UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Studio della percezione della complessità linguistica da parte di madrelingua russi apprendenti l'italiano come L2

Candidato: *Ekaterina Chernysheva*

Relatore: *Felice Dell'Orletta*

Correlatori: *Alessandro Lenci*

Dominique Brunato

Anno Accademico 2020-2021

Indice

1 Introduzione 3
2 Risorse e strumenti dello studio 6
2.1 Il corpus 6
2.2 Selezione degli annotatori e creazione del questionario 8
2.3 Analisi delle risposte 11
2.4 Accordo tra gli annotatori 16
3 Monitoraggio linguistico e analisi delle caratteristiche linguistiche 20
3.1 Fondamenti della metodologia 20
3.1.1 Analisi linguistica automatica 21
3.2 Analisi del profilo linguistico del corpus 23
3.2.1 Caratteristiche di base 24
3.2.2 Caratteristiche lessicali 25
3.2.3 Caratteristiche morfo-sintattiche 26
3.2.4 Caratteristiche sintattiche 32
4 Analisi dei giudizi di complessità linguistica 43
4.1 Calcolo dei giudizi medi degli annotatori 43
4.2 Metriche per il calcolo della correlazione 46
4.3 Analisi della correlazione fra giudizi e caratteristiche linguistiche 49
4.4 Influenza del livello di competenza della L2 54
4.5 Confronto con i risultati di madrelingua italiani 62
5 Conclusione 66
6 Bibliografia 69
7 Appendice 72

1 Introduzione

Lo studio riportato in questa tesi nasce dalla seguente domanda: *quali caratteristiche rendono una frase in lingua italiana più o meno complessa dal punto di vista linguistico per un soggetto non madrelingua?*

Il concetto di complessità linguistica, sebbene oggetto di indagine da ormai diverso tempo, è tuttora difficile da definire. Pallotti (2014), infatti, nell'osservare che la nozione di complessità linguistica è affrontata in numerosi studi che indagano il processo di acquisizione della lingua seconda (in inglese, *Second Language Acquisition*, generalmente abbreviato SLA), pone l'attenzione sul fatto che una definizione dettagliata e universalmente condivisa di tale concetto è ancora assente. Ancora Pallotti, nel medesimo studio, propone una definizione di 'complessità' adottando una prospettiva linguistica che ha l'obiettivo di valorizzare la natura multidimensionale di questo concetto. Si fa infatti riferimento ad una *complessità strutturale,* che ha a che vedere con proprietà formali della lingua, ad una *complessità cognitiva,* relativa al costo di elaborazione delle strutture linguistiche, e infine ad una *complessità evolutiva*, che determina l'ordine di apprendimento delle strutture linguistiche. Questa definizione proposta da Pallotti fornisce ulteriore prova del fatto che il concetto di complessità linguistica può essere indagato da diverse prospettive.

L'ambito di ricerca del Trattamento Automatico del Linguaggio (TAL o, più frequentemente, NLP dal nome inglese *Natural Language Processing*), per esempio, cerca di individuare misure per quantificare la complessità linguistica dei testi e di usare questa informazione nello sviluppo di approcci di insegnamento di una lingua seconda (L2) o dell'insegnamento scolastico. Secondo Miestamo (2008) la complessità linguistica può essere considerata dal punto di vista assoluto o relativo. Studiare la complessità linguistica adottando un approccio assoluto significa valutare la complessità di una lingua rispetto alle sue proprietà formali, mentre adottare un approccio relativo vuol dire definirla in termini di difficoltà per i parlanti. In questo lavoro è usato un approccio relativo per studiare la percezione della complessità linguistica da parte di apprendenti L2.

Più precisamente, lo scopo di questo studio è definire quali caratteristiche linguistiche determinano la percezione della complessità linguistica di frasi in italiano da parte di soggetti madrelingua russi che parlano l'italiano come lingua seconda. Come dimostrato in vari studi (si veda, per esempio, Norris e Ortega, 2009), la complessità linguistica e sintattica delle produzioni in una lingua seconda si sviluppa gradualmente con una competenza crescente. Per questo motivo, nel presente studio, il campione degli annotatori è stato diviso in tre gruppi a seconda del livello di competenza: principianti, intermedi e avanzati. L'analisi è stata fatta sia per l'insieme di tutti gli annotatori, sia singolarmente per ciascun gruppo. Questo ci permetterà di verificare se il livello di competenza della L2 influisce sulle caratteristiche che determinano la complessità percepita della frase.

In questo lavoro è stato adottato un approccio di tipo "*crowdsourcing*" per la raccolta dei giudizi sulla complessità linguistica. Per lo studio è stato preso in esame un corpus di frasi in italiano già oggetto di valutazione dal punto di vista della complessità linguistica da parte di parlanti madrelingua italiani, adottando al contempo un test di significatività statistica per studiare la qualità dei dati raccolti e il grado di accordo tra gli annotatori. Il corpus è stato esplorato attraverso strumenti di annotazione linguistica automatica e di monitoraggio linguistico del testo al fine di individuare quali proprietà linguistiche, a diversi livelli di granularità, correlino maggiormente con la percezione della complessità della frase e se esistano differenze legate al livello di competenza in ingresso dell'annotatore L2. Lo studio tiene conto di un ampio set di caratteristiche linguistiche per indagare la percezione della complessità linguistica di una frase da parte di apprendenti l'italiano come L2.

Il resto della tesi è organizzato come descritto qui di seguito. Nel capitolo 2 di questo elaborato sarà descritta la costruzione del corpus delle frasi impiegato nello studio, sarà mostrata la creazione del questionario in modalità sondaggio per la raccolta dei dati dello studio e saranno analizzati i risultati ottenuti in seguito alla valutazione delle frasi da parte di annotatori madrelingua russi apprendenti l'italiano come L2. Un'attenzione particolare sarà dedicata alla qualità dei dati raccolti la quale sarà verificata con l'ausilio del coefficiente d'accordo tra gli annotatori.

Il capitolo 3 introdurrà i concetti dell'analisi automatica del testo e lo strumento per il monitoraggio linguistico Profiling-UD. Successivamente verranno descritte le caratteristiche linguistiche estratte dalle frasi oggetto dello studio usando Profiling-UD. Inoltre, saranno visualizzate alcune delle caratteristiche più significative per mostrare la loro interazione e come si influenzano vicendevolmente nel testo.

Nel capitolo 4 sarà effettuata l'analisi delle correlazioni tra le caratteristiche linguistiche e i giudizi di complessità assegnati dagli annotatori. L'obiettivo di questo capitolo è quello di determinare le caratteristiche più significative per la percezione della complessità linguistica nella lingua italiana da parte di madrelingua russi. Verranno individuate le caratteristiche rilevanti sia per tutto l'insieme degli annotatori, sia per i gruppi di annotatori individuati sulla base della loro competenza in ingresso sulla lingua italiana. Infine, i risultati del presente elaborato verranno confrontati con i dati dello studio sulla complessità linguistica condotto coinvolgendo i madrelingua italiani.

Il capitolo 5 discute le conclusioni della tesi riportanti i principali risultati ottenuti.

2 Risorse e strumenti dello studio

In questo capitolo sarà introdotta la metodologia, la quale si articola in diverse fasi. Nella sezione 2.1 verrà descritta la fase di raccolta del set di frasi in italiano oggetto dello studio. La sezione 2.2 invece discuterà la fase della creazione del questionario e del reclutamento degli annotatori madrelingua russi per la valutazione della complessità linguistica percepita delle frasi in italiano. Successivamente, nella sezione 2.3 descriveremo i giudizi raccolti attraverso il questionario con misure quantitative. L'attendibilità dei giudizi raccolti verrà valutata per mezzo di metriche volte a misurare l'accordo fra gli annotatori, come descritto nella sezione 2.4.

2.1 Il corpus

Un corpus è una collezione di testi selezionati, annotati e organizzati in maniera tale da supportare analisi linguistiche. I corpora testuali rappresentano la principale fonte di dati della Linguistica Computazionale, la quale ne ha favorito l'analisi con metodi statistici. Anche se i corpora sono precedenti all'avvento del computer, la rivoluzione informatica ha sicuramente dato una forte spinta verso la raccolta di maggiori quantità di dati testuali, facilitandone la gestione, l'ottimizzazione e l'elaborazione, e favorendo lo sviluppo di modelli computazionali della lingua. Il ruolo dei computer nell'analisi dei corpora occupa una posizione così cruciale che oggi, quando si parla di corpora, si fa generalmente riferimento a corpora in formato digitale.
I corpora si dividono in tipologie sulla base dei seguenti parametri (Lenci et al., 2005):

- *Generalità*: descrive la trasversalità dei testi facenti parte del corpus rispetto alla varietà possibili della lingua. I corpora si dividono in corpora specialistici (o verticali) oppure corpora generali a seconda del grado di generalità;
- *Modalità*: designa la modalità di produzione e può essere scritta oppure orale;
- *Cronologia*: dipende dall'asse temporale in cui sono stati prodotti i testi. Secondo il criterio di modalità i corpora si dividono in corpora sincronici e corpora diacronici;

- *Lingua*: dipende dalla quantità di lingue presentate in un corpus e si dividono in corpora monolingue, bilingue o multilingue. I corpora multilingue sono distinti ulteriormente in corpora paralleli e corpora comparabili;
- *Integrità dei testi*: indica se un corpus contiene testi interi oppure porzioni di testi di una lunghezza prefissata;
- *Codifica digitale dei testi*: dipende dal modo in cui sono rappresentati i testi digitali.

La scelta dei dati linguistici appropriati per gli scopi dello studio è un fattore essenziale per la qualità del lavoro. Prima di passare all'analisi computazionale dei dati, è necessario valutare con attenzione se il corpus contiene testi adeguati alla ricerca dei fenomeni che si vuole indagare. I corpora permettono di sviluppare modelli e applicazioni sulla base di dati linguistici "ecologici", ovvero direttamente ricavati dai testi prodotti da parlanti nativi della lingua. I parametri che determinano la loro conformazione sono la quantità e la qualità dei prodotti della lingua che registrano. È importante usare come punto di riferimento la valutazione del grado di rappresentatività dei dati per costruire i modelli della lingua.

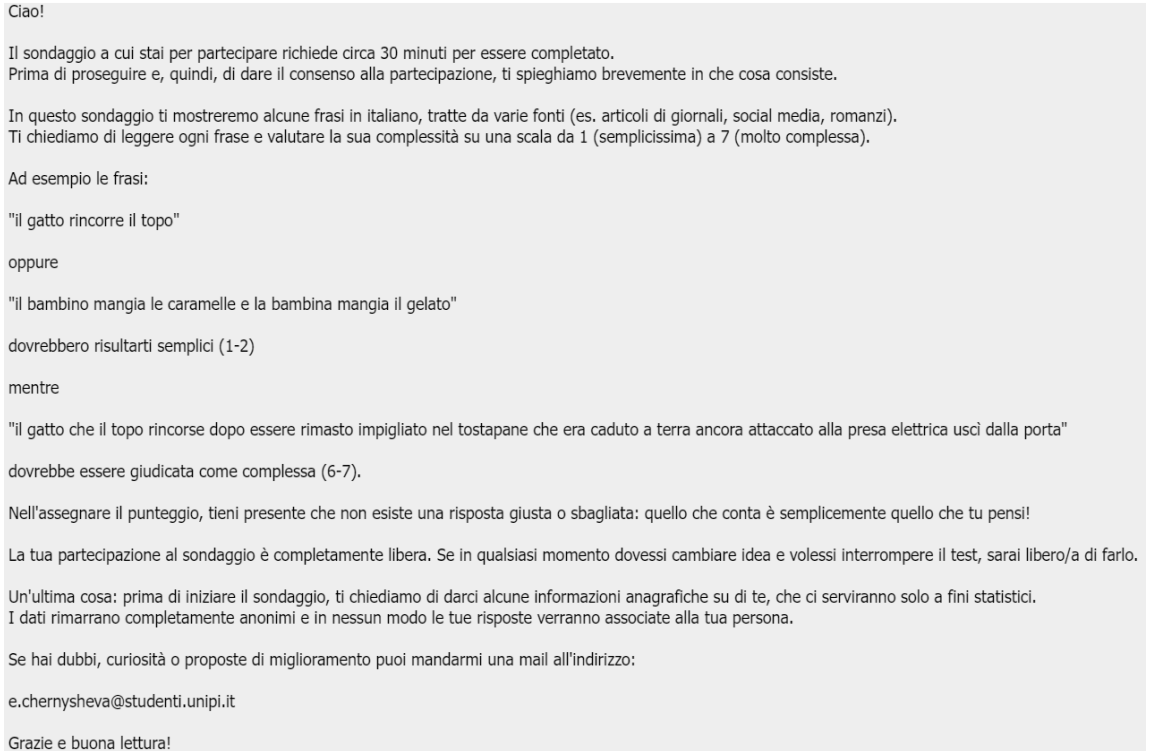
Per condurre lo studio descritto in questa tesi è stato utilizzato un corpus di 184 frasi in italiano estratte dal corpus descritto in (Brunato et al., 2018), originariamente creato per svolgere un'indagine in crowdsourcing volta ad indagare la percezione della complessità a livello di frase da parte di parlanti madrelingua. Le 184 frasi provengono dalla sezione giornalistica della Italian Universal Dependency Treebank (IUDT) annotata secondo lo schema delle *Universal Dependencies* (UD). UD è un'iniziativa internazionale che ha lo scopo di fornire un inventario multilingue per l'annotazione dei fenomeni linguistici a livello morfosintattico. Lo schema proposto da UD è volto a produrre un'annotazione a dipendenze. IUDT annotata secondo lo schema UD è stata ottenuta con la conversione semi-automatica dalla Italian Stanford Dependency Treebank (ISDT) (Simi et al., 2014). Il corpus utilizzato nel presente studio è scritto, sincronico, monolingua e specializzato, in quanto contiene le frasi scritte in una solo lingua che appartengono a un solo genere giornalistico testuale.

2.2 Selezione degli annotatori e creazione del questionario

Al fine di raccogliere giudizi sulla complessità delle frasi del corpus, abbiamo reclutato 15 annotatori per svolgere un questionario nel quale veniva a loro chiesto di esprimere un giudizio circa la complessità linguistica percepita di alcune frasi in italiano. Sono stati selezionati 15 madrelingua russi parlanti l'italiano come L2. Una parte di loro ha ricevuto una formazione speciale in lingua italiana, l'altra parte ha vissuto in Italia per un periodo prolungato e ha svolto la sua principale attività lavorativa o formativa in lingua italiana.

Per la creazione del questionario è stata utilizzata la piattaforma Questbase[1] che consente di creare sondaggi, quiz e questionari, pubblicarli online ed estrarre successivamente i risultati. Il corpus delle frasi dello studio è stato inserito in un unico questionario per permette agli annotatori di valutarne la complessità. Nella prima pagina del questionario (figura 1) viene riportata una breve descrizione e le istruzioni per lo svolgimento.

1 https://story.questbase.com/


Ciao!

Il sondaggio a cui stai per partecipare richiede circa 30 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 alcune frasi in italiano, tratte da varie fonti (es. articoli di giornali, social media, romanzi).
Ti chiediamo di leggere ogni frase e valutare la sua complessità su una scala da 1 (semplicissima) a 7 (molto complessa).

Ad esempio le frasi:

"il gatto rincorre il topo"

oppure

"il bambino mangia le caramelle e la bambina mangia il gelato"

dovrebbero risultarti semplici (1-2)

mentre

"il gatto che il topo rincorse dopo essere rimasto impigliato nel tostapane che era caduto a terra ancora attaccato alla presa elettrica uscì dalla porta"

dovrebbe essere giudicata come complessa (6-7).

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 rimarrano completamente anonimi e in nessun modo le tue risposte verranno associate alla tua persona.

Se hai dubbi, curiosità o proposte di miglioramento puoi mandarmi una mail all'indirizzo:

e.chernysheva@studenti.unipi.it

Grazie e buona lettura!

Figura 1. Il messaggio della pagina iniziale del questionario con le istruzioni per lo svolgimento.

Prima di passare al questionario, all'annotatore sono richieste informazioni anagrafiche che permettono di descrivere il campione di annotatori sulla base di alcune variabili di sfondo (figura 2). Le informazioni personali richieste sono:

1. Età;
2. Sesso;
3. Titolo di studio;
4. Livello di competenza in italiano;
5. Modalità di apprendimento della lingua italiana.

In particolare, ci interessa raccogliere l'autovalutazione di ciascun annotatore sul proprio livello di competenza della lingua italiana. Questo ci permette di studiare il grado di accordo sui giudizi espressi tra gruppi di annotatori omogenei per competenza e indagare se gruppi diversi attribuiscono lo stesso livello di complessità linguistica a frasi caratterizzate da diversi costrutti linguistici .

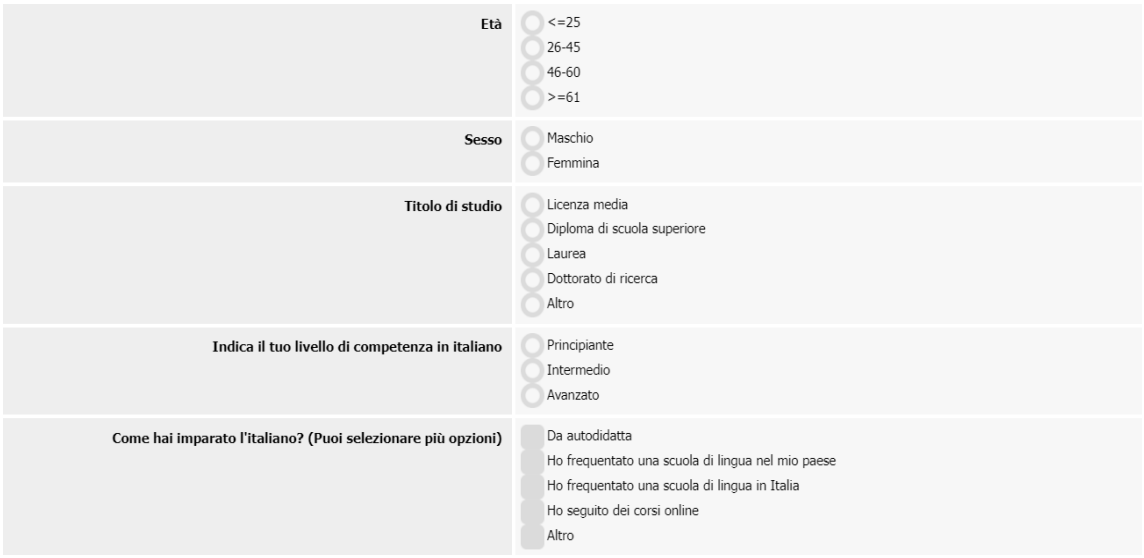


Figura 2. La pagina del questionario dove sono richieste le informazioni personali degli annotatori.

Dopo aver inserito i dati anagrafici, un annotatore può iniziare a svolgere il questionario. Le 184 domande sono divise 20 per pagina secondo un ordinamento predefinito. Per ogni domanda è richiesto di valutare la complessità della frase proposta sulla base di una scala valutativa con un punteggio da 1 a 7, dove il valore 1 designa le frasi molto facili e il valore 7 le frasi molto difficili. L'attribuzione di un punteggio basso indica che la frase è semplice mentre l'assegnazione di un punteggio alto indica che la frase assume una configurazione più complessa e quindi più difficile (figura 3). Non esiste una risposta giusta o sbagliata, tutte le risposte rispecchiano la percezione individuale di un annotatore che può non coincidere con le scelte di altri in quanto ogni soggetto ha una propria percezione di complessità.

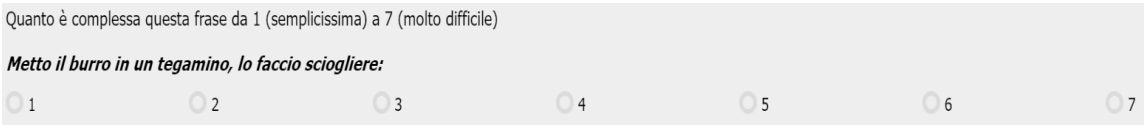


Figura 3. Esempio della frase dal questionario.

Nella fase di creazione del questionario è stato fondamentale indicare l'obbligatorietà di tutte le domande in modo che nessuno tra gli annotatori possa trascurarne alcuna. Questo ha consentito di svolgere correttamente le fasi d'analisi successive.

2.3 Analisi delle risposte

Questbase permette di estrarre le risposte dalla piattaforma in vari formati. Per i nostri scopi è adatto il formato csv (*comma-separated values*) che rappresenta i dati nel formato di una tabella di dati. Nel file scaricato sono riportate le informazioni anagrafiche di ogni annotatore, il suo indirizzo IP, il tempo totale impiegato per il completamento del questionario (figura 4) e i giudizi che ha attribuito ad ogni frase.

Indirizzo IP	Tempo totale	Eta	Sesso	Titolo	Livello_competenza_L2	Apprendimento_1	Apprendimento_2	Apprendimento_3	Apprendimento_4	Apprendimento_5
131.114.192.146	0:49:54	1	2	2	3		1	1		
151.29.176.123	0:39:28	1	2	3	2			1		
151.29.176.123	0:38:00	2	1	4	3			1		
79.27.205.106	0:25:58	2	2	5	3					1
83.220.239.34	1:27:38	2	1	5	1		1			
85.26.232.6	0:28:32	2	2	3	1		1			
2.45.2.43	0:48:11	1	2	3	2	1				
109.252.73.120	0:29:29	2	2	4	3		1			1
90.154.72.176	1:01:13	2	2	3	2		1			
37.116.149.141	0:53:44	2	2	3	3					1
93.57.254.24	1:28:47	1	2	3	3	1	1	1		
37.163.62.252	1:04:58	2	2	3	2			1		
131.114.215.221	0:34:21	1	2	3	2	1				
5.171.25.23	0:42:40	1	2	2	2		1			
37.162.54.186	0:41:41	1	2	3	2	1			1	

Figura 4. Una parte della tabella estratta da Questbase contenente i dati anagrafici degli annotatori.

Le categorie di dati anagrafici sono tradotte in numeri. Le tabelle seguenti riportano la traduzione di questi valori.

Tabella 1. Età.

Codice	Dato anagrafico
1	<=25
2	26-45
3	46-60
4	>=61

Tabella 2. Sesso.

Codice	Dato anagrafico
1	Maschio
2	Femmina

Tabella 3. Livello di apprendimento dell'italiano.

Codice	Dato anagrafico
1	Principiante
2	Intermedio
3	Avanzato

Tabella 4. Titolo di studio.

Codice	Dato anagrafico
1	Licenza media
2	Diploma di scuola superiore
3	Laurea
4	Dottorato di ricerca
5	Altro

Tabella 5. Il modo di apprendimento dell'italiano.

Codice	Dato anagrafico
1	Da autodidatta
2	Ho frequentato una scuola di lingua nel mio paese
3	Ho frequentato una scuola di lingua in Italia
4	Ho seguito dei corsi online
5	Altro

Uno dei modi per controllare la qualità delle risposte date dagli annotatori e valutare il grado di attenzione con cui hanno svolto il questionario è verificare il tempo impegnato per lo svolgimento. Per elaborare i dati del file di tipo csv è stato usato Microsoft Excel, un programma dedicato alla produzione ed alla gestione dei fogli di calcolo. Il tempo medio impegnato per lo svolgimento del questionario da parte di tutti gli annotatori calcolato con una funzione di Excel è 0:48:58: il tempo minimo è stato di 0:25:58 e il tempo massimo di 1:28:47. Il grafico seguente (Grafico 1) riporta la distribuzione del tempo impiegato da ogni annotatore a compilare il questionario rispetto alla media.

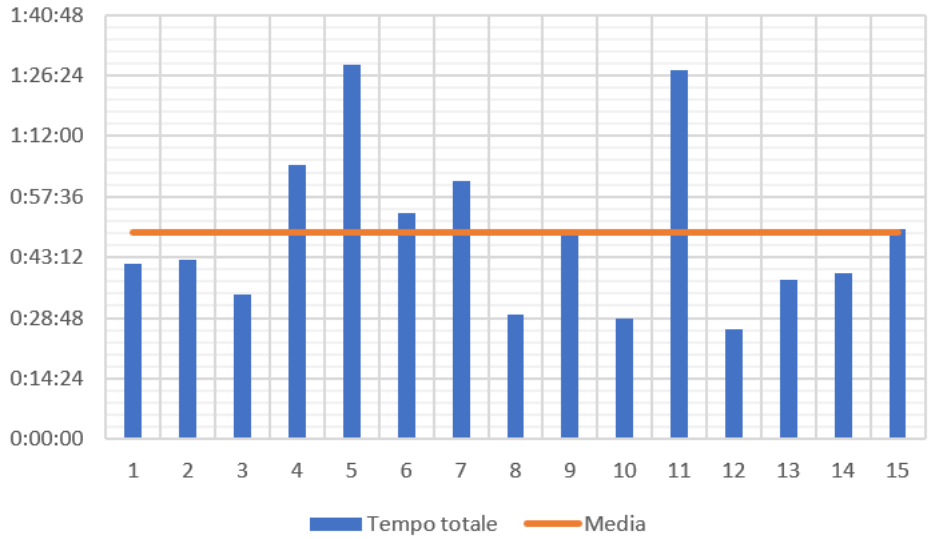


Grafico 1. Il tempo impiegato da ogni annotatore a compilare il questionario e la media del tempo.

Dopo lo studio del grafico si può concludere che nessun tentativo di compilazione deve essere eliminato e che tutti i dati raccolti sono accettabili in questa fase del procedimento perché tutti gli annotatori hanno impiegato un lasso di tempo adeguato nella compilazione del questionario.

Il Grafico 2, che mostra la distribuzione degli annotatori rispetto al sesso, rivela che la maggior parte degli annotatori sono di sesso femminile. In particolare, l'87% dei giudizi sono stati assegnati da soggetti di sesso femminile e il 13% da soggetti del sesso opposto.

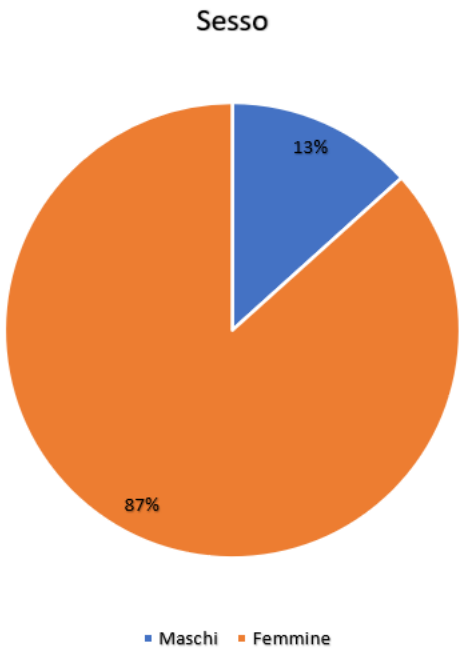


Grafico 2. Il sesso degli annotatori

Nel Grafico 3, che mostra la distribuzione degli annotatori rispetto all'età, si osserva la prevalenza di soggetti appartenenti a gruppi di età <=25 e 26-45. Al primo gruppo appartengono 7 annotatori e al secondo 8 annotatori. I gruppi di età 46-60 e >=60 invece non sono rappresentati.

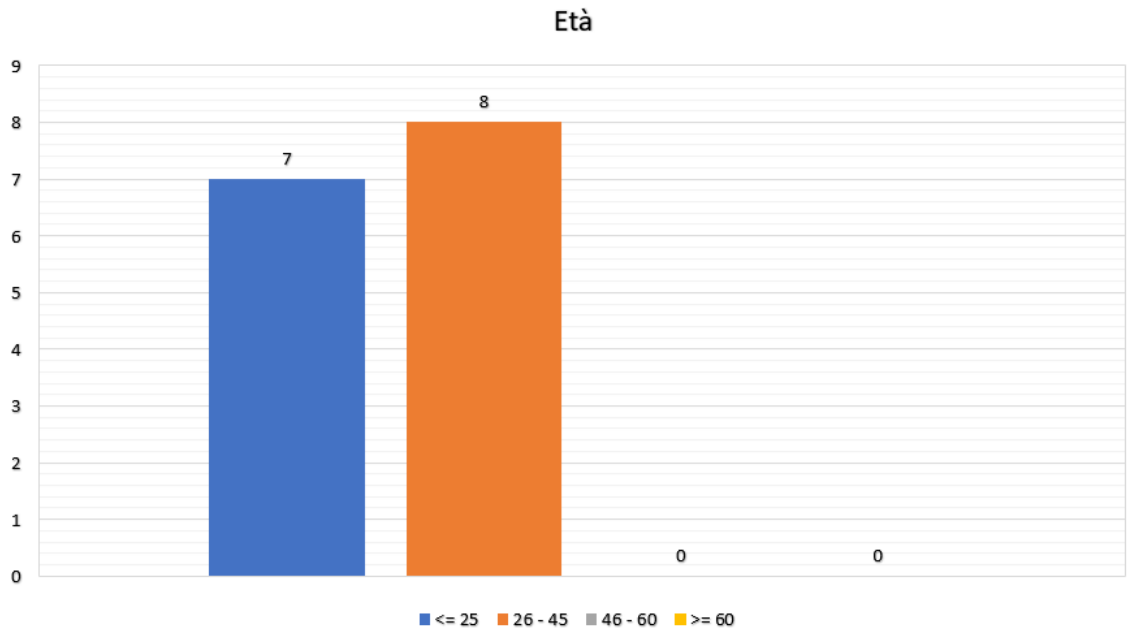


Grafico 3. L'età degli annotatori.

Dal Grafico 4, che mostra la distribuzione degli annotatori a seconda del titolo di studio, risulta evidente la prevalenza degli annotatori con una laurea rispetto agli altri gruppi, sono 9 in totale. I gruppi con i soggetti aventi un diploma di scuola media, un dottorato di ricerca o altro sono rappresentati meno e contengono solo 2 persone ciascuno. Non compare nessuno con la sola licenza media.

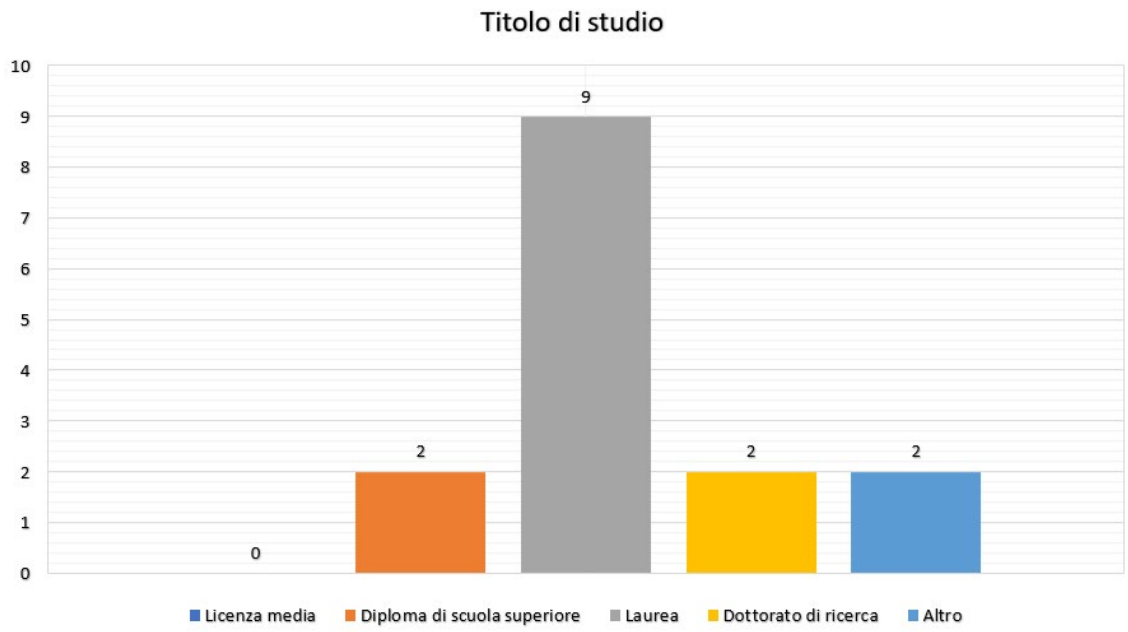


Grafico 4. Il titolo di studio degli annotatori.

Dal grafico 5, che mostra la distribuzione degli annotatori rispetto al livello di competenza in italiano, si osserva la prevalenza dei gruppi di annotatori con livelli di competenza intermedia e avanzata e corrispondono a 7 e 6. Il gruppo di principianti è rappresentato da 2 annotatori.

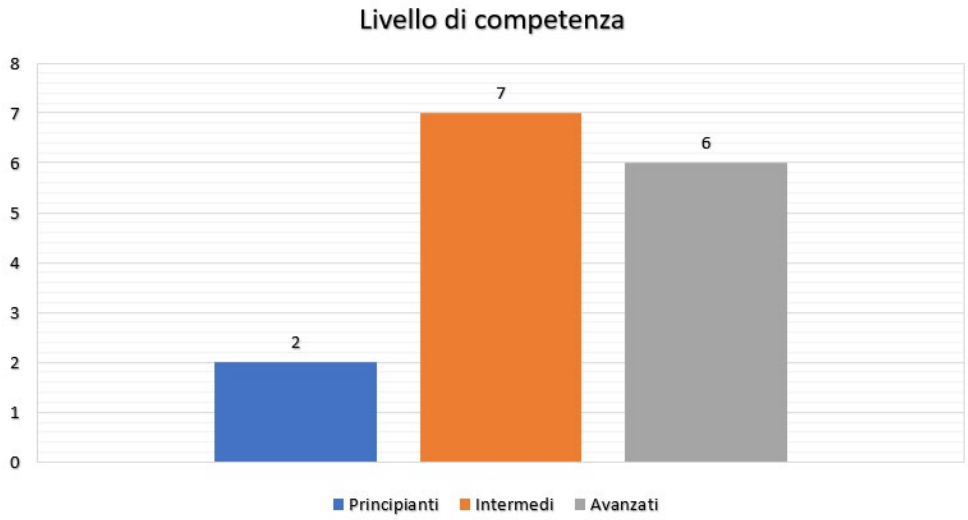


Grafico 5. Il livello di competenza in italiano degli annotatori.

Il questionario permetteva di selezionare più di un modo attraverso cui è stata appresa la lingua italiana. Per questo motivo i dati sono di un numero maggiore rispetto al totale degli annotatori. Diversamente da quanto è stato osservato per i dati anagrafici precedenti, dal Grafico 6 non risulta una prevalenza significativa di nessun gruppo. Il gruppo di soggetti che ha imparato la lingua italiana seguendo dei corsi online è composto da un singolo soggetto, mentre gli altri sono rappresentati da un numero maggiore di persone. La maggioranza degli annotatori ha frequentato una scuola di lingua nel proprio paese d'origine.

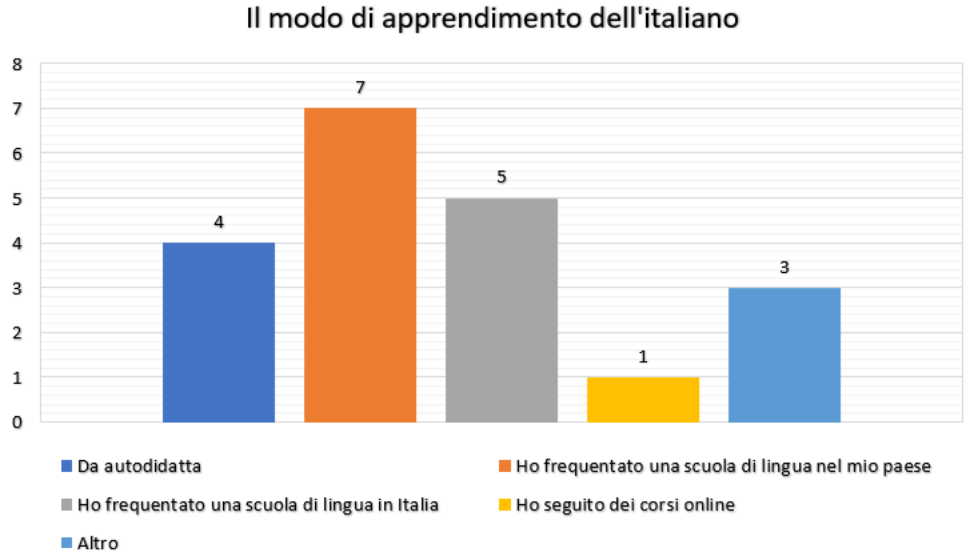


Grafico 6. Il modo di imparare l'italiano.

Dopo l'analisi delle informazioni anagrafiche fornite dagli annotatori si può concludere che il campione degli annotatori è rappresentato prevalentemente da donne, l'età massima degli annotatori non supera 45 anni, la maggioranza degli annotatori ha una laurea e gli annotatori hanno appreso la lingua L2 in diversi modi senza la prevalenza di un modo in particolare.

2.4 Accordo tra gli annotatori

Il grado di accordo fra gli annotatori si può calcolare usando delle metriche apposite che consentono di misurare l'omogeneità dei giudizi espressi dai soggetti. Un alto grado di accordo tra gli annotatori è indice del fatto che gli annotatori hanno interiorizzato le istruzioni di annotazione correttamente e producono annotazioni coerenti con tali istruzioni. Questo è garanzia della qualità dei dati raccolti e dei risultati ottenuti con l'esperimento.

Come riassunto in Gagliardi (2018), il calcolo dell'accordo fra annotatori tiene conto delle seguenti informazioni:

- il numero totale di annotatori;
- il numero di unità da annotare;
- le possibili categorie da assegnare alle unità.

Nel presente studio gli annotatori sono le persone che hanno svolto il questionario mentre le unità sono le frasi e le categorie corrispondono ai giudizi sulla complessità linguistica che gli annotatori hanno assegnato ad ogni frase.

Esistono varie metriche per misurare un accordo tra gli annotatori a seconda dei tipi di dati considerati. I principali sono:

- k di Cohen: misura un accordo tra due annotatori e considera la possibilità di accordo dovuto al caso;

- k di Fleiss: estensione di k di Cohen con cui si possono prendere in considerazione più di due annotatori;
- π di Scott: assume che gli annotatori hanno la stessa distribuzione delle risposte;
- α di Krippendorff: prende in considerazione qualsiasi numero di annotatori e ammette i dati mancanti.

Ai fini dello studio è stata utilizzata l'alpha di Krippendorff perché permette di coinvolgere un alto numero di annotatori. Alpha di Krippendorff è un coefficiente di attendibilità introdotto negli anni '70 dal professore Klaus Krippendorff per misurare un accordo tra gli annotatori.
Alpha (α) di Krippendorff si calcola con l'utilizzo della seguente formula generale:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Formula 1. Alpha di Krippendorff.

dove D_o corrisponde al disaccordo osservato tra i valori calcolato secondo la seguente formula:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\ metric} \delta^2_{ck}$$

Formula 2. Do.

e D_e è un disaccordo dovuto al caso e si calcola grazie alla seguente formula:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k\ metric} \delta^2_{ck}$$

Formula 3. De.

I valori estremi di alpha sono:

- α = 1 significa un accordo perfetto;
- α = 0 indica assenza di un accordo;
- α < 0 indica il disaccordo sistematico che supera quello casuale.

Più grande è il valore di alpha, più attendibili sono i dati ricevuti dagli annotatori (Krippendorff, 2007).

L’accordo è stato calcolato per ciascuno dei tre gruppi di annotatori raggruppati a seconda del grado di competenza della lingua italiana. I dati di ciascun gruppo sono riportati all’interno di una tabella. Ciascuna tabella contiene i giudizi assegnati dagli annotatori. In ogni tabella le righe rappresentano un annotatore e le colonne le frasi del questionario.

Per calcolare l'alpha di Krippendorff è stato scritto un programma nel linguaggio di programmazione ad alto livello Python modificando l’implementazione di Thomas Grill disponibile su GitHub[2]. Prima di calcolare alpha, il programma accede ai dati contenuti nelle tabelle salvate in un file csv grazie alla libreria csv e la sua funzione *reader*. Il seguente codice contiene un frammento del programma che apre il file csv e trasforma i dati per l'elaborazione successiva:

```
import csv
data = []
with open(sys.argv[1]) as csvfile:
      r = csv.reader(csvfile)
      for row in r:
            row = ''.join(row).replace(';;', ';*;').replace(';', '
').replace('п»ї', '')
            data.append(row)
data = tuple(data)
missing = '*'
array = [d.split() for d in data]
```

Successivamente il programma calcola alpha come dimostrato nelle Formule 1,2 e 3 dopo aver adattato un array al formato di calcolo. Il codice che segue riporta il calcolo di alpha:

2 https://pypi.org/project/krippendorff/

```
Do = 0.
for grades in units.values():
    if np_metric:
        gr = np.asarray(grades)
        Du = sum(np.sum(metric(gr, gri)) for gri in gr)
    else:
        Du = sum(metric(gi, gj) for gi in grades for gj in grades)
    Do += Du/float(len(grades)-1)
Do /= float(n)
if Do == 0:
    return 1.
De = 0.
for g1 in units.values():
    if np_metric:
        d1 = np.asarray(g1)
        for g2 in units.values():
            De += sum(np.sum(metric(d1, gj)) for gj in g2)
    else:
        for g2 in units.values():
            De += sum(metric(gi, gj) for gi in g1 for gj in g2)
De /= float(n*(n-1))
return 1.-Do/De if (Do and De) else 1.
```

I risultati del calcolo dell'accordo per ciascun gruppo di annotatori sono i seguenti:

- gruppo di principianti: $\alpha = 0.360$;
- gruppo di intermedi: $\alpha = 0.445$;
- gruppo di avanzati: $\alpha = 0.369$;
- il risultato di tutti i gruppi uniti: $\alpha = 0.395$;

Il coefficiente di accordo più alto è stato ottenuto dal gruppo con un livello intermedio di conoscenza dell'italiano, i principianti e gli avanzati hanno conseguito un risultato simile. Anche se in Linguistica Computazionale solitamente si tende a giudicare valori di accordo come questi troppo bassi per garantire l'affidabilità dei dati, la difficoltà del compito e la soggettività dei giudizi ci porta a considerarli come valori di accordo accettabili per i fini dello studio e quindi si può proseguire con i prossimi passaggi d'analisi.

3 Monitoraggio linguistico e analisi delle caratteristiche linguistiche

In questo capitolo verranno introdotti i principi della metodologia di monitoraggio linguistico del testo e come questa è stata applicata ai testi usati in questo studio. Nella sezione 3.1 verranno spiegati i fondamenti della metodologia. In particolare, la sezione 3.1.1 descrive lo strumento utilizzato per effettuare il monitoraggio linguistico e il suo funzionamento. La ricostruzione del profilo linguistico dei testi è stato eseguito tramite Profiling-UD[3] un tool sviluppato dall'ItaliaNLP Lab, un laboratorio dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) (Brunato et al., 2020). Nella sezione 3.2 verranno descritte le caratteristiche estratte dalle frasi oggetto dello studio grazie allo strumento di monitoraggio linguistico. Queste caratteristiche sono raggruppate in 4 categorie che verranno descritte ciascuna in una sezione separata. La sezione 3.2.1 riporterà le caratteristiche linguistiche di base. Nella sezione 3.2.2 verranno descritte le caratteristiche lessicali. La sezione 3.2.3 invece sarà dedicata alle caratteristiche linguistiche morfo-sintattiche. Infine, nella sezione 3.2.4 verranno riportate le caratteristiche sintattiche. Nel descrivere le caratteristiche monitorate dallo strumento, verranno anche riportate i valori di tali caratteristiche per il corpus utilizzato nello studio descritto in questa tesi.

3.1 Fondamenti della metodologia

Usando tecnologie linguistico-computazionali è possibile ricostruire il profilo linguistico di un testo monitorando un vasto spettro di parametri linguistici. Grazie a queste tecnologie è infatti possibile individuare la struttura linguistica del testo, rappresentarla in modo esplicito e accedere al contenuto informativo del testo stesso.

Il monitoraggio linguistico è una ricostruzione del profilo linguistico di un testo e consiste in un processo di analisi che si basa sull'uso di tecnologie per l'annotazione linguistica automatica al fine di monitorare un ampio spettro di parametri che riguardano

3 http://www.italianlp.it/demo/profiling-ud/

i diversi livelli di descrizione linguistica (Montemagni, 2013). Prerequisito del monitoraggio linguistico è l'annotazione linguistica a livello morfo-sintattico e, grazie ad essa, possiamo identificare specifici costrutti sintattici, morfosintattici e informazioni relative alle strutture semantiche dei testi che stiamo analizzando.

3.1.1 Analisi linguistica automatica

Per gli scopi di questo studio, le frasi del corpus sono state elaborate con l'applicazione Profiling-UD (Brunato, 2020) che consente di effettuare il monitoraggio linguistico di un testo oppure di una grande collezione di testi. Con questo strumento è possibile estrarre circa 130 caratteristiche linguistiche appartenenti a diversi livelli di analisi linguistica. Profiling-UD è basato sul modello di Universal Dependencies che permette di fare delle analisi in varie lingue. Lo strumento prevede due fasi: l'annotazione linguistica e il monitoraggio linguistico. La prima fase preliminare è supportata da UDPipe, un software che effettua l'identificazione della struttura linguistica del testo.
L'identificazione della struttura linguistica del testo prevede la successione dei passaggi di analisi linguistica a livelli di complessità crescente:

- segmentazione del testo in frasi;
- tokenizzazione: segmentazione delle frasi in unità minime di analisi chiamate "token";
- lemmatizzazione e analisi morfo-sintattica: consiste nel ricondurre un token al relativo esponente lessicale chiamato "lemma" e successivamente assegnare a questo token un'informazione relativa alla categoria grammaticale che ha in un dato contesto;
- analisi sintattica: consiste nel analizzare una frase secondo le relazioni di dipendenza tra le parole;

I risultati dell'analisi della struttura linguistica sono disponibili in un formato tabellare CoNLL in cui ogni riga rappresenta un token e contiene il numero della sua posizione e ogni colonna rappresenta le proprietà di questo token ai diversi livelli di analisi linguistica. Un esempio dell'analisi automatica di una frase è riportato nella Figura 5.

	ID	Token	Lemma	C-POS	F-POS	Morphosyntactic feats	HEAD	DEP
1	1	Mi	mi	P	PC	num:s gen:n per:1	2	clit
	2	congratulo	congratularsi	V	V	num:s mod:i per:1 ten:p	0	ROOT
	3	con	con	E	E		2	comp
	4	lui	lui	P	PE	num:s gen:m per:3	3	prep
	5	per	per	E	E		2	comp
	6	l'	il	R	RD	num:s gen:n	8	det
	7	eccellente	eccellente	A	A	num:s gen:n	8	mod
	8	relazione	relazione	S	S	num:s gen:f	5	prep
	9	.	.	F	FS		2	punc

Figura 5. Esempio di una rapresentazione tabellare CoNLL-U di una frase.

Nell'esempio sopra è possibile notare il risultato di un processo di analisi. Nella prima colonna si trova un identificatore univoco per ogni token rappresentato da un numero che cresce progressivamente. Nella seconda colonna si trova un token della frase e nella terza colonna si trova il lemma corrispondente. Ad ogni token è associata informazione relativa alla categoria grammaticale. Questa informazione è integrata nella sesta colonna con le specificazioni morfologiche riguardanti le categorie flessionali. Le ultime due colonne riportano una descrizione della frase dal punto di vista delle relazioni binarie di dipendenza tra le parole (di solito si tratta di relazioni binarie asimmetriche tra una testa e un dipendente). Nella settima colonna si trova un identificatore univoco della testa da cui dipende il token e nell'ottava colonna è specificato il tipo di dipendenza.

La Figura 6 riporta una rappresentazione grafica dell'albero di dipendenze sintattiche in Figura 5, all'interno della quale gli archi segnalano la dipendenza sintattica tra la testa e il dipendente.

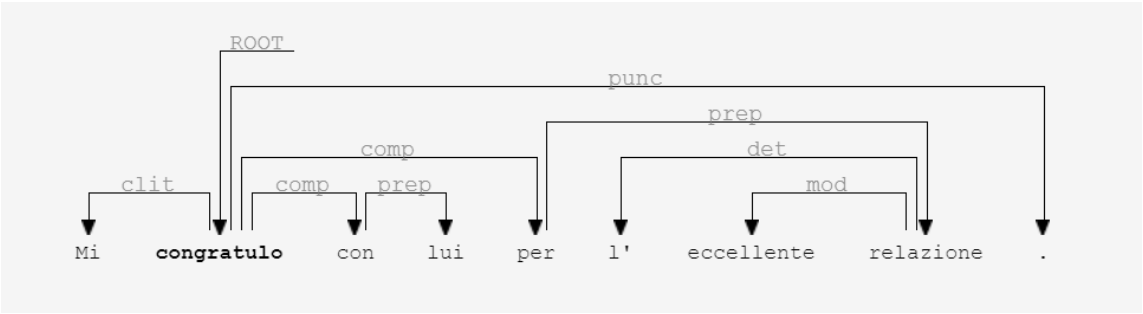


Figura 6. Rappresentazione ad albero a dipendenze della frase dalla figura 5.

Le frasi, arricchite con le informazioni descritte precedentemente, sono pronte per la successiva elaborazione automatica che consentirà di identificare un'ampia collezione di parametri utili per i compiti di monitoraggio linguistico.

3.2 Analisi del profilo linguistico del corpus

Nella seconda fase di analisi chiamata "monitoraggio linguistico" sono state estratte dalle frasi del corpus, con lo strumento Profiling-UD, 126 caratteristiche linguistiche rappresentative di diversi livelli di annotazione linguistica. Queste caratteristiche rappresentano un'ampia collezione di fenomeni linguistici rilevanti per lo studio. Il risultato è estratto nel formato tabellare dove ogni caratteristica si trova in una colonna separata e ad ogni colonna è associato un valore corrispondente. Le caratteristiche ottenute possono essere raggruppate in quattro categorie principali:

- caratteristiche di base;
- caratteristiche lessicali;
- caratteristiche morfo-sintattiche;
- caratteristiche sintattiche.

Nelle prossime sezioni saranno descritte le caratteristiche di ciascuna categoria. Per ogni caratteristica estratta verrà riportato il valore ottenuto dall'analisi del corpus oggetto dello studio.

3.2.1 Caratteristiche di base

Le caratteristiche linguistiche di base si riferiscono al livello di analisi più semplice e basilare e sono chiamate anche "*le caratteristiche del testo grezzo*". Queste caratteristiche sono quattro:

- *n_sentences*: la lunghezza del testo in termini di numero totale delle frasi che lo compongono;
- *n_tokens*: la lunghezza del testo in termini di numero di tokens;
- *tokens_per_sent*: la lunghezza media di una frase del testo calcolata in termini di numero di tokens per una frase;
- *char_per_tok*: la lunghezza media di un token calcolata in termini di numero medio dei caratteri per un token eccetto i segni di punteggiatura.

La lunghezza di una frase e la lunghezza di una parola sono solitamente considerate degli indicatori di complessità sintattica e di complessità lessicale di una frase. Sono spesso utilizzate per il calcolo tradizionale della leggibilità.

Nel Grafico 7 si osservano i valori medi di caratteristiche di base calcolati su tutte le frasi del corpus.

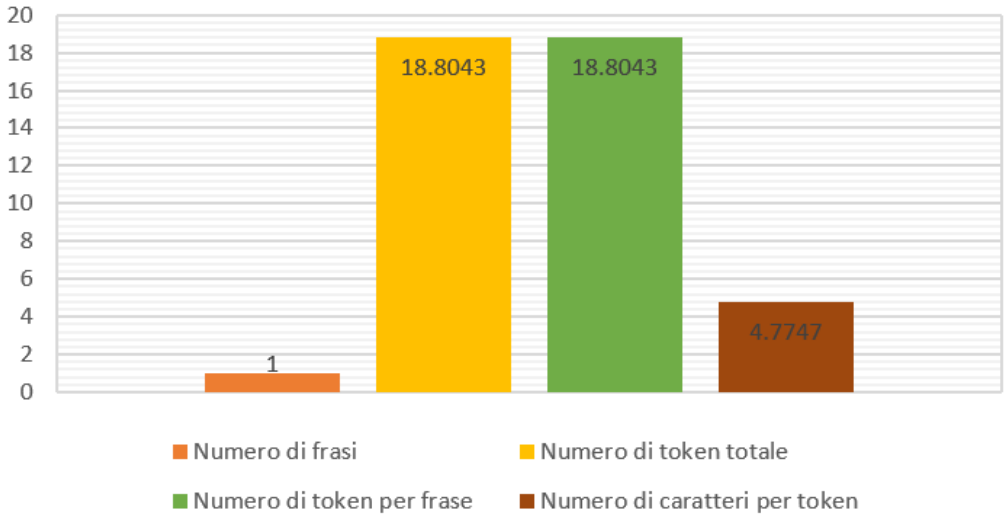


Grafico 7. Le medie dei valori delle caratteristiche di base estratti dalle frasi del corpus.

Il corpus utilizzato per il presente studio è composto dalle singole frasi quindi il numero totale medio delle frasi è uno. Per lo stesso motivo il numero di token totale è uguale al numero di token per frase.

3.2.2 Caratteristiche lessicali

La varietà lessicale è solitamente calcolata con il rapporto tipo/unità (type/token ratio) che può essere considerato un indice della ricchezza lessicale di un testo. I token sono le unità dell'analisi linguistica mentre la parola tipo è una unità distinta. L'indice si calcola con la seguente formula:

$$\frac{|Vt|}{|T|}$$

dove |Vt| è la cardinalità di un vocabolario di un testo definito come l'insieme delle parole tipo che ricorrono in un testo e |T| è la lunghezza di questo testo. I valori di rapporto oscillano tra 0 e 1. Nei casi in cui il valore tende a 0, questo sta ad indicare un testo il cui vocabolario è poco vario. Quando invece il valore tende a 1 il vocabolario risulta molto ricco. Se il rapporto è pari a 1 allora il testo è formato interamente dalle parole *hapax*, parole che ricorrono solo una volta nel testo. A causa della sensibilità del TTR alla lunghezza del testo, Profiling-UD calcola il rapporto per due porzioni di testo che contengono i primi 100 e i primi 200 tokens e restituisce le seguenti caratteristiche:

- ttr_lemma_chunks_100: TTR calcolato rispetto ai lemmi nei primi 100 tokens del testo;
- ttr_lemma_chunks_200: TTR calcolato rispetto ai lemmi nei primi 200 tokens del testo;
- ttr_form_chunks_100: TTR calcolato rispetto alle forme delle parole nei primi 100 tokens del testo;
- ttr_form_chunks_200: TTR calcolato rispetto alle forme delle parole nei primi 200 tokens del testo;

Nel presente studio sono utilizzate solamente le frasi distinte che non raggiungono a 100 tokens ciascuna e per questo motivo non è stato possibile estrarre le caratteristiche lessicali con il Profiling-UD.

3.2.3 Caratteristiche morfo-sintattiche

La prima caratteristica morfo-sintattica estratta con il Profiling-UD è la distribuzione di categorie grammaticali. Al livello di annotazione morfo-sintattica a ogni token vengono assegnati le informazioni relative alla categoria grammaticale (detta anche "parte del discorso") che il token ha nel contesto specifico.
Profiling-UD calcola la distribuzione percentuale nel testo di 17 principali categorie grammaticali definiti nell'Universal Dependencies Part-of-Speech tagset[4] che è ulteriormente diviso in tre gruppi:

1. Classi aperte di parole:
    - ADJ: aggettivi (*upos_dist_ADJ*);
    - ADV: avverbi (*upos_dist_ADV*);
    - INTJ: interiezioni (*upos_dist_INTJ*);
    - NOUN: nomi (*upos_dist_NOUN*);
    - PROPN: nomi propri (*upos_dist_PROPN*);
    - VERB: verbi (*upos_dist_VERB*);
2. Classi chiuse di parole:
    - ADP: apposizioni (*upos_dist_ADP*);
    - AUX: ausiliari *(upos_dist_AUX*);
    - CCONJ: congiunzioni coordinanti (*upos_dist_CCONJ*);
    - DET: articoli determinativi (*upos_dist_DET*);
    - NUM: numerali (*upos_dist_NUM*);
    - PART: particelle grammaticali (*upos_dist_PART*);
    - PRON: pronomi (*upos_dist_PRON*);
    - SCONJ: congiunzioni di subordinazione (*upos_dist_SCONJ*);

4 https://universaldependencies.org/u/pos/index.html

3. Altre:
    - PUNCT: punteggiatura (*upos_dist_PUNCT*);
    - SUM: simboli (*upos_dist_SUM*);
    - X: altro (*upos_dist_X*);

Per ogni categoria grammaticale viene calcolata la distribuzione statistica in percentuali. Il Grafico 8 riporta la distribuzione media delle principali categorie grammaticali nelle frasi del corpus.

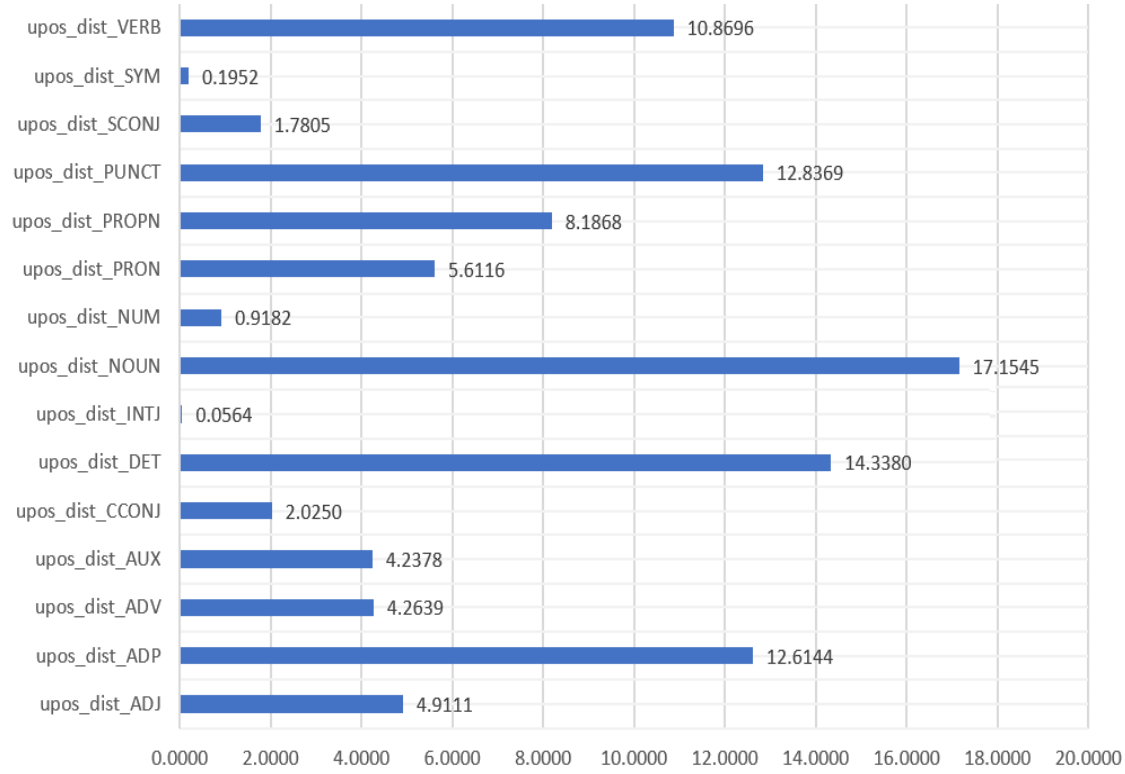


Grafico 8. La distribuzione media delle categorie grammaticali nelle frasi.

La notevole prevalenza si osserva per la classe di nomi che costituiscono in media il 17.1545% di una frase. I valori alti si presentano anche nelle classi di articoli determinativi (14.3380%), segni di punteggiatura (12.8369%), apposizioni (12.6144%), verbi (10.8696%) e nomi propri (8.1868%). Sono invece meno presenti le classi di pronomi (5.6116%), aggettivi (4.9111%), avverbi (4.2639%), ausiliari (4.2378%), congiunzioni coordinanti (2.0250%) congiunzioni di subordinazione (1.7805%) e

numerali (0.9182%). Le classi di simboli e di interiezioni non sono quasi rappresentati e costituiscono 0.1952% e 0.0564% di una frase in media rispettivamente.

La seconda caratteristica estratta con il Profiling-UD è la densità lessicale. Il suo valore è calcolato come il rapporto tra le parole lessicalmente piene (nomi, nomi propri, verbi, aggettivi, avverbi) e il numero totale di parole in un documento. Il suo valore oscilla tra 0 e 1. Se il valore tende a 0 significa che il testo è poco informativo. Se invece il valore si avvicina a 1, significa che il testo è ricco di contenuti. La densità media delle frasi del corpus è pari a 0.5221. È stato ottenuto un valore così alto perché il corpus oggetto di analisi consiste solo in frasi distinte la cui lunghezza massima è pari a 61. La densità lessicale, così come il type-token ratio, è molto sensibile alla lunghezza del testo su cui è calcolata. Il valore alto di densità lessicale è spiegato dalla lunghezza bassa delle frasi analizzate.

Per ogni verbo lessicale e ausiliario sono state analizzate le categorie flessionali sulla base di specifiche dell'Universal Dependencies. Sono state calcolate le distribuzioni delle forme, dei tempi, dei numeri, delle persone, dei modi e dei generi verbali raggruppati secondo il seguente schema delle etichette proposta dalle Universal Dependencies per la lingua italiana:

1. Il tempo:
    - Fut: tempo futuro (*verbs_tense_dist_Fut*);
    - Imp: tempo imperfetto (*verbs_tense_dist_Imp*);
    - Past: tempo passato (*verbs_tense_dist_Past*);
    - Pres: tempo presente (*verbs_tense_dist_Pres*);
2. Il modo:
    - Cnd: modo condizionale (*verbs_mood_dist_Cnd*);
    - Imp: modo imperativo (*verbs_mood_dist_Imp*);
    - Ind: modo indicativo (*verbs_mood_dist_Ind*);
    - Sub: modo congiuntivo (*verbs_mood_dist_Sub*);
3. La forma verbale:
    - Fin: verbo finito (*verbs_form_dist_Fin*);

- Ger: gerundio (*verbs_form_dist_Ger*);
- Inf: verbo infinito (*verbs_form_dist_Inf*);
- Part: participio (*verbs_form_dist_Part*);

4. Il genere:
    - Masc: genere maschile (*verbs_gender_dist_Masc*);
    - Fem: genere femminile (*verbs_gender_dist_Fem*);
5. Il numero e la persona:
    - Plur: plurale:
        1. prima persona (*verbs_num_pers_dist_Plur+1*);
        2. seconda persona (*verbs_num_pers_dist_Plur+2*);
        3. terza persona (*verbs_num_pers_dist_Plur+3*);
    - Sing: singolare:
        1. prima persona (*verbs_num_pers_dist_Sing+1*);
        2. seconda persona (*verbs_num_pers_dist_Sing+2*);
        3. terza persona (*verbs_num_pers_dist_Sing+3*);

Il Grafico 9 mostra le distribuzioni medie delle caratteristiche morfo-sintattiche riguardanti il tempo verbale nelle frasi del corpus. Si osserva la prevalenza del tempo verbale presente rispetto ad altri tempi (42.9620%) seguito dal tempo passato (28.3605%). Il tempo futuro e il tempo imperfetto sono invece in minoranza e sono rappresentati in 7.3098% e 3.9764% rispettivamente.

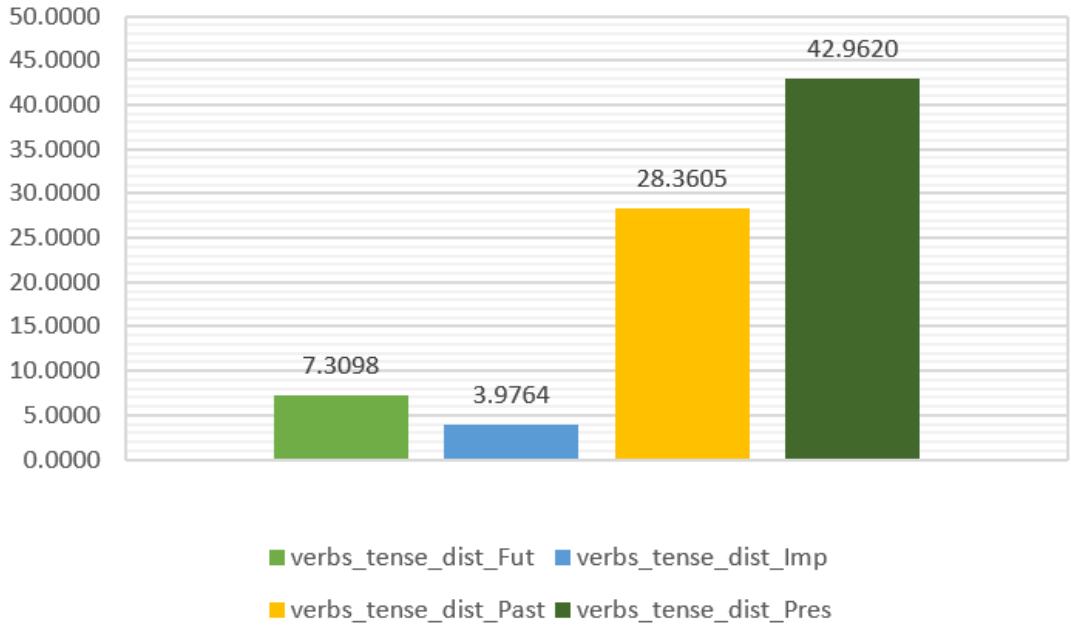


Grafico 9. La distribuzione media dei tempi verbali nelle frasi.

Nel Grafico 10 si osservano le distribuzioni medie delle caratteristiche linguistiche connesse al modo verbale nelle frasi del corpus. La maggior parte delle frasi contiene i verbi nel modo indicativo (58.1522%) che è il modo più diffuso. I modi imperativo (4.3478%), condizionale (0.8605%) e congiuntivo (2.4004%) si trovano in poche frasi.

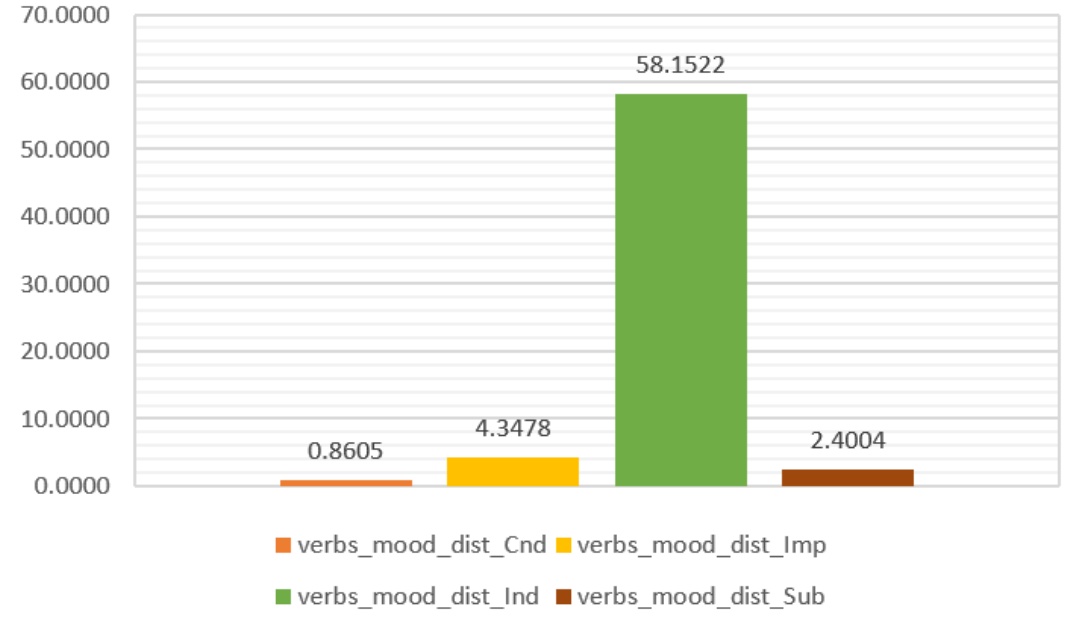


Grafico 10. La distribuzione media dei modi verbali nelle frasi.

Il Grafico 11 mostra le distribuzioni medie delle caratteristiche linguistiche che riguardano le forme verbali nelle frasi. La maggior parte dei verbi nelle frasi sono nella forma finita (47.7355%). I participi (19.6299%) e i verbi infiniti (17.7704%) occupano posizioni vicine nella scala. I gerundi sono in minoranza (2.9076%).

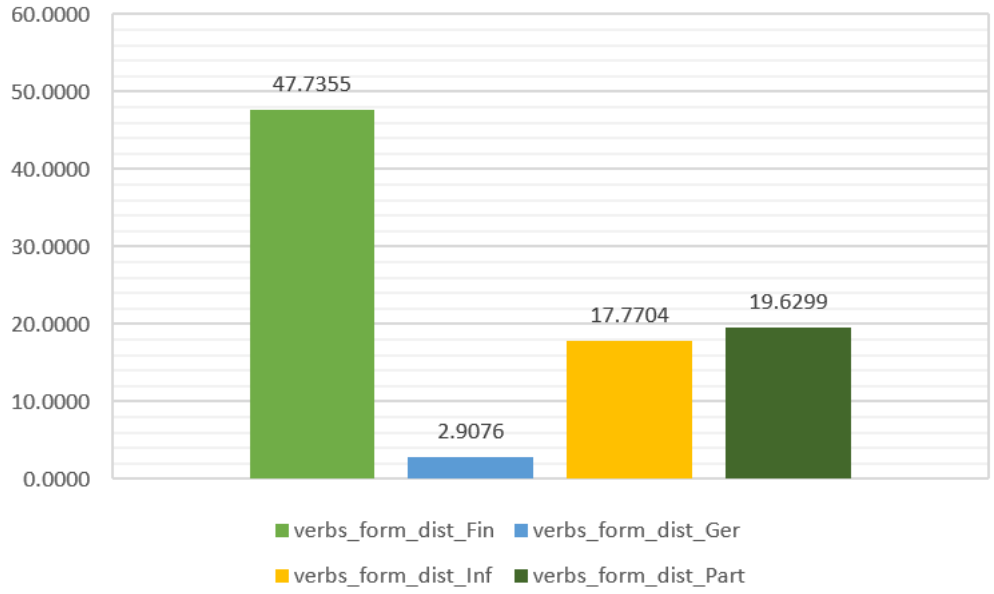


Grafico 11. La distribuzione media delle forme verbali nelle frasi.

Il Grafico 12 riporta la distribuzione delle caratteristiche che riguardano il numero e la persona dei verbi. Vi è una prevalenza notevole dei verbi in terza persona singolare (38.9946%). La minor parte dei verbi stanno in seconda persona plurale (1.9928%). La distribuzione media di altri gruppi dei verbi oscilla tra 4 e 11%.

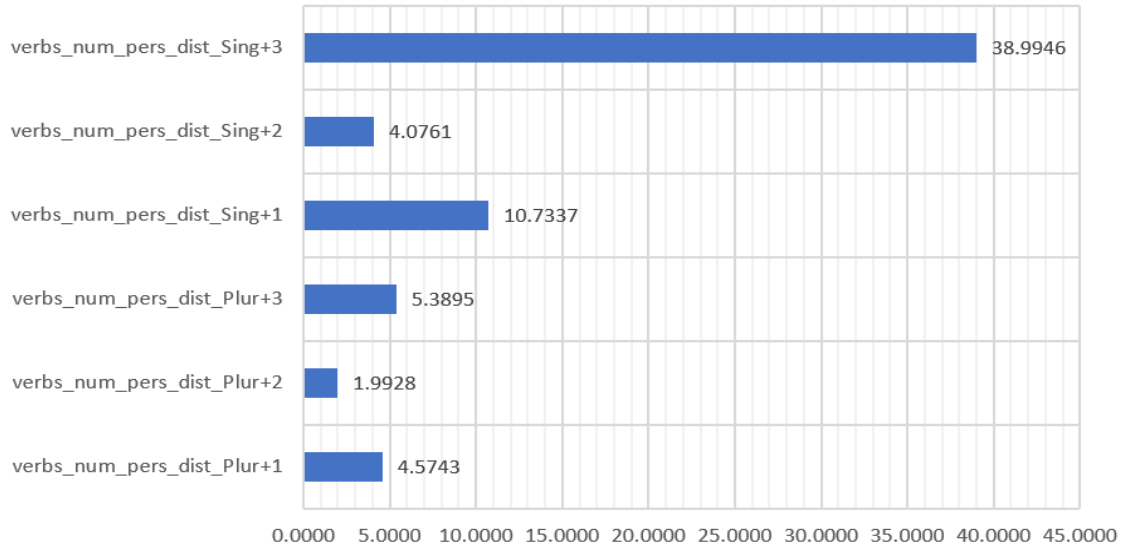


Grafico 12. La distribuzione media delle caratteristiche legate al numero e la persona nelle frasi.

3.2.4 Caratteristiche sintattiche

Le caratteristiche linguistiche sintattiche si riferiscono alle informazioni relative all'analisi sintattica delle frasi e sono ulteriormente suddivise nelle categorie che riguardano la struttura del predicato verbale, la struttura dell'albero sintattico, le relazioni sintattiche e i fenomeni di subordinazione. L'annotazione sintattica rende il testo adatto a molte forme di esplorazione e analisi avanzata. Esistono due principali approcci teorici alla sintassi che consentono di rappresentare la struttura sintattica di una frase (Lenci et al., 2006):

- rappresentazioni a costituenti: si basano sull'identificazione di costituenti sintattici e delle loro relazioni gerarchiche;
- rappresentazioni a dipendenze o funzionali: si basano sulle relazioni grammaticali binarie di dipendenza tra le parole di una frase.

Profiling-UD utilizza un approccio basato sulla rappresentazione a dipendenze per esaminare la struttura grammaticale delle frasi. Questo metodo consente di costruire un albero sintattico di una frase con le rispettive relazioni di dipendenza tra le parole. La Figura 7 riporta la resa grafica di un esempio di questo tipo di rappresentazione costruita

partendo dal file in formato CoNLL-U. La frase di esempio è "Signora presidente, se il verbale rispecchierà correttamente il voto del mio gruppo non avrò alcuna obiezione."

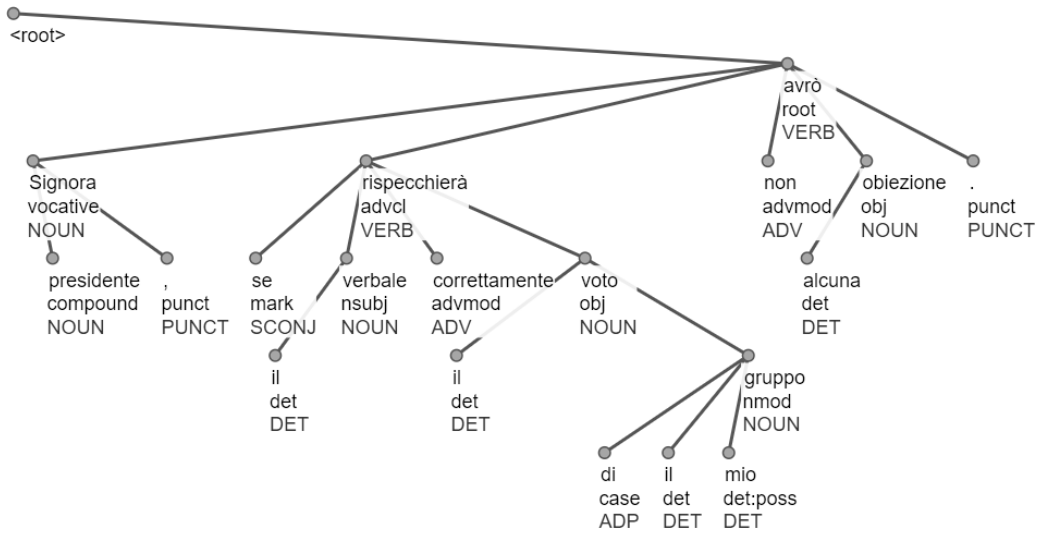


Figura 7. Esempio della rappresentazione a dipendenze della struttura sintattica di una frase.

Il gruppo di caratteristiche sintattiche relative alla struttura del predicato verbale contiene le seguenti caratteristiche:

- distribuzione media di teste verbali per frase in un testo (*verbal_head_per_sent*);
- distribuzione media di radici verbali calcolata in percentuali (*verbal_root_perc*);
- arità dei verbi, ovvero una caratteristica che cattura il numero di dipendenti di una testa verbale e che viene calcolata in questo caso come il numero medio di collegamenti di dipendenza istanziati (*avg_verb_edges*);
- distribuzione dei verbi per le classi di arità (ovvero, verbi raggruppati in base al numero di dipendenti diretti che presentano nella frase). (*verb_edges_dist_**).

Nelle frasi del corpus si osservano i valori vicini alla distribuzione media di teste verbali (2.2935%) e dell'arità dei verbi (2.3530%). La distribuzione di radici verbali (85.8696%) non è significativa per il presente studio perché la maggior parte del corpus consiste principalmente in frasi brevi con una sola testa verbale.

Il Grafico 13 mostra la distribuzione media dei verbi per le classi di arità.

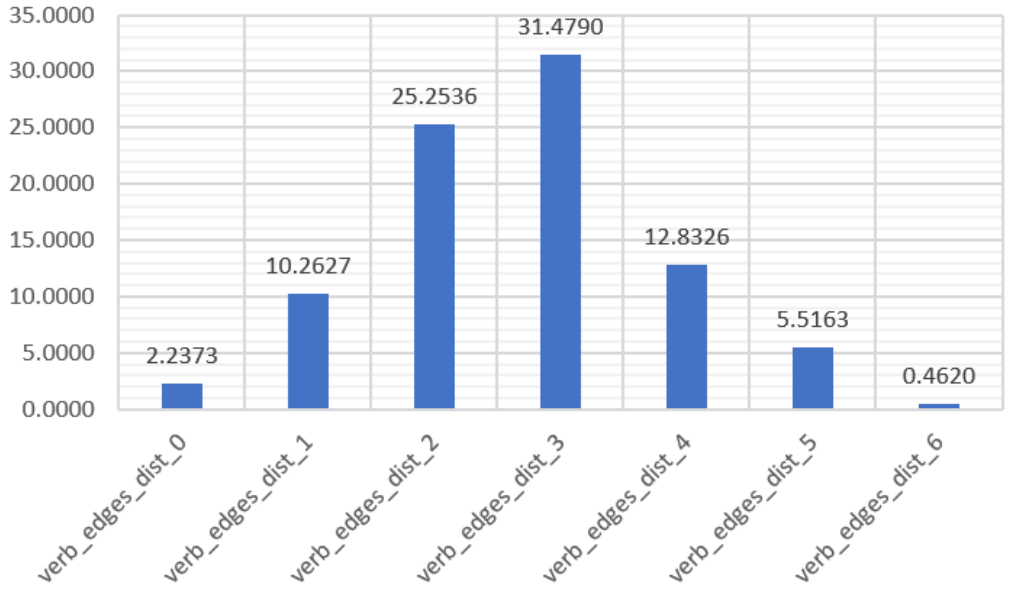


Grafico 13. La distribuzione media delle arità dei verbi delle frasi.

Nel grafico 13 si osserva che i valori delle distribuzioni medie delle classi di arità nel corpus si incrementano avvicinandosi all'arità media che nel caso presente è pari a tre e che ha un valore più grande rispetto alle altre classi di arità (31.4790%). I valori più bassi si registrano per le classi dei verbi con arità zero (2.2373%) e sei (0.4620%).

Il gruppo di caratteristiche sintattiche che riguardano la struttura dell'albero sintattico contiene le seguenti caratteristiche:

- media di profondità massima degli alberi sintattici estratta da ogni frase di un testo (*avg_max_depth*). La profondità massima di una frase si calcola come un percorso massimo, in termini di collegamenti a dipendenza, da una radice di un albero sintattico a un suo foglio;
- lunghezza media di una proposizione calcolata in termini di numero medio di tokens per una proposizione dove il numero delle proposizioni corrisponde a un ratio tra il numero di tokens in una frase e il numero delle loro teste verbali (*avg_token_per_clause*);
- distanza media tra una testa e un suo dipendente calcolata come il numero di parole che stanno tra la testa sintattica e i suoi dipendenti (*avg_links_len*);
- il valore medio della lunghezza massima di collegamenti di dipendenza è l'informazione complementare alla caratteristica precedente e corrisponde alla

lunghezza media di un collegamento più lungo per ogni frase di un testo (*avg_max_links_len*);

- il valore di lunghezza di collegamento di dipendenza massima in un testo calcolato in numero di tokens (*max_links_len*);
- lunghezza media delle catene preposizionali calcolata come il numero di complementi preposizionali dipendenti da testa nominale di una frase (*avg_prepositional_chain_len*);
- numero totale delle catene preposizionali estratti da tutte le frasi di un testo (*n_prepositional_chains*);
- distribuzione delle catene preposizionali per la profondità (*prep_dist_**).

Nel grafico 14 si osservano i valori medi delle caratteristiche sintattiche legati ai collegamenti di dipendenze nelle frasi del corpus. La media di profondità massima degli alberi sintattici delle frasi è pari a 4.2174. La lunghezza media delle clausole nelle frasi ha un valore alto (9.2708). Le due caratteristiche *avg_max_links_len* e *max_links_len* sono pari (9.0054) nel presente studio perché i testi di analisi sono rappresentati principalmente dalle singole frasi. La distanza media tra una testa e un suo dipendente nelle frasi è pari a 2.4432.

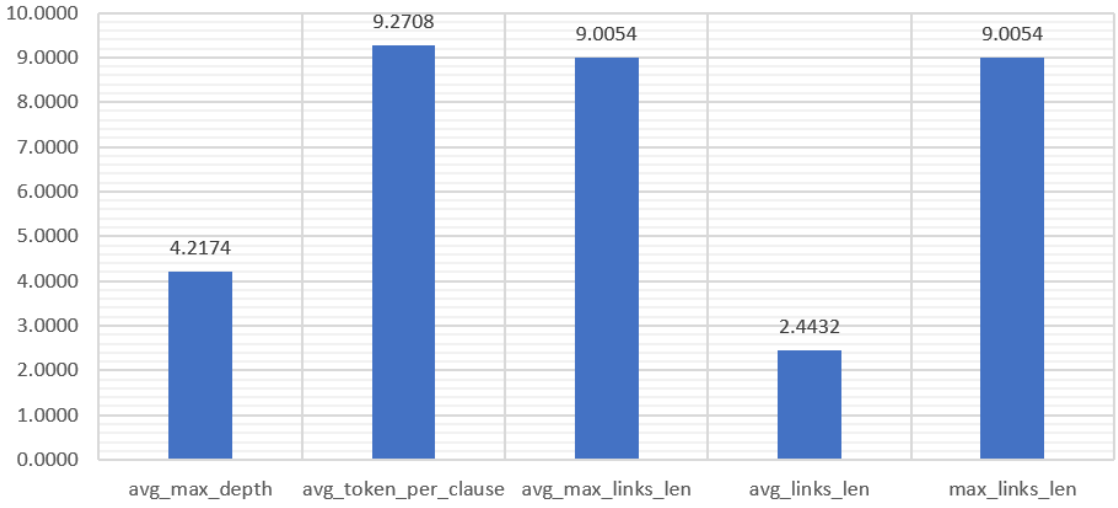


Grafico 14. Le caratteristiche sintattiche che riguardano i collegamenti di dipendenza nelle frasi.

La lunghezza media delle catene preposizionali nelle frasi del corpus è pari a 0.7917 e il valore medio del numero totale delle catene preposizionali estratti da tutte le frasi è pari

a 0.9728. Il Grafico 15 riporta le distribuzioni medie delle catene preposizionali per la profondità. Si nota la tendenza alla diminuzione del valore al crescere del numero di complementi. Le catene con un solo complemento mostrano una netta prevalenza e hanno un valore medio di 54.6196%. Registrano un valore più basso le catene con tre complementi (2.2645%).

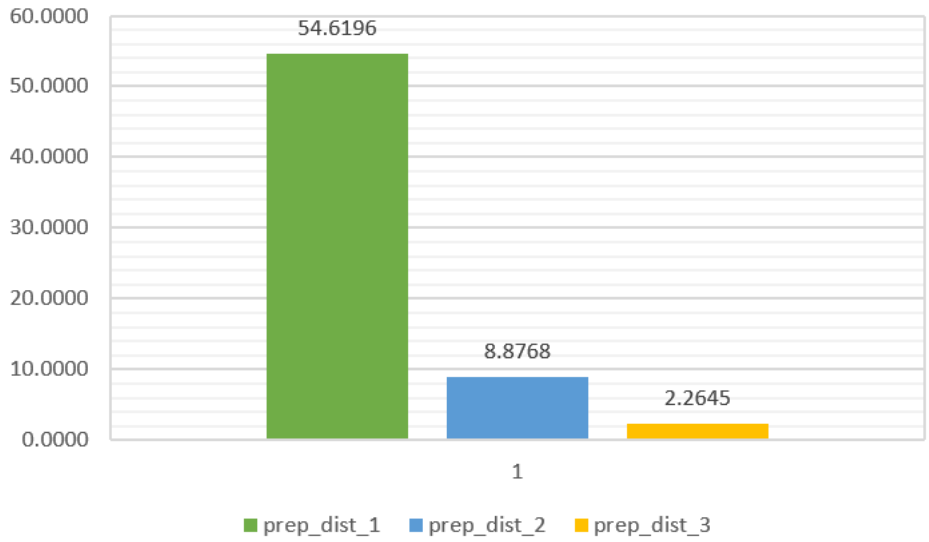


Grafico 15. Le caratteristiche sintattiche che si riferiscono alla distribuzione delle catene preposizionali nelle frasi per profondità.

Profiling-UD analizza il fenomeno di ordine di parole dal punto di vista della posizione del soggetto e dell'oggetto che costituiscono gli elementi principali di una frase. Lo strumento cattura la variazione dell'ordine delle parole in varie lingue e anche nelle diverse varietà di una stessa lingua. Con Profiling-UD è possibile estrarre le seguenti caratteristiche relative all'ordine delle parole:

- distribuzione di oggetti preverbali (*obj_pre*);
- distribuzione di oggetti postverbali (*obj_post*);
- distribuzione di soggetti preverbali (*subj_pre*);
- distribuzione di soggetti postverbali (*subj_post*);

Nel Grafico 16 si osservano le distribuzioni medie di caratteristiche linguistiche relative ai fenomeni dell'ordine delle parole nelle frasi. L'italiano come le altre lingue romanze (ad esempio, francese, spagnolo, portoghese) è la lingua in cui l'ordine principale delle parole è SVO (soggetto-verbo-oggetto). Questo particolare della lingua italiana è rispecchiato nel grafico dove la distribuzione media di soggetti che precedono il verbo è

pari a 52.8080% e la distribuzione media di oggetti che seguono il verbo è pari a 45.1359%. Comunque ci sono anche delle eccezioni in cui i soggetti stanno nella posizione postverbale (23.8225%) e gli oggetti si trovano nella posizione preverbale (14.1033%).

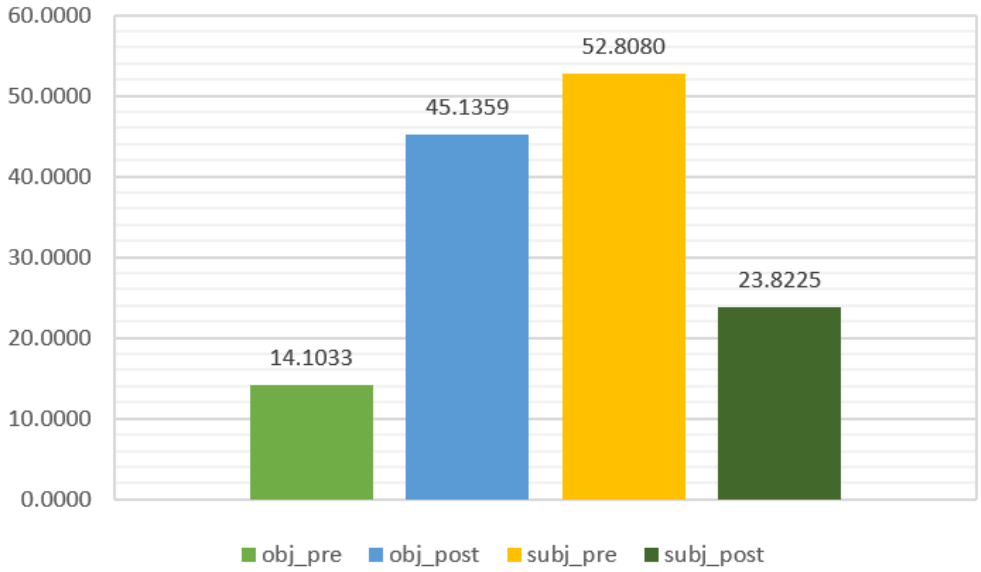


Grafico 16. La distribuzione media delle caratteristiche sintattiche legate all'ordine di soggetto e oggetto nelle frasi.

Le caratteristiche linguistiche estratte con lo strumento Profiling-UD per il corpus oggetto dello studio che descrivono le relazioni sintattiche dentro le frasi si riferiscono alla distribuzione in percentuale di 39 relazioni (*dep_dist_**) secondo la schema di annotazione a dipendenze delle Universal Dependencies[5]:

acl: proposizione che modifica nome	acl:relcl: proposizione relativa che modifica nome
advcl: proposizione avverbiale che modifica nome	advmod: modificatore avverbiale
amod: modificatore aggettivale	appos: modificatore apposizionale
aux: ausiliare	aux:pass: ausiliare passivo
case: marcatore del caso	cc: congiunzione coordinante
ccomp: proposizione complementare	compound: composto

5 https://universaldependencies.org/u/dep/index.html

conj: congiunzione	cop: copula
csubj: soggetto proposizionale	det: determinante
det:poss: determinante possessivo	det:predet: predeterminante
discourse: elemento discorsivo	dislocated: elementi dislocati
expl: espletivo	expl:impers: espletivo impersonale
fixed: espressione fissa	flat: espressione uniforme
flat:name: specificazione di *flat* per nomi	iobj: oggetto indiretto
mark: marcatore che introduce una frase subordinata	nmod: modificatore nominale
nsubj: soggetto nominale	nsubj:pass: soggetto nominale passivo
nummod: modificatore numerico	obj: oggetto
obl: funzione nominale obliqua	obl:agent: agente modificatore
parataxis: paratassi	punct: punteggiatura
root: radice	vocative: vocativo
xcomp: proposizione complementare aperta	

Il grafico 17 riporta le caratteristiche sintattiche relative agli argomenti nucleari di un predicato. Nelle frasi del corpus si osserva la prevalenza di soggetti nominali (5.1127%) e di oggetti (4.6149%). Le proposizioni complementari (1.0609%) e le proposizioni complementari aperte (1.0750%) hanno valori simili alle distribuzioni medie. Gli oggetti indiretti sono meno diffusi (0.5067%). I soggetti proposizionali, nella distribuzione, risultano numericamente inferiori rispetto a tutto il gruppo avendo un valore dello 0.1179%.

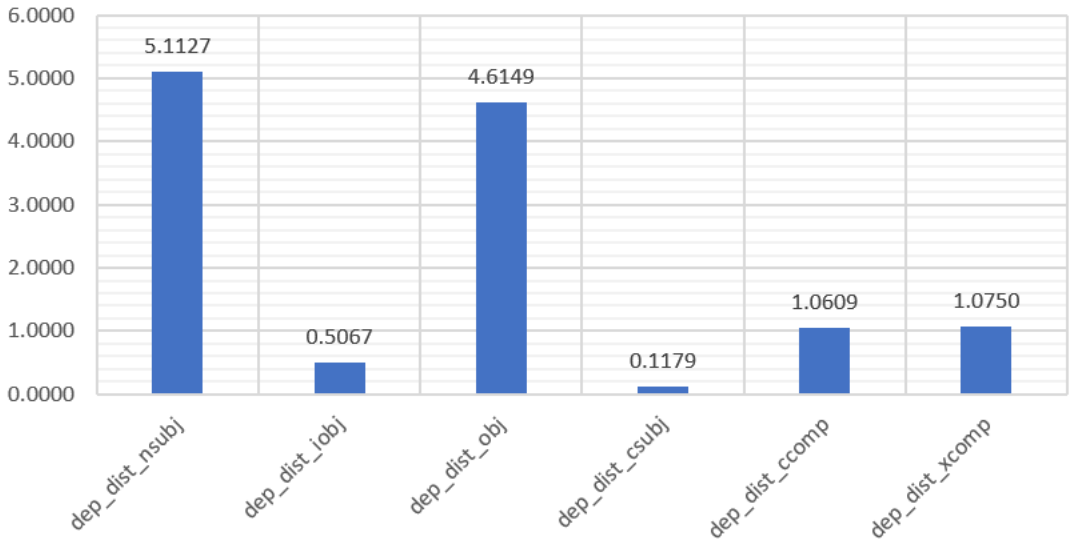


Grafico 17. Le distribuzioni medie delle caratteristiche relative agli argomenti nucleari di un predicato.

Il grafico 18 mostra le distribuzioni medie delle caratteristiche sintattiche che riguardano gli argomenti non nucleari di un predicato nelle frasi. La distribuzione media più alta del gruppo ha la funzione nominale obliqua (5.4121%) seguita dal modificatore avverbiale (4.2720%). Gli elementi dislocati (0.0155%) e gli elementi discorsivi (0.0564%) sono i più rari.

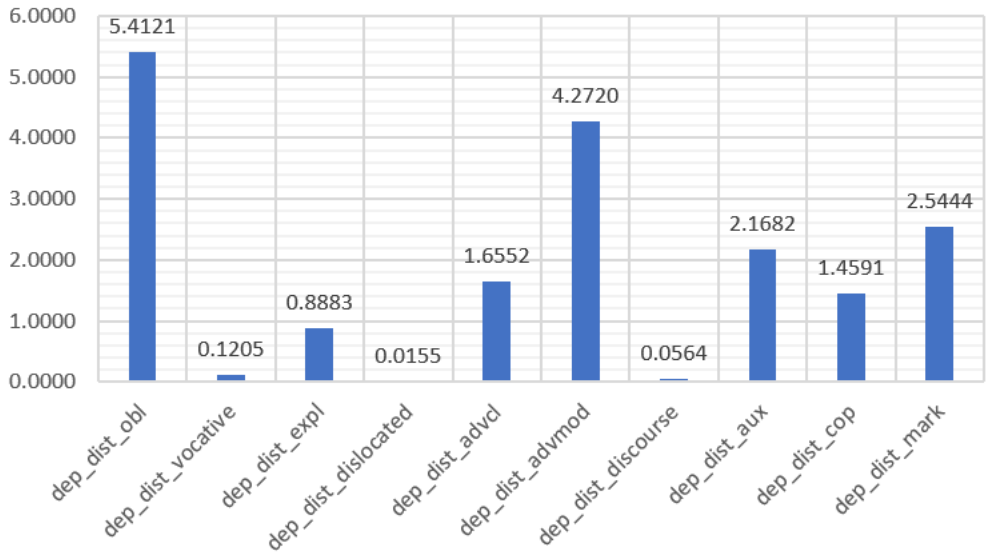


Grafico 18. Le distribuzioni medie delle caratteristiche relative agli argomenti non nucleari di un predicato.

Nel grafico 19 sono rappresentate le distribuzioni medie delle caratteristiche sintattiche che riguardano i dipendenti nominali nelle frasi. Le più frequenti di queste caratteristiche sono i determinanti (13.4517%) e i marcatori del caso (11.5092%). I meno frequenti invece sono i modificatori apposizionali (0.3593%), i numerali (0.4306%) e le proposizioni che modificano il nome (0.4329%).

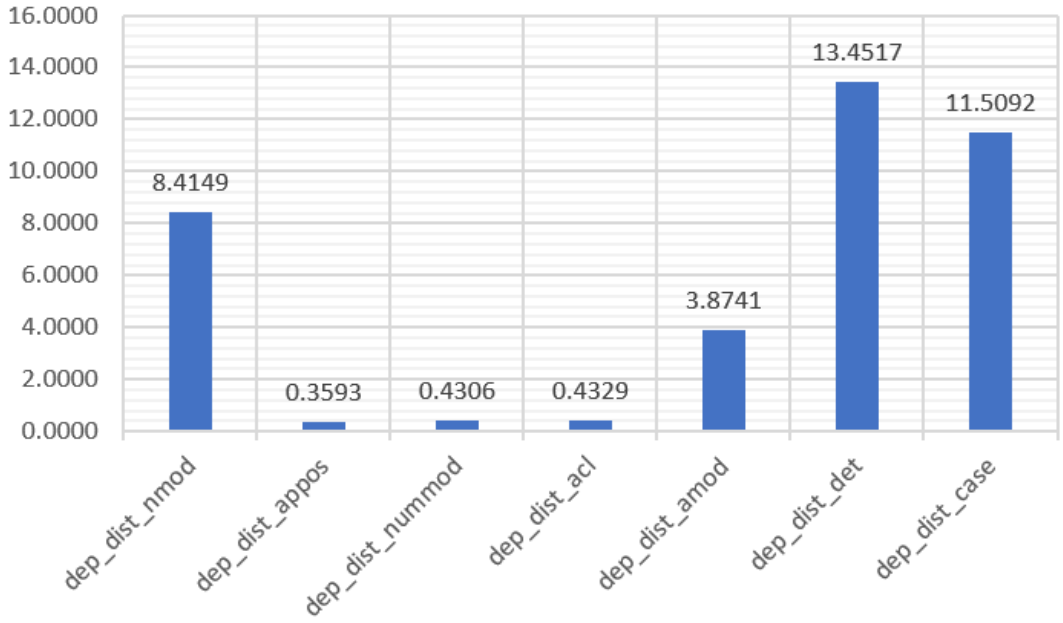


Grafico 19. Le distribuzioni medie delle caratteristiche relative ai dipendenti nominali.

Nel grafico 20 si osservano le distribuzioni medie di altre caratteristiche sintattiche che non ricadono nei gruppi precedenti. Dal grafico si evince che i segni di punteggiatura (12.8369%) sono, come atteso, i più diffusi dal momento che ogni frase presenta quanto meno un segno di punteggiatura che corrisponde al punto di fine frase. Le radici delle frasi, similmente, hanno un valore alto nella distribuzione media (6.4720%). Le relazioni di paratassi (0.4849%), i sintagmi composti (0.2985%), i sintagmi ”piatti”, ovvero senza struttura interna a dipendenze, (0.2744%) e le espressione fisse (0.2371%) sono i meno frequenti del gruppo.

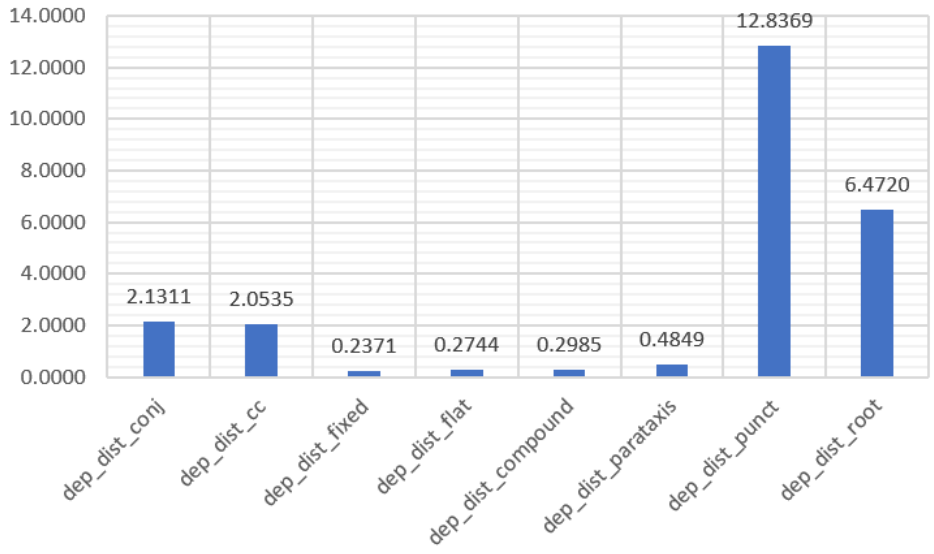


Grafico 20. Le distribuzioni medie delle caratteristiche relative alle relazioni coordinative.

L'ultimo gruppo di caratteristiche linguistiche sintattiche estratte da Profiling-UD contiene le caratteristiche che riguardano l'uso di subordinazioni nelle frasi:

- distribuzione delle proposizioni principali (*principal_proposition_dist*);
- distribuzione delle proposizioni subordinate (*subordinate_proposition_dist*);
- distribuzione delle proposizioni subordinate successive alla proposizione principale (*subordinate_post*);
- distribuzione delle proposizioni subordinate precedenti alla proposizione principale (*subordinate_pre*);
- lunghezza media delle catene di proposizioni subordinate (*avg_subordinate_chain_len*);
- distribuzione delle proposizioni subordinate per la profondita (*subordinate_dist_* *).

Il grafico 21 mostra le distribuzioni medie delle proposizioni subordinate per la profondità estratte con il Profiling-UD per le frasi del corpus. Si osserva che i valori delle distribuzioni diminuiscono con l'incremento della profondità quindi le distribuzioni più alte hanno le catene proposizionali con profondità 1 (40.0815%). Le distribuzioni più basse hanno le catene proposizionali subordinate con profondità 2 (10.9149%), tre

(3.0797%) e cinque (0.2717%). Nelle frasi analizzate non sono presenti le proposizioni subordinate con profondità 4.

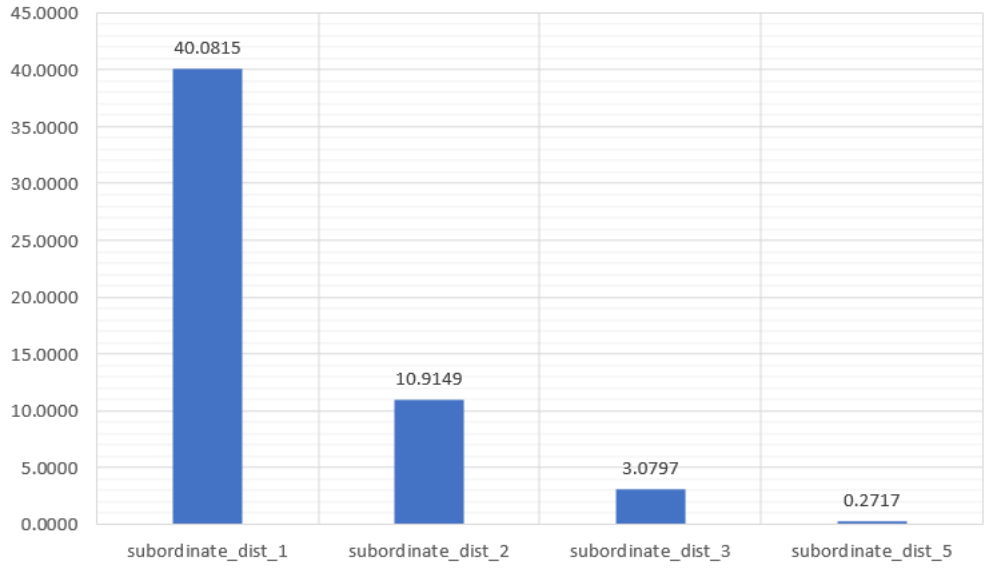


Grafico 21. Le distribuzioni medie delle proposizioni subordinate per la profondita.

Le proposizioni principali hanno una distribuzione media di 56.7171% e le proposizioni subordinate hanno invece una distribuzione media di 36.7611%. Le subordinate successive e precedenti alle proposizioni principali sono distribuite rispettivamente con i valori 45.6884% e 8.6594%.

4 Analisi dei giudizi di complessità linguistica

In questo capitolo verranno analizzati i giudizi raccolti mediante il questionario descritto nel capitolo precedente. Nella sezione 4.1 verranno riportati i giudizi medi espressi dai vari gruppi di annotatori, omogenei per livello di competenza della L2. Nella sezione 4.2 saranno introdotte le metriche per il calcolo della correlazione. La sezione 4.3 sarà invece dedicata ad analizzare quali caratteristiche linguistiche hanno avuto un impatto nel determinare il giudizio di complessità linguistica assegnato a ciascuna frase. Nello specifico, valuteremo se è possibile trovare correlazione tra le caratteristiche linguistiche descritte nel capitolo precedente e i giudizi sulla complessità linguistica delle frasi assegnati dagli annotatori. Per farlo, useremo le metriche per il calcolo della correlazione descritte nella sezione precedente. Grazie a questa analisi sarà possibile definire quali caratteristiche influenzano maggiormente la percezione della complessità linguistica di una frase da parte di madrelingua russi apprendenti l'italiano a diversi livelli di competenza della L2. Infine, nella sezione 4.5 i risultati ottenuti grazie ai questionari svolti dai madrelingua russi saranno confrontati con i giudizi di complessità linguistica assegnati alle stesse frasi da madrelingua italiani per lo studio descritto in Brunato et al. (2018). Questa ultima analisi ci permetterà di indagare se esistono differenze per quanto riguarda la percezione della complessità linguistica da parte di madrelingua russi e i parlanti nativi l'italiano.

4.1 Calcolo dei giudizi medi degli annotatori

Per trovare le caratteristiche rilevanti per la percezione della complessità linguistica vanno anzitutto calcolate le medie aritmetiche dei giudizi degli annotatori su ogni frase per ciascun gruppo di conoscenza della lingua italiana. La media aritmetica è un indice numerico comunemente utilizzato in statistica per descrivere un insieme di dati. Si ottiene sommando tutti i valori in questione e dividendo il risultato per il numero complessivo dei dati.

Per calcolare la media di giudizi di ciascun gruppo di competenza è stato scritto un programma in Python che ricorre al suo interno alla libreria *pandas* e la sua funzione *read_csv* per accedere al file csv che contiene la tabella con i dati di giudizi degli annotatori dove ogni colonna rappresenta una frase del corpus e ogni riga rappresenta un annotatore. Successivamente il programma accede ad ogni colonna della tabella e calcola la media aritmetica dei valori della colonna con la funzione *mean* della libreria *numpy*. Dopo il calcolo, il programma salva i risultati in un file separato. Il seguente codice contiene il programma che è stato chiamato su ogni file con i giudizi degli annotatori di ogni livello di competenza in italiano:

```
import pandas as pd
import numpy
import sys
df = pd.read_csv(sys.argv[1], header=None)
medie = []
for i in range (184):
    punt = df.iloc[:,i]
    punt = punt.to_numpy()
    punt = punt.tolist()
    m = numpy.mean(punt)
    medie.append(m)
data = pd.DataFrame(medie)
data.to_csv('med.csv',header = 'Media', index = False)
```

Nel Grafico 22 sono mostrati i giudizi medi assegnati alle frasi dagli annotatori. Le frasi hanno registrato un grado di facilità maggiore tra gli annotatori con un livello avanzato di competenza in italiano i quali hanno assegnato un giudizio medio di 2.1313 alle frasi. Gli annotatori appartenenti al gruppo dei principianti hanno considerato le suddette frasi più difficili rispetto agli altri gruppi (2.7799). Agli annotatori con il livello intermedio di conoscenza dell'italiano le frasi sono risultate leggermente più facili rispetto ai principianti (2.7166). Il giudizio medio assegnato da tutti gli annotatori senza divisione in gruppi è pari a 2.4909.

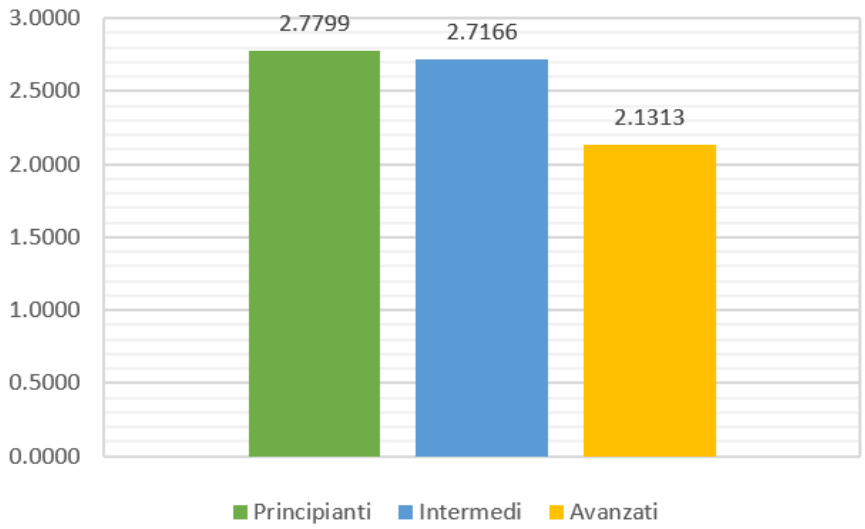


Grafico 22. I giudizi medi sulla complessità delle frasi di ogni gruppo di annotatori.

Dopo aver eseguito i calcoli sulla complessità media, è stato possibile determinare i differenti gradi di difficoltà delle frasi per ogni gruppo degli annotatori, trarre le prime osservazioni e trovare delle somiglianze tra i risultati.

Alcune frasi che sono risultate più facili per i principianti sono:

"Anche allo stadio si canta #Grillo uno di noi"

"Dove si trova l'aeroporto di Heathrow?"

Un esempio di una frase facile per gli intermedi:

"Quale città italiana è la sede della Cattedrale di Santa Maria del Fiore o Duomo?"

Alcune frasi che sono risultate più facili per gli avanzati sono:

"Metto il burro in un tegamino, lo faccio sciogliere:"

"@user l'Italia è la culla dell'ipocrisia"

"Che cosa viene esposto nel Vitra Design Museum?"

Le frasi più facili in assoluto per tutti i gruppi:

"Quale animale è simbolo della Namibia?"

"In quale anno è stata creata la Banca Mondiale?"

"Chi è il ministro del commercio in Cina?"

Negli esempi sopra riportano casi di frasi brevi che non fanno uso della subordinazione. La maggior parte di esse presentano la forma interrogativa con soggetti e predicati espliciti e contengono vocaboli di uso comune nella lingua italiana.

Qui di seguito sono riportati esempi di frasi che sono risultate più difficili per i principianti:

> "Vagabondi incontrati per le pianure mi dicevano che i confini non erano lontani."
>
> "attenzione xché se #Grillo è furbo e attenua un po' giustizialismo x cavalcare questione fiscale sono dolori per tutti #bluffitalia"
>
> "Consigliere comunale M5S sorpreso a rubare negli armadietti della palestra. Li apriva come una scatoletta di tonno. [@user]"

Gli intermedi hanno considerato la seguente frase la più difficile del corpus:

> "Non sappiamo cosa stia succedendo e quindi mi chiedo perché si debba permettere ai produttori di armi dell'Ue di trarne profitto a scapito di persone innocenti."

Per gli avanzati la seguente è stata la frase più difficile:

> "Un manifestante #NoTav chiede a Cicchitto cosa farebbe il #PDL se il governo #Monti toccasse le TV di Berlusconi ... KABOOM ! #PiazzaPulita"

La frase più difficile in assoluto per tutti i gruppi risulta essere:

> "non che io tema nulla, ma non vorrei che qualche lettore scarabocchiatore, consultando l'elenco telefonico, trovasse il mio indirizzo e per divertimento venisse a imbrattare di nuovo la nostra casa."

Tutte queste frasi presentano diversi fenomeni che ne aumentano il grado di difficoltà. Ad esempio, alcune frasi contengono verbi al modo congiuntivo e particelle pronominali, solitamente difficili da apprendere per i non madrelingua. Inoltre, sono presenti hashtag e onomatopeiche, tipiche del linguaggio dei social media, che ostacolano la lettura. Le frasi sono ulteriormente complicate dalla presenza di vocaboli di uso poco comune (ad esempio, scarabocchiatore).

Per definire le caratteristiche che determinano la percezione della complessità da parte dei madrelingua russi apprendenti l'italiano deve essere eseguita un'analisi statistica più approfondita che sarà discussa nel prossimo paragrafo.

4.2 Metriche per il calcolo della correlazione

Per individuare le caratteristiche linguistiche che hanno influenzato maggiormente l'assegnazione dei giudizi sulla complessità linguistica delle frasi deve essere stabilita una relazione tra la complessità e le caratteristiche linguistiche. Per fare questo è stata

calcolata la correlazione[6] tra la media delle valutazioni da parte degli annotatori e le caratteristiche linguistiche estratte per ogni frase del corpus descritte nel 3 capitolo. Nel presente studio è stato utilizzato il coefficiente di correlazione di Spearman per trovare il legame tra la complessità e le caratteristiche linguistiche. Il coefficiente di correlazione di Spearman, chiamato anche *rho* (ρ) di Spearman, è una misura statistica non parametrica di correlazione tra due insiemi di dati R e S e viene calcolato con il seguente formula:

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

dove *N* è il numero totale di osservazioni e *Di* si ottiene come:

$$D_i = r_i - s_i$$

dove *ri* e *si* rappresentano gli elementi che fanno parte degli insiemi R e S.

Il coefficiente di Spearman assume i valori compresi tra -1 e 1 che vengono interpretati come segue:

- quando ρ si avvicina a 1 significa che esiste una forte correlazione positiva tra i due fenomeni;
- quando ρ assume un valore vicino a 0 significa che non c'è nessuna dipendenza tra i due fenomeni;
- quando ρ si avvicina a -1 significa che i due fenomeni sono correlati negativamente;

L'indice di correlazione di Spearman è definito come non parametrico perché stabilisce una perfetta correlazione se due insiemi R e S sono correlati da qualsiasi funzione monotona e perché non richiede la conoscenza della distribuzione di probabilità congiunta di R e S.

Per verificare che i risultati ottenuti con l'indice di correlazione di Spearman non siano casuali viene calcolato il *p-value*. Il p-value misura quanto è probabile che la correlazione osservata sia dovuta al caso e assume i valori da 0 a 1. Quando il p-value

6 La correlazione è un rapporto di reciproca dipendenza tra i due fenomeni. La correlazione dipende dalla tendenza di una variabile a cambiare in funzione di un'altra.

tende a 1 significa che non esiste una correlazione tra i valori osservati oltre quella casuale. Un p-value vicino a 0 indica una forte correlazione tra due insiemi di dati. Generalmente si considerano rilevanti i p-value minori di 0.05 i quali evidenziano la probabilità che la correlazione identificata sia dovuta al caso è minore del 5%.

Per calcolare il coefficiente di correlazione di Spearman e il p-value per ogni caratteristica linguistica estratta per le frasi oggetto dello studio è stato scritto un programma in Python che utilizza la libreria *SciPy*, in particolare il suo pacchetto *stats*, che mette a disposizione gli strumenti di calcolo matematico. Il modulo *spearmanr* viene invocato su due vettori di dati e restituisce l'indice di correlazione e il p-value. In prima battuta è necessario accedere ad un file contenente i valori delle caratteristiche linguistiche estratte tramite il monitoraggio linguistico descritte nel capitolo 3. Successivamente, dopo aver ricevuto un file contenente i giudizi medi assegnati dagli annotatori di uno dei gruppi di competenza in italiano sulla complessità delle frasi, il programma esegue un ciclo in cui per ogni caratteristica viene creato un vettore con i valori che calcola la correlazione con il vettore di giudizi di complessità. Di seguito è riportato un frammento di codice che dimostra il calcolo del coefficiente di correlazione e del p-value:

```
import pandas as pd
from scipy import stats
from scipy.stats import spearmanr
cor = []
pv = []
for i in range(125):
    colonna = df.iloc[:, i+2]
    colonna = colonna.to_numpy()
    colonna = colonna.tolist()
    correlation, pval = spearmanr(punt, colonna)
    cor.append(correlation)
    pv.append(pval)
```

Per confrontare i gruppi di annotatori è stato utilizzato il coefficiente di correlazione di Pearson che valuta il rapporto lineare tra due insiemi di dati. Il rapporto è lineare quando il cambiamento di una variabile è associato al cambiamento proporzionale di un'altra variabile. Il coefficiente di Pearson si calcola mediante la seguente formula:

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

dove

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t, \overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$$

.

Il coefficiente di Pearson è stato calcolato in Python ricorrendo alla funzione *pearsonr* della libreria *scipy.stats* applicato ai dati delle coppie dei gruppi di competenza in italiano.

4.3 Analisi della correlazione fra giudizi e caratteristiche linguistiche

In questa sezione saranno riportati i risultati ottenuti dopo l'applicazione del programma di calcolo del coefficiente di Spearman sui giudizi di complessità linguistica assegnati dagli annotatori.

Dopo aver eseguito il codice sul file con i giudizi sulla complessità di tutti gli annotatori è stata trovata la caratteristica linguistica più significativa che ha influenzato la valutazione. Questa caratteristica è legata al numero totale dei tokens in una frase e ha il coefficiente di correlazione pari a 0.775954569 e il p-value pari a 2.88E-38. I valori ottenuti indicano che all'aumentare il numero dei tokens nelle frasi aumenta la complessità delle frasi. Il valore basso di p-value dimostra che la correlazione calcolata non è dovuta al caso. Il grafico 23 mostra la distribuzione del numero dei token nelle frasi rispetto ai giudizi sulla complessità di queste frasi dagli annotatori. Si osserva che il punteggio della complessità aumenta al crescere del numero dei tokens.

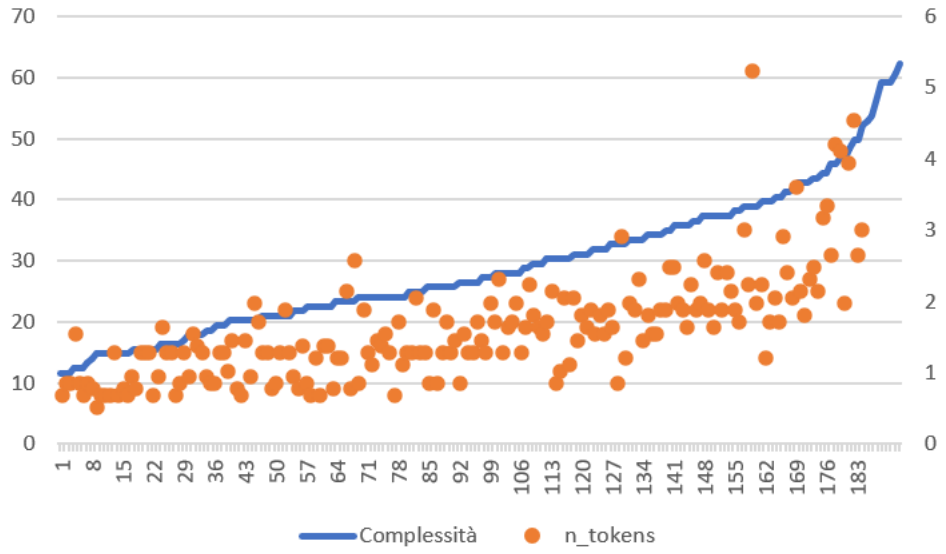


Grafico 23. La distribuzione del numero dei token nelle frasi rispetto ai giudizi sulla complessità.

La caratteristica che è risultata il meno rilevante per la percezione della complessità è la distribuzione di modificatori nominali. Il coefficiente di correlazione calcolato per la distribuzione di modificatori nominali è estremamente basso ed è pari a 0.002522646. Il p-value di questa caratteristica è pari a 0.972888561 che indica che la correlazione incontrata è solo casuale. Il Grafico 24 dimostra la distribuzione dei modificatori nominali nelle frasi rispetto ai giudizi sulla complessità assegnati dagli annotatori. Dal grafico si evince che non c'è alcuna correlazione tra i due fenomeni.

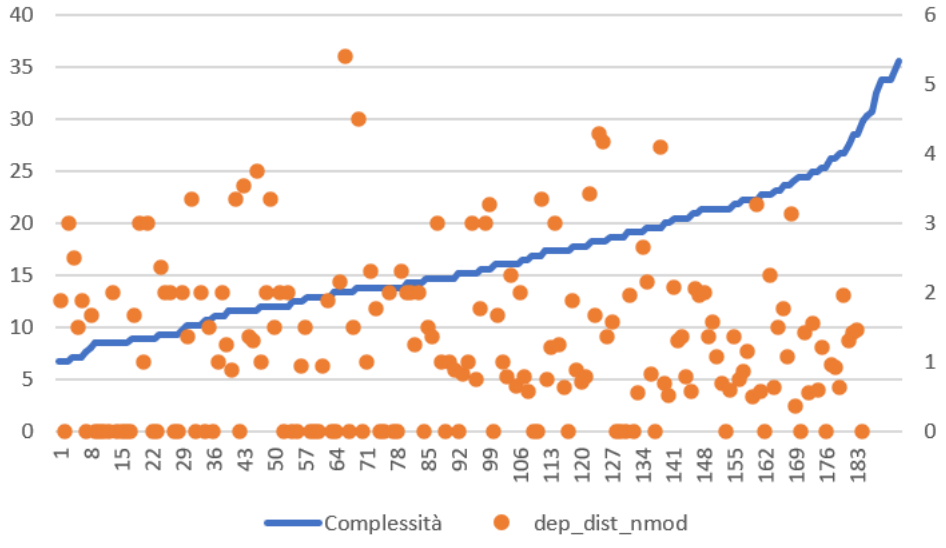


Grafico 24. La distribuzione dei modificatori nominali rispetto ai giudizi sulla complessità.

Dopo il calcolo del coefficiente di correlazione di Spearman è stato possibile ottenere un ordinamento delle caratteristiche rilevanti per ogni gruppo degli annotatori. Per creare gli ordinamenti, le caratteristiche sono state ordinate per il valore di correlazione decrescente. Nella tabella 6 sono riportate le caratteristiche significative che correlano in maniera significativa (p-value >= 0.05) con i giudizi di complessità espressi dagli annotatori a prescindere dal gruppo di appartenenza. Con il colore arancione sono evidenziate le caratteristiche che hanno una correlazione positiva. Il colore blu indica, invece, una correlazione di tipo negativo. Gli ordinamenti delle caratteristiche linguistiche significative di ogni gruppo di annotatori divise sulla base della loro competenza nella lingua italiana saranno riportati nell'appendice del presente lavoro.

Caratteristica	Correlazione	P-value
n_tokens	0.775954569	2.88E-38
subordinate_post	0.666124449	5.84E-25
dep_dist_mark	0.585485808	2.57E-18
dep_dist_advcl	0.582802507	3.98E-18
upos_dist_ADV	0.582802507	3.98E-18
obj_pre	0.533529332	6.30E-15
subordinate_dist_1	0.518301847	4.86E-14
dep_dist_acl:relcl	0.513004166	9.66E-14
verbs_num_pers_dist_Sing+3	0.438378069	4.85E-10
avg_max_depth	0.435896271	6.23E-10
subordinate_dist_3	0.374606173	1.62E-07
verbs_tense_dist_Pres	0.373238064	1.81E-07
dep_dist_obj	0.366809198	3.03E-07
verbs_mood_dist_Ind	0.364153396	3.74E-07
dep_dist_root	0.361351947	4.67E-07
dep_dist_advmod	0.356282875	6.91E-07
avg_verb_edges	0.328347387	5.36E-06
dep_dist_iobj	0.306991351	2.25E-05
upos_dist_SCONJ	0.285197801	8.71E-05

subordinate_dist_2	0.278527471	1.29E-04
avg_links_len	0.270612246	2.03E-04
verb_edges_dist_5	0.266860476	0.0002504
max_links_len	0.263996676	2.93E-04
subordinate_proposition_dist	0.263683352	2.98E-04
principal_proposition_dist	0.26324713	3.06E-04
verbs_num_pers_dist_Plur+3	0.262899024	0.0003116
verbal_head_per_sent	0.260706311	3.51E-04
obj_post	0.25785085	4.10E-04
avg_subordinate_chain_len	0.252763655	5.37E-04
dep_dist_parataxis	0.247893634	0.0006918
verbs_form_dist_Fin	0.244273927	0.0008327
verb_edges_dist_4	0.237745908	0.0011553
verbs_mood_dist_Cnd	0.233864333	0.0013978
n_prepositional_chains	0.229189704	0.0017513
dep_dist_xcomp	0.228139762	0.0018411
subordinate_pre	0.219085008	0.0028086
aux_num_pers_dist_Plur+3	0.216082956	0.0032191
verbs_mood_dist_Sub	0.209947611	0.0042309
verbs_tense_dist_Imp	0.20817315	0.0045727
subj_pre	0.206017283	0.0050212
dep_dist_conj	0.20566194	0.0050988
char_per_tok	0.204987057	0.0052491
aux_mood_dist_Sub	-0.466595218	2.46E-11
verbs_num_pers_dist_Sing+1	-0.775954569	2.88E-38

Tabella 6. Le caratteristiche linguistiche significative per tutti gli annotatori.

Come già discusso precedentemente, la caratteristica più significativa per la percezione della complessità linguistica da parte di madrelingua russi apprendenti l'italiano è il numero totale dei token nella frase. Segue la caratteristica legata alle subordinate postverbali. La posizione alta di questo tipo di subordinata può essere spiegata dal fatto che questa è la collocazione più tipica delle subordinate nella lingua italiana. Un'altra caratteristica importante è legata alla distribuzione dei marcatori che introducono una

frase subordinata. Il fatto che questa caratteristica è connessa alla precedente spiega la loro vicinanza nell'ordinamento. La sesta caratteristica che rende una frase difficile è la distribuzione degli oggetti nella posizione preverbale. Questo tipo di collocazione dell'oggetto è il meno usato nella lingua italiana e il meno distribuito nelle frasi oggetto di studio e quindi complica una frase. Al 7° posto si trova la distribuzione delle subordinate con profondità 1. La sua posizione più alta rispetto alla distribuzione delle subordinate con profondità 3 e 2 che occupano le posizioni 11° e 20° nell'ordinamento, per le quali si aspetta che influiscano maggiormente sulla complessità delle frasi, è spiegata dal fatto che le ultime due sono meno rappresentate nel corpus com'è stato dimostrato nel Grafico 21. Si può supporre che eseguendo lo studio su un corpus più ampio, le caratteristiche legate alla distribuzione delle subordinate occuperanno le posizioni seguendo un grado di profondità decrescente. La caratteristica legata alla distribuzione dei verbi in terza persona singolare occupa il 9° posto: questo può essere spiegato a causa del loro alto grado di diffusione nelle frasi rispetto alle altre forme verbali come dimostrato nel Grafico 12. Il 12° posto è occupato dalla distribuzione dei verbi al tempo presente: come si evince dal Grafico 9 è il tempo più diffuso nelle frasi del corpus. Al 14° posto si trova la distribuzione dei verbi coniugati al modo indicativo: è infatti il periodo verbale maggiormente diffuso nel corpus oggetto dello studio come si evince dal Grafico 10. La caratteristica legata alla distribuzione delle radici nelle frasi occupa il 15° posto e dipende fortemente dal numero di tokens presenti all'interno delle frasi. Al 19° posto è posizionata la distribuzione delle congiunzioni subordinate che, nonostante la bassa rappresentazione nelle frasi del corpus come si osserva nel Grafico 2, ha dimostrato una forte correlazione con la complessità. Alla posizione 22° è collocata la distribuzione dei verbi con arità 5: occupa una posizione rilevante nonostante la rappresentanza ridotta nelle frasi com'è stato dimostrato nel Grafico 13. Al 28° posto si trova la distribuzione degli oggetti postverbali che, com'è stato dimostrato nel Grafico 16, sono i più diffusi nella lingua italiana. Al posto 31° è collocata la distribuzione dei verbi nella forma finita che è la più diffusa nelle frasi del corpus come si può osservare nel Grafico 11. Al posto 32° è collocata la distribuzione dei verbi con arità 4 che ha una posizione alta nonostante la scarsa rappresentanza nelle frasi come si evince dal Grafico 13. Il 33° posto è occupato dalla distribuzione dei verbi nel modo condizionale che ha la

più bassa diffusione nelle frasi dispetto agli altri com'è stato dimostrato nel Grafico 10 quindi è una caratteristica molto rilevante per lo studio. Al 36° posto si trova la distribuzione delle subordinate pre-verbali che è significativa perché è meno diffusa nella lingua italiana rispetto alla posizione post-verbale. Il 38° posto è occupato dalla distribuzione dei verbi in modo congiuntivo che è poco diffuso nelle frasi del corpus com'è stato dimostrato nel Grafico 10. La distribuzione dei verbi nel tempo imperfetto ha dimostrato una forte correlazione e ha occupato la 39° posizione nell'ordinamento nonostante la scarsa distribuzione nel corpus come si evince dal Grafico 9 quindi è particolarmente significativa. Al 40° posto è posizionata la distribuzione dei soggetti preverbali che è un tipo di collocazione più diffuso nella lingua italiana.

Le correlazioni negative significative si osservano per due caratteristiche: la distribuzione dei verbi in prima persona singolare e la distribuzione dei verbi ausiliari nel modo congiuntivo. Il valore di correlazione negativa indica che al crescere del valore di queste caratteristiche diminuisce la percezione della complessità

4.4 Influenza del livello di competenza della L2

Sulla base degli ordinamenti che sono riportati interamente in Appendice, è stato possibile analizzare quanto siano correlati i gruppi degli annotatori tra di loro. Per confrontare i gruppi di annotatori è stato utilizzato il coefficiente di correlazione di Pearson descritto nella sezione 4.2. La tabella seguente riporta i risultati ottenuti dopo esecuzione del programma di calcolo.

Coppia	Pearson correlazione	Pearson P-value
Principianti-intermedi	0.968794954	2.87E-39
Principianti-avanzati	0.965871775	1.01E-35
Intermedi-avanzati	0.992747986	6.48E-57

Tabella 7. Le correlazioni tra le coppie dei gruppi degli annotatori.

Dalla tabella 7 si evince che tutte le coppie dei gruppi sono fortemente correlati. Tra la coppia dei gruppi di annotatori intermedi e avanzati esiste una maggiore correlazione. Al contrario, i principianti e gli avanzati sono correlati di meno tra loro.

Per costruire gli ordinamenti con le differenze tra le posizioni delle coppie dei gruppi è stato scritto un programma in Python. Il frammento seguente mostra la parte del codice che confronta le caratteristiche di due ordinamenti, calcola lo scarto tra le posizioni e salva le informazioni in un nuovo ordinamento:

```
ranking = [[0 for c in range(cols)] for r in range(rows)]
for i in range(rows):
    for k in range(rows):
        if gruppo1[i][1] == gruppo2[k][1]:
            ranking[i][0] = i-k
            ranking[i][1] = i+1
            ranking[i][2] = k+1
            ranking[i][3] = gruppo1[i][1]
```

Anzitutto saranno analizzate le differenze nella percezione della complessità della coppia dei gruppi degli annotatori principianti e intermedi. Per questa analisi le caratteristiche sono state ordinate per il p-value crescente. Nella Tabella 8 sono riportate le differenze tra le posizioni delle caratteristiche che hanno correlazione significativa coi giudizi per entrambi gruppi. Con il colore blu sono evidenziate le caratteristiche più rilevanti per il gruppo degli intermedi rispetto al gruppo dei principianti. Con il colore verde sono evidenziate le caratteristiche più significative per il gruppo di principianti. Con il colore grigio sono marcate le caratteristiche che hanno la stessa rilevanza per entrambi i gruppi. Nella prima colonna sono riportate le differenze tra le posizioni relative alle caratteristiche del gruppo dei principianti rispetto al gruppo degli intermedi.

Differenza pr - int	**Poizione principianti**	**Posizione intermedi**	**Caratteristica**
20	56	36	verb_edges_dist_4
16	54	38	n_prepositional_chains
15	49	34	dep_dist_xcomp
13	26	13	obj_post
13	39	26	dep_dist_parataxis
12	36	24	aux_mood_dist_Sub
12	43	31	verbs_mood_dist_Cnd
10	32	22	verbs_num_pers_dist_Sing+1

9	34	25	verbs_num_pers_dist_Sing+3
9	44	35	subordinate_dist_3
9	46	37	aux_num_pers_dist_Plur+3
9	50	41	dep_dist_conj
8	29	21	avg_verb_edges
8	31	23	verb_edges_dist_5
5	19	14	subordinate_post
5	20	15	dep_dist_acl:relcl
5	33	28	verbs_form_dist_Fin
5	38	33	verbs_num_pers_dist_Plur+3
5	45	40	verbs_mood_dist_Sub
4	11	7	subordinate_proposition_dist
4	16	12	subordinate_dist_2
4	23	19	dep_dist_obj
3	7	4	verbal_head_per_sent
3	9	6	avg_subordinate_chain_len
3	13	10	upos_dist_SCONJ
3	57	54	prep_dist_1
1	3	2	dep_dist_root
1	4	3	avg_max_depth
1	6	5	max_links_len
1	10	9	principal_proposition_dist
1	12	11	dep_dist_mark
0	1	1	n_tokens
0	8	8	avg_links_len
0	18	18	dep_dist_advmod
-1	15	16	upos_dist_ADV
-1	28	29	subordinate_dist_1
-3	14	17	dep_dist_advcl
-3	17	20	verbs_mood_dist_Ind
-4	40	44	dep_dist_case
-4	47	51	dep_dist_expl:impers
-6	41	47	upos_dist_CCONJ

-8	22	30	dep_dist_iobj
-11	21	32	obj_pre
-12	30	42	subordinate_pre
-14	25	39	verbs_tense_dist_Pres
-14	42	56	dep_dist_ccomp
-15	35	50	dep_dist_cc
-18	27	45	verbs_tense_dist_Imp
-25	24	49	char_per_tok

Tabella 8. Differenza tra gli ordinamenti delle caratteristiche linguistiche significative per i gruppi degli annotatori principianti e intermedi.

Dalla Tabella 8 si evince che gli annotatori intermedi sono molto più influenzati dalla distribuzione dei verbi con arità 4, dal numero delle catene preposizionali, dalla distribuzione dei complementi proposizionali aperti, dalla posizione degli oggetti post-verbali, dalle relazioni di paratassi e dalle distribuzioni dei verbi ausiliari al modo congiuntivo e dei verbi al modo condizionale all'interno delle frasi rispetto agli annotatori intermedi. Il gruppo degli annotatori principianti ha dimostrato, invece, la più alta sensibilità alle caratteristiche linguistiche legate al numero medio di caratteri per un token, alla distribuzione dei verbi in tempo imperfetto, alla distribuzione delle congiunzioni coordinative, dei complementi proposizionali, dei verbi nel tempo presente, delle subordinate preverbali e oggetti preverbali rispetto al gruppo dei principianti.

La Tabella 9 mostra la differenza nella percezione della complessità tra gli annotatori intermedi e avanzati. Con il colore arancione sono evidenziate le caratteristiche che hanno influenzato maggiormente il gruppo di annotatori avanzati rispetto al gruppo di annotatori intermedi. Il colore blu evidenzia le caratteristiche più rilevanti per il gruppo degli intermedi rispetto al gruppo degli avanzati. Con il colore grigio sono identificate le caratteristiche che occupano la stessa posizione negli ordinamenti di entrambi gruppi. Nella prima colonna si trovano le differenze tra le posizioni delle caratteristiche del gruppo degli annotatori intermedi rispetto al gruppo degli annotatori avanzati.

Differenza int - av	Posizione intermedi	Posizione avanzati	Caratteristica
18	39	21	verbs_tense_dist_Pres
17	54	37	prep_dist_1
14	56	42	dep_dist_ccomp
9	36	27	verb_edges_dist_4
6	35	29	subordinate_dist_3
5	38	33	n_prepositional_chains
5	57	52	dep_dist_compound
3	50	47	dep_dist_cc
2	28	26	verbs_form_dist_Fin
1	41	40	dep_dist_conj
0	1	1	n_tokens
0	15	15	dep_dist_acl:relcl
0	17	17	dep_dist_advcl
0	20	20	verbs_mood_dist_Ind
0	45	45	verbs_tense_dist_Imp
0	49	49	char_per_tok
-1	2	3	dep_dist_root
-1	3	4	avg_max_depth
-1	5	6	max_links_len
-1	11	12	dep_dist_mark
-1	30	31	dep_dist_iobj
-1	33	34	verbs_num_pers_dist_Plur+3
-2	6	8	avg_subordinate_chain_len
-2	7	9	subordinate_proposition_dist
-2	8	10	avg_links_len
-2	9	11	principal_proposition_dist
-2	12	14	subordinate_dist_2
-2	14	16	subordinate_post
-2	16	18	upos_dist_ADV
-2	21	23	avg_verb_edges
-2	23	25	verb_edges_dist_5

-3	4	7	verbal_head_per_sent
-3	10	13	upos_dist_SCONJ
-3	40	43	verbs_mood_dist_Sub
-3	47	50	upos_dist_CCONJ
-3	52	55	upos_dist_PUNCT
-3	53	56	dep_dist_punct
-4	18	22	dep_dist_advmod
-4	24	28	aux_mood_dist_Sub
-4	31	35	verbs_mood_dist_Cnd
-4	42	46	subordinate_pre
-5	19	24	dep_dist_obj
-5	34	39	dep_dist_xcomp
-6	13	19	obj_post
-6	32	38	obj_pre
-7	25	32	verbs_num_pers_dist_Sing+3
-7	29	36	subordinate_dist_1
-8	22	30	verbs_num_pers_dist_Sing+1
-8	43	51	dep_dist_acl
-11	37	48	aux_num_pers_dist_Plur+3
-15	26	41	dep_dist_parataxis
-26	27	53	subj_pre

Tabella 9. Differenza tra gli ordinamenti delle caratteristiche linguistiche significative per i gruppi degli annotatori intermedi e avanzati.

Nella Tabella 9 si osserva che le caratteristiche che hanno influenzato particolarmente gli annotatori avanzati rispetto agli annotatori intermedi riguardano la distribuzione dei verbi al tempo presente, la distribuzione delle proposizioni complementari e la distribuzione dei verbi con arità 4. Il gruppo degli annotatori intermedi si è dimostrato più sensibile alla distribuzione dei soggetti preverbali e delle relazioni di paratassi.

L'analisi delle differenze nella percezione della complessità tra i gruppi degli annotatori principianti e avanzati è dimostrata nella Tabella 10. Il colore arancione evidenzia le

caratteristiche che hanno avuto il maggior impatto sul gruppo degli annotatori avanzati. Con il colore verde sono evidenziate le caratteristiche più significative per il gruppo dei principianti rispetto al gruppo degli avanzati. Con il colore grigio sono evidenziate le caratteristiche che occupano la stessa posizione negli ordinamenti di entrambi gruppi. Nella prima colonna si trovano le differenze tra le posizioni di caratteristiche del gruppo degli annotatori principianti rispetto al gruppo degli annotatori avanzati.

Differenza pr - av	**Posizione principianti**	**Posizione avanzati**	**Caratteristica**
29	56	27	verb_edges_dist_4
21	54	33	n_prepositional_chains
20	57	37	prep_dist_1
15	44	29	subordinate_dist_3
10	49	39	dep_dist_xcomp
10	50	40	dep_dist_conj
8	36	28	aux_mood_dist_Sub
8	43	35	verbs_mood_dist_Cnd
7	26	19	obj_post
7	33	26	verbs_form_dist_Fin
6	29	23	avg_verb_edges
6	31	25	verb_edges_dist_5
5	20	15	dep_dist_acl:relcl
4	25	21	verbs_tense_dist_Pres
4	38	34	verbs_num_pers_dist_Plur+3
3	19	16	subordinate_post
2	11	9	subordinate_proposition_dist
2	16	14	subordinate_dist_2
2	32	30	verbs_num_pers_dist_Sing+1
2	34	32	verbs_num_pers_dist_Sing+3
2	45	43	verbs_mood_dist_Sub
1	9	8	avg_subordinate_chain_len
1	55	54	verb_edges_dist_1

0	1	1	n_tokens
0	2	2	tokens_per_sent
0	3	3	dep_dist_root
0	4	4	avg_max_depth
0	5	5	avg_max_links_len
0	6	6	max_links_len
0	7	7	verbal_head_per_sent
0	12	12	dep_dist_mark
0	13	13	upos_dist_SCONJ
0	42	42	dep_dist_ccomp
-1	10	11	principal_proposition_dist
-1	23	24	dep_dist_obj
-2	8	10	avg_links_len
-2	39	41	dep_dist_parataxis
-2	46	48	aux_num_pers_dist_Plur+3
-3	14	17	dep_dist_advcl
-3	15	18	upos_dist_ADV
-3	17	20	verbs_mood_dist_Ind
-4	18	22	dep_dist_advmod
-7	37	44	upos_dist_AUX
-8	28	36	subordinate_dist_1
-9	22	31	dep_dist_iobj
-9	41	50	upos_dist_CCONJ
-12	35	47	dep_dist_cc
-16	30	46	subordinate_pre
-17	21	38	obj_pre
-18	27	45	verbs_tense_dist_Imp
-25	24	49	char_per_tok

Tabella 10. Differenza tra gli ordinamenti delle caratteristiche linguistiche significative per i gruppi degli annotatori principianti e avanzati.

La Tabella 10 dimostra che le caratteristiche linguistiche più rilevanti per la percezione della complessità del gruppo degli annotatori avanzati rispetto al gruppo degli annotatori

principianti sono la distribuzione dei verbi con arità 4, il numero medio delle catene preposizionali, la distribuzione delle catene preposizionali con profondità 1, la distribuzione delle catene subordinate con la profondità 3, la distribuzione dei complementi proposizionali aperti e la distribuzione delle congiunzioni. Il gruppo di principianti, invece, è più propenso ad attribuire il giudizio più alto in termini di complessità alle frasi che registrano valori alti circa le caratteristiche legate al numero medio di caratteri per token, alla distribuzione dei verbi nel tempo imperfetto, agli oggetti nella posizione preverbale, alla posizione preverbale di subordinate e alla distribuzione delle congiunzioni coordinative.

4.5 Confronto con i risultati di madrelingua italiani

In questo paragrafo saranno confrontati i risultati ottenuti nel presente studio con i risultati dello studio sulla percezione della complessità linguistica condotto coinvolgendo madrelingua italiani. I dettagli di questo studio sono riportati in due lavori precedenti a questa tesi, ovvero Brunato et al. (2018); e Iavarone (2017). Il confronto ha lo scopo di individuare quali caratteristiche influenzano maggiormente la percezione di complessità da parte di madrelingua russi apprendenti l'italiano come L2 rispetto a madrelingua italiani.

La Tabella 11 mostra le caratteristiche linguistiche che influenzano significativamente (p-value < 0.05) la percezione di complessità nella lingua italiana per i soggetti madrelingua russi e madrelingua italiani. L'ordinamento delle caratteristiche significative è stato reperito fra i dati riportati nello studio sulla complessità percepita (Iavarone, 2017). Il colore arancione indica le caratteristiche correlate positivamente con la complessità e il colore blu evidenzia la correlazione negativa. Le liste sono state ordinate seguendo una scala di correlazione decrescente.

Madrelingua russi	Madrelingua italiani
n_tokens	avg_max_depth
subordinate_post	n_tokens
dep_dist_mark	n_prepositional_chains
dep_dist_advcl	dep_freq_nmod

upos_dist_ADV	dep_freq_case
obj_pre	n_subordinate_proposition
subordinate_dist_1	total_subordinate_chain_len
dep_dist_acl:relcl	verbal_head
verbs_num_pers_dist_Sing+3	verbal_head_per_sent
avg_max_depth	n_subordinate_chain
subordinate_dist_3	dep_freq_det
verbs_tense_dist_Pres	avg_subordinate_chain_len
dep_dist_obj	subordinate_proposition_dist
verbs_mood_dist_Ind	dep_freq_amod
dep_dist_root	avg_links_len
dep_dist_advmod	subordinate_pre
avg_verb_edges	subordinate_post
avg_max_links_len	dep_freq_punct
dep_dist_iobj	prep_freq_1
upos_dist_SCONJ	cpos_dist_NUM
subordinate_dist_2	cpos_dist_NOUN
avg_links_len	dep_dist_nsubj
verb_edges_dist_5	dep_freq_vocative
max_links_len	dep_dist_vocative
subordinate_proposition_dist	in_dict_types
principal_proposition_dist	cpos_dist_PUNCT
verbs_num_pers_dist_Plur+3	dep_dist_punct
verbal_head_per_sent	cpos_dist_DET
obj_post	dep_dist_det
avg_subordinate_chain_len	in_AD_types
dep_dist_parataxis	max_links_len
verbs_form_dist_Fin	principal_proposition_dist
aux_mood_dist_Sub	ttr_lemma
verbs_num_pers_dist_Sing+1	ttr_form
	dep_dist_root

Tabella 11. Il confronto delle caratteristiche significative per i madrelingua italiani e le caratteristiche significative per i madrelingua russi.

Dalla tabella 11 si evince che tra i madrelingua russi e i madrelingua italiani esistono delle differenze per quanto riguarda le caratteristiche linguistiche con un impatto maggiore sulla percezione della complessità. Per dimostrare la disuguaglianza degli ordinamenti riportati nella Tabella 11 è stato utilizzato il programma descritto nella sezione precedente per calcolare lo scarto tra gli ordinamenti delle caratteristiche significative sia per i madrelingua italiani che per i madrelingua russi. I risultati sono riportati nella tabella 12 dove con il colore verde sono evidenziate le caratteristiche più significative per i madrelingua italiani rispetto ai madrelingua russi mentre con il colore rosso sono marcate le caratteristiche che hanno influenzato particolarmente i madrelingua russi rispetto ai madrelingua italiani:

Differenza ru - it	Posizione russi	Posizione italiani	Caratteristica
19	28	9	verbal_head_per_sent
18	30	12	avg_subordinate_chain_len
12	25	13	subordinate_proposition_dist
9	10	1	avg_max_depth
7	22	15	avg_links_len
-1	1	2	n_tokens
-6	26	32	principal_proposition_dist
-7	24	31	max_links_len
-15	2	17	subordinate_post
-20	15	35	dep_dist_root

Tabella 12. La differenza tra le caratteristiche significative per i madrelingua italiani e le caratteristiche significative per i madrelingua russi.

Dalla Tabella 12 si evince che ci sono poche caratteristiche comuni tra due ordinamenti. Entrambi campioni di annotatori concordano sulla rilevanza del numero di tokens nelle frasi. Altre caratteristiche comuni occupano comunque delle posizioni piuttosto lontane negli ordinamenti. Gli annotatori italiani hanno considerato la distribuzione delle teste verbali, la lunghezza media delle catene subordinate e la distribuzione delle subordinate più rilevanti nella valutazione delle frasi. Per gli annotatori russi invece le

caratteristiche legate alla posizione di subordinate post-verbale e il numero medio della lunghezza massima di una catena di dipendenza hanno la maggior rilevanza rispetto agli annotatori italiani.

5 Conclusione

Nel presente elaborato è stato presentato uno studio sulla percezione della complessità linguistica di frasi in italiano da parte di madrelingua russi apprendenti l'italiano come L2. L'obbiettivo dello studio è stato quello di determinare quali caratteristiche linguistiche delle frasi del corpus oggetto di studio hanno una maggiore influenza sulla complessità linguistica percepita dal campione di soggetti preso in esame.

Lo studio è stato organizzato in diverse fasi. Come primo passo, è stato creato un corpus di 184 frasi estratte da testi relativi al dominio giornalistico. In seconda battuta, con l'utilizzo dello strumento di analisi automatica Prodiling-UD è stato effettuato il monitoraggio linguistico delle frasi che ha consentito di estrarre dal corpus 126 caratteristiche appartenenti a quattro livelli d'analisi linguistica (di base, lessicale, morfo-sintattica e sintattica).

Per la valutazione delle frasi in termini di complessità linguistica sono stati reclutati 15 madrelingua russi che possiedono conoscenze di lingua italiana a diversi livelli di competenza (classificati come principianti, intermedi e avanzati). Per raccogliere i giudizi di complessità linguistica percepita è stato creato un questionario da sottoporre a ciascun soggetto utilizzando la piattaforma Questbase. Ogni annotatore ha assegnato un giudizio sulla complessità da lui percepita per ciascuna frase usando una scala di valutazione con punteggi che vanno da 1 a 7. Con 1 sono state indicate le frasi considerate molto facili e con 7 le frasi considerate molto difficili. Per verificare la validità dei dati raccolti è stato calcolato il grado d'accordo tra gli annotatori con l'alpha di Krippendorff che ha confermato la qualità delle risposte rese.

Con i risultati ottenuti in seguito alla valutazione del corpus da parte degli annotatori è stato possibile, da una parte, valutare la percezione della complessità linguistica da parte di madrelingua russi su frasi in italiano, e dall'altro verificare se esiste un rapporto fra la complessità linguistica percepita e le caratteristiche linguistiche estratte dalle frasi. Per quanto riguarda la prima indagine, sono stati calcolati i giudizi di complessità media per ciascuna frase sia considerando l'intero set di annotatori reclutati, sia separatamente per ciascuno dei tre gruppi creati sulla base del livello di competenza della lingua italiana. Dal calcolo dei giudizi medi è emerso che, come atteso, i principianti hanno considerato

le frasi del corpus mediamente più difficili rispetto ai gruppi di intermedi e avanzati. Viceversa, il gruppo degli avanzati ha giudicato le frasi mediamente più facili rispetto agli altri annotatori. Per quanto riguarda l'interazione fra complessità percepita e caratteristiche linguistiche, questa è stata studiata calcolando la correlazione tra le caratteristiche linguistiche e i giudizi sulla complessità assegnati dagli annotatori ricorrendo all'indice di correlazione di Spearman. Questa analisi ha consentito di determinare le quali caratteristiche linguistiche hanno influenzato maggiormente la percezione della complessità linguistica e ordinarle secondo la rilevanza. Dall'analisi è emerso che i soggetti sono stati principalmente influenzati nel giudizio dalla lunghezza delle frasi osservate. Accanto a questa caratteristica di base, i dati mostrano anche il forte impatto di caratteristiche che catturano la struttura sintattica della frase, in particolare tutti gli annotatori hanno mostrato una sensibilità alla profondità dell'albero sintattico e alla presenza di subordinate nella frase.

Grazie alla divisione degli annotatori in gruppi per il livello di competenza nella lingua italiana è stato possibile svolgere un'ulteriore analisi per individuare quali caratteristiche influiscono sulla percezione degli annotatori a seconda del loro livello di conoscenza dell'italiano. Il gruppo degli annotatori principianti, per esempio, ha mostrato una maggiore sensibilità alla lunghezza in caratteri dei tokens e ai tempi verbali. Le caratteristiche linguistiche più distintive per il gruppo degli intermedi sono la distribuzione delle relazioni di paratassi, e, per quanto riguarda i verbi, la persona e il numero. Il gruppo degli avanzati si è dimostrato più sensibile di altri gruppi alla presenza di subordinate. Per quanto riguarda il confronto generale fra i tre gruppi, notiamo una maggior vicinanza fra i gruppi di intermedi e avanzai, che infatti mostrano ordinamenti di caratteristiche linguistiche fra loro più simili rispetto a quello ottenuto dal gruppo dei principianti.

Infine, le caratteristiche linguistiche significative per la percezione della complessità da parte dei madrelingua russi sulle frasi in italiano sono state messe a confronto con i risultati ottenuti nello studio sulla percezione di complessità condotto sulle stesse coinvolgendo i madrelingua italiani (Iavarone, 2017). Dall'analisi parallela condotta sulle due liste relative alle caratteristiche linguistiche è emerso che i fenomeni che influenzano la percezione della complessità sono molto diversi per gli insiemi di

madrelingua italiani e di madrelingua russi apprendenti l'italiano come L2. In particolare, le caratteristiche relative alla struttura sintattica della frase sembrano influire maggiormente sui giudizi degli italiani.

Per quanto riguarda i futuri sviluppi di questo studio, una possibile direzione consisterebbe nell'ampliare il corpus delle frasi permettendo una più ampia rappresentazione di fenomeni linguistici. Inoltre, aumentando il numero di annotatori madrelingua russi coinvolti nello studio potremmo aumentare la rappresentatività del campione e valutare la percezione della complessità considerando la significatività non più solo rispetto alla competenza della L2 ma anche su altre variabili di sfondo, come ad esempio l'età o la conoscenza di altre L2. La metodologia e le analisi descritte nel presente studio possono infatti essere riprodotte per estendere l'indagine sulla percezione della complessità linguistica a parlanti nativi di altre lingue oltre al russo. Quest'ultima direzione è particolarmente promettente dal momento che ci permetterebbe di indagare le differenze nella percezione dei parlanti madrelingua di lingue diverse e verificare, per esempio, se madrelingua di lingue fra loro vicine sono influenzati dagli stessi fenomeni quando devono valutare il livello di complessità linguistica di frasi italiane.

6 Bibliografia

Bosco C., Lombardo V., Lesmo L. and Vassallo D. (2000). *Building a treebank for italian: a data-driven annotation schema.* In: *Proceedings of LREC'00*, Athens, Greece.

Bosco C., Montemagni S., Mazzei A., Lombardo V., Dell'Orletta F., and Lenci A. (2009). *Evalita'09 parsing task: comparing dependency parsers and treebanks.* In: *Proceedings of Evalita'09*, Reggio Emilia, Italy.

Bosco C., Simi M., and Montemagni S. (2012). *Harmonization and merging of two italian dependency treebanks.* In: *Proceedings of the LREC 2012 Workshop on Language Resource Merging*, Istanbul, Turkey.

Bosco C., Montemagni S., Simi M. (2013). *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank*. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, ISBN 978-1-937284-58-9, 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, 8-9 August 2013, pp. 61-69.

Brunato D., De Mattei L, Dell'Orletta F., Iavarone B., Venturi G. (2018). *Is this Sentence Difficult? Do you Agree?* In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Bruxelles, pp. 31-41

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.

Dell'Orletta F., Marchi S., Montemagni S., Plank B., Venturi G. (2012) *The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts.* In: *Proceedings of the 4th*

Workshop on Semantic Processing of Legal Texts (SPLeT 2012), held in conjunction with LREC 2012, Istanbul, Turkey, 27th May, pp. 42-51.

Gagliardi G. (2018) *Inter-annotator agreement in linguistica: una rassegna critica*. In: *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Aachen, CEUR-WS, 2018, 2253, pp. 206 - 212 (atti di: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, Torino, 10-12 dicembre 2018) [Contributo in Atti di convegno]

Iavarone B. (2017). *Indagine multilingue sulla complessit`a della frase: confronto tra complessit`a percepita e analisi automatica*. Pisa, Italia.

Krippendorff K. (2007). *Computing Krippendorff's Alpha-Reliability*.

Lenci A., Montemagni S. e Pirrelli V. (2005). *Testo e computer.* Carocci, Roma, 2005.

Miestamo M. (2008). *Grammatical complexity in a cross–linguistic perspective.* In: Miestamo M., Sinnemäki K. and Karlsson F. *Language Complexity: Typology, Contact, Change.* Amsterdam, Benjamins, pp. 23–41

Montemagni S. (2013). *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*. In: *Studi Italiani di Linguistica Teorica e Applicata (SILTA),* Numero 1, Anno XLII, pp. 145-172

Norris J. M. , Ortega L. (2009). *Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity.* In: *Applied Linguistics*, 30, pp. 555-578

Pallotti G. (2014). *A simple view of linguistic complexity.* In: *Second language research*, 31, pp. 117-134

Simi M., Bosco C. e Montemagni S. (2014). *Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies.* In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* ISBN 978-2-9517408-8-4, Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26-31 May 2014, Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. e Piperidis S. (a cura di), edito da European Language Resources Association ELRA, Parigi, Francia.

Vedder I. (2019). *La valutazione della complessità sintattica nella classe di lingua. Un'analisi delle riflessioni di insegnanti di italiano L2*. In: *EL.LE*, 8(3), pp. 551-566.

7 Appendice

Caratteristica	Correlazione	P-value
n_tokens	0.586838381	2.06E-18
avg_max_depth	0.502189	3.79E-13
max_links_len	0.468081675	2.08E-11
verbal_head_per_sent	0.456800099	7.14E-11
avg_links_len	0.444034971	2.73E-10
avg_subordinate_chain_len	0.439480545	4.34E-10
subordinate_proposition_dist	0.431404752	9.73E-10
dep_dist_mark	0.381904888	8.83E-08
upos_dist_SCONJ	0.360576406	0.000000496
dep_dist_advcl	0.341737127	0.00000206
upos_dist_ADV	0.323241563	0.00000762
subordinate_dist_2	0.303248169	0.0000286
verbs_mood_dist_Ind	0.294192305	0.0000505
dep_dist_advmod	0.29056343	0.000063
subordinate_post	0.288170922	0.0000728
dep_dist_acl:relcl	0.284101137	0.0000929
obj_pre	0.280046395	0.000118034
dep_dist_iobj	0.272551493	0.000181846
dep_dist_obj	0.271373698	0.000194
char_per_tok	0.267633628	0.000239862
verbs_tense_dist_Pres	0.267294911	0.000244433
obj_post	0.24715779	0.000719
verbs_tense_dist_Imp	0.242331371	0.000918788
subordinate_dist_1	0.229543532	0.001721934
avg_verb_edges	0.227239246	0.00192
subordinate_pre	0.226718675	0.001969425
verb_edges_dist_5	0.22530515	0.002105053
verbs_num_pers_dist_Sing+1	0.222520738	0.002397349
verbs_form_dist_Fin	0.221702698	0.002489969
verbs_num_pers_dist_Sing+3	0.216227451	0.003198166
dep_dist_cc	0.210777258	0.004079165
aux_mood_dist_Sub	0.208896871	0.004430427
verbs_num_pers_dist_Plur+3	0.204693364	0.005315771
upos_dist_AUX	-0.207171077	0.004776488
principal_proposition_dist	-0.434617478	7.08E-10
dep_dist_root	-0.586838381	2.06E-18

Tabella 13. Le caratteristiche linguistiche significative per il gruppo degli annotatori principianti.

Caratteristica	Correlazione	P-value
n_tokens	0.732496229	3.24E-32
avg_max_depth	0.631313341	7.49E-22
verbal_head_per_sent	0.558099819	1.86E-16
max_links_len	0.544814659	1.3E-15
avg_subordinate_chain_len	0.496378979	7.74E-13
subordinate_proposition_dist	0.485835728	2.74E-12
avg_links_len	0.475145052	9.42E-12
upos_dist_SCONJ	0.403478636	1.35E-08
dep_dist_mark	0.400646316	1.75E-08
subordinate_dist_2	0.36268724	4.2E-07
obj_post	0.359276018	5.49E-07
subordinate_post	0.35153812	9.93E-07
dep_dist_acl:relcl	0.351138833	1.02E-06
upos_dist_ADV	0.342177621	1.99E-06
dep_dist_advcl	0.333670672	3.68E-06
dep_dist_advmod	0.311353873	1.69E-05
dep_dist_obj	0.295817191	4.56E-05
verbs_mood_dist_Ind	0.293022479	5.42E-05
avg_verb_edges	0.287752767	7.47E-05
verbs_num_pers_dist_Sing+1	0.27267873	0.000181
verb_edges_dist_5	0.26747591	0.000242
aux_mood_dist_Sub	0.263189398	0.000307
verbs_num_pers_dist_Sing+3	0.262826458	0.000313
dep_dist_parataxis	0.254826935	0.000481
subj_pre	0.251821319	0.000564
verbs_form_dist_Fin	0.248231359	0.00068
subordinate_dist_1	0.239029806	0.001084
dep_dist_iobj	0.236032948	0.001257
verbs_mood_dist_Cnd	0.235128393	0.001314
obj_pre	0.235125879	0.001314
verbs_num_pers_dist_Plur+3	0.233795573	0.001403
dep_dist_xcomp	0.230143034	0.001673
subordinate_dist_3	0.229830616	0.001698
verb_edges_dist_4	0.229667505	0.001712
aux_num_pers_dist_Plur+3	0.210403732	0.004147
n_prepositional_chains	0.209072	0.004397
verbs_tense_dist_Pres	0.206580002	0.0049
verbs_mood_dist_Sub	0.206137954	0.004995
dep_dist_conj	0.193603725	0.008458
subordinate_pre	0.192652741	0.008792
dep_dist_acl	0.180136753	0.014409
verbs_tense_dist_Imp	0.174975434	0.017517

verb_edges_dist_6	0.174556015	0.017794
upos_dist_CCONJ	0.17450952	0.017825
dep_dist_fixed	0.172343012	0.019316
char_per_tok	0.1705354	0.020642
dep_dist_cc	0.168551808	0.022187
dep_dist_expl:impers	0.161940034	0.028077
prep_dist_1	0.161136655	0.028877
upos_dist_VERB	0.157186847	0.033096
dep_dist_ccomp	0.151406869	0.040206
dep_dist_compound	0.145144985	0.049316
upos_dist_PUNCT	-0.161425294	0.028587
dep_dist_punct	-0.161425294	0.028587
dep_dist_case	-0.176159453	0.016757
principal_proposition_dist	-0.405854979	1.09E-08
dep_dist_root	-0.732496229	3.24E-32

Tabella 14. Le caratteristiche linguistiche significative per il gruppo degli annotatori intermedi.

Caratteristica	Correlazione	P-value
n_tokens	0.769998978	2.32E-37
avg_max_depth	0.662533517	1.28E-24
max_links_len	0.564981603	6.59E-17
verbal_head_per_sent	0.562432541	9.71E-17
avg_subordinate_chain_len	0.513544825	9.01E-14
subordinate_proposition_dist	0.495482621	8.63E-13
avg_links_len	0.489019939	1.88E-12
dep_dist_mark	0.423292034	2.14E-09
upos_dist_SCONJ	0.42079835	2.72E-09
subordinate_dist_2	0.371696486	0.000000205
dep_dist_acl:relcl	0.363800197	0.000000385
subordinate_post	0.360756202	0.000000489
dep_dist_advcl	0.34739092	0.00000136
upos_dist_ADV	0.335318773	0.00000327
obj_post	0.332104282	4.11321E-06
verbs_mood_dist_Ind	0.314956162	0.0000133
verbs_tense_dist_Pres	0.302587737	2.98223E-05
dep_dist_advmod	0.300023424	0.0000351
avg_verb_edges	0.276528765	0.0001448
dep_dist_obj	0.269023237	0.000221928
verb_edges_dist_5	0.26477413	0.000281085
verbs_form_dist_Fin	0.262826754	0.000312829
verb_edges_dist_4	0.262211905	0.000323523
aux_mood_dist_Sub	0.261262495	0.000340704

subordinate_dist_3	0.25932526	0.000378419
verbs_num_pers_dist_Sing+1	0.256675515	0.000436292
dep_dist_iobj	0.256459973	0.000441342
verbs_num_pers_dist_Sing+3	0.245375758	0.000787273
n_prepositional_chains	0.244280691	0.000832441
verbs_num_pers_dist_Plur+3	0.238370154	0.001120128
verbs_mood_dist_Cnd	0.232531793	0.001491284
subordinate_dist_1	0.228764865	0.001787156
prep_dist_1	0.226733486	0.001968047
obj_pre	0.220097565	0.0026812
dep_dist_xcomp	0.212261825	0.003819773
dep_dist_conj	0.209436978	0.00432685
dep_dist_parataxis	0.206036913	0.005016933
principal_proposition_dist	-0.483779986	3.48E-12
dep_dist_root	-0.769998978	2.32E-37

Tabella 15. Le caratteristiche linguistiche significative per il gruppo degli annotatori avanzati.