Token Ratio (TTR),* che rappresenta un indice di ricchezza lessicale e misura la varietà di parole diverse contenute in un testo. È il rapporto tra il numero di parole tipo e il totale di occorrenze di unità del vocabolario. Può assumere valori compresi tra 0 e 1, dove valori prossimi allo 0 indicano un testo con un vocabolario poco variegato, mentre

valori tendenti a 1 contraddistinguono testi lessicalmente variegati (se è pari a 1 il testo è formato esclusivamente delle parole hapax - parole che ricorrono solo una volta). Quindi nel caso della mia ricerca la *TTR* rappresentata non è un indicatore affidabile, perché i testi analizzati sono molto brevi (circa due frasi).

3.2.3 Analisi delle caratteristiche morfo-sintattiche

Le caratteristiche morfosintattiche monitorate derivano dal livello di annotazione morfosintattica che, come detto nel paragrafo precedente, ha lo scopo di fornire informazioni che riguardano le categorie grammaticali (dette *Part of Speech*) di ogni unità minima (*token*) del testo nel contesto specifico in cui la parola compare. Pertanto, per ogni *token*, viene prima creata la categoria che corrisponde alla caratteristica morfo-sintattica, assegnata in base ai tag *dell'Universal Dependencies Part-of-Speech tagset*[16]. La distribuzione statistica di *Universal POS tag* è calcolata sia rispetto alle categorie contemplate nello *Universal Part of Speech tagset* sia rispetto alle categorie morfo-sintattiche definite dal precedente schema di annotazione usato per la lingua italiana, che si basa sul *tagset* definito nell'ambito del progetto *TANL (Text Analytics and Natural Language processing) (ILC-CNR e Università di Pisa).*

In seguito, vengono mostrate le *Part of Speech*[17] contemplate nello *Universal Dependencies tag* cui indici iniziano con "*upos"*-).

[16] Il tagset dell'Universal Part of Speech https://universaldependencies.org/u/pos/index.html

[17] La lista completa delle Part of Speech http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

- Le classi aperte di parole solo le seguenti:
  - **ADJ**: aggettivi;
  - **ADV:** avverbi;
  - **INTJ**: interiezioni (o esclamazioni);
  - **NOUN**: nomi;
  - **PROPN**: nomi propri;
  - **VERB**: verbi.
- Invece, le classi chiuse di parole sono le seguenti:
  - **ADP**: apposizioni;
  - **AUX**: ausiliari;
  - **CCONJ**: congiunzioni di coordinazione;
  - **DET:** articoli determinativi;
  - **NUM**: numerali;
  - **PART**: particelle grammaticali;
  - **PRON**: pronomi;
  - **SCONJ**: congiunzioni di subordinazione.
- Altre classi presentate sono le seguenti:
  - **PUNCT**: punteggiatura;
  - **SYM**: simboli;
  - **X**: altro.

Nella tabella seguente *(Tabella 6)* sono riportate *le medie* e *le deviazioni standard* delle principali caratteristiche morfo-sintattiche prese in considerazione per il corpus. L'etichetta *(upos_dist_*)* rappresenta la distribuzione delle 11 principali categorie di *part-of-speech* definite nel *Universal POS tags.*

Features morfo-sintattiche	Primo gruppo di 60 testi		Secondo gruppo di 60 testi	
	Media aritmetica	**Deviazione standard**	**Media aritmetica**	**Deviazione standard**
upos_dist_NOUN	15.3	3.9	15.3	3.9
upos_dist_ADP	14.7	3.1	14.7	3.1
upos_dist_DET	14.5	2.6	14.5	2.6
upos_dist_NUM	1.05	1.2	1.1	1.2
upos_dist_PUNCT	9.9	2.7	9.5	2.7
upos_dist_VERB	7.55	3.03	8.5	3.1
upos_dist_ADJ	3.2	1.1	6.3	3.1
upos_dist_PROPN	5.5	3.1	6.3	4.1
upos_dist_ADV	5.7	2.8	5.7	2.8
upos_dist_AUX	4.2	0.05	3.5	2.1
upos_dist_PRON	3.8	2.3	4.4	2.9

Tabella 6. Medie e deviazioni standard delle principali features morfo-sintattiche nel corpus di blog di viaggio.

Il valore della *media aritmetica* è più elevata in alcuni parametri. Per esempio, la media in assoluto più elevata (15.3448) è la distribuzione percentuale di nomi (*upos_dist_NOUN*). Non sono presenti casi in cui la deviazione standard è più elevata della media aritmetica. Quest'ultima può essere considerata come il valore più alto presentato nella tabella. Un'altra caratteristica importante che viene monitorata grazie al livello di annotazione morfosintattica è la distribuzione dei verbi rispetto ai tratti di tempo, modo e persona. Queste informazioni sono rese disponibili nel *tagset* di annotazione morfosintattico *UD* in una colonna dedicata del formato *CoNLL-U*, che si riferisce alle "*features*" morfosintattiche. Di seguito è riportata la legenda del *set* di *tag* della lingua italiana fornita *dall'Universal Dependencies* per queste caratteristiche morfo-sintattiche necessarie.

- Forma verbale:
  - **Fin**: finito,
  - **Inf**: infinito,
  - **Part**: participio,
  - **Ger**: gerundio.

- Modo verbale:
  - **Ind**: indicativo,
  - **Imp**: imperativo,
  - **Cnd**: condizionale,
  - **Sub**: congiuntivo.

- Tempo verbale:
  - **Past**: passato (prossimo e remoto),
  - **Pres**: presente,
  - **Fut**: futuro,
  - **Impo**: imperfetto.

- Genere:
  - **Masc**: maschile,
  - **Fem**: femminile.

La tabella seguente *(Tabella 7)* mostra i valori delle medie e delle deviazioni standard delle principali features che riguardano il modo, il tempo, il genere e il numero dei verbi:

Features (tempo, genere e il numero dei verbi)	Primo gruppo di 60 testi		Secondo gruppo di 60 testi	
	Media aritmetica	**Deviazione standard**	**Media aritmetica**	**Deviazione standard**
v_mood_dist_Ind	84.9	29.4	84.4	**31.3**
v_tense_dist_Past	44.7	29.8	42.5	**29.3**
v_form_dist_Part	31.8	21.8	28.2	**20.6**
v_tense_dist_Pres	46.1	28.4	84.4	**31.3**
v_form_dist_Fin	40.5	23.4	39.5	**20.8**
v_num_pers_dist_+Ger	2.7	5.8	1.8	**4.3**

v_num_pers_ dist_Sing+1	9.8	25.2	5.7	**14.7**
v_num_pers_ dist_Plur+3	17.1	27.1	16.1	**24.7**
v_num_pers_ dist_Sing+3	50.1	37.7	55.8	33.1

Tabella 7. Medie e deviazioni standard delle principali features riguardanti il modo, il tempo, il genere e il numero dei verbi.

Caratteristiche significative sono presentate con i valori di media ed alta deviazione standard (*v_mood_dist_Ind)*, che rappresenta i valori relativi della distribuzione percentuale dei verbi coniugati con il modo *Ind*: indicativo. Invece, per quanto riguarda i tempi verbali sono risultati più frequenti il presente, ma non il passato (*v_tense_dist_Pres)*. Invece, nella seconda porzione dei testi media aritmetica (84,4344) e deviazione standard sono più alti della prima (31,3689), proprio perché sono frequenti frasi che riportano informazioni su eventi espressi al tempo presente come:

> «*… Mentre ti inerpichi per i sentieri in salita e il cane della casa accanto ti viene incontro per farsi accarezzare non puoi far altro che respirare questo senso di pace, insieme al profumo verde delle piante che divorano il paesaggio*».[18]

I viaggiatori stanno raccontando la loro esperienza in uno spazio post-sovietico al presente affinché i lettori siano immersi nella storia e si crei un'impressione di presenza scenica grazie ai tempi reali utilizzati nella frase.

Un'ultima caratteristica monitorata a questo livello è la densità lessicale (*lexical_density*), che esprime la relazione tra parole semanticamente complete (cioè *nomi*, *verbi*, *aggettivi*), chiamate anche "*parole di contenuto*", e il numero totale di token nel testo. Il suo valore varia da 0 a 1: maggiore è la densità lessicale, più il testo sarà informativo e talvolta più complesso.

In seguito, viene presentato il risultato ottenuto con i valori calcolando la media aritmetica e la deviazione standard della densità lessicale di entrambe le porzioni dei testi *(Tabella 8).*

[18] Trattato da: https://www.scorcidimondo.it/russia-benvenuti-sochi-trote-fiumi-ghiacciati-foreste-mowgli/

Densità lessicali	Primo gruppo di 60 testi		Secondo gruppo di 60 testi	
	Media aritmetica	**Deviazione standard**	**Media aritmetica**	**Deviazione standard**
lexical_density	0,4	0,04	0,5	0,04

Tabella 8. Medie aritmetiche e deviazioni standard rispetto la densità lessicale in due gruppi di testi.

Come previsto, questo parametro è piuttosto alto nella seconda porzione dei testi, quindi una deviazione standard molto bassa (0,0414).

3.2.4 Analisi delle caratteristiche linguistiche sintattiche

L'ultima fase del monitoraggio linguistico è stata fatta a partire dall'annotazione linguistica sintattica del testo, utilizzando le tecnologie linguistiche e computazionali descritte.
L'obiettivo dell'annotazione sintattica è analizzare i parametri più complessi e informativi della struttura grammaticale del testo, ad esempio:

• **Struttura interna del periodo:**
- il numero di proposizioni per il periodo;
- la distribuzione di frasi di base, principali;
- la distribuzione di frasi subordinate.

• **Formulazione interna della proposizione Struttura dell'albero sintattico:**
- il numero di parole in una frase;
- la struttura profondità dell'albero sintattico;
- il numero medio di dipendenti per testa verbale;
- la distribuzione di diversi tipi di dipendenze sintattiche.

Nella tabella seguente *(Tabella 9)* vengono presentate le medie e le deviazioni standard di alcune features sintattiche per le due gruppi di corpus.

Features sintattiche	Primo gruppo di 60 testi		Secondo gruppo di 60 testi	
	Media aritmetica	Deviazione standard	Media aritmetica	Deviazione standard
Verbal head per sent	2.3	1.1	2.8	1.5
Verbal root perc	85.0	20.5	83.3	20.5
Avg token per clause	11.1	4.2	10.1	3.9
Dep dist case	13.6	3.2	13.2	3.3
Dep dist det	14.1	2.6	13.5	2.6
Dep dist punct	9.9	2.7	9.4	2.7
Dep dist obl	5.5	2.5	5.2	2.2
Dep dist root	3.3	1.2	3.4	1.5
Dep total case	13.6	3.2	13.2	3.3
Dep total nmod	6.9	3.7	7.1	3.1
Dep total amod	6.6	3.2	5.1	2.7
Dep total nsubj	3.6	1.6	3.5	1.8
Dep total obj	2.8	1.9	3.1	1.6
Dep total conj	3.8	2.4	3.5	2.1
Dep total mark	1.6	1.6	2.2	1.9
Dep dist cc	2.9	1.7	3.1	1.6
Subj pre	84.9	19.5	90.4	13.3
Subj post	14.8	19.3	2.3	0.4
n. prep chains	4.8	1.8	4.9	2.05
Avg prep chain len	1.0	0.18	1.0	0.13
Prep dist 1	83.2	19.1	74.4	24.7

Prep dist 2	3.3	8.1	20.9	22.1
Principal proposition dist	41.4	19.2	37.4	17.7
Subordinate post	80.1	27.3	84.1	22.5
Subordinate pre	14.6	20.0	13.9	19.5
Subordinate proposition dist	57.9	19.1	61.9	17.6
Avg subordinate chain len	0.9	0.2	1.03	0.2
Avg links len	2.2	0.4	2.2	0.5
Avg max links len	12.8	5.6	13.03	7.9
Max links len	20.9	8.4	20.4	9.9
Avg max depth	4.9	1.3	5.1	1.9

Tabella 9. Medie aritmetiche e deviazioni standard di alcune features sintattiche nei tre corpora.

Le caratteristiche presentate mostrano una media aritmetica decisamente più elevata rispetto alla deviazione standard calcolata sui valori ottenuti per ogni testo di quel corpus. Le *features sintatticihe* più importanti nella tabella sono state presentate anche nella legenda di *Profiling-UD* e sono della struttura verbale del predicato.

(verbal_head_per_sent): distribuzione media delle teste verbali nel documento, sul totale delle teste. La prima riga della tabella che mostra una testa verbale è un verbo che rappresenta il nucleo di un sintagma verbale. I valori deviazione standard e media aritmetica per entrambe le porzioni di testo sono poche. La deviazione standard per la prima porzione dei testi è 1.1299, mentre la media aritmetica 2.3275, per la seconda invece la deviazione standard è 1.5944 mentre la media aritmetica 2.8103. In entrambi i casi la media aritmetica è molto elevata, non come la deviazione standard, poiché presenta uno scarto con la relativa deviazione standard più elevato alla seconda porzione dei testi rispetto a quanto si verifica nel primo caso.

(verbal_root_perc): distribuzione media delle radici guidate verbali da un lemma etichettato come verbo, sul totale delle radici della frase (ad esempio frasi di tipo

nominale). La deviazione standard per la prima porzione dei testi è 20.5753, mentre la media aritmetica 85.0862; per la seconda la deviazione standard è 20.5040 e la media 83.3793.

(avg token per clause): lunghezza media (*average*) delle clausole (o proposizioni), calcolata in termini di numero medio di token per clausola, nonché il rapporto tra il numero di *token* in una frase e il numero di teste verbali. Per questa feature sintattica, i risultati più importanti sono stati ottenuti per il primo blocco di testi (media 11.0862), mentre per la seconda media (10.1379) e deviazione standard (3.9755).

È rilevante anche l'analisi delle varie tipologie di relazioni di dipendenza, assegnate anche sulla basedel *Universal Dependency tagset*[19] per le relazioni di dipendenza universale (*dep*), anche dal punto di vista del numero di occorrenze dei diversi tipi (*dep_total_**) che di quello delle percentuali (*dep_dist_**).

Per il passo successivo sono state calcolate anche:

- La media della loro lunghezza (*avg links len*), che calcola il numero medio delle parole che separano si verificano linearmente tra ciascuna testa sintattica e dai suoi dipendenti (escluse le dipendenze di punteggiatura).
- La media delle lunghezze massime per frase (*avg max links len*), la media dei link di dipendenza più lunghi estratti da ogni frase di un documento. Invece *(max_links_len)*, calcola il valore del collegamento di dipendenza più lungo nel documento, calcolato in numero di token (unita minima).

Le *features* sintattiche rimaste nell'elenco dei *Profiling-UD* riguardano l'ordinamento degli elementi: la posizione nella frase del soggetto che può essere pre-verbale *(subj pre)* o post-verbale *(subj post).*

La percentuale di soggetti *pre-verbali* è elevata: si approssima a 84.9827 nel caso del primo gruppo di 60 testi, e 90.4482 nel caso di altro. Le deviazioni standard per questa caratteristica sono basse rispetto alla media aritmetica., quindi questi testi si dimostrano molto conformi all'ordine canonico degli elementi in italiano. Inoltre, le medie aritmetiche sono sempre più elevate rispetto alla deviazione standard, a dimostrazione del fatto che il soggetto posto dopo il verbo da cui dipende è una caratteristica tipica delle due fasi di analisi.

[19] Universal Dependency tagset (dep) https://universaldependencies.org/u/dep/index.html

Un'altra importante caratteristica dell'analisi del testo riguarda le catene di dipendenza a testa nominale. Riguardo questo aspetto sono state evidenziate alcune caratteristiche importanti, tra le quali si mostrano nella *Tabella 9.*

• Il numero totale di catene di dipendenza a testa nominale (*n. prepositional chains*).

Con questo parametro e con il valore medio delle catene preposizionali estratte per tutte le frasi nel documento (*avg prepositional chain len*), viene misurata la frequenza di strutture nominali complesse con modificatori. In generale, il valore medio per la prima caratteristica, nonostante i valori siano abbastanza elevati (attorno 5 in entrambi i casi): la media per i primi 60 testi è 4.8965 e la sua deviazione standard 1.8889; per la seconda porzione la media è 4.9137 mentre la deviazione standard 2.0543. Quest'ultimo è molto elevato, poiché indica una significativa variabilità dei dati. La lunghezza media delle catene a teste nominali (*avg prepositional chain len*) è simile per entrambi i corpus: per il primo 1.0 e per il secondo 1.0172.

• Il numero di catene di dipendenza a testa nominale suddivise per livello di profondità (*prep dist **).

Il monitoraggio ha tenuto conto di catene fino al livello 5 della catena di dipendenza della testa nominale: è certamente quando scendendo in profondità i valori si diminuiscono. Per questi casi i valori sono stati dati fino solo al secondo livello della profondità della catena di dipendenza della testa nominale, e per i livelli successivi i valori sono stati prossimi o uguali a 0. È possibile fare un confronto tra la percentuale di frasi principali nei testi (*principal proposition dist*) e quella di frasi subordinate (*subordinate proposition dist*) nelle due porzioni dei testi. La media delle subordinate è maggiore di quella delle principali: la media delle subordinate nel primo gruppo di 60 testi è 57.9827, nel secondo vale 61.9655. In percentuale, la media delle subordinate per il primo è 17.7647 mentre per il secondo è 19.2099.

4. Analisi statistica delle correlazioni

In questo capitolo viene indagata la relazione fra le caratteristiche linguistiche dei testi del corpus e i giudizi di interesse espressi dagli annotatori. Per farlo, verifichiamo la presenza di correlazione fra le caratteristiche linguistiche e i giudizi. L'obiettivo di questa analisi è indagare quali caratteristiche abbiano contribuito a coinvolgere maggiormente l'attenzione del lettore e motivarne il giudizio espresso.

4.1 Gli indici di correlazione

Come prima analisi è stata calcolata la *correlazione* tra la media dei voti dati dagli annotatori e le caratteristiche relative a diversi livelli di descrizione linguistica estratte da ciascun testo. Tale analisi è stata effettuata calcolando il coefficiente di correlazione di rango proposto da *Charles Spearman.*

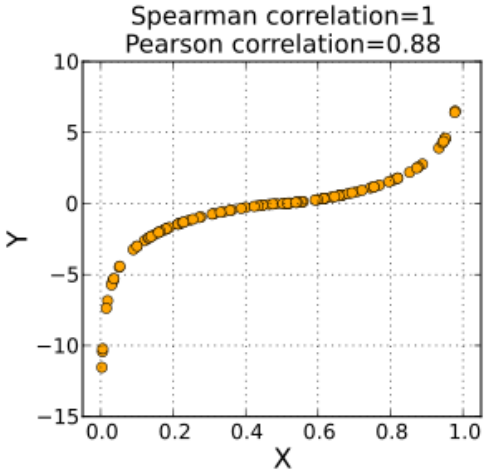

Figura 23. I dati della correlazione ch'è calcolata con
La correlazione di Spearman

L'analisi della correlazione è un metodo statistico che consente di utilizzare i coefficienti di correlazione per determinare se esiste una relazione tra le variabili e quanto questa sia forte. La Figura 23 riporta un esempio nel quale la correlazione fra due variabili è calcolata utilizzando il metodo di correlazione *Spearman*. Nel grafico presentato la correlazione *Spearman* è uguale a 1 che corrisponde ad una correlazione perfetta.
In particolare, per calcolare l'indice di correlazione del rango di *Spearman*, (indicata con ρs o rs) non è necessario che la distribuzione delle variabili sia un distribuzione normale.

L'indice di *Spearman* ha come unico vincolo che le variabili siano ordinabili (e calcola il coefficiente di correlazione sfruttando gli ordinamenti dei dati). Di seguito, riportiamo la formula del calcolo del coefficiente di *Spearman* (Formula 2).

$$\rho_s = 1 - \frac{6 \sum_i D_i^2}{N(N^2 - 1)}$$

Formula 2. L'indice di correlazione del rango di Spearman.

Nella formula 2, ***N*** rappresenta il numero di caratteristiche classificate (indicatori, soggetti);

$\boldsymbol{D_i}$ è la differenza tra i ranghi di due variabili;

$\sum_i \mathbf{D^2_i}$ è la somma dei quadrati delle differenze di rango.

In sostanza, il coefficiente di correlazione di *Spearman* è una misura della forza e della direzione di una relazione lineare tra due variabili quantitative.
Il coefficiente di correlazione di *Spearman* può assumere i valori compresi fra -1 e +1.

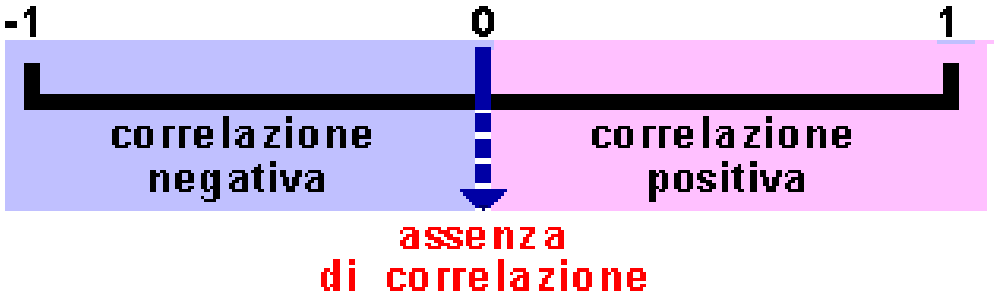

Figura 24. Intervallo dei valori di coefficiente di correlazione.

Come illustrato in Figura 24 e 25:

- Se la correlazione è pari a 1, esiste una *perfetta* correlazione lineare positiva tra le due variabili (ovvero al crescere dei valori di una variabile, crescono anche quelli dell'altra);
- Se la correlazione è pari a 0, *non esiste* nessuna correlazione tra le due variabili (esse sono indipendenti);

- Se la correlazione è pari a -1, allora *esiste una perfetta correlazione negativa* tra le due variabili (al crescere dei valori di una variabile, quelli dell'altra decrescono).

Per quanto riguarda i valori intermedi, consideriamo:

a) I valori di correlazione inferiori a +/- 0.3 come indicatori di una bassa correlazione fra le variabili;

b) I valori di correlazione maggiori di +/- 0.7 come indicatori di una forte correlazione fra le variabili;

c) I valori di correlazione compresi fra +/- 0,3 e +/- 0.7 come indicatori di una moderata correlazione fra le variabili.

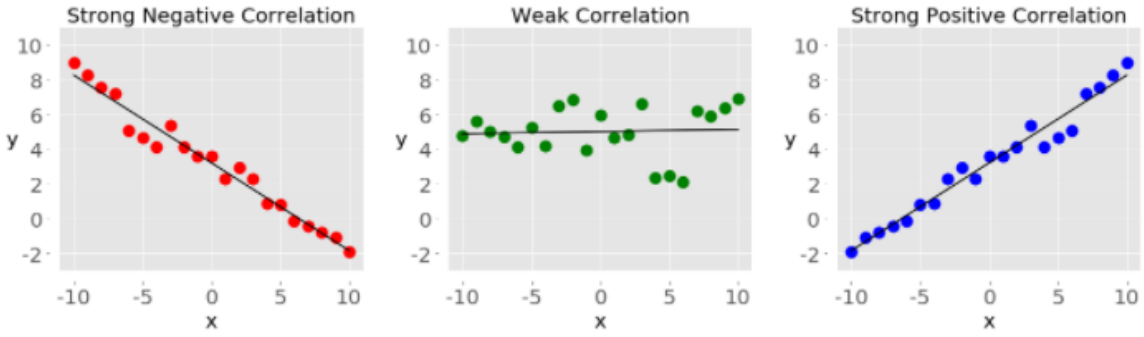

Tabella 25. Esempio di una correlazione strettamente negativa, correlazione a zero e strettamente positiva.

Per verificare che la correlazione tra le due variabili sia statisticamente significativa e non data dal caso o da un effetto del campionamento dei dati, tradizionalmente viene calcolato il *valore di probabilità p (p-value)* associato alla correlazione*:* quando il valore *p* è inferiore o uguale a 0,05 (≤ 0,05, si trova ovvero nella *zona di incertezza* o in quella di *importanza* (Figura 26), questo indica una forte evidenza contro l'ipotesi nulla e suggerisce che i risultati non sono dovuti al caso.

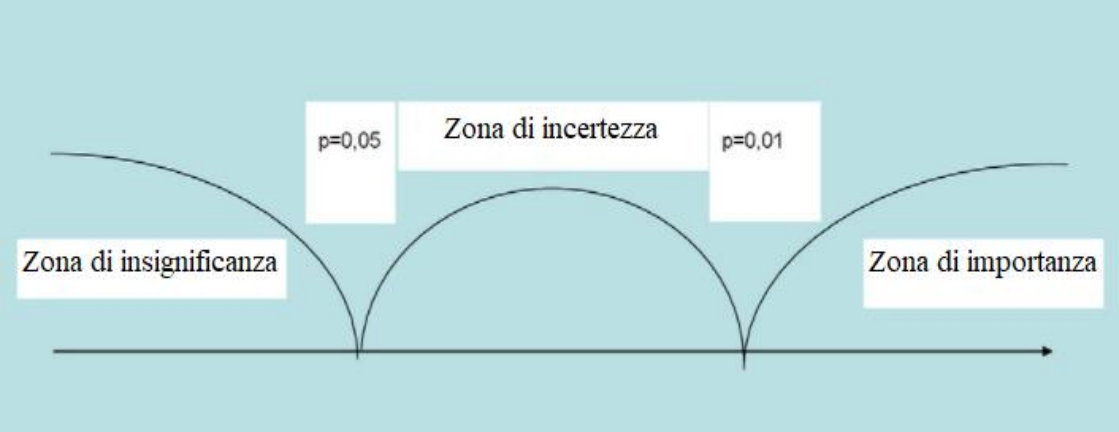

Figura 26. Intervallo delle zone significative e non significative di ps.

4.1.1 Calcolo degli indici di correlazione

L'indice di *Spearman* descritto sopra è stato utilizzato per verificare la presenza di correlazione tra la media dei voti dati dagli annotatori e le caratteristiche linguistiche dei testi. Per questa analisi sono stati presi in esame tutti i 120 testi rappresentativi dei blog di viaggi.

Dal punto di vista operativo, l'indice di correlazione *Spearman* è stato calcolato per mezzo di uno script implementato nel linguaggio *Python* che prende come input i dati riportati in due file .xlsx, uno contente i risultati del monitoraggio linguistico di ciascun testo e l'altro contente la media dei giudizi di interesse. In particolare, ho usato la funzione *spearmanr*[20] della libreria *scipy.stats.*

Lo script restituisce:

- Il valore di *correlazione di Spearman* tra la media dei voti assegnati dagli annotatori e le features linguistiche che sono state estratte e analizzate per ciascun testo;
- Il *p-value* relativo a ogni indice di correlazione.

[20] Funzione stats.spearmanr della libreria scipy
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

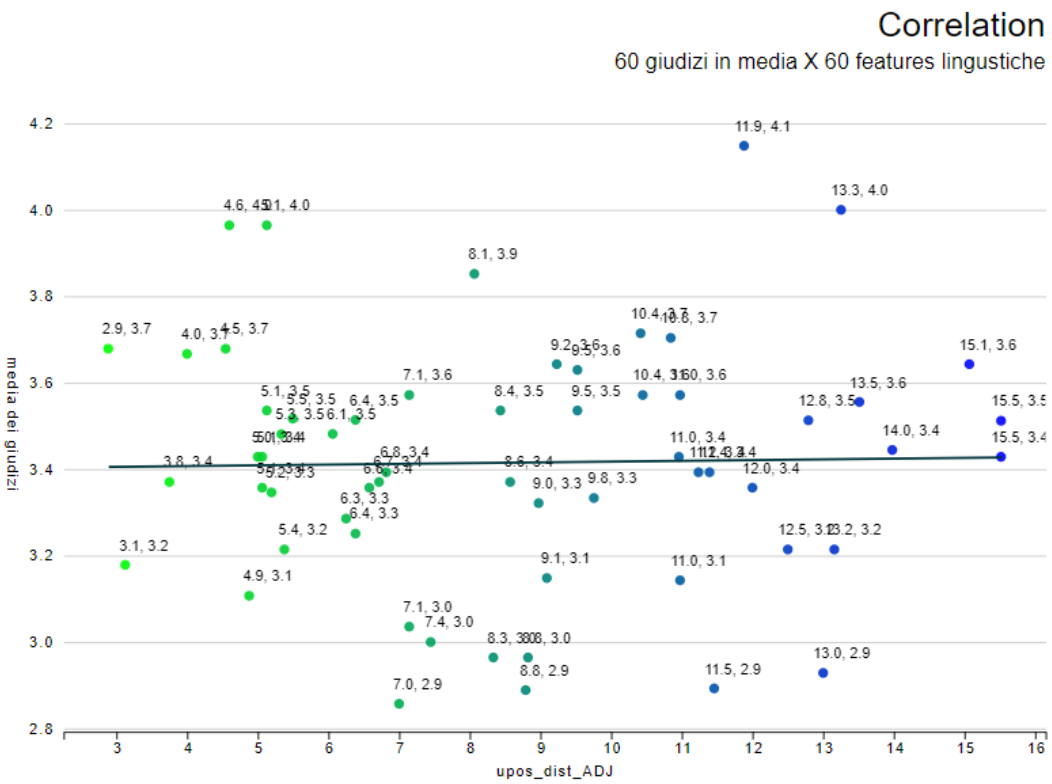

Figura 27. Correlazione di Spearman fra la media dei giudizi di interesse e la caratteristica linguistica che cattura la distribuzione degli aggettivi nel testo.

Un esempio di visualizzazione dei risultati ottenuti con questa analisi è riportata in Figura 27. In particolare, nel grafico a dispersione in figura è stato visualizzato il primo gruppo di 60 testi del corpus sulla base del giudizio medio di interesse ottenuto e il valore di una feature specifica usata come esempio: la *distribuzione degli aggettivi* nel testo. Questo tipo di popolare visualizzazione rende evidente la presenza o l'assenza di relazione fra le variabili osservate. Nell'esempio nella Figura 26, per esempio, possiamo osservare che le due variabili osservate sono indipendenti dal momento che al crescere dei valori dell'una non si osserva nessuna regolarità nelle variazioni dell'altra. Questa osservazione che emerge al grafico a dispersione può essere confermata dal calcolo del indice di *Spearman*, che ottiene, per i dati considerati in questo esempio specifico, il valore di 0.0002.

4.2 Analisi dei risultati

Qui di seguito riportiamo l'analisi delle correlazioni fra le caratteristiche linguistiche dei 120 testi considerati nello studio ed i giudizi di interesse assegnati dagli annotatori. Nello specifico, il calcolo dei valori di correlazione è stato eseguito considerando sia i valori medi (sezione 4.2.2) che le deviazioni standard (sezione 4.2.3) fra i giudizi collezionati in fase di raccolta dati. Si includono nelle analisi solo le caratteristiche linguistiche per le quali è stata ottenuta una correlazione significative (valore di $p \leq 0.05$) e si riportano i valori di correlazione ottenuti nelle tabelle sottostanti. Fra queste, si distinguono le caratteristiche che ottengono una correlazione positiva e negativa colorando i valori riportati nelle tabelle in nero e rosso rispettivamente.

4.2.2 Correlazione tra caratteristiche linguistiche e media dei giudizi di interesse

Nella Tabella 10 sono riportati il coefficiente di correlazione di *Spearman* delle caratteristiche morfologiche che maggiormente correlano con le medie dei giudizi di interesse (per tutte le caratteristiche, $p \leq 0.05$). Come possiamo notare dai risultati riportati nella tabella, le caratteristiche morfologiche che correlano più fortemente con i giudizi medi riguardano il tempo e la forma dei verbi. In particolare, il giudizio degli annotatori sembra essere correlato, seppur debolmente, con la scelta di usare verbi al tempo presente o passato e con l'uso di participi.

Features (la forma verbale, il tempo di verbo)	**120 testi**	
	Spearman correlazione	**P-value < 0,05**
verbs_tense_dist_Pres	0.1846	**0.0435**
verbs_form_dist_Part	-0.1794	**0.0498**
verbs_tense_dist_Past	-0.2367	**0.0092**

Tabella 10. Il coefficiente di correlazione Spearman e p-value delle features *la forma verbale e il tempo di verbo* nel corpus di blog di viaggio.

Le *feature* la forma verbale e il tempo di verbo fanno riferimento alla morfologia verbale. Quindi i risultati suggeriscono che i testi più brevi e che usano verbi coniugati al tempo presente catturano maggiormente l'interesse dei lettori, mentre i tempi composti, come ad esempio alcune forme al passato, suscitano minor interesse.

Features sintattiche	120 testi	
	Spearman correlazione	P-value < 0,05
avg_subordinate_chain_len	-0.1828	**0.0456**
dep_dist_aux	-0.2137	**0.0190**
verbal_root_perc	-0.1934	**0.0342**

Tabella 11. Il coefficiente di correlazione Spearman e p-value delle *features sintattiche* nel corpus di blog di viaggio.

La Tabella 11 riporta le caratteristiche sintattiche che correlano più fortemente coi giudizi di interesse. Fra queste troviamo caratteristiche che descrivono l'uso nei testi di proposizioni subordinate e di verbi ausiliari. Complessivamente, i risultati ottenuti con questa analisi suggeriscono che i testi caratterizzati da una maggiore complessità linguistica (dovuta, per esempio, all'uso consistente di proposizioni subordinate) sono anche quelli che suscitano minor interesse nei lettori. Viceversa, testi più semplici catturano maggiormente l'interesse dei lettori.

4.2.3 Correlazioni tra le caratteristiche linguistiche e la deviazione standard dei giudizi

Dopo aver considerato la correlazione fra le caratteristiche linguistiche e la media dei giudizi, passiamo ad una seconda analisi volta ad esplorare se esiste una correlazione le caratteristiche monitorate e la deviazione standard fra i giudizi raccolti. Questa analisi ci permette di indagare se alcune proprietà linguistiche influiscono non solo sui giudizi espressi, ma anche sulla variabilità osservata fra i diversi annotatori coinvolti nello studio.

Come fatto precedentemente, riportiamo nella Tabella 12 le caratteristiche morfologiche, nella Tabella 13 le caratteristiche morfo-sintattiche, e nella Tabella 14 le

caratteristiche sintattiche che correlano maggiormente ($p \leq 0.05$) con la deviazione standard dei testi calcolata sui giudizi degli annotatori.

Features (la forma verbale, il tempo di verbo)	120 testi	
	Spearman correlazione	**P-value < 0,05**
verbs_form_dist_Fin	0.2891	**0.0013**
verbs_tense_dist_Fut	0.2108	**0.0207**
verbs_tense_dist_Pres	0.1829	**0.0231**
verbs_tense_dist_Past	-0.1967	**0.0312**

Tabella 12. Il coefficiente di correlazione Spearman e p-value delle delle *features la forma verbale e il tempo di verbo* nel corpus di blog di viaggio.

Features morfo-sintattiche	120 testi	
	Spearman correlazione	**P-value < 0,05**
upos_dist_PUNCT	0.2717	**0.0026**
upos_dist_NOUN	0.2121	**0.0200**
upos_dist_VERB	-0.2237	**0.0140**
upos_dist_AUX	-0.2719	**0.0026**

Tabella 13. Il coefficiente di correlazione Spearman e p-value delle *features morfo-sintattiche* nel corpus di blog di viaggio.

Features sintattiche	120 testi	
	Spearman correlazione	**P-value < 0,05**
dep_dist_punct	0.2615	**0.0039**
dep_dist_mark	-0.2268	**0.0127**
dep_dist_xcomp	-0.2043	**0.0251**
verbal_head_per_sent	-0.2860	**0.0015**
verbal_root_perc	-0.3183	**0.0003**
subordinate_dist_3	-0.1953	**0.0325**
prep_dist_3	0.2397	**0.0083**
avg_token_per_clause	0.2078	**0.0226**
aux_num_pers_dist_Sing+3	0.1816	**0.0470**
aux_form_dist_Fin	0.1797	**0.0494**
aux_num_pers_dist_Plur+1	-0.2603	**0.0040**

Tabella 14. Il coefficiente di correlazione Spearman e p-value delle *features. sintattiche* nel corpus di blog di viaggio.

In questa analisi notiamo una maggiore correlazione fra le proprietà linguistiche e la variazione dei giudizi rispetto a quanto osservato considerando le medie. Questo da una parte suggerisce che la valutazione espressa dagli annotatori è piuttosto influenzata da giudizi soggettivi, che però complessivamente riflettono un legame con la complessità dei testi presi in analisi. Più precisamente, come già osservato nella sezione precedente, anche in questo caso la presenza di correlazione con caratteristiche che descrivono proprietà legate al tempo verbale e alla struttura proposizionale dei testi suggerisce che più un testo è complesso e articolato più le opinioni dei lettori su di esso tendono ad essere in disaccordo. Per mostrare più nel dettaglio questa osservazione, possiamo notare il caso della feature *subordinate_dist_3*, la quale descrive la distribuzione nei testi di proposizioni subordinate piuttosto distanti dal verbo principale della frase (3 indica infatti che durante l'analisi linguistica sono state individuate 2 subordinate che intercorrono fra quella osservata e il verbo della principale). La correlazione negativa

riscontrata fra la deviazione standard e questa caratteristica suggerisce che all'aumentare del valore della feature (indicando per tanto una maggiore presenza di questo tipo di subordinate), la deviazione standard tende a diminuire. Una minor deviazione standard è indice di maggior accordo fra i giudizi, pertanto possiamo concludere che frasi particolarmente complesse e articolate risultano poco interessanti per tutti gli annotatori. La correlazione positiva osservata invece con altre caratteristiche, come ad esempio il numero di tokens per proposizione (*avg_tokens_per_clause*), suggerisce un minor accordo fra i giudizi espressi per testi che ottengono valori alti per questa feature. In effetti, sebbene la presenza di un alto numero di tokens in una frase sia generalmente identificato come una proprietà associata alla complessità di un testo, questa caratteristica dovrebbe essere analizzata alla luce di altre caratteristiche (ad esempio, lessicali) per poter affermare con certezza che si tratta di un indicatore di complessità linguistica. Questo tipo di analisi, che rimandiamo a future analisi, permetterebbe di fare chiarezza sul motivo per il quale i giudizi degli annotatori tendono a divergere all'aumentare del valore di questa feature.

5. Conclusioni

In questa tesi è stato presentato uno studio volto ad indagare l'influenza della forma linguistica di un testo sul grado di intrattenimento e interesse suscitato dal testo stesso. Per condurre questa indagine ho preso in considerazione un corpus composto di 120 testi in italiano estratti da blog di viaggio dedicati nello specifico a descrizioni di viaggi in Russia e territori limitrofi. Per quanto riguarda l'analisi della forma linguistica dei testi, ho estratto un ampio insieme di proprietà che descrivono il profilo linguistico dei testi, ovvero catturano fenomeni legati alla struttura di base, ad aspetti morfologici, morfo-sintattici e sintattici. Per quanto riguarda l'analisi del grado di interesse suscitato dai testi, ho definito un questionario per collezionare giudizi circa il grado di intrattenimento di ciascuno dei 120 testi del corpus. Gli annotatori madrelingua italiana che hanno svolto il questionario sono stati reclutati in modalità crowdsourcing.

I dati ottenuti attraverso il questionario e le caratteristiche linguistiche estratte dai testi sono stati come prima cosa analizzati da un punto di vista quantitativo. Da questa analisi è emerso che i soggetti dello studio hanno espresso un interesse genericamente più alto verso testi che riportano giudizi personali dell'autore sui luoghi visitati. Viceversa, i testi che contengono meno dettagli e descrivono i luoghi con uno stile distaccato e oggettivo catturano meno l'attenzione dei lettori. In una seconda fase, allo scopo di quantificare la relazione tra profilo linguistico dei testi e interesse suscitato, ho svolto un test di indagine statistica per misurare il coefficiente di correlazione fra giudizi di interesse e caratteristiche linguistiche estratte. Questa seconda analisi ha messo in luce un interessante risultato, ovvero che i lettori tendono ad essere maggiormente attratti da testi semplici. Questa osservazione emerge dal fatto che i nostri dati suggeriscono una correlazione significativa fra i giudizi di interesse e caratteristiche linguistiche genericamente associate alla complessità dei testi. Più precisamente, ho osservato che più un testo presenta proprietà linguistiche che lo caratterizzano come semplice (come ad esempio l'uso di verbi al presente o di una struttura proposizionale caratterizzata da frasi brevi e poche subordinate), più alto sarà il grado di interesse associato a quel testo. Questo interessante risultato suggerito dalle nostre analisi è molto promettente, seppur ancora preliminare vista la ridotta dimensione del corpus. Pertanto, proponiamo qui di seguito diverse direzioni nelle quali lo studio descritto in questa tesi potrebbe essere ampliato:

- Le analisi potrebbero essere riproposte su un corpus di testi più ampio o su testi più lunghi.
- L'indagine potrebbe essere estesa ad altri generi testuali oltre al genere blog esplorato in questa tesi. Questo permetterebbe di indagare come le variazioni dovute al genere testuale influiscono sui giudizi di interesse espressi dai soggetti e quali strategie comunicative sono più efficaci per catturare l'attenzione dei lettori.
- Sebbene il numero di annotatori reclutati per questo studio è stato sufficiente per ottenere delle analisi statisticamente significative, la modalità crowdsourcing permette di collezionare i giudizi di campioni molto ampi di popolazione. Questa possibilità potrebbe essere esplorata in future analisi per corroborare la solidità dei risultati ottenuti.
- Questo stesso studio potrebbe essere proposto utilizzando testi di lingue diverse per verificare se, al variare della lingua, variano anche le caratteristiche che determinano il grado di interesse dei lettori. In questo caso si potrebbero coinvolgere sia annotatori madrelingua che annotatori non madrelingua per verificare se le due popolazioni sono attratte dalle stesse caratteristiche linguistiche.

6. Bibliografia

Lenci, Alessandro, Montemagni Simonetti, Pirrelli Vito. 2005. "*Testo e computer*". *Elementi di linguistica computazionale*. Carocci, Roma;

Simonetta Montemagni (Istituto di Linguistica Computazionale "Antonio Zampolli" - ILC-CNR) "*Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*" In Studi Italiani di Linguistica Teorica e Applicata (SILTA) Anno XLII, Numero 1, 2013, pp. 145-172;

Brunato D., Cimino A., Dell'Orletta F., Montemagni S., Venturi G. (2020) "*Profiling-UD: a Tool for Linguistic Profiling of Texts*". In Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020), 11-16 May 2020, Marseille, France;

Lenci A., S. Montemagni, V. Pirrelli. 2009. "*Annotazione sintattica di corpora: aspetti metodologici.*" In: Corpora di italiano L2: tecnologie, metodi, spunti teorici, Guerra Edizioni, Perugia;

Barbagli, A., Dell'Orletta, F., Venturi, G., Lucisano, P., & Montemagni, S. (2015). "*Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati*." IJCoL. Italian Journal of Computational Linguistics, 1(1-1), 105-123;

F. Dell'Orletta, "*Ensemble system for Part-of-Speech tagging. In Proceedings of Evalita '09*" (Evaluation of NLP and Speech Tools for Italian), Reggio Emilia, 2009.

7. Sitografia

1. Il blog di viaggio “Miprendoemiportovia”:
https://www.miprendoemiportovia.it/2013/06/05/viaggio-in-russia-fra-mosca-e-san-pietroburgo

2. Il blog di viaggio “Viaggi di Gusto”:
https://viaggidigusto.wordpress.com/2018/02/19/russia/

3. Piattaforma “QuestBase”:
https://questbase.com/

4. Pagina di Questionario sulla piattaforma QuestBase:
https://my.questbase.com/take.aspx?pin=2566-9954-0778

5. Una porzione del testo di blog di viaggio di Simona Sacri:
https://www.simonasacri.com/emozioni/frammenti-san-pietroburgo.php

6. Una porzione del testo di blog di viaggio di “Scorcidi mondo”:
https://www.scorcidimondo.it/russia-benvenuti-sochi-trote-fiumi-ghiacciati-foreste-mowgli/

7. Una porzione del testo di blog di viaggio di “Pimp my trip”:
hhttps://www.pimpmytrip.it/transiberiana-mosca-pechino-di-enrico/

8. Una porzione del testo di blog di viaggio di “Evaneos”:
https://www.evaneos.it/polonia/viaggio/destinazioni/3800-auschwitz/

9. Una porzione del testo di blog di viaggio di “Il passeggero”:
https://www.ilpasseggero.eu/belovezhskaya-pushcha-bielorussia/

10. Una porzione del testo di blog di viaggio di “Nomavic”:
https://www.nomavic.it/lettonia-cosa-vedere/

11. Una porzione del testo di blog di viaggio di “Beborghi”:
https://www.beborghi.com/tagikistan-pamir-viaggio-asia-centrale/

12. Una porzione del testo di blog di viaggio di “Evaneos”:
https://www.evaneos.it/ucraina/viaggio/destinazioni/5678-leopoli/

13. Tool di Profiling-UD:
http://linguistic-profiling.italianlp.it

14. Lo strumento di Profiling UD:
http://www.italianlp.it/demo/profiling-UD/

15. Dipendenze universali (UD):
http://lindat.mff.cuni.cz/services/udpipe/

16. Il tagset dell’Universal Part of Speech:
https://universaldependencies.org/u/pos/index.html

17. Lista completa delle Part of Speech:
http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

18. Una porzione del testo di blog di viaggio di “Scorci di mondo”:
https://www.scorcidimondo.it/russia-benvenuti-sochi-trote-fiumi-ghiacciati-foreste-mowgli/

19. Universal Dependency tagset (dep):
https://universaldependencies.org/u/dep/index.html

20. Funzione stats.spearmanr della libreria scipy:
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Al mio marito…

per avermi trasmesso la sua immensa forza e il suo coraggio. Grazie per tutto il tempo che mi hai dedicato. Grazie perché ci sei sempre stato con me. È grazie a te che ho superato i momenti più difficili. Senza i tuoi consigli, non ce l'avrei mai fatta.

Я тебя люблю.

Ai miei genitori…

за то, что всегда были рядом, даже на расстоянии поддерживали меня во всех моих начинаниях и свершениях, без вас не было бы меня! Спасибо за все жизненные уроки, теплоту и любовь.

Я вас люблю, мама и папа!

Un GRAZIE sincero…

…al mio relatore **Felice Dell'Orletta** e relatrice **Dominique Brunato** che sono sempre stati con me molto gentili, presenti, puntuali e disponibili nei tutti i momenti importanti del mio lavoro. Grazie al percorso intrapreso insieme ho sviluppato maggiormente la mia capacità di analisi e di problem solving. Anche un grazie sincero a tutti i lavoratori di laboratorio ItaliaNLP.

Grazie di cuore.

…ad **Alina** che stata sempre come un sole durante i giorni del buio. È sempre stata molto gentile con me e mi aiutava in tutto. Sei arrivata nella mia vita all'improvviso, e questo «improvviso» è migliore di tutto quello che è successo con me in quest'anno!

Люблю тебя!

…alla mia collega e amica **Ekaterina** per essermi stata sempre accanto me in questo periodo lungo e difficile.

Спасибо тебе за поддержку!

…a tutti altri **amici** che hanno aiutato e contribuito durante questo incredibile periodo della mia vita!