UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Indagine sulla distribuzione di fenomeni linguistici all'interno del testo in generi diversi

Candidato: *Francesco Gracci*

Relatore: *Felice Dell'Orletta*

Correlatore: *Alessandro Lenci*

Correlatore: *Dominique Brunato*

Anno Accademico 2017-2018

Alla mia famiglia

Sommario

INTRODUZIONE

I fenomeni linguistici sono da sempre l'oggetto di base della linguistica, una disciplina umanistica che studia il linguaggio umano nel passato e nel presente, e nelle varie parti del mondo. L'interesse per la suddetta scienza ha dato origine a una serie di sottodiscipline che hanno permesso di portare avanti studi diversi a seconda delle varie componenti che costituiscono la lingua. Tra queste discipline si è venuta a formare anche la linguistica computazionale, che si occupa di sviluppare i formalismi descrittivi del funzionamento di una lingua naturale, di modo che essi possano essere tradotti in programmi eseguibili al computer. Tale tecnologia ha fatto sì che fosse possibile individuare determinati fenomeni linguistici all'interno di un testo in modo del tutto automatico, garantendo inoltre la possibilità di eseguire confronti tra un testo e l'altro. Sono proprio i confronti fra i testi il punto di partenza della seguente relazione. Da quale prospettiva però vengono distinti tra loro i testi? Le possibilità sono molteplici, ma in questo caso si parla di differenze di genere testuale. Testi appartenenti ad un genere linguistico come quelli narrativi confrontati con documenti di natura scientifica o testi con finalità didattiche, e così via. L'obbiettivo primario diventa quindi determinare quali fenomeni linguistici caratterizzino un genere rispetto ad un altro.

La novità delle seguente analisi rispetto a quanto visto in altri studi sta però nel considerare l'andamento dei fenomeni linguistici all'interno del testo, per ciascun genere considerato. A questo proposito i documenti sono stati divisi in fasce che permettono di determinare i fenomeni linguistici caratteristici per ciascuna di esse. Non si parla quindi solo di un'analisi globale ma di uno studio ben più specifico. Una caratteristica linguistica potrebbe caratterizzare un insieme di documenti solo per la prima porzione e non essere discriminante nella porzione successiva. La divisione dei documenti in porzioni non si limita però soltanto al confronto fra generi, ma permette inoltre di determinare l'andamento e il grado di caratterizzazione che un determinato fenomeno linguistico assume in un dato genere. Ciascuna coppia di fasce diventa infatti un intervallo considerabile, per esempio quello tra la prima e la seconda fascia, ed è quindi possibile calcolare il grado di correlazione di una data feature in quell'intervallo e determinare se inoltre sia per esso caratterizzante.

Nel capitolo 1 sarà descritto uno stato dell'arte del monitoraggio linguistico con un approfondimento sugli studi comparativi tra varietà di lingua, generi e livelli di complessità. Il capitolo 2 descriverà i corpora utilizzati e i fenomeni linguistici su di essi monitorati. Inoltre verranno presi in esame due insiemi di fenomeni linguistici particolarmente interessanti nel corso dell'analisi. Nel capitolo 3 avranno spazio le funzioni statistiche e gli strumenti di analisi utilizzati per l'estrazione de dati. Il capitolo 4 descriverà nel dettaglio i quattro studi, eseguiti in merito all'indagine sui fenomeni linguistici in generi testuali differenti. E infine nel capitolo 5 saranno discussi i principali risultati dello studio.

CAPITOLO 1

1. Monitoraggio linguistico

1.1. Cos'è il monitoraggio linguistico

Il termine "Monitoraggio Linguistico" indica lo studio di una determinata lingua nelle sue varietà diamesiche, diafasiche e diastratiche, nonché sull'asse diacronico. Nel dettaglio la diamesia comprende le variazioni dipendenti dal mezzo fisico-ambientale (scritto/parlato), la diafasia le variazioni dipendenti dal contesto situazionale (registro formale/registro informale) e la diastratia le variazioni riconducibili all'estrazione sociale dei parlanti (età, sesso, livello di istruzione, ecc.). I fenomeni linguistici possono essere anche monitorati vengono analizzati da un punto di vista diacronico, ovvero vengono considerati secondo il loro divenire nel tempo.

Mediante il ricorso alle tecnologie linguistico-computazionali è oggi possibile monitorare in modo affidabile un ampio spettro di parametri, che spaziano tra i diversi livelli di descrizione linguistica, in relazione a corpora testuali di sempre più vaste dimensioni (*Montemagni*, 2013).

In questo capitolo ci occuperemo di descrivere le tecnologie linguistico-computazionali e la loro applicazione allo studio della variazione linguistica (1.1.1.) e di come sia cambiato il concetto di "Monitoraggio Linguistico" nel corso degli anni (1.1.2.).

1.1.1. Analisi linguistica

Le tecnologie linguistico-computazionali consentono di determinare la struttura linguistica sottostante a un testo e di accedere al contenuto informativo di quest'ultimo. Nello specifico esse realizzano un processo incrementale, attraverso analisi linguistiche a livelli di complessità crescente. A partire dalla rappresentazione esplicita è possibile segmentare il testo in frasi (sentence splitting) e successivamente in parole ortografiche ("tokens") mediante la fase di "tokenizzazione". Seguono le fasi di analisi morfo-sintattica e lemmatizzazione del testo "tokenizzato" per culminare nell'analisi della struttura sintattica della frase in termini di relazioni di dipendenza. Ogni "token" può essere visto come un'occorrenza di forma di parola e per ogni livello di analisi essa presenta determinate proprietà. All'interno di una frase

ogni forma è univocamente identificata da un numero progressivo. Il livello di lemmatizzazione è caratterizzato dal lemma a cui ogni forma è associata. Al livello di annotazione morfo-sintattica, a ogni "token" viene associata una categoria grammaticale (V=verbo, R=articolo, S=sostantivo, A=aggettivo, ecc.), eventuali sottocategorie (EA=preposizione articolata, RD= articolo determinativo, ecc.) e specificazioni morfologiche (persona, genere, numero ecc.). Il livello di annotazione sintattica o *parsing* fornisce una descrizione della frase in termini di relazioni binarie di dipendenze tra parole (soggetto, oggetto, modificatore, ecc.). Nello specifico tale livello presenta l'identificatore univoco (0=radice, generalmente il verbo) della forma che costituisce la testa da cui il token dipende e il tipo di dipendenza (det, subj, mod, obj, ROOT).

In linguistica computazionale per la rappresentazione delle analisi linguistiche viene utilizzato il formato tabellare CoNLL, in cui ogni riga rappresenta un token e ogni colonna un livello di annotazione ottenuto mediante l'utilizzo degli strumenti automatici (v. tabella 1).

		Lemmatizzazione	Annotazione morfo-sintattica			Annotazione a dipendenze	
Id	Forma	Lemma	CaGra1	CaGra2	Tratti	Testa	Tipo di relazione
1	Le	il	R	RD	num=p gen=f	2	det
2	tecnologie	tecnologia	S	S	num=p gen=f	4	subj
3	linguistiche	linguistico	A	A	num=p gen=f	2	mod
4	rappresentano	rappresentare	V	V	num=p per=3 mod=i ten=p	0	ROOT
5	un	un	R	RI	num=s gen=m	6	det
6	ausilio	ausilio	S	S	num=s gen=m	4	obj
7	importante	importante	A	A	num=s gen=n	6	mod
8	per	per	E	E	_	6	comp
9	il	il	R	RD	num=s gen=m	10	det
10	monitoraggio	monitoraggio	S	S	num=s gen=m	8	prep
11	della	di	E	EA	num=s gen=f	10	comp
12	lingua	lingua	S	S	num=s gen=f	11	prep
13	italiana	italiano	A	A	num=s gen=f	12	mod
14	.	.	F	FS	_	4	punc

Tabella 1: esempio di rappresentazione tabellare del testo annotato linguisticamente

Dalla tabella 1 è possibile ottenere una rappresentazione esplicita dell'albero di dipendenze sintattiche, all'interno del quale gli archi marcano le relazioni fra la testa e un dipendente (v. figura 1).

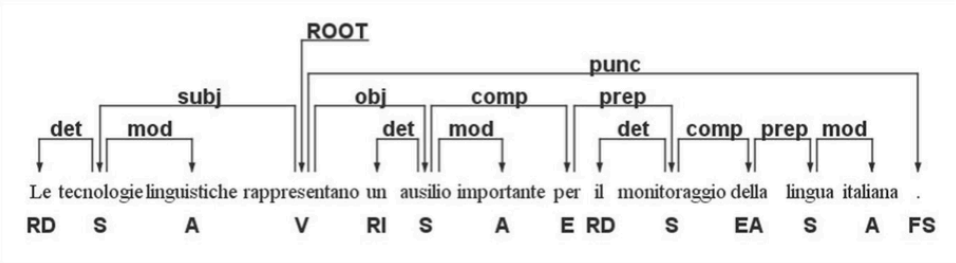

Figura 1: rappresentazione grafica dell'annotazione linguistica dell'esempio in Tabella 1

Tutte le informazioni sopra riportate possono essere sfruttate per ulteriori elaborazioni automatiche, volte a individuare una vasta tipologia di parametri che possano essere sfruttati in compiti di monitoraggio linguistico.

Al giorno d'oggi, nel campo della linguistica computazionale, l'annotazione linguistica viene eseguita mediante sistemi basati su algoritmi di apprendimento automatico supervisionato. L'annotazione linguistica diventa una sorta di classificazione probabilistica in cui il sistema sceglie l'annotazione più probabile data la parola in input, il contesto, i suoi tratti caratteristici, ecc. Il modello probabilistico per l'annotazione del testo viene costruito a partire da un corpus di addestramento.

1.1.2. Evoluzione degli studi di monitoraggio

Per quanto riguarda la lingua italiana il monitoraggio era in origine eseguito sui corpora (collezioni di testi selezionati e organizzati per le analisi linguistiche) che venivano sottoposti a processi di lemmatizzazione e annotazione morfo-sintattica condotti in modo manuale o semiautomatico. Successivamente un'analisi linguistica condotta in modo completamente automatico ha fatto sì che l'unica fase del processo rimasta prettamente manuale fosse la revisione conclusiva, non sempre eseguita. Gli studi di Voghera (2004, 2005) o quelli di Cresti (2005), per esempio, si sono basati su risorse del suddetto tipo e presentano come differenza sostanziale la mancanza della fase di revisione manuale in Cresti. I suddetti studi si limitano però a fornire una lista di lemmi senza indicare in alcun modo la loro funzione sintattico-semantica. L'unica eccezione nel panorama italiano è il corpus Penelope, una risorsa di dimensioni contenute - poco più di 30.000 parole -concepita a supporto dello studio delle

differenze a livello sintattico tra diverse varietà d'uso della lingua italiana. Su di essa si sono basati tutti gli studi e ricerche di tipo sintattico condotti durante i suoi venticinque anni di vita.

Per altre lingue (inglese, somalo, coreano, taiwanese, spagnolo) vi sono studi basati su corpora di maggiori dimensioni con l'analisi linguistica condotta in modo completamente automatico fin dalla metà degli anni Ottanta. In tali studi si utilizzano tecniche di statistica multivariata finalizzate a identificare i tratti caratterizzanti di una varietà di lingua rispetto ad un'altra. Tutto ciò mostra come l'uso di corpora di vaste dimensioni e di strumenti di analisi linguistica automatica permetta di condurre un monitoraggio linguistico ad ampio spettro.

Grazie alle tecnologie linguistiche-computazionali il monitoraggio linguistico oggi può essere condotto su corpora di vaste dimensioni, può basarsi su informazioni della struttura sintattica (originariamente attingibili solo tramite un accurato lavoro manuale) e non è circoscritto a limitate porzioni di testo o a ristretti repertori linguistici.

1.2. Parametri monitorati

1.2.1. Tipologie di caratteristiche (features) linguistiche

Tramite l'analisi linguistica di un testo è possibile estrarre da quest'ultimo diversi parametri (features), riconducibili a quattro categorie linguistiche (caratteristiche di base, lessicali, morfo-sintattiche e sintattiche).

- **Caratteristiche di base**

Le caratteristiche di base di un testo costituiscono la prima categoria di parametri ricavabile mediante un'analisi linguistica. Esse sono informazioni per lo più di tipo quantitativo che si limitano ad indicare il numero di componenti di cui un determinato testo è costituito (numero di frasi, numero di token). Alcune informazioni vengono inoltre ottenute mediante indagini statistiche come per esempio il numero medio di token per frase o il numero medio di caratteri per parola.

- **Caratteristiche lessicali**

La caratteristiche lessicali descrivono la varietà di parole diverse contenuta in un testo. Il valore percentuale che descrive tale varietà prende il nome di Type/Token Ratio (TTR) e si calcola dividendo il numero delle parole diverse usate in un testo (types) per il numero di parole di complessive del testo stesso (tokens) moltiplicato

per 100. Tanto più sarà alto il valore percentuale, maggiore sarà la varietà del vocabolario usato nel testo.

- **Caratteristiche morfo-sintattiche**

Le caratteristiche morfo-sintattiche si distinguono in due categorie. La prima è quella maggiormente generica e comprende aggettivi, articoli, preposizioni, nomi, verbi, ecc. L'altra è più specifica e comprende le cosiddette sottocategorie morfo-sintattiche che offrono un ulteriore livello di distinzione fra le categorie generiche sopra riportate. Per esempio gli articoli determinativi e gli articoli indeterminativi, le preposizioni semplici e le preposizioni articolate e così via.

- **Caratteristiche sintattiche**

Le caratteristiche sintattiche comprendono tutte le informazioni ottenibili dall'albero sintattico di una frase. Per esempio l'altezza, l'ampiezza, il numero di subordinate, il numero di oggetti, ecc. Tali informazioni diventano materia di analisi all'interno di indagini statistiche che forniscono un ulteriore livello di conoscenza mediante informazioni quali la media delle altezze degli alberi, la media di archi entranti in teste verbali, il numero di token per clausola ecc.

- **Features**

Il risultato del monitoraggio è una sequenza o vettore di numeri, in cui ogni numero definisce il valore di una feature, che a sua volta definisce una caratteristica linguistica del testo preso in esame (v. tabella 2).

Frasi	Tokens	Caratteri	Media Caratteri X Parola
1	20	74	3,7
2	14	61	4,36
3	18	67	3,72
4	7	33	4,71
5	8	34	4,25
6	10	51	5,1
7	13	54	4,15
8	14	64	4,57
9	13	57	4,38
10	20	94	4,7

Tabella 2: tokens, caratteri e media caratteri per parola nelle prime 10 frasi di un testo tipo

Dalla tabella 1 è possibile ottenere una rappresentazione grafica dell'andamento dei vari parametri all'intero del testo preso in esame (v. figura 2).

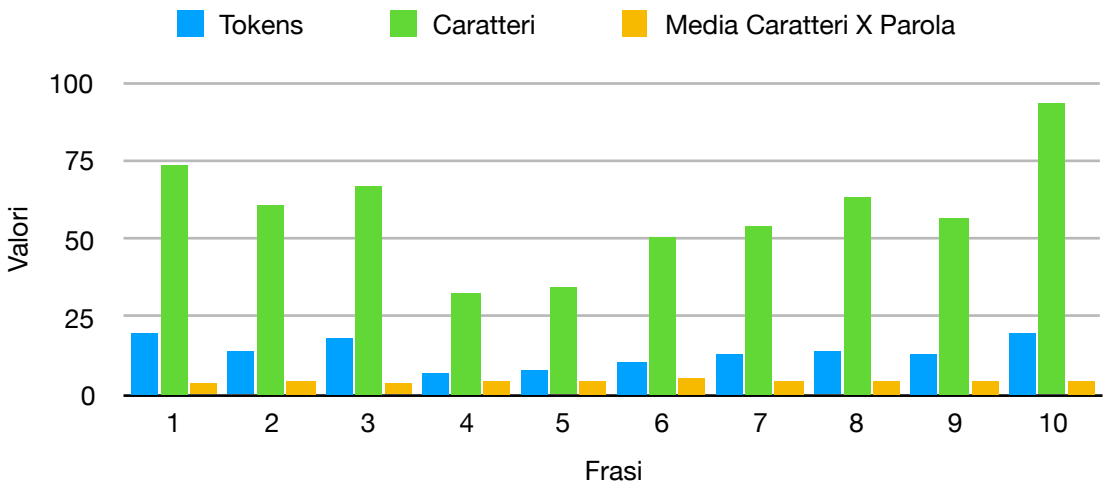

Figura 2

Tale operazione fa sì che ogni indagine o studio eseguito sui dati possa essere accompagnato da una o più rappresentazioni grafiche.

1.2.2. Estrazione delle features

La prima operazione da compiere nel processo di estrazione delle features consiste nel convertire il corpus di testi preso in esame nel formato CoNLL. L'ItaliaNLP Lab dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) ha realizzato il programma LinguA che permette di compiere in modo automatico l'analisi linguistica (sentence splitting, part of speech tagging, syntactic parsing, syntactic trees) di un testo e di ottenere il relativo file in formato tabellare (o CoNLL).

Un corpus di testi verrà quindi in prima analisi suddiviso in frasi, che a loro volta saranno suddivise in token. Per ogni token così ottenuto LinguA eseguirà l'analisi morfo-sintattica e sintattica, restituendo inoltre anche la rappresentazione grafica dell'albero di dipendenze sintattiche.

	ID	Token	Lemma	C-POS	F-POS	Morphosyntactic feats	HEAD	DEP
1	1	Gli	il	R	RD	num:p gen:m	2	det
	2	aspetti	aspetto	S	S	num:p gen:m	4	subj
	3	organizzativi	organizzativo	A	A	num:p gen:m	2	mod
	4	sono	essere	V	V	num:p mod:i per:3 ten:p	0	ROOT
	5	di	di	E	E		4	pred
	6	fondamentale	fondamentale	A	A	num:s gen:n	7	mod
	7	importanza	importanza	S	S	num:s gen:f	5	prep
	8	.	.	F	FS		4	punc

Tabella 3: analisi sintattica ottenuta mediante "LinguA"

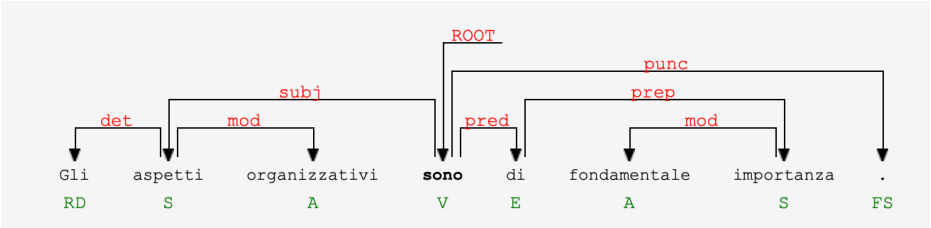

Figura 3: albero di dipendenze sintattiche ottenuto mediante “LinguA”

Per esempio dalla frase “Gli aspetti organizzativi sono di fondamentale importanza” LinguA restituirà la divisione in frasi (in questo caso superflua in quanto analisi di una singola frase), l’analisi morfo-sintattica, l’analisi sintattica (v. tabella 3) e l’albero di dipendenze sintattiche (v. figura 3). Ad operazione conclusa sarà possibile scaricare la frase o le frasi analizzate in formato .CoNLL, che potrà essere visualizzata tramite un editor di testo:

```
1       Gli     il      R       RD      num:p gen:m   2       det
2       aspetti         aspetto         S       S       num:p gen:m   4       subj
3       organizzativi   organizzativo   A       A       num:p gen:m   2       mod
4       sono    essere  V       V       num:p mod:i per:3 ten:p     0       ROOT
5       di      di      E       E               4       pre
6       fondamentale    fondamentale    A       A       num:s gen:n   7       mod
7       importanza      importanza      S       S       num:s gen:f   5       prep
8       .       .       F       FS              4       punc
```

Nel caso sopra presentato il token “sono” occupa la quarta posizione all’interno della frase, ha come lemma “essere”, è un verbo (V), non presenta una sottocategoria grammaticale, è una terza persona plurale del modo indicativo al tempo presente, ha come testa sintattica 0 e come relazione di dipendenza ROOT ed è quindi la radice della frase. Le stesse osservazioni possono essere eseguite sulle altre componenti della frase e fanno sì che tutti gli aspetti morfo-sintattici e sintattici della frase vengano rivelati, in modo totalmente automatico e con un alto grado di precisione.

Il passo successivo consiste nell’utilizzare le suddette informazioni (ottenute su tutte le frasi presenti nel testo o nei corpus di testi) per determinare i parametri (features) di nostro interesse. L’estrazione della features viene eseguita mediante l’ausilio di strumenti informatici come per esempio il linguaggio di programmazione “Python”. Uno studio di questo tipo consente infatti di scrivere algoritmi che saranno di volta in volta applicati a tutti i corpus di testi considerati utili per il tipo di analisi prefissata. Tra gli studi più comuni abbiamo il conteggio dei token di una frase, la

media dei caratteri per frase, il numero di sostantivi pre verbali, il numero di subordinate di primo grado, la percentuale di articoli determinativi rispetto a quelli indeterminativi, ecc. Tutti questi esempi diventano parametri da monitorare e vanno a definire l'insieme delle features sulle quali sarà possibile eseguire ulteriori analisi automatiche e/o osservazioni di natura statistica o simili. Per esempio se il numero di subordinate presenti in un corpus di testi risultasse maggiore rispetto a quello presente in un altro corpus analizzato potremmo dire che il primo corpus è costituito da testi con periodi più lunghi rispetto al secondo.

1.2.3 Genere e complessità

Le features estratte mediante il monitoraggio linguistico permettono di eseguire studi di diversa natura. Il semplice confronto fra i dati ottenuti da un testo e quelli derivanti da un altro mostra, come visto nelle righe precedenti, come i due testi differiscano l'uno rispetto all'altro. Possono esserci differenze sintattiche ben visibili, si può individuare un vocabolario più ampio in uno rispetto ad un altro, i periodi possono essere più brevi o più lunghi e via dicendo.

Lo studio di maggior interesse si ha però quando i confronti non vengono più eseguiti tra due testi o tra piccoli gruppi di essi, ma quando si inizia a lavorare su grandi corpora, aumentando quindi il numero di testi su cui eseguire l'operazione di monitoraggio. Testi appartenenti a diversi generi testuali e con una diversa complessità offrono molte più informazioni rispetto a un semplice confronto uno a uno. Lo studio degli andamenti e della caratteristiche linguistiche di corpus di testi appartenenti a una dato genere porta inevitabilmente a individuare delle regole/leggi fisse per il genere stesso. Il confronto con altri generi permette poi di determinare se le cosiddette regole individuate possano essere considerati tali o se trovino riscontro anche in altre analisi, perdendo quindi valore e non definendo più il genere nello specifico.

Nello studio di *Dell'Orletta*, *Montemagn*i e *Venturi*, 2013 vengono confrontati testi appartenenti a 4 categorie di generi testuali differenti (letterario, giornalistico, didattico e scientifico). Da ognuno dei documenti appartenenti ai corpora presi in esame vengono estratte una serie di features, che vengono poi confrontate tra di loro in modo da determinare le caratteristiche strutturali tipiche di un genere rispetto ad un altro. Per esempio viene evidenziato come nei testi letterari la lunghezza media delle frasi per documento o la lunghezza media dei caratteri per frasi siano piuttosto

inferiori rispetto agli altri generi. Ciò significa che i testi altamente informativi presentano parole e periodi più lunghi rispetto ai testi appartenenti alla fiction. Per quanto riguarda le caratteristiche lessicali i testi letterari sono invece quelli con la maggiore varietà lessicale. Sul piano delle differenze morfo-sintattiche abbiamo invece un maggior numero di pronomi e di verbi rispetto agli altri generi, caratteristiche tipiche della conversazione e riconducibili alle parti dialogiche e a quelle prettamente narrative dei testi presi in esame. E tante altre deduzioni su cui non mi soffermerò.

Ciò che quindi risulta particolarmente interessante è come ogni genere testuale presenti delle caratteristiche strutturali che lo distinguono rispetto a tutti gli altri e che permettono di identificare un documento o un testo ad esso appartenente senza necessariamente conoscerne il contenuto testuale.

Tra i generi testuali più comuni troviamo i testi narrativi, gli articoli di natura scientifica, i testi giornalistici, i materiali scolastici, i testi giuridici e la letteratura. Gli esperimenti trattati nei capitoli successivi avranno proprio come materia di studio e analisi i suddetti generi divisi rispettivamente in due livelli di complessità (facile e difficile). Un articolo scientifico con un livello di complessità basso è quello pubblicato su siti come Wikipedia, mentre un articolo scientifico con un livello di complessità alto è quello che si può trovare all'interno di una rivista specializzata o su siti specializzati, strettamente legati all'ambito della ricerca e del progresso scientifico.

CAPITOLO 2

2. Corpora e raccolta dei dati

2.1. Corpora

Un corpus è una collezione di testi selezionati e organizzati in maniera tale da soddisfare specifici criteri che li rendono funzionali per le analisi linguistiche (*Testo e computer*, Lenci e altri, 2016).

I corpora sono la fonte più importante di dati in linguistica computazionale e possono essere facilmente conservati grazie al computer e alla capienza che esso offre. Vi sono casi specifici di corpus, come per esempio i corpus elettronici che possono assumere più di un significato. Essi si riferiscono infatti sia a collezioni di testi di grandi dimensioni sia a testi digitalizzati e tradotti in un formato machine-readable (Nesselhauf, 2005).

I corpora possono essere utilizzati per la progettazione di strumenti intelligenti, dotati di conoscenze linguistiche, oppure, per indagare determinati fenomeni linguistici. In questa analisi noi ci occuperemo del secondo caso.

La classificazione di un corpus si basa sui seguenti parametri (*Testo e computer*, Lenci e altri, 2016):

- Generalità: un corpus può essere generale o specialistico, a seconda del grado di specificità con cui viene descritta una lingua. Nel primo caso si parla di una collezione di testi selezionati per descrivere la lingua nel suo complesso. I corpus specialistici si concentro invece su una specifica varietà linguistica.
- Modalità: la modalità di trasmissione di un corpus fa sì che si distinguano rispettivamente corpus che si rifanno a informazioni scritte, parlate o miste.
- Cronologia: un corpus diacronico comprende testi appartenenti a periodi diversi e descrive uno o più mutamenti linguistici nel tempo. Un corpus sincronico, contenente testi appartenenti a una particolare lasso di tempo, si concentra su una particolare fase della lingua.
- Lingua: un corpus può essere monolingue o plurilingue a seconda dei testi in esso contenuti. Nel secondo caso si ha un'ulteriore distinzione:
    - Corpus parallelo: contiene testi sia nelle loro versione originale sia tradotti in un'altra lingua.

    - Corpus comparabile: contiene testi appartenenti a due lingue ma relativi allo stesso argomento o dominio.
- Integrità dei testi: un corpus può contenere testi interi o porzioni di essi di lunghezza prefissata.
- Codifica dei testi:
    - Corpus codificato: i testi sono arricchiti con etichette (codici) che forniscono informazioni aggiuntive di tipo strutturale.
    - Corpus annotato: variante dei corpus codificati in cui le informazioni rese esplicite sono di natura linguistica.

2.2. Corpora analizzati

Per le analisi trattate nei capitoli successivi sono stati utilizzati 8 corpora appartenenti e 4 generi testuali diversi (2 per genere). Distinguiamo i corpora giornalistici (2.2.1.), i corpora scientifici (2.2.2.), i corpora dei materiali didattici (2.2.3.) e i corpora narrativi (2.2.4.).

2.2.1. I corpora giornalistici

I corpora di genere giornalistico utilizzati per l'analisi sono *Repubblica* e *Due Parole*.

- **La Repubblica**: corpus contenente le annate del quotidiano "la Repubblica". La prima versione, sviluppata dall'Università degli Studi di Bologna, contiene tutti gli articoli pubblicati tra il 1958 e il 2000 per un totale di circa 380 milioni di tokens. La collezione che verrà invece analizzata nei capitolo successivi è quella comprendente gli articoli scritti tra il 2000 e il 2005, suddivisi in 321 documenti per un totale di circa 232.000 token.
- **2 Parole**: corpus che trae il nome dall'omonimo quotidiano "Due Parole" (Due-Parole, 2002), un giornale italiano d'informazione di facile lettura studiato e scritto da linguisti esperti in semplificazione dei testi utilizzando un linguaggio controllato per un pubblico adulto con un livello di alfabetizzazione primitivo o con lievi disabilità intellettuali (Piemontese 2006, citato in Brunato *e altri* (2015)). Il corpus prende il nome dall'omonimo giornale di facile lettura fondato nel 1989 dall'iniziativa di un gruppi di ricerca dell'Università di Roma "La Sapienza". Il giornale si rivolge ad un pubblico con uno scarso livello di alfabetizzazione e con difficoltà linguistiche che diventano la base per la creazione di criteri di scrittura controllata. «I criteri principali della scrittura controllata sono: la brevità dei testi,

la semplicità delle frasi, la scelta di parole più comuni della lingua italiana e perciò note alla quasi totalità dei parlanti» (www.dueparole.it). Il corpus comprende gli articoli pubblicati tra il 2001 e il 2006, suddivisi in 322 documenti per un totale di circa 73.000 token.

2.2.2. I corpora scientifici

I corpora di genere scientifico utilizzati per l'analisi sono *Articoli scientifici* e *Wikipedia*.

- **Articoli scientifici**: corpus costituito da documenti tratti da rivisti scientifiche specialistiche relative a diversi ambiti da indagine quali i cambiamenti climatici e la linguistica. Collezione di 84 documenti per un totale di circa 471.000 token.
- **Wikipedia**: corpus contenente documenti tratti dal portale italiano dell'omonima enciclopedia online sul tema di "Ecologia e Ambiente". Collezione di 293 documenti per un totale di circa 205.000 token.

2.2.3. I corpora dei materiali didattici (scolastici)

I corpora di genere scolastico utilizzati per l'analisi sono *Materiali Didattici per la Scuola Elementare* e *Materiali Didattici per la Scuola Superiore*.

- **Materiali Didattici per la Scuola Elementare**: corpus di testi di natura scolastica indirizzati agli studenti dei scuola elementare. Testi di natura piuttosto semplici visto il target di riferimento. Collezione di 127 documenti per un totale di circa 48.000 token.
- **Materiali Didattici per la Scuola Elementare**: corpus di testi di natura scolastica ma con target differente rispetto al corpus per la Scuola Elementare. Testi indirizzati a studenti di scuola superiore e quindi dotati di un grado di complessità lievemente superiore. Collezione di 70 documenti per un totale di circa 48.00 token.

2.2.4. I corpora narrativi

I corpora di genere narrativo utilizzati per l'analisi sono *Terence* e *Teacher*, rispettivamente in versione originale e semplificata.

- **Terence**: corpus costituito da 32 racconti brevi per bambini in lingua italiana, con le rispettive versioni semplificate manualmente. Il nome del corpus deriva dal "Progetto Terence", un progetto dell'Unione Europea (Teren- ce_Corsortium, 2012) ideato nel 2007 e progettato nel 2011, con obbiettivo il potenziamento della

comprensione del testo scritto in bambini di età compresa tra i sette e dieci anni afflitti da deficit o da particolari difficoltà di comprensione.

- **Teacher**: corpus che comprende 24 documenti provenienti da siti web educativi specializzati che forniscono risorse gratuite per gli insegnanti (*Brunato e altri*, 2015). Nella raccolta sono presenti diversi generi testuali oltre ai testi narrativi, com per esempio testi scolastici di storia e di geografia. La semplificazione dei testi per la realizzazione del corpus semplificato è stata eseguita secondo la strategia "intuitiva". Ogni testo è stato semplificato indipendentemente, senza tenere conto di regole o gerarchie ma basandosi esclusivamente su diversi livello linguistici.

2.3. Features e creazione dei file in formato tabellare

In questa sezione verrano trattati gli algoritmi e gli script utilizzati nell'analisi dei corpus. Il linguaggio di programmazione utilizzato per la scrittura dei programmi di analisi è il Python[1], accompagnato da alcune librerie che tratteremo meglio nei capitoli successivi. Di seguito viene descritta l'estrazione delle feature (2.3.1.) e la fase della creazione dei file che verrano utilizzati nella analisi di tipo statistico (2.3.2.).

2.3.1. Features

Per ogni documento presente nei corpora analizzati vengono estratte 89 features. Nel dettaglio abbiamo 4 features riconducibili alle categorie linguistiche di base, 28 alle categorie morfo-sintattiche e le restanti 67 alle categorie sintattiche.

- **Features di base**: indice della frase, numero di tokens per frase (esclusa la punteggiatura), numero di caratteri per frase (esclusa la punteggiatura), media caratteri per token.
- **Features morfo-sintattiche**: categorie generiche e sottocategorie rappresentate in forma percentuale. Per la distinzione delle varie feature è stato utilizzato l'ISST-TANL *morpho-syntactic tagset* sviluppato dall'istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) che permette di interrogare i file in formato .CoNLL.
- **Features sintattiche**: informazioni ottenibili dell'albero di dipendenze sintattiche. Le dipendenze (anch'esse in formato percentuale) sono state distinte tramite ISST-

[1] **Python** è un linguaggio di programmazione ad alto livello, orientato agli oggetti, adatto, tra gli altri usi, a sviluppare applicazioni distribuite, scripting, computazione numerica e system testing.

TANL *dependency tagset*, una nuova versione dell'ISST-CoNLL *dependency tagset* che permette di determinare il tipo di dipendenza di un dato token.

Di seguito vengono presi in esame due dei casi più interessanti considerati durante il monitoraggio.

2.3.2. Subordinate

Nell'estrazione delle feature particolare attenzione è stata riservata allo studio delle subordinate, divise rispettivamente in subordinate di primo grado, ovvero dipendenti dalla proposizione principale, e subordinate di grado superiore al primo, ovvero dipendenti da un'altra subordinata. Di seguito la funzione in linguaggio Python utilizzata nel suddetto studio:

```
def stampaSUB(listaSUB):
	listaSUB1=0
	elencoSUB1Pre=[]
	elencoSUB1Post=[]
	altezzaSUB1=0
	ampiezzaSUB1=0
	listaSUB2=0
	elencoSUB2Pre=[]
	elencoSUB2Post=[]
	altezzaSUB2=0
	ampiezzaSUB2=0
	Max=0
	subID=[]
	if listaSUB:
		Max=listaSUB[0].head
	for tok in listaSUB:
		h=1
		massimo=[]
		altezza=Altezza(tok,h,massimo)
		ampiezza=trovaAmpiezzaMassima(tok,altezza)
		distanzaSintatticaDallaTesta=-(tok.head-tok.id)
		for figlio in tok.children:
			subID.append(trovaDipendenze(figlio))
			subID=flatten(subID)
		if tok.head in subID:
			listaSUB2+=1
			altezzaSUB2+=altezza
			ampiezzaSUB2+=ampiezza

			elencoSUB2Pre,elencoSUB2Pre=prepost(tok,distanzaSintatti
			caDallaTesta,elencoSUB2Pre,elencoSUB2Post,listaSUB)

			elencoSUB2Pre,elencoSUB2Post=prepost(tok,distanzaSintatt
			icaDallaTesta,elencoSUB2Pre,elencoSUB2Post,listaSUB)

		elif tok.head not in subID:
			listaSUB1+=1
			altezzaSUB1+=altezza
			ampiezzaSUB1+=ampiezza

			elencoSUB1Pre,elencoSUB1Post=prepost(tok,distanzaSintatt
			icaDallaTesta,elencoSUB1Pre,elencoSUB1Post,listaSUB)
```

```
return listaSUB1,altezzaSUB1,ampiezzaSUB1,len(elencoSUB1Pre)
,len(elencoSUB1Post),listaSUB2,altezzaSUB2,ampiezzaSUB2,
len(elencoSUB2Pre),len(elencoSUB2Post)
```

Dove `listaSUB` è una lista contenente caratteristiche morfo-sintattiche e i valori di dipendenza che un token deve presentare per introdurre una subordinata, e `trovaDipendenze` una funzione che restituisce l'id dei tokens considerati, di modo da controllare la dipendenze delle subordinate e stabilire il grado di appartenenza.

Oltre al numero di subordinate di primo grado e di grado superiore al primo, e alle relative medie di altezza (massima profondità raggiunta dalle foglie dell'albero) e ampiezza (larghezza massima dell'albero) sono stati restituiti i valori, in formato percentuale, del numero di subordinate di primo grado che precedono la principale, del numero di subordinate di primo grado che seguono la principale, del numero di subordinate di grado superiore al primo che precedono la subordinata reggente e del numero di subordinate di grado superiore al primo che seguono la subordinata reggente.

2.3.3. Feature relative all'ordine degli elementi

Nel paragrafo 1.1.1. abbiamo visto come l'analisi sintattica di una frase ci permetta di rappresentare graficamente quest'ultima come un albero di derivazione sintattica, in cui i nodi sono i token e gli archi le relazioni di dipendenza che esistono tra essi. Ogni albero presenta un nodo radice (ROOT) che è generalmente rappresentato dal verbo principale. Da esso partono poi tutte le diramazioni e i possibili sviluppi della frase. Ogni nodo ha una relativa testa sintattica, ovvero il token da cui esso dipende. Nell'analisi sintattica ottenuta mediante LinguA, la penultima colonna del file ottenuto in formato CoNLL è quella della testa sintattica. Per esempio, nella frase "Mario ha difficoltà nell'uso del computer":

```
1       Mario   Mario   S       SP      _       2       subj
2       ha      avere   V       V       num=s|per=3|mod=i|ten=p     0       ROOT
3       difficoltà      difficoltà      S       S       num=n|gen=f 2       obj
4       nell'   in      E       EA      num=s|gen=n     3       comp
5       uso     uso     S       S       num=s|gen=m     4       prep
6       del     di      E       EA      num=s|gen=m     5       comp
7       computer        computer        S       SW      num=s|gen=m 6       prep
```

- "ha" è la radice ed ha testa sintattica = 0, ovvero non dipende da un altro token.

- “Mario” ha la testa sintattica = 2 in quanto ha come testa sintattica il token “ha” che si trova in posizione 2.
- “difficoltà” ha la testa sintattica = 2 in quanto ha come testa sintattica il token “ha” che si trova in posizione 2.
- “nell’” ha la testa sintattica = 3 in quanto ha come testa sintattica il token “difficoltà” che si trova in posizione 3.
- E così via.

Come abbiamo potuto osservare diversi token presentano la medesima testa sintattica, per esempio nel caso sopra descritto i token “Mario” e “difficoltà” hanno entrambi come testa sintattica il token “ha”. Gli archi di un albero partono da pochi nodi comuni che aumentano all’aumentare dell’altezza dell’albero stesso. Ciò che però diventa per noi materia di analisi è la posizione del dipendente rispetto alla testa, che è calcolata come la differenza tra l’id di un nodo (la posizione all’interno della frase) e l’id della sua testa sintattica. Se consideriamo la frase precedente, i token “Mario” e “difficoltà” hanno entrambi la testa sintattica = 2, ma presentano un diverso id, rispettivamente id = 1 e id = 3. Calcolando la differenza tra l’id e la testa sintattica otteniamo rispettivamente il valore -1 per Mario e 1 per difficoltà. Questo significa che “Mario” si trova in una posizione precedente la testa sintattica mentre “difficoltà” si trova in una posizione successiva alla testa sintattica. Più precisamente diremo che il token “Mario” occupa una posizione pre-verbale mentre il token “difficoltà” occupa una posizione post-verbale.

La seguente funzione restituisce rispettivamente gli elementi precedenti e successivi alla testa sintattica selezionata:

```
def prepost(pcd, distanza, elencoPRE, elencoPOST, elenco):
    for tipo in elenco:
        if pcd==tipo:
            if distanza<0:
                elencoPRE.append(pcd)
            else:
                elencoPOST.append(pcd)
    return elencoPRE, elencoPOST
```

Dove pcd è il valore del token che ci interessa considerare (tok.dep = “obj” per oggetti, tok.pos = “A” per gli aggettivi, ecc.), distanza è il valore della distanza dalla testa sintattica, elencoPRE ed elencoPOST sono due elenchi inizialmente vuoti che

vengono arricchiti ogni volta che si trovi un elemento rispettivamente pre o post testa sintattica, ed elenco è un lista delle occorrenze che ci interessa studiare (nei casi esaminati successivamente saranno aggettivi, avverbi, soggetti e oggetti).

L'analisi sopra riportata può essere applicata su ogni feature estratta e permette di determinare degli aspetti caratteristici dei corpus analizzati. Un gruppo di testi con una maggiore percentuale di soggetti precedenti la testa sintattica ci fornisce un'informazione differente rispetto al caso opposto e complementare, ovvero quello dei soggetti successivi la testa sintattica.

2.3.4. Creazione delle tabelle

Ai fini dell'indagine i dati estratti durante il processo di analisi linguistica sono stati suddivisi in diversi documenti in formato tabellare (.csv) di modo da poter considerare le righe e le colonne nelle successivi analisi statistiche.

Sentences	Tokens	Chars	mCxT	CPOS_A(%)	...
3.0	13.6	64.8	4.82	4.04	...
8.0	27.6	130.8	4.4	4.2	...
13.0	38.0	186.0	4.83	6.07	...
18.0	39.0	205.4	5.24	10.75	...
23.5	30.67	155.17	4.99	8.88	...
29.5	27.33	127.17	4.59	7.48	...

Tabella 1: tabella da 6 righe ottenuta eseguendo la media aritmetica sui dati.

Per ogni corpous analizzato otteniamo un file in formato .csv relativo al corrispettivo documento in formato CoNLL. Ogni documento .csv è una tabella definita da un numero di righe pari al numero di frasi presenti nel documento e un numero di colonne pari al numero di feature estratte, quindi 89. In questo modo otteniamo una serie di tabelle con lo stesso numero di colonne ma con diverso numero di righe. Tale aspetto non ci permette di eseguire uno studio per fasce, ovvero andare a determinare come certe caratteristiche possano essere significative nell'introduzione, nella parte centrale e nella conclusione di un documento. Da ciò la necessità di uniformare tutte le tabelle ad uno specifico numero di righe che da qui in avanti chiameremo fasce. Il numero di fasce stabilito è 6 poiché ci permette di mantenere gran parte dei documenti e di escluderne un numero piuttosto esiguo (i file .csv con un numero di fasce < 6 sono stati esclusi dall'indagine).

Tutte le tabelle vengono quindi uniformate e sei fasce (v. tabella 1) mediante le seguente funzione:

```
df=pd.read_csv(file, sep="\t")
righe,colonne=df.shape
rows=np.arange(0,1)
cols=list(df)
fasce=int(round(righe*1.0/6))
x=0
step=fasce
numFasce=6
df1=pd.DataFrame()
if righe>=numFasce:
       fasceVuote=numFasce
    righeRestanti=righe
    for i in range(numFasce):
              if(math.ceil(righeRestanti*1.0/fasceVuote))==fasce:
                     df_fascia=popolaRighe()
                     df1=df1.append(df_fascia)
              else:
              fasce=righeRestanti/fasceVuote
              step=x+fasce
              df_fascia=popolaRighe()
              df1=df1.append(df_fascia)
                     mediaTOT=[]
                     valori=[]
                     x=step
                     step=x+fasce
                     fasceVuote-=1
                     righeRestanti=righe-x

def popolaRighe():
       mediaTOT=[]
    valori=[]
    media = df.iloc[x:step].mean(axis=0)
    for elem in media:
              elem = round(elem,2)
              valori.append(elem)
       mediaTOT.append(valori)
    df_fascia = pd.DataFrame(mediaTOT,index=rows,columns=cols)
    return df_fascia
```

La suddetta funzione permette di eseguire in modo uniforme la divisione in fasce tramite l'utilizzo del modulo math[2] del linguaggio Python e in particolare della funzione math.ceil(x) che restituisce il limite massimo di x, ovvero il più piccolo intero maggiore o uguale a x.

Ogni tabella così ottenuta viene poi sommata alle altre per creare una tabelle definitiva, contenente tutte le informazioni di ogni documento presente nel corpus analizzato.

Per ogni tabella da sei righe viene creata la versione normalizzata, che consiste nel dividere ogni elemento di una colonna per il maggiore tra essi. Per esempio nella versione normalizzata della tabella 3.1, in merito alla colonna denominata "Sentences", ogni valore sarà diviso per 29.5 (v. tabella 2).

Sentences	Tokens	Chars	mCxT	CPOS_A(%)	...
0.1	0.35	0.32	0.92	0.38	...
0.27	0.71	0.64	0.84	0.39	...
0.44	0.97	0.91	0.92	0.56	...
0.61	1.0	1.0	1.0	1.0	...
0.8	0.79	0.76	0.95	0.83	...
1.0	0.7	0.62	0.88	0.7	...

Tabella 2: versione normalizzata della tabella 1.

La stessa operazione eseguita per la tabelle non normalizzate (creazione della tabella definitiva) viene eseguita anche in questo caso.

2.3.5. Tabelle per singola feature

Ogni feature estratta è stata considerata individualmente e presenta la rispettiva tabella in formato .cdv. Ognuna della suddette tabelle presenta un numero di colonne

fascia 1	fascia 2	fascia 3	fascia 4	fascia 5	fascia 6
2.0	12.0	67.0	38.0	39.5	28.5
5.5	31.5	53.0	13.5	26.0	15.0
20.67	76.33	29.67	45.33	51.0	25.67
8.5	40.5	63.0	16.0	46.0	25.0
16.0	26.0	13.0	15.0	25.5	14.75
...	...	...	...	...	...

Tabella 3: tabella per i Token del genere "giornalistico".

[2] Modulo che fornisce l'accesso alle funzioni matematiche definite dallo standard C.

pari al numero di fasce, ovvero 6, e un numero di righe pari al numero di documenti monitorati (v. Tabella 3).

In questo modo ogni genere presenta, oltre alle rispettive tabelle per documento e alle tabelle definitive ottenute mediante esse, 89 tabelle corrispondenti alle feature estratte durante il monitoraggio.

Riassumendo avremmo quindi una tabella definitiva contenente ogni tabella da sei fasce, una tabella definitiva contenente ogni tabella da sei fasce dopo averle sottoposte al processo di normalizzazione e una tabella per ognuna delle feature estratte.

CAPITOLO 3

3. Strumenti di analisi

3.1. Generi testuali

Nel paragrafo 1.2.3 sono stati discussi alcune tipologie di confronto che si possono fare tra testi, ad esempio il genere e la complessità. Le analisi che seguiranno si concentrano esclusivamente sulla distinzione in base al genere. Gli 8 corpus di testi presentati nel paragrafo 2.1 vengono quindi ridotti a 4 in base al genere di appartenenza (giornalistico, scientifico, scolastico e narrativo).

Le tabelle viste nel capitolo 2 vengono a loro volta accoppiate di modo da avere rispettivamente una tabella definitiva e una tabella definitiva normalizzata per ogni genere. Inoltre viene eseguita nuovamente l'operazione di normalizzazione sulla tabella definitiva di modo da poter distinguere i dati normalizzati per i singoli documenti da quelli normalizzati in modo globale.

Abbiamo quindi, per ogni genere, una tabella (definitiva.csv) creata dall'unione di tutte le tabelle ottenute da ogni documento appartenente al suddetto genere, una tabella (definitiva_norm.csv) che contiene tutte le tabelle normalizzate ottenute dalle singoli tabelle per documento, una tabella per ogni feature estratta (per esempio "Tokens.csv") e una tabella (definitiva_norm_TOTALE.csv) che non è altro che la tabella "definitiva.csv" ma sottoposta al processo di normalizzazione.

3.2. Funzioni statistiche

Per l'analisi statistica dei dati sono state utilizzati tre algoritmi di correlazione, denominati rispettivamente Pearson correlation coefficient (3.2.1), Spearman's rank correlation (3.2.2) test e Wilcoxon rank-sum test (3.2.3).

Tutti e tre i test statistici sono stati utilizzati per il calcolo del p-value, necessario per determinare la significatività dei confronti eseguiti.

Per l'applicazione dei tre algoritmi è stata utilizzata la libreria open source "Scipy" (3.3).

3.2.1 Pearson correlation coefficient

In statistica, l'indice di correlazione di Pearson esprime un'eventuale relazione di linearità tra due variabili statistiche.

Date X e Y, l'indice di correlazione di Pearson è definito come la loro covarianza diviso per il prodotto delle deviazioni standard delle due variabili:

$$\rho xy = \frac{\sigma xy}{\sigma x \sigma y}$$

dove σxy è la covarianza tra X e Y e σx e σy sono le due deviazioni standard.

Il coefficiente assume sempre valori compresi tra -1 e 1 che ne definiscono il tipo. Vi sono infatti tre tipologie di correlazione:

- Se $\rho xy > 0$, le variabili X e Y si dicono correlate positivamente, o direttamente correlate;
- Se $\rho xy = 0$, le variabili X e Y si dicono incorrelate;
- Se $\rho xy < 0$, le variabili X e Y si dicono correlate negativamente, o inversamente correlate.

La correlazione diretta (e analogamente quella inversa) si distingue a sua volta in tre categorie:

- Se $0 < \rho xy < 0{,}3$ se ha correlazione debole;
- Se $0{,}3 < \rho xy < 0{,}7$ se ha correlazione moderata;
- Se $0{,}7 < \rho xy < 1$ se ha correlazione forte;

3.2.2 Spearman rank-correlation coefficient

Il coefficiente di correlazione per ranghi di Spearman è un caso particolare del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente. Inoltre esso non misura necessariamente una relazione lineare ma anche qualora vengano usate misure intervallari.

- I valori di X e i valori di Y vengono rispettivamente ordinati da 1 a n e vengono assegnati ad essi i ranghi, tenendo conto anche dei valori uguali (in questi casi viene considerata la media del rango);
- Per ogni coppia si calcola la differenza d_i del rango di Y - il rango di X, elevandola poi al quadrato;
- Si calcola la somma del d_i al quadrato $\sum d_i^2$

La formula utilizzata è la seguente:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

3.2.3 Wilcoxon rank sum test

Il Wilcoxon rank-sum test è un test statistico non parametrico per due campioni indipendenti. Verifica, tramite la presenza di valori provenienti da una distribuzione continua, se i due campioni provengono da una stessa popolazione.

Il test comporta il calcolo della statistica comunemente nota come U, che ha una distribuzione nota sotto l'ipotesi nulla[3]. Una statistica a essa equivalente è quella della somma dei ranghi.

Tutte le osservazioni vengono disposte in una singola serie di rango, indifferentemente dal campione in cui esse si trovino. I ranghi provenienti dal primo campione vengono sommati.

Di conseguenza, la somma di tutti i ranghi vale:

$$R_1 + R_2 = \frac{N(N+1)}{2}$$

dove "N" è il numero totale delle osservazioni.

La statistica U è invece fornita da:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

dove n1 è la dimensione del primo campione e R1 è la somma dei ranghi del primo campione.

La stessa formula è applicabile anche per il secondo campione:

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

[3] Generalmente l'ipotesi nulla è indicata con H_0. Nel caso in cui non venga accettata viene sostituita dall'alternativa H_1.

Sommando U1 e U2 otteniamo:

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}$$

Sapendo che $R_1 + R_2 = N(N+1)/2$, si ha che la somma corrisponde a:

$$U_1 + U_2 = \frac{N(N+1)}{2} - \frac{n_1(n_1+1)}{2} - \frac{n_2(n_2+1)}{2}$$

$$U_1 + U_2 = \frac{N^2+N}{2} - \frac{n_1(n_1+1)}{2} - \frac{n_2(n_2+1)}{2}$$

Inoltre, sapendo che $N=n_1+n_2$:

$$U_1 + U_2 = \frac{n_1^2 + n_2^2 + 2n_1n_2}{2} + \frac{n_1+n_2}{2} - \frac{n_1^2+n_2^2}{2} - \frac{n_1+n_2}{2}$$

Semplificando si ottiene la formula:

$$U_1 + U_2 = n_1n_2$$

Il valore massimo di U è quindi il prodotto delle dimensioni dei due campioni.

3.2.4. P-value e significatività

Il p-value di un test statistico indica la probabilità di ottenere un risultato uguale o più estremo rispetto a quello osservato, supposta vera l'ipotesi nulla. Spesso viene definito livello di significatività osservato.

Supponiamo di avere due campioni A e B contenenti rispettivamente n_a e n_b informazioni. Voglio determinare se la distribuzione di X valori sia la stessa in A e in B. Per i testi unilaterali l'ipotesi nulla H_0 vale che A = B. Le altre due possibilità sono H_1: A > B e H2: A < B. Se A > B è vera, allora il $p - value = pr(X \geq R1)$ dove X è una variabile aleatori e R1 è la somma dei ranghi del primo campione. Se invece è vera A < B, allora il $p - value = pr(X \leq R1)$.

Per i test bilaterali l'ipotesi nulla H_0 vale che A = B e la sua alternativa è H_1: A ≠ B. In questo secondo caso viene calcolata la probabilità che i valori finiscano più spostati verso destra o verso sinistra e viene raddoppiato il valore precedentemente calcolato per i test unilaterali. Da ciò $p - value = 2pr(X \geq R1)$ o $p - value = 2pr(X \leq R1)$.

Se il p-value è molto piccolo significa che la significatività è alta. Si distinguono quattro categorie:

- P-value ≥ 0.05, non c'è significatività;
- P-value < 0.05, c'è significatività statistica;
- P-value < 0.01, variazione molto significativa;
- P-value < 0.001, variazione estremamente significativa.

3.3. Scipy

Scipy è una libreria open source di algoritmi e strumenti matematici per il linguaggio di programmazione Python. Tra i vari pacchetti che essa offre vi è il modulo "stats" contenente numerose funzioni statistiche, tra cui "pearsonr", "spearmanr" e "ranksums". Di seguito la sintassi per ognuna delle funzioni proposte[4], dove *x* e *y* sono due variabili statistiche:

- `stats.pearsonr(x,y)`, restituisce il coefficiente di correlazione di Pearson e il p-value.
- `stats.spearmanr(x,y)`, restituisce il coefficiente di correlazione per ranghi di Spearman e il p-value.
- `stats.ranksums(x,y)`, restituisce la statistica della somma dei ranghi e il p-value.

Pearson e Spearman vengono utilizzati per determinare se due variabili statistiche variano secondo la medesima funzione. Ciò dipende dal valore assunto dal coefficiente di correlazione. Nei casi trattati in seguito il valore minimo che può essere assunto dal coefficiente di correlazione è 0,3; tutti i casi con un valore inferiore non vengono considerati.

Wilcoxon indica invece se esiste una variazione significativa tra le variabili prese in esame.

[4] Necessaria l'installazione della libreria scipy e l'importazione del modulo stats.

CAPITOLO 4

4. Gli esperimenti

In questo capitolo saranno illustrati 4 differenti studi eseguiti su corpora appartenenti a generi testuali differenti. Le tabelle utilizzate sono quelle presentate nel secondo capitolo (v. Paragrafo 2.3.4.), che verranno sottoposte ad analisi statistiche tramite l'utilizzo delle funzioni descritte nel capitolo 3.

4.1. Primo esperimento

Il primo studio presentato nella seguente relazione descrive un primo confronto tra generi. Per ciascun corpus viene utilizzata la tabella "definitiva" non sottoposta al processo di normalizzazione. Abbiamo quindi un file .csv per ogni genere analizzato (giornalistico, scientifico, narrativo e didattico) che ci porterà a un totale di 6 confronti. L'obbiettivo primario è stabilire quali feature sono particolarmente significative in un genere rispetto all'altro. Per farlo utilizzo il Wilcoxon rank sum test e il calcolo del relativo p-value come visto nei paragrafi 3.2.3. e 3.2.4.

4.1.1. Analisi statistica dei confronti fra generi

Come abbiamo visto ogni tabella definitiva presenta un numero di colonne corrispondenti al numero di feature estratte e un numero di righe corrispondenti al numero di fasce analizzate. Ogni colonna di un genere ha quindi una corrispettiva controparte per tutti i generi considerati. Per esempio, la colonna "Tokens" (lunghezza media della frase) della tabella "definitiva.csv" nel genere giornalistico può essere confrontata con la colonna "Tokens" della tabella "definitiva.csv" per il genere scientifico (v. figura 1).

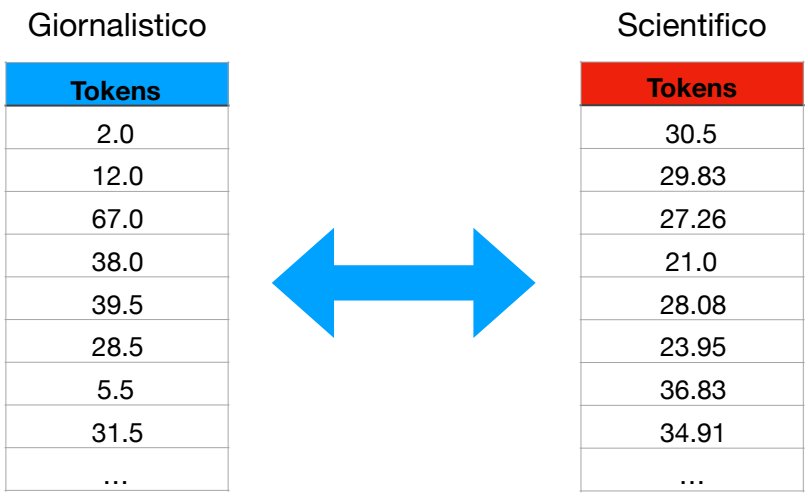

Figura 1: Confronto fra il genere Giornalistico e il genere Scientifico in merito alla feature "Tokens"

Così facendo, per ciascuna coppia di corpora analizzati otterremo 89 confronti, corrispondenti alle feature estratte.

Ciascuna colonna viene quindi confrontata con la corrispettiva in un altro genere grazie al Wilcoxon rank sum test. Ciò che otteniamo è quindi il p-value di tale confronto che ci permette di determinare l'eventuale presenza di una variazione statisticamente significativa per la feature considerata (v. tabella 1).

features	pvalue
Tokens	7.499523534094972e-12

Tabella 1: risultato del test di Wilcoxon del confronto illustrato in figura 1

Così facendo ciascuno confronto risulta caratterizzato da una tabella di 89 righe (una per feature) che vengono inoltre ordinate in base al p-value, di modo che le prime righe della tabelle ottenuta siano quelle degne di nota e interessanti per il confronto eseguito.

4.1.2. Confronti fra generi

Come detto nel paragrafo precedente, i confronti possibili fra 4 generi testuali diversi sono 6 e corrispondono a Giornalistico vs Scientifico, Giornalistico vs Narrativo, Giornalistico vs Didattico, Didattico vs Scientifico, Narrativo vs Didattico, e Narrativo vs Scientifico. Per ciascuno di essi vengono analizzate le feature estratte durante il monitoraggio, suddivise in base alla categoria linguistica di appartenenza. Tra questa feature vengono considerate idonee per l'analisi solo quelle in formato percentuale o risultanti di un'operazione di media aritmetica. Le tabelle descritte nelle righe successive presentano in colonna i 6 confronti presi in esame, mentre le righe sono occupate dalla feature. In merito alla simbologia, per descrivere l'eventuale significatività statistica, in questo e negli studi successivi, vengono utilizzati 4 simboli corrispondenti a 4 livelli di significatività, come visto nel paragrafo dedicato al p-value (v. paragrafo 3.2.4.). Al simbolo ❌ si associa l'assenza di significatività, a ✓ corrisponde una variazione statisticamente significativa, il simbolo ✔ indica una variazione molto significativa, mentre ✅ è indice di una variazione estremamente significativa.

Feature	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Aggettivi	✅	✅	✅	✅	✅	✅
Avverbi	✅	✅	✅	✅	✅	✅
Congiunzioni	✅	✅	✅	✅	✅	✅
Determinanti	✅	✅	❌	✅	✅	✅
Preposizioni	❌	✅	✅	✅	✅	✅
Numeri	✔	✅	✓	✅	✓	✅
Pronomi	✅	✅	✅	✅	✅	✅
Articoli	✅	✅	✔	✅	✅	✅
Nomi	✅	✅	✅	✅	✅	✅
Verbi	✅	✅	✅	✅	✅	✅

Tabella 3: calcolo del p-value nel confronto tra corpora di diverso genere per singola part of speech (POS).

Nelle tabelle 3 e 4, vengono prese in esame le part of speech (POS)[5] analizzate in merito al confronto tra generi.

Feature	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Congiunzioni coordinanti	✅	✅	✅	✅	✅	✅
Congiunzioni subordinanti	✅	✔️	✓	❌	❌	✓
Pronomi dimostrativi	✅	✅	❌	✅	✅	✅
Pronomi personali	❌	✅	✅	❌	❌	❌
Pronomi indefiniti	✅	❌	❌	❌	❌	❌
Pronomi possessivi	❌	❌	❌	❌	❌	❌
Pronomi interrogativi	✔️	❌	❌	✅	❌	✔️
Pronomi relativi	✅	✅	✔️	❌	✅	✅
Pronomi clitici	✅	✅	✔️	✅	✅	✔️
Articoli determinativi	✅	✅	❌	✅	✅	✅
Articoli indeterminativi	❌	✅	❌	❌	✅	✅
Abbreviazioni	✅	❌	❌	✅	❌	✅
Sostantivi propri	✅	✅	✅	✅	✅	❌
Verbi ausiliari	✅	✅	✅	❌	✅	✅
Verbi modali	✅	❌	❌	✅	❌	✅

Tabella 4: calcolo del p-value nel confronto tra corpora di diverso genere per singola part of speech (POS).

Appare subito chiaro il numero piuttosto considerevole di variazioni estremamente significative in merito alle feature di natura morfo-sintattica. Ci sono comunque alcuni casi da segnalare come per esempio l'assenza di variazione significativa per le preposizioni tra i testi giornalistici e i testi scientifici e per i

[5] Selezionate secondo l'ISST-TANL morpho-syntactic tagset

determinanti fra i testi giornalisti e i testi didattici, oppure la variazione molto significativa degli articoli tra i testi giornalistici e i testi didattici.

In merito alle caratteristiche appartenenti alle sottocategorie morfo-sintattiche le considerazioni da fare aumentano. Ciò che salta subito all'occhio è come la variazione relativa alle congiunzioni coordinanti sia estremamente significativa per tutti i confronti presi in esame, mentre i pronomi possessivi non presentano alcun tipo di variazione statisticamente significativa, in nessuno dei casi analizzati.

Le congiunzioni subordinanti presentano una variazione statisticamente significativa sia nel confronto fra testi giornalistici e didattici, sia in quello fra testi narrativi e scientifici.

Vi sono poi diverse variazioni molto significative come per esempio il caso dei pronomi interrogativi nei confronti Giornalistico vs Scientifico e Narrativo vs Scientifico o dei pronomi relativi nel confronto Giornalistico vs Didattico.

Per concludere, in merito agli articoli, osserviamo che i determinativi presentano una variazione estremamente significativa in gran parte dei casi considerati, mentre quelli indeterminativi presentano tre casi caratterizzanti e tre che non lo sono.

Feature	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
arg	✅	✅	❌	✅	✅	✔️
aux	✅	✅	✅	❌	✅	✅
clit	✓	✓	✅	✅	✅	❌
comp	✅	✅	✅	✅	✅	✅
comp_ind	✔️	✅	❌	✓	✅	✅
comp_loc	✅	❌	✅	✅	✅	✅
comp_temp	✅	✅	✅	✅	❌	✅
con	❌	✅	✅	✅	✅	✅

Tabella 5: calcolo del p-value nel confronto tra corpora di diverso genere per singola dipendenza.

concat	✅	❌	❌	✅	❌	✅
conj	❌	✅	✅	✅	✅	✅
det	✅	✅	✔	✅	✅	✅
dis	✅	✔	✓	✅	✅	✅
disj	✅	❌	✓	✅	✅	✅
mod	✅	✅	❌	✅	✅	✅
mod_loc	✔	✓	❌	✅	❌	✓
mod_rel	❌	✅	✅	✅	✅	✅
mod_temp	✅	❌	✅	✓	✅	✅
modal	✅	❌	❌	✅	❌	✅
neg	✅	❌	❌	✔	✔	✅
obj	✅	✅	❌	✅	✅	✅
pred	✓	✅	✅	✅	✅	✅
prep	✓	✅	✅	✅	✅	✅
punc	❌	✅	✅	❌	✅	✅
ROOT	✔	✅	✅	✅	✅	✅
sub	✅	✅	❌	✅	✅	✅
subj	✅	✅	✅	✅	✅	✅
subj_pass	✅	✅	❌	✅	✅	✅

Tabella 5: calcolo del p-value nel confronto tra corpora di diverso genere per singola dipendenza.

Vediamo ora che risultati (v. tabella 5) si ottengono nell'analisi delle dipendenze sintattiche (dependency[6]).

Tra i casi più interessanti segnaliamo la dipendenza denominata "clit", che lega un pronome clitico al verbo, indicativo della presenza di strutture impersonali, e che presenta una variazione statisticamente significativa nei confronti fra testi giornalisti e scientifici e fra testi giornalistici e narrativi, mentre nei casi Didattico vs Scientifico e Narrativo vs Didattico risulta caratterizzata da una variazione estremamente

[6] Selezionate secondo l'ISST-TANL dependency tagset

significativa. Osservazioni simili possono essere fatte anche in merito ai casi "comp_ind" e "mod_temp" in cui per la maggior parte dei confronti risulta esserci una variazione significativa.

Casi con "concat", "mod_loc" e "modal" presentano invece diversi confronti in cui la feature presa in esame non è caratterizzante in quel caso specifico.

Da segnalare infine i casi delle dipendenze soggetto e complemento ("subj" e "comp") che presentano una variazione estremamente significativa in tutti i casi presi in esame.

Feature	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Tokens pre_testa sintattica	✅	✅	❌	✅	✅	✅
Tokens post_testa sintattica	✅	✅	❌	✅	✅	✅
Soggetti pre_testa sintattica	❌	✅	❌	✅	✅	✅
Soggetti post_testa sintattica	✅	✅	✓	❌	✅	✅
Oggetti pre_testa sintattica	✅	✓	❌	❌	✓	✓
Oggetti post_testa sintattica	✅	✅	❌	❌	✅	✅
Aggettivi pre_testa sintattica	✅	✅	❌	✔	✅	✅
Aggettivi post_testa sintattica	✅	✅	✅	✔	✅	✅
Avverbi pre_testa sintattica	✅	✅	✔	✅	✅	✅
Avverbi post_testa sintattica	✅	✅	✓	❌	✅	✅

Tabella 6: calcolo del p-value nel confronto tra corpora di diverso genere per caratteristiche sintattiche.

Nella tabella 6 sono descritti i casi relativi alla posizione di un token rispetto alla testa sintattica da cui dipende (v. paragrafo 2.3.3.). Sono stati considerati sia tutti i tipi di dipendenza sia i casi specifici che interessano i soggetti pre e postverbali, gli oggetti pre e postverbali, gli aggettivi pre e postnominali, e gli avverbi pre e postverbali.

Tra i confronti più interessanti vi sono quello tra i testi giornalistici e i testi didattici e quello tra i testi didattici e i testi scientifici. In questi casi, alcune delle feature statisticamente significative per gli altri confronti non risultano in alcun modo caratterizzanti, come per esempio gli oggetti in entrambi i casi, i tokens per quanto riguarda il primo dei due confronti e i soggetti, rispettivamente nel caso pre_testa sintattica per il primo e post test_sintattica per il secondo.

Gli oggetti preverbali sono caratterizzati da una variazione estremamente significativa solo nel primo dei sei casi presi in esame, mentre gli aggettivi postnominali e gli avverbi preverbali risultano caratterizzanti in tutti i confronti eseguiti.

Oltre al caso dell'ordine sintattico degli elementi, nel capitolo 2, è stato introdotto il fenomeno delle subordinate (come caratteristica monitorata), con particolare attenzione alla distinzione fra quelle di primo grado e quelle di grado superiore al primo (v. paragrafo 2.3.2.).

Feature	**Giornalistico vs Scientifico**	**Giornalistico vs Narrativo**	**Giornalistico vs Didattico**	**Didattico Vs Scientifico**	**Narrativo Vs Didattico**	**Narrativo Vs Scientifico**
Altezza dell'albero sintattico	✅	✅	❌	✅	✅	✅
Ampiezza dell'albero sintattico	✅	✅	✅	❌	✅	✅
Numero medio di figli per token	✅	✅	✅	✅	✓	✅
Numero medio di figli per sostantivo	✅	✅	✅	❌	✅	✅
Numero medio di figli per verbo	✅	✅	✅	✅	✅	✅

Tabella 7: calcolo del p-value nel confronto tra corpora di diverso genere per caratteristiche sintattica.

Media delle altezze degli alberi delle subordinate	✔	✅	❌	✔	✅	✅
Media delle ampiezze degli alberi delle subordinate	❌	✅	✅	❌	✅	✅
Subordinate di primo grado	❌	✅	❌	✔	✅	✅
Media delle altezze degli alberi delle subordinate di primo grado	✔	✅	❌	✅	✅	✅
Media delle ampiezze degli alberi delle subordinate di primo grado	✓	✅	✅	❌	✅	✅
Subordinate di primo grado che precedono la principale	❌	❌	❌	✅	❌	✔
Subordinate di primo grado che seguono la principale	❌	❌	✓	✔	✅	✅
Subordinate di grado superiore al primo	❌	✅	❌	✔	✅	✅
Media delle altezze degli alberi delle subordinate di grado superiore al primo	❌	✅	❌	✅	✅	✅
Media delle ampiezze degli alberi delle subordinate di grado superiore al primo	❌	✅	❌	✅	✅	✅

Tabella 7: calcolo del p-value nel confronto tra corpora di diverso genere per caratteristiche sintattica.

Subordinate di grado superiore al primo che precedono la principale	❌	❌	❌	✔️	❌	✓
Subordinate di grado superiore al primo che seguono la principale	❌	✅	❌	❌	✅	✅

Tabella 7: calcolo del p-value nel confronto tra corpora di diverso genere per caratteristiche sintattica.

Le considerazioni relative alla suddetta feature si rifanno naturalmente alle caratteristiche legate all'albero sintattico, che vengono prese in esame nella tabella 7.

Possiamo subito notare come l'ultimo dei confronti presi in esame, ovvero quello fra testi narrativi e testi scientifici, è caratterizzato dalla presenza di variazione significativa per tutte le feature considerate. Una situazione simile la si ritrova anche nel confronto fra testi narrativi e testi didattici, ad eccezione delle subordinate di primo grado o di grado superiore al primo che precedono la principale, parametri la cui variazione non è in alcun modo caratterizzante nel confronto analizzato.

Per quanto riguarda le feature, troviamo che le informazioni relative all'altezza e all'ampiezza dell'albero sintattico sono caratterizzanti per la maggior parte dei confronti, ad eccezione dei casi Giornalistico vs Didattico e Didattico vs Scientifico, il primo per quanto riguarda l'altezza mentre il secondo in merito all'ampiezza. Anche il numero medio di figli presenta una variazione in quasi tutti i casi considerati, eccetto per il confronto fra il genere didattico e il genere scientifico nel caso del numero medio di figli per sostantivi.

Lo studio delle subordinate, sia di primo grado che di grado di superiore al primo, e di tutto ciò che risulta ad esso legato, presenta meno feature che hanno una variazione statisticamente significativa. I casi che vale la pena segnalare sono l'altezza media degli alberi delle subordinate di primo grado, che presenta una variazione molto significativa per il confronto Giornalistico vs Scientifico, e una variazione estremamente significativa nei casi Giornalistico vs Narrativo, Didattico vs Scientifico, Narrativo vs Didattico, Narrativo vs Scientifico, e la media dell'ampiezza degli alberi delle subordinate di primo grado, feature che presenta una variazione statisticamente significativa per il confronto Giornalistico vs Scientifico, e

una variazione estremamente significativa nei casi Giornalistico vs Narrativo, Giornalistico vs Didattico, Narrativo vs Didattico e Narrativo vs Scientifico.

4.2. Secondo esperimento

La seconda analisi presentata in questa relazione consiste nel determinare le feature significative nei confronti tra generi per ognuna delle 6 fasce analizzate. Per fare ciò, sono state create 89 tabelle (una per feature) a partire dalla tabella di 6 fasce per singolo documento (v. paragrafo 2.3.4.). Ogni genere presenta quindi un file in formato .csv per ciascuna delle feature estratte, in cui le colonne rappresentano le fasce, mentre le righe rappresentano i documenti (v. tabella 8).

fascia 1	fascia 2	fascia 3	fascia 4	fascia 5	fascia 6
0.0	0.0	0.0	100.0	50.0	25.0
50.0	100.0	0.0	100.0	0.0	100.0
22.22	60.0	33.33	66.67	33.33	33.33
50.0	50.0	50.0	0.0	75.0	100.0
33.33	58.57	40.0	75.0	66.67	50.0
0.0	0.0	66.67	0.0	50.0	0.0
50.0	75.0	56.67	50.0	73.33	70.0
…	…	…	…	…	…

Tabella 8: estratto relativo alla percentuale degli aggettivi post nominali nei testi giornalistici in base alle fasce.

Per ottenere i file sopra descritti ho estratto da ciascuna delle tabelle in 6 fasce, per singolo genere, i valori della colonna (feature) ricercata e li ho inseriti come nuova riga in un DataFrame creato ad hoc, che è stato poi trasformato in un file in formato .csv. Questa operazione è stata eseguita per ciascuno dei generi di interesse. Di seguito la funzione in Python per ottenere le tabelle per singola feature:

```
for i in range(0,89):
    df1=pd.DataFrame()
    for file in sorted_aphanumeric(os.listdir(os.getcwd())):
          if file.endswith("table.csv"):
                      df=pd.read_csv(file, sep="\t")
                      colonna=df.ix[:,i]
                      df1=df1.append(colonna)
```

In seguito, ciascuna tabella così ottenuta è stata confrontata con le corrispondenti negli altri generi tramite il Wilcoxon rank sum test e il conseguente calcolo del p-

value. Per ciascuno confronto (6 totali), vi sono 89 tabelle corrispondenti alle feature, con in colonna il p-value calcolato tramite Wilcoxon e in riga le 6 fasce considerate (v. tabella 9).

adjPost	pvalue
fascia 1	✅
fascia 2	✅
fascia 3	✅
fascia 4	✅
fascia 5	✅
fascia 6	✅

Tabella 9: p-value calcolato sugli aggettivi post nominali nel confronto Giornalistico vs Narrativo in base alle fasce.

A partire dai file così ottenuti è quindi possibile eseguire un'indagine sui dati volta a determinare, per ciascuno dei confronti eseguiti, quali feature siano rispettivamente più significative, e quindi caratterizzanti, per ognuna delle fasce considerate. Per esempio, nella tabella 9 il p-value indica una variazione estremamente significativa per tutte le fasce prese in esame. Questo dato ci dice che nel confronto fra testi giornalisti e testi narrativi la variazione tra gli aggettivi post nominali è caratterizzante in ognuna delle fasce in cui i vari documenti sono stati divisi.

4.2.1. Confronto di genere per singola feature in base alle fasce

Come descritto nel paragrafo precedente i possibili confronti tra ciascun corpus sono 89, ovvero il numero delle feature monitorate per ciascuno dei generi presi in esame. In questo caso però, rispetto a quanto visto in altre analisi presenti in questa relazione, verrano presentate esclusivamente le feature di natura morfo-sintattica e sintattica in quanto le caratteristiche di base sono risultate estremamente significative in tutte e 6 le fasce analizzate, per ciascuno dei 6 confronti descritti. Sono state considerate 5 feature di natura morfo-sintattica e 5 feature di natura sintattica.

Caratteristiche morfo-sintattiche

Tra le caratteristiche morfo-sintattiche sono state selezionate la percentuale di nomi propri, di verbi, di congiunzioni, di pronomi dimostrativi e di articoli determinativi.

- **Sostantivi Propri (POS_SP)**

POS_P	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	✅	✅	✅	✓	❌
Fascia 2	✅	✅	✅	✔	✓	❌
Fascia 3	✅	✅	✅	✔	✓	❌
Fascia 4	✅	✅	✅	✅	✓	❌
Fascia 5	✅	✅	✅	✅	❌	❌
Fascia 6	✅	✅	✅	✅	✓	❌

Tabella 10: tabella rappresentante la variazione significativa dei sostantivi propri nelle fasce per ciascuno dei confronti realizzati.

Possiamo notare che nei confronti Giornalistico vs Narrativo, Giornalistico vs Scientifico e Giornalistico vs Didattico i sostantivi propri presentano una variazione estremamente significativa in ciascuna delle 6 fasce analizzate. Una situazione differente è invece quella rappresentata dai tre casi successivi, in quanto nel caso Didattico vs Scientifico le fasce 2 e 3 presentano una variazione molto significativa mentre nel caso Narrativo vs Didattico le variazioni sono statisticamente significative in tutte le fasce eccetto nella quinta che invece non presenta alcun tipo di variazione. Per finire, l'ultimo caso (Narrativo vs Scientifico) non presenta una variazione significativa in nessuna delle 6 fasce considerate.

Tra le possibili considerazioni in merito al seguente confronto vi è senz'altro la variazione estremamente significativa che i sostantivi propri hanno in tutte le fasce dei confronti riguardanti il genere giornalistico. Oppure possiamo sottolineare come negli ultimi due confronti, entrambi comprendenti il genere narrativo, i casi di variazioni siano tutti statisticamente significativi o addirittura assenti.

- **Verbi (CPOS_V)**

CPOS_V	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	✅	✅	✅	✅	❌
Fascia 2	✅	✅	❌	✅	✅	❌
Fascia 3	✅	✅	❌	✅	✅	✔️
Fascia 4	✅	✅	❌	✅	✅	✓
Fascia 5	✅	✅	✓	✅	✅	✓
Fascia 6	✅	✅	❌	✅	✅	✔️

Tabella 11: tabella rappresentante la variazione significativa dei verbi nelle fasce per ciascuno dei confronti realizzati.

In merito alla percentuale dei verbi osserviamo che quattro casi su sei presentano una variazione estremamente significativa in ciascuna delle fasce considerate. I casi particolari, in questo caso, sono i confronto fra testi giornalistici e didattici e fra testi narrativi e scientifici. Nel primo caso troviamo una variazione estremamente significativa nella prima fascia, corrispondente alle prime righe del testo, e una variazione statisticamente significativa nella quinta fascia. Nel secondo caso le fasce 3 e 6 presentano una variazione estremamente significativa, mentre le fasce 4 e 5 presentano una variazione statisticamente significativa. È quindi lecito pensare che i testi giornalistici non siano troppo diversi nell'uso dei verbi rispetto ai testi didattici, se non per la prime righe di testo, e che tra i testi narrativi e i testi scientifici le variazioni non siano evidenti come nei confronti Giornalistico vs Scientifico e Didattico vs Scientifico.

- **Congiunzioni (CPOS_C)**

La terza feature descritta in questa analisi è la percentuale di congiunzioni presenti nel testo. Come possiamo notare nella tabella 12, i confronti che presentano il maggior numero di casi di variazione estremamente significativa sono quelli fra testi didattici e scientifici e quelli fra i testi narrativi e didattici. Nello specifico nel primo caso abbiamo una variazione estremamente significativa in ciascuna delle fasce considerate, eccetto la fascia 3 che presenta invece una variazione molto significativa. Il secondo caso (Narrativo vs Didattico) vede invece le congiunzioni

come feature caratterizzante per tutte le fasce considerate. È interessante notare come entrambi i confronti presentino i materiale didattici come uno dei due genere controllati.

CPOS_C	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	❌	❌	✅	✅	✅	✓
Fascia 2	❌	✔	✅	✅	✅	✅
Fascia 3	❌	✔	✓	✔	✅	✅
Fascia 4	❌	✔	✅	✅	✅	✓
Fascia 5	✔	✔	✔	✅	✅	✔
Fascia 6	✔	✅	✔	✅	✅	✔

Tabella 12: tabella rappresentante la variazione significativa delle congiunzioni nelle fasce per ciascuno dei confronti realizzati.

Gli altri casi sono piuttosto singolari in quanto presentano delle differenze notevoli gli uni dagli altri. In primis, il confronto fra testi giornalistici e testi scientifici si limita ad una variazione molto significativa nelle fasce 5 e 6, ovvero nella terza parte del documento. Gli altri due casi riguardanti i testi giornalistici (Giornalistico vs Narrativo e Giornalistico vs Didattico) presentano rispettivamente casi differenti di variazione significativa. Il primo per esempio vede le congiunzioni come un aspetto caratterizzante delle fascia 6 ma totalmente privo di rilievo nella fascia 1; le fasce 2, 3, 4, 5 presentano una variazione molto significativa. Il secondo caso vede invece presenti tutti e tre i gradi di significatività, proprio come l'ultimo dei confronti non ancora discusso (Narrativo vs Scientifico). In entrambi i casi abbiamo infatti una variazione statisticamente significativa, fascia 3 nel primo e fasce 1 e 4 nel secondo, casi di variazione molto significativa, fasce 5 e 6 in entrambi i casi, e variazioni estremamente significative, fasce 1, 2 e 4 nel primo e fasce 2 e 3 nel secondo.

- **Pronomi Dimostrativi (POS_PD)**

In merito alla percentuale dei pronomi dimostrativi nel testo troviamo dei casi piuttosto differenti l'uno dall'altro. Procedendo in ordine di confronti osserviamo che nel primo caso le fasce 2, 3, 4 e 5 presentano una variazione estremamente significativa, mentre la fascia 1 la fascia 6 sono rispettivamente statisticamente significative e molto significative. I due casi successivi, che coinvolgono entrambi i testi giornalistici, non hanno quasi niente da segnalare se non una variazione estremamente significativa nella prima fascia del primo confronto. Possiamo quindi concludere che i pronomi dimostrativi risultano caratterizzanti nel confronto Giornalistico vs Sscientifico nelle prime righe del testo.

POS_D	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	✓	❌	✔	✅	✅
Fascia 2	❌	✅	❌	✓	✔	✅
Fascia 3	❌	✅	❌	✔	✔	✅
Fascia 4	❌	✅	❌	❌	✅	✅
Fascia 5	❌	✅	❌	❌	✅	✅
Fascia 6	❌	✔	❌	✅	❌	✅

Tabella 13: tabella rappresentante la variazione significativa dei pronomi dimostrativi nelle fasce per ciascuno dei confronti realizzati.

Il quarto confronto non presenta alcun caso di variazione estremamente significativa, e inoltre nelle fasce 4 e 5 i pronomi dimostrativi non sono in alcun modo caratteristici. Le altre 4 fasce sono rispettivamente molto significative (fascia 1, 3 e 6) e statisticamente significativi (fascia 2).

I due casi conclusivi sono invece caratterizzati da numerosi casi di variazione estremamente significativa, eccetto la fascia 1 la fascia 2 nel confronto fra testi narrativi e materiali didattici.

- **Articoli Determinativi (POS_RD)**

Gli articoli determinativi in tre dei confronti considerati (Giornalistico vs Narrativo, Narrativo vs Didattico, Narrativo vs Scientifico) presentano una variazione estremamente significativa in tutte le fasce considerate. È interessante notare come tutti e tre i suddetti confronti siano quelli comprendenti il genere narrativo.

POS_RD	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	✅	❌	✅	✅	✅
Fascia 2	✅	✅	❌	✔️	✅	✅
Fascia 3	❌	✅	❌	✔️	✅	✅
Fascia 4	✓	✅	❌	✓	✅	✅
Fascia 5	❌	✅	❌	❌	✅	✅
Fascia 6	❌	✅	❌	❌	✅	✅

Tabella 14: tabella rappresentante la variazione significativa degli articoli determinativi nelle fasce per ciascuno dei confronti realizzati.

Gli altri tre casi considerati descrivono situazioni singolari. Il confronto tra i testi giornalistici e i testi scientifici individua una variazione estremamente significativa nelle fasce 1 e 2, mentre nella fascia 4 vi è una variazione statisticamente significativa. Le altre fasce considerate non presentano invece alcun tipo di variazione significativa. Nel confronto fra i testi giornalistici e i materiali didattici non vi è invece niente da segnalare. Per concludere il confronto fra genere didattico e genere scientifico presenta tutti i tre casi di variazione significativa presi in esame: la fascia 1 presenta infatti una variazione estremamente significativa, le fasce 2 e 3 sono caratterizzate da variazioni molto significativa, e la fascia 4 presenta una variazione statisticamente significativa. Le fasce 5 e 6 non sono invece in alcun modo caratterizzanti per quanto concerne agli articoli determinativi. L'aspetto interessante di quest'ultimo caso considerato è come una feature diventi mano a mano sempre meno caratteristica durante l'analisi di un documento. Abbiamo infatti una variazione che tende a essere sempre meno caratterizzante fino a scomparire del tutto.

Caratteristiche sintattiche

Tra caratteristiche sintattiche sono state selezionati le dipendenze DEP_aux, DEP_obj e DEP_subj, gli aggettivi precedenti la testa sintattica e le subordinate di grado superiore al primo.

- **DEP_aux**

  La dipendenza "Aux" descrive la relazione tra un verbo e il suo ausiliare.

DEP_aux	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✓	❌	❌	❌	✔	❌
Fascia 2	✅	✅	✔	❌	✓	❌
Fascia 3	❌	✅	✅	✔	❌	❌
Fascia 4	✓	✅	✅	✓	✔	✅
Fascia 5	❌	✅	✓	❌	✅	✅
Fascia 6	❌	✅	✅	❌	❌	✅

Tabella 15: tabella rappresentante la variazione significativa della dipendenza "Aux" nelle fasce per ciascuno dei confronti realizzati.

Nei 6 confronti considerati, in merito alla dipendenza "Aux", si evidenzia uno scenario piuttosto differente (v. Tabella 15). Nel caso Giornalistico vs Narrativo troviamo una variazione significativa per tutte le fasce considerate ad eccezione delle prima. Il confronto fra testi giornalisti e testi narrativi sottolinea come l'uso delle dipendenza utilizzata sia caratteristica per il confronto, in gran parte dei casi considerati. Il secondo confronto (Giornalistico vs Scientifico) presenta una variazione statisticamente significativa nella prima e nella quarta fascia, mentre la fascia 2 è caratterizzata da una variazione estremamente significativa. Il confronto fra testi giornalistici e testi didattici presenta una variazione estremamente significativa in tre delle fasce considerate (3, 4 e 6), una variazione statisticamente significativa in fascia 5 e una variazione molto significativa in fascia 2. Il quarto caso considerato (Didattico vs Scientifico) non offre molto da segnalare, se non una variazione molto significativa in fascia 3, e una variazione statisticamente significativa in fascia 4.

L'uso degli ausiliari non è quindi particolarmente caratteristico nel confronto fra testi Didattici e testi Scientifici, se non per alcune eccezioni. Il confronto fra testi narrativi e testi didattici presenta una variazione statisticamente significativa in fascia due, un variazione molto significativa in fascia 1 e in fascia 4 e una variazione estremamente significativa in fascia 5. Per concludere, il confronto Narrativo vs Scientifico, presenta una variazione estremamente significativa nelle fasce 3, 4 e 5. Questo dato ci dice che l'uso dell'ausiliare risulta caratteristico per il confronto solo nella seconda parte dei documenti, ad eccezione della parte conclusiva.

- **DEP_obj**

La dipendenza "Obj" descrive la relazione tra una testa sintattica a il suo oggetto diretto.

DEP_obj	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	✓	❌	✅	✅	❌
Fascia 2	✅	✅	❌	❌	✓	❌
Fascia 3	✅	✅	❌	✅	✅	✅
Fascia 4	✅	✅	❌	✅	✅	✔
Fascia 5	✅	✅	❌	✅	✅	❌
Fascia 6	✅	✅	❌	✅	✅	✓

Tabella 16: tabella rappresentante la variazione significativa della dipendenza "Obj" nelle fasce per ciascuno dei confronti realizzati.

In merito alla dipendenza "Obj" i primi due confronti considerati (Giornalistico vs Narrativo e Giornalistico vs Scientifico) sono piuttosto simili, entrambi hanno infatti una variazione estremamente significativa in tutte le fasce considerate, ad eccezione del primo caso che vede la fascia 1 caratteristica ma in modo molto più debole rispetto alle altre (variazione statisticamente significativa). Il terzo confronto non presenta invece alcune tipo di variazione significativa. I due casi successivi (Didattico vs Scientifico e Narrativo vs Didattico) presentano a loro volta una variazione estremamente significativa in gran parte delle fasce considerate eccetto che nella fascia 2, che nel primo caso non presenta alcun tipo di variazione significativa e nel secondo presenta una variazione statisticamente significativa. Il

confronto fra genere narrativo e genere scolastico presenta una variazione estremamente significativa in fascia 3, una variazione molto significativa in fascia 4 e una variazione statisticamente significativa in fascia 6. Parliamo nuovamente di un caso in cui l'aspetto caratteristico di una fratture diminuisce gradualmente con lo scorrere dei documenti.

- **DEP_subj**

La dipendenza "Subj" descrive la relazione fra un verbo attivo e il suo soggetto.

DEP_subj	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	✓	✔	✅	✅	❌
Fascia 2	✔	✔	❌	✅	✔	✔
Fascia 3	✅	✅	✓	✅	✅	✅
Fascia 4	✅	✅	✔	✅	✅	✔
Fascia 5	✅	✅	✓	✅	✅	✅
Fascia 6	✅	✅	❌	✅	✅	✔

Tabella 17: tabella rappresentante la variazione significativa della dipendenza "Subj" nelle fasce per ciascuno dei confronti realizzati.

I confronti che comprendono i testi didattici e i testi giornalistici presentano numerosi variazioni estremamente significative nelle fasce considerate, ad eccezione del confronto che avviene proprio tra i suddetti genere, nel quale le fasce 1 e 4 sono caratterizzate da una variazione molto significativa e le fasce 3 e 5 sono caratterizzate da una variazione statisticamente significativa. Tutto ciò sottolinea come tra i vari casi considerati quello fra il genere giornalistico e il genere didattico sia il confronto in cui la dipendenza "Subj" risulta in tutto e per tutto caratterizzante in maniera minore. Il caso Narrativo vs Scientifico presenta infine una variazione estremamente significativa nelle fasce 3 e 5, e una variazione molto significativa nelle fasce 2, 4 e 6.

- **Aggettivi precedenti la testa sintattica (adjPre)**

Il confronto fra i testi giornalistici e i testi narrativi, in merito alla percentuale di aggettiva precedenti la testa sintattica, presenta una variazione estremamente significativa in tutte le fasce considerate eccetto che nella fascia 1, che presenta

adjPre	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✔️	✓	✅	✅	✅	✓
Fascia 2	✔️	✅	❌	❌	✅	✅
Fascia 3	✅	✅	❌	❌	✅	✅
Fascia 4	✅	❌	❌	❌	✅	✅
Fascia 5	✅	✅	❌	❌	✅	✅
Fascia 6	✅	❌	❌	❌	✅	✅

Tabella 18: tabella rappresentante la variazione significativa degli adjPre nelle fasce per ciascuno dei confronti realizzati.

invece una variazione statisticamente significativa. Il caso Giornalistico vs Scientifico presenta invece una variazione molto significativa nelle prime due fasce e una variazione estremamente significativa nelle fasce 3 e nelle fasce 5. I casi Giornalistico vs Didattico e Didattico vs Scientifico presentano una variazione significativa esclusivamente nella fascia 1, nel primo caso estremamente significativa mentre nel secondo molto significativa. Questo dato ci dice come la posizione degli aggettivi precedente la testa sintattica sia caratteristica esclusivamente per le prime righe dei testi. I due casi conclusivi presentano invece una variazione estremamente significativa in tutte le fasce considerate per quanto riguarda il primo caso, ed è una situazione analoga per il secondo ad eccezione della fascia 1, che presenta invece una variazione statisticamente significativa.

- **Subordinate di grado superiore al primo (subMinor)**

Tra i confronti considerati, per quanto riguarda le subMinor, i casi con il maggior numero di variazioni significative sono il confronto fra testi giornalisti e testi narrativi e quello fra testi narrativi e testi scientifici. Il primo presenta infatti una variazione estremamente significativa in tutte le fasce eccetto che nella prima, mentre il secondo è caratterizzato da una variazione estremamente significativa in tutte le fasce considerate. I confronti Giornalistico vs Scientifico, Giornalistico vs Didattico e Didattico vs Scientifico hanno invece meno da offrire e si limitano a poche variazioni esclusivamente nella prima parte di testi (fascia 1, 2 e 3). Il primo caso presenta infatti una variazione estremamente nella fascia 1, il secondo una variazione

molto significativa nella fascia 1 e variazioni statisticamente significative delle fasce 2 e 3, mentre il terzo è caratterizzato da una variazione statisticamente significative nella fascia 1 e da una variazione molto significativa nella fascia 3.

subMinor	Giornalistico vs Scientifico	Giornalistico vs Narrativo	Giornalistico vs Didattico	Didattico Vs Scientifico	Narrativo Vs Didattico	Narrativo Vs Scientifico
Fascia 1	✅	❌	✔️	✓	✅	✅
Fascia 2	❌	✅	✓	❌	✔️	✅
Fascia 3	❌	✅	✓	✔️	❌	✅
Fascia 4	❌	✅	❌	❌	✅	✅
Fascia 5	❌	✅	❌	❌	✅	✅
Fascia 6	❌	✅	❌	❌	✅	✅

Tabella 19: tabella rappresentante la variazione significativa delle subMinor nelle fasce per ciascuno dei confronti realizzati.

Il confronto fra genere narrativo e genere didattico è infine caratterizzato da variazioni estremamente significative in tutta la seconda porzione del testo (fascia 4, 5 e 6), da una variazione estremamente significativa nella fascia 1 e da una variazione molto significativa nella fascia 2.

4.1. Terzo esperimento

Il terzo studio si discosta leggermente dai precedenti in quanto propone di individuare le caratteristiche che variano in maniera significativa all'interno del singolo genere. Come abbiamo visto in precedenza (v. paragrafo 1.2.1.) le feature si dividono in diverse categorie categorie linguistiche e per ognuna di esse sono state selezionate le più interessanti ai fini dello studio. La divisione in 6 fasce delle tabelle per singola feature (v. paragrafo 2.3.5.) permette un confronto tra fasce consecutive e non, consentendo di determinare quali variazioni, e soprattuto la significatività di esse, si manifestano tra una fascia e l'altra.

4.1.1. Analisi statistica delle tabelle per singola feature

Per ogni genere considerato sono state create 89 tabelle, corrispondenti alle feature estratte durante il monitoraggio linguistico (v. paragrafo 2.3.5.). Ogni tabella presenta una divisione in colonna in base alle fasce e permette uno studio basato proprio sugli intervalli tra le fasce stesse. Nello specifico sono stati considerati gli intervalli tra la prima e la seconda fascia, tra la seconda e la terza fascia, tra la terza e la quarta fascia, tra la quarta e la quinta fascia, tra la quinta e la sesta fascia e tra la sesta e la prima fascia. Parliamo quindi di 6 intervalli di fascia formati ognuno da una coppia di valori.

Per esempio nell'intervallo tra la terza e la quarta fascia sono considerati i valori della colonna "fascia 3" e i valori della colonna fascia "4" (v. tabella 20).

fascia 1	fascia 2	fascia 3	fascia 4	fascia 5	fascia 6
2.0	12.0	67.0	38.0	39.5	28.5
5.5	31.5	53.0	13.5	26.0	15.0
20.67	76.33	29.67	45.33	51.0	25.67
8.5	40.5	63.0	16.0	46.0	25.0
16.0	26.0	13.0	15.0	25.5	14.75
…	…	…	…	…	…

Tabella 20: tabella per i Token del genere "giornalistico" con l'intervallo di fascia 3-4 evidenziato.

Ogni intervallo di fascia è quindi costituito da due gruppi di dati che diventano le variabili *x* e *y* per le funzioni statistiche descritte nel capitolo 4. Per ognuno degli intervalli considerati otteniamo quindi il coefficiente di correlazione di Pearson (r_p), il coefficiente di correlazione per ranghi di Spearman (rho_s) e i p-value relativi, compreso quello ottenuto tramite Wilcoxon. I risultati, per essere analizzati, vengono poi inseriti in un file in formato .csv (v. tabella 21, tabella 22).

In questo modo otteniamo 89 tabelle che contengono le analisi statistiche, interne al genere, per ognuna delle feature considerate.

Lunghezza Media	pvalue_p	r_p
fascia 1/2	❌	0.06269996091261125
fascia 2/3	✅	0.29658496699117975
fascia 3/4	✅	0.43442183209103225
fascia 4/5	✅	0.42450629772707205
fascia 5/6	✅	0.4440100075931976
fascia 1/6	✓	-0.09699833979123942

Tabella 21: tabella dei risultati statistici per i Tokens nei testi giornalistici.

Lunghezza Media	pvalue_s	rho_s	pvalue_w
fascia 1/2	✔	0.1147606867540016	✅
fascia 2/3	✅	0.3650767372002572	✔
fascia 3/4	✅	0.48263820740216057	❌
fascia 4/5	✅	0.46356190141699843	❌
fascia 5/6	✅	0.4502421010507468	❌
fascia 1/6	✔	-0.10858413598425848	✅

Tabella 22: tabella dei risultati statistici per i Tokens nei testi giornalistici.

Per tutti i generi presi in esame sono state considerate le feature di base, mentre quelle legate alle caratteristiche morfo-sintattiche e sintattiche sono state selezionate di modo da presentare i casi più interessanti e degni di nota.

4.1.2. Genere giornalistico

Caratteristiche di base

- **Lunghezza media delle frasi (tokens)**

Nel caso della lunghezza media delle frasi possiamo notare come gli intervalli di fascia 3/4, 4/5 e 5/6 siano correlati positivamente e in modo moderato per quanto riguarda sia Pearson che Spearman (v. tabella 23, tabella 24). Quest'ultimo presenta una correlazione positiva anche nell'intervallo di fascia 2/3.

Per quanto riguarda Wilcoxon (v. tabella 24) troviamo una variazione estremamente significativa negli intervalli di fascia 1/2, 2/3 e 1/6.

Tokens	pvalue_p	r_p
fascia 1/2	❌	0.06269996091261125
fascia 2/3	✅	0.29658496699117975
fascia 3/4	✅	0.43442183209103225
fascia 4/5	✅	0.42450629772707205
fascia 5/6	✅	0.4440100075931976
fascia 1/6	✓	-0.09699833979123942

Tabella 23: tabella dei risultati statistici per la lunghezza media della frasi (tokens) nei testi giornalistici.

Tokens	pvalue_s	rho_s	pvalue_w
fascia 1/2	✔	0.1147606867540016	✅
fascia 2/3	✅	0.3650767372002572	✅
fascia 3/4	✅	0.48263820740216057	❌
fascia 4/5	✅	0.46356190141699843	❌
fascia 5/6	✅	0.4502421010507468	❌
fascia 1/6	✔	-0.10858413598425848	✅

Tabella 24: tabella dei risultati statistici per la lunghezza media della frasi (tokens) nei testi giornalistici.

Possiamo quindi dedurre che gli articoli di giornale considerati presentano nella parte centrale una tendenza positiva a variare assieme (co-variare) per quanto riguarda la lunghezza media delle frasi. Se quindi in un determinato articolo di giornale la parte centrale sarà costituita da frasi più lunghe rispetto alla parte iniziale

e a quella conclusiva, la stessa variazione sarà presente anche negli altri articoli considerati. Vediamo inoltre che le differenze maggiori, per quanto riguarda la lunghezza media della frasi, sono individuabili tra la parte iniziale e quella centrale del testo e tra la parte iniziale e quella finale, com'è lecito aspettarsi. In un articolo, infatti, la parte iniziale funge da introduzione all'argomento trattato e generalmente presenta frasi piuttosto brevi, mentre la parte conclusiva tira la fila di quanto detto spesso tramite periodi più complessi. Se infatti le prime righe hanno come finalità quella di far sì che il lettore provi interesse per l'argomento trattato e non abbandoni l'articolo, allo stesso modo il termine del tutto deve necessariamente concludere quanto detto e non può farlo in maniera immediata e quindi tramite frasi brevi e concise.

- **Lunghezza media delle parole (mCxT)**

mCxT	pvalue_p	r_p
fascia 1/2	✅	0.3306173811966396
fascia 2/3	✅	0.37171816585424694
fascia 3/4	✅	0.3248334450950354
fascia 4/5	✅	0.35894421639841934
fascia 5/6	✅	0.3134687285582494
fascia 1/6	✅	0.19858520985700467

Tabella 25: tabella dei risultati statistici per l'mCxT nei testi giornalistici.

mCxT	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.39497827010124437	❌
fascia 2/3	✅	0.3916920877054668	❌
fascia 3/4	✅	0.3295978224917563	❌
fascia 4/5	✅	0.3698193002972479	❌
fascia 5/6	✅	0.3356489388248706	❌
fascia 1/6	✅	0.21213181226522143	✓

Tabella 26: tabella dei risultati statistici per l'mCxT nei testi giornalistici.

Per quanto riguarda la lunghezza media delle parole vi è una correlazione positiva e moderata in tutti gli intervalli consecutivi considerati, sia per Pearson sia per Spearman (v. tabella 25, tabella 26).

Per Wilcoxon (v. tabella 26) invece l'unica variazione significativa è quella nell'intervallo 1/6 ovvero tra la porzione iniziale e quella finale dell'articolo.

La lunghezza media delle parole è quindi una grandezza che varia secondo lo stesso andamento negli articoli considerati e che presenta una variazione nell'intervallo di fascia compreso fra l'inizio e la fine dell'articolo.

Caratteristiche morfo-sintattiche

Tra le feature estratte di natura morfo-sintattica sono stati selezionati due casi considerati particolarmente interessanti rispetto agli altri.

- **Pronomi Personali (PE)**

PE	pvalue_p	r_p
fascia 1/2	✓	0.09564125004729249
fascia 2/3	✅	0.16786261247158243
fascia 3/4	✔	0.11115379857161763
fascia 4/5	❌	0.07104002880609289
fascia 5/6	✅	0.21374909059458422
fascia 1/6	❌	0.015568959190437208

Tabella 27: tabella dei risultati statistici per i PE nei testi giornalistici.

PE	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.35095606222667136	❌
fascia 2/3	✅	0.37255050650990257	❌
fascia 3/4	✅	0.3140594812456572	❌
fascia 4/5	✅	0.3560843442995083	❌
fascia 5/6	✅	0.3095796294397318	❌
fascia 1/6	✅	0.22732439158910606	✓

Tabella 28: tabella dei risultati statistici per i PE nei testi giornalistici.

L'aspetto di maggior interesse in questo caso è l'assenza di correlazione per Pearson (v. tabella 27) rispetto alla correlazione positiva e moderata in gran parte degli intervalli analizzati per Spearman (v. tabella 28). Ciò significa che le assunzioni per il modello di correlazione parametrica (coefficiente r Pearson) non sono soddisfatte. Tali assunzioni consistono nel considerare la distribuzione di x e y

normale e supporre la medesima varianza per le sottopopolazioni dei valori assunti dalle due variabili. Un caso simile, necessita che la misura dell'associazione venga eseguita mediante formule su scale diverse da quella quantitativa, come per esempio il livello di scala ordinale utilizzata dalla correlazione a ranghi di Spearman. In questo modo diventa possibile calcolare la correlazione fra due variabili e studiarne il grado di significatività.

Da tutto ciò si deduce che i pronomi personali negli articoli di giornale seguono il medesimo andamento per gran parte degli intervalli considerati, ad eccezione dell'intervallo 1/6 che sottolinea come l'utilizzo dei pronomi personali nella parte introduttiva dell'articolo rispetto a quella conclusiva non presenti alcun tipo di correlazione.

- **Sostantivi propri (SP)**

Oltre ai pronomi personali ho selezionato il caso dei sostantivi propri, un'altra feature piuttosto comune nel genere giornalistico.

SP	pvalue_p	r_p
fascia 1/2	✅	0.42701987263205937
fascia 2/3	✅	0.4049227228878055
fascia 3/4	✅	0.44340644446527533
fascia 4/5	✅	0.49091387529750663
fascia 5/6	✅	0.46499403536222744
fascia 1/6	✅	0.30816278620214277

Tabella 29: tabella dei risultati statistici per i SP nei testi giornalistici.

Osserviamo che vi è una correlazione positiva e moderata per tutti gli intervalli considerati per Pearson (v. Tabella 29). Aspetto che viene confermato anche dalla correlazioni per ranghi di Spearman (v. tabella 30).

Come abbiamo visto nelle precedenti analisi, un simile grado di correlazione evidenzia come i sostantivi propri varino secondo la medesima funzione in tutti gli intervalli di fascia considerati.

In merito a Wilcoxon (v. Tabella 4.30) le variazioni significative si limitano agli intervalli di fascia 1/2 e 1/6, corrispondenti agli intervalli fra l'inizio e la parte centrale dell'articolo e fra l'inizio e la parte conclusiva.

SP	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.5007036379805866	✅
fascia 2/3	✅	0.5315045055545158	❌
fascia 3/4	✅	0.4938035777169979	❌
fascia 4/5	✅	0.529590145948426	❌
fascia 5/6	✅	0.5265322980658059	❌
fascia 1/6	✅	0.36145861565223764	✅

Tabella 30: tabella dei risultati statistici per i SP nei testi giornalistici.

Caratteristiche sintattiche

Tra le caratteristiche sintattiche sono state estratte la percentuale di tokens che precedono la testa sintattica e le subordinate.

- **Percentuale di tokens che precedono la testa sintattica (PreHead)**

PreHead	pvalue_p	r_p
fascia 1/2	❌	0.0782724680546881
fascia 2/3	✅	0.13644303826359686
fascia 3/4	✅	0.18382206936894382
fascia 4/5	✅	0.2630373425638272
fascia 5/6	✅	0.295068325065332
fascia 1/6	❌	0.04063834660669806

Tabella 31: tabella dei risultati statistici per i PreHead nei testi giornalistici.

PreHead	pvalue_s	rho_s	pvalue_w
fascia 1/2	✔️	0.11541423301200712	❌
fascia 2/3	✅	0.19002255974919632	❌
fascia 3/4	✅	0.20054028124953746	❌
fascia 4/5	✅	0.2823970374070254	❌
fascia 5/6	✅	0.2755230597432891	❌
fascia 1/6	❌	0.058387132600408065	❌

Tabella 32: tabella dei risultati statistici per i PreHead nei testi giornalistici.

In questo caso avrei potuto trattare anche la percentuale di tokens che seguono la testa sintattica in quanto i risultati statistici sono i medesimi. L'aspetto interessante di

questi due casi è infatti l'assoluta assenza di correlazione sia per Pearson sia per Spearman, e la non presenza di anche una sola variazione significativa per Wilcoxon (v. tabella 31, tabella 32).

La percentuale di tokens che precedono o che seguono la testa sintattica varia quindi in maniera libera e sottolinea come la struttura sintattica dei periodi nei testi giornalistici non sia soggetta a funzioni comuni su più intervalli di fascia.

- **Subordinate (SubTOT)**

SubTOT	pvalue_p	r_p
fascia 1/2	✅	0.1797652612003511
fascia 2/3	✅	0.297020951887865
fascia 3/4	✅	0.31171648218776415
fascia 4/5	✅	0.3259599541730096
fascia 5/6	✅	0.2692105454651059
fascia 1/6	✔️	0.12068407711324793

Tabella 33: tabella dei risultati statistici per le SubTot nei testi giornalistici.

Con il termine "subordinata" mi riferisco a tutte le proposizioni che dipendono da un'altra proposizione. Non vengono distinte in questo caso le subordinate di primo grado da quelle di grado superiore, ma vengono tutte considerate nello stesso modo. Per quanto riguarda Pearson troviamo una correlazione positiva e moderata negli intervalli di fascia 3/4 e 4/5, mentre Spearman aggiunge ai suddetti casi anche l'intervallo 2/3.

SubTOT	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.24501476011553014	✅
fascia 2/3	✅	0.36437246763603914	✓
fascia 3/4	✅	0.34493155603775266	❌
fascia 4/5	✅	0.34247165209955266	❌
fascia 5/6	✅	0.28093615748562106	❌
fascia 1/6	✅	0.14587690171160037	✅

Tabella 34: tabella dei risultati statistici per le SubTot nei testi giornalistici.

Wilcoxon indica invece la presenza di una variazione significativa nell'intervallo du fascia 2/3, e una variazione estremamente significativa negli intervalli di fascia 1/2 e 1/6, caratteristica che abbiamo individuato in gran parte dei casi finire considerati.

Com'era quindi lecito aspettarsi, le subordinate negli articoli di giornale variano in modo piuttosto simile nella porzione centrale del testo, mentre nell'apertura e nella chiusura dell'articolo si trovano le variazioni estremamente significative e quindi non determinate dal caso. Valgono quindi tutte le considerazioni che abbiamo visto in merito alle caratteristiche di base e che diventano quindi degli aspetti piuttosto comuni negli articoli di giornale considerati.

4.1.3. Genere scientifico

Caratteristiche di base

- **Lunghezza media delle frasi (tokens)**

In merito alla lunghezza media delle frasi troviamo una correlazione positiva e moderata per l'intervallo di fascia 5/6 per Pearson (v. tabella 35), mentre per Spearman una correlazione positiva e moderata si ha negli intervalli di fascia 4/5 e 5/6 (v. tabella 36).

Tokens	pvalue_p	r_p
fascia 1/2	✅	0.23351228279249792
fascia 2/3	✅	0.26273220311145074
fascia 3/4	✅	0.23464676721239006
fascia 4/5	✅	0.2062978380080489
fascia 5/6	✅	0.3530776143336875
fascia 1/6	✅	0.28911701824853053

Tabella 35: tabella dei risultati statistici per la lunghezza media delle frasi (tokens) nei testi scientifici.

Per Wilcoxon, gli intervalli estremamente significativi sono l'intervallo 1/2 e l'intervallo 1/6 (v. tabella 36).

La tendenza a co-variare della lunghezza media delle frasi nei testi scientifici è quindi caratteristica della parte conclusiva dei testi stessi. È lecito pensare che la descrizione di un determinato processo scientifico o di una ricerca compiuta non segua uno sviluppo coerente in tutti i casi selezionati. La lunghezza media delle frasi

sembra quindi variare secondo lo stesso andamento solo nella chiusura del documento, o al massimo poco prima, che corrisponde alle considerazioni finali in merito a quanto trattato.

Tokens	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.2297638146151292	✅
fascia 2/3	✅	0.2894631581557976	❌
fascia 3/4	✅	0.2731258358642893	❌
fascia 4/5	✅	0.3754190849490417	❌
fascia 5/6	✅	0.3945243879937794	❌
fascia 1/6	✅	0.2542090901365836	✅

Tabella 36: tabella dei risultati statistici per la lunghezza media delle frasi (tokens) nei testi scientifici.

Wilcoxon sottolinea invece come gli intervalli di fascia 1/2 e 1/6 presentino una variazione significativa e siano quindi caratterizzanti per i corpora analizzati. I testi scientifici hanno quindi nella lunghezza media delle frasi un elemento caratterizzante nella porzione iniziale del documento e nel rapporto fra porzione iniziale e porzione conclusiva.

- **Lunghezza media delle parole (mCxT)**

mCxT	pvalue_p	r_p
fascia 1/2	✅	0.19936779529242615
fascia 2/3	✅	0.2651777892363111
fascia 3/4	✅	0.2787642382921736
fascia 4/5	✅	0.2794945229704885
fascia 5/6	✔️	0.17866418205395942
fascia 1/6	✅	0.18495461110725597

Tabella 37: tabella dei risultati statistici per l'mCxT nei testi scientifici.

Per quanto riguarda la lunghezza media delle parole non c'è molto da segnalare. Vi è infatti una correlazione positiva e moderata per Spearman (v. tabella 38) negli intervalli di fascia 3/4 e 4/5, e una variazione significativa per Wilcoxon (v. tabella 38) negli intervalli di fascia 1/2 e 1/6.

mCxT	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.27583372862764355	✅
fascia 2/3	✅	0.24736622938141067	❌
fascia 3/4	✅	0.3275018720807155	❌
fascia 4/5	✅	0.3041636789474282	❌
fascia 5/6	✅	0.25541975529102123	❌
fascia 1/6	✅	0.2660948404677597	✔️

Tabella 38: tabella dei risultati statistici per l'mCxT per parola nei testi scientifici.

La lunghezza media delle parole segue lo stesso andamento nella parte centrale del documento, corrispondente all'intervallo di fascia fra la 3 e la 5, corrispondente al contenuto vero e proprio del documento in cui vengono illustrati esperimenti, analisi di ricerca, caratteristiche estratte ecc.

In merito alla significatività troviamo nuovamente gli intervalli 1/2 e 1/6 come caratterizzanti, proprio come nel caso della lunghezza media delle frasi.

Caratteristiche morfo-sintattiche

Anche nel caso dei testi Scientifici sono stati considerati, in merito alle caratteristiche morfo-sintattiche, due casi di particolare rilievo.

- **Numeri (N)**

N	pvalue_p	r_p
fascia 1/2	✅	0.5837797341267353
fascia 2/3	✅	0.5801040403718549
fascia 3/4	✅	0.5890620275712981
fascia 4/5	✅	0.6927952103015363
fascia 5/6	✅	0.5326717975372153
fascia 1/6	✅	0.27758837430729233

Tabella 39: tabella dei risultati statistici per N nei testi scientifici.

I numeri, che comprendono sia i cardinali che gli ordinali, presentano una correlazione positiva e moderata in tutti gli intervalli di fascia considerati nel caso di Spearman (v. tabella 40), e tutti gli intervalli escluso quello tra la prima e la sesta fascia nel caso di Pearson (v. Tabella 39). Per quanto riguarda Wilcoxon (v. Tabella

40), l'unico intervallo che presenta una variazione significativa è quello tra la prima e la sesta fascia, quindi un intervallo di fasce non consecutive ma distanti.

N	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.4779869226970775	❌
fascia 2/3	✅	0.4165642670581913	❌
fascia 3/4	✅	0.5255756732053242	❌
fascia 4/5	✅	0.5871556593803527	❌
fascia 5/6	✅	0.5102036682612794	❌
fascia 1/6	✅	0.3791918037260926	✓

Tabella 40: tabella dei risultati statistici per N nei testi scientifici.

I numeri, nel genere scientifico, tendono dunque a co-variare in tutti gli intervalli di fascia considerati mentre risultano caratterizzanti soltanto tra la prima e la sesta fascia dei testi presi in esame.

- **Abbreviazioni (SA)**

Nell'analisi delle abbreviazioni appare piuttosto evidente quanto il test di Pearson debba necessariamente sottostare a determinati vincoli e non offra tutte le informazioni che sono visibili tramite Spearman. Basti notare come ben tre intervalli di fascia non trovino riscontro nel caso di Pearson (v. tabella 41) rispetto a quanto accade in Spearman (v. tabella 42). La correlazione di Spearman è meno forte

SA	pvalue_p	r_p
fascia 1/2	❌	-0.013337517129058398
fascia 2/3	❌	0.002325911612920119
fascia 3/4	✅	0.4640449955128897
fascia 4/5	✅	0.31713663420385874
fascia 5/6	✅	0.3732414029696046
fascia 1/6	❌	0.0139662808067952

Tabella 41: tabella dei risultati statistici per SA nei testi scientifici.

rispetto a quello di Pearson, ma quest'ultima non può essere sempre utilizzata.
Ciò che emerge in ogni caso è l'andamento tipico delle abbreviazioni in ogni fascia considerata.

Per Wilcoxon in questo caso non c'è nulla da segnalare, se non che le abbreviazioni non sono in alcun modo caratterizzanti negli intervalli di fascia considerati (v. Tabella 42)

SA	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.39323581652648504	❌
fascia 2/3	✅	0.3881179850159965	❌
fascia 3/4	✅	0.4326956129892601	❌
fascia 4/5	✅	0.39084807947817346	❌
fascia 5/6	✅	0.38834533634001117	❌
fascia 1/6	✅	0.38622415893612677	❌

Tabella 42: tabella dei risultati statistici per SA nei testi scientifici.

Caratteristiche sintattiche

Tra le caratteristiche sintattiche sono state selezionate le concatenazioni e il numero medio di figli per token.

- **Concatenazioni (concat)**

Le concatenazioni sono relazioni fra tokens che portano a forme di parole complesse. Nei testi di natura scientifica, secondo Spearman, esse presentano un indice di correlazione positivo e moderato per gran parte degli intervalli considerati. Nello specifico esse variano secondo lo stesso andamento negli intervalli di fascia 1/2, 3/4, 4/5, 5/6 e 1/6 (v. tabella 44).

Concat	pvalue_p	r_p
fascia 1/2	✅	0.2500963244852144
fascia 2/3	❌	-0.017356501247709294
fascia 3/4	❌	0.04430192826399905
fascia 4/5	❌	0.08605119834149755
fascia 5/6	❌	0.06946832923823643
fascia 1/6	❌	0.06433738584749672

Tabella 43: tabella dei risultati statistici per le concat nei testi scientifici.

Osservando i risultati di Wilcoxon sappiamo inoltre che esse non sono caratterizzanti per i corpora analizzati (v. Tabella 44).

Concat	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.35481171300033976	❌
fascia 2/3	✅	0.21459951058060106	❌
fascia 3/4	✅	0.3409660039942867	❌
fascia 4/5	✅	0.3496377706165833	❌
fascia 5/6	✅	0.3331467928006571	❌
fascia 1/6	✅	0.343057365659222	❌

Tabella 44: tabella dei risultati statistici per le concat nei testi scientifici.

- **Numero medio di figli per token (mChildren)**

Il numero medio di figli per token presenta una correlazione positiva e moderata negli intervalli 2/3, 3/4 e 4/5 per Pearson e per Spearman (v. tabella 45, tabella 46), a

mChildren	pvalue_p	r_p
fascia 1/2	❌	0.0216323262330448
fascia 2/3	✅	0.3649457475829155
fascia 3/4	✅	0.31335660546110106
fascia 4/5	✅	0.31637536481122275
fascia 5/6	✅	0.2513790408567764
fascia 1/6	❌	0.001983183557717785

Tabella 45: tabella dei risultati statistici per mChildren nei testi scientifici.

cui si aggiunge l'intervallo di fascia 5/6. Tale caratteristica di natura sintattica segue quindi lo stesso andamento nella parte centrale dei documenti selezionati. Rimangano escluse dalla suddetta considerazione le prime e le ultime fasce del testo, in cui l'andamento del numero medio di figli per token non è lo stesso per tutti i documenti selezionati. Interessanti, da questo punto di vista, sono i valori assunti dai coefficienti di correlazione (Pearson e Spearman) per il primo e l'ultimo intervallo di fascia, molto più piccoli rispetto a quanto osservato negli altri intervalli.

Wilcoxon indica invece come intervalli significativi quelli compresi tra la prima e la seconda fascia, tra la quinta e la sesta e tra la prima e la sesta (v. tabella 46). Il

primo e il secondo in particolar modo sono estremamente significativi. Abbiamo quindi tre intervalli di fascia in cui il numero medio di figli per token è caratterizzante per il corpus analizzato.

mChildren	pvalue_s	rho_s	pvalue_w
fascia 1/2	❌	0.0500387418647463	✅
fascia 2/3	✅	0.39858133614767854	❌
fascia 3/4	✅	0.32330133301925756	❌
fascia 4/5	✅	0.3421403588629379	❌
fascia 5/6	✅	0.33191575967643916	✓
fascia 1/6	❌	0.07503482183044032	✅

Tabella 46: tabella dei risultati statistici per mChildren nei testi scientifici.

4.1.4. Genere narrativo

Caratteristiche di base

- **Lunghezza media delle frasi (tokens)**

Tokens	pvalue_p	r_p
fascia 1/2	✅	0.7673027626136589
fascia 2/3	✅	0.8333579464483706
fascia 3/4	✅	0.8412451980042889
fascia 4/5	✅	0.7954132621402626
fascia 5/6	✅	0.7847396821237737
fascia 1/6	✅	0.7959747032896748

Tabella 47: tabella dei risultati statistici per la lunghezza media delle frasi nei testi narrativi.

Nei testi narrativi, la lunghezza media delle frasi gode di una correlazione forte (coefficiente di correlazione > 0,7) per tutti gli intervalli di fascia considerati, sia per Pearson (v. tabella 47) sia per Spearman (v. tabella 48). La presenza di una forte correlazione dimostra quanto gli andamenti negli intervalli di fascia si somiglino, molto più rispetto a quanto visto nei casi precedenti per gli altri generi.

Per Wilcoxon c'è un solo intervallo che presenta una variazione significativa, quello tra la prima e la sesta fascia. È quindi lecito considerare la lunghezza media delle frasi caratterizzante per l'intervallo tra l'inizio e la conclusione del testo.

Tokens	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.8000529024184634	❌
fascia 2/3	✅	0.7663740922184642	❌
fascia 3/4	✅	0.787966394121673	❌
fascia 4/5	✅	0.7645619025818962	❌
fascia 5/6	✅	0.7703759040678744	❌
fascia 1/6	✅	0.769680148944372	✓

Tabella 48: tabella dei risultati statistici per la lunghezza media delle frasi nei testi narrativi.

- **Lunghezza media delle parole (mCxT)**

mCxT	pvalue_p	r_p
fascia 1/2	✅	0.971608109698067
fascia 2/3	✅	0.9796082901999635
fascia 3/4	✅	0.9846760220614157
fascia 4/5	✅	0.9755732724041483
fascia 5/6	✅	0.9705188312297637
fascia 1/6	✅	0.9698831965167058

Tabella 49: tabella dei risultati statistici per l'mCxT nei testi narrativi.

mCxT	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.7909523474066693	❌
fascia 2/3	✅	0.8249390282261243	❌
fascia 3/4	✅	0.8413560249418626	❌
fascia 4/5	✅	0.7877311529148158	❌
fascia 5/6	✅	0.7705958518200529	❌
fascia 1/6	✅	0.7870173251678082	❌

Tabella 50: tabella dei risultati statistici per la media dei caratteri parola nei testi narrativi.

Rispetto a quanto visto per la lunghezza media delle frasi, la lunghezza media delle parole presenta una correlazione estremamente forte per Pearson (v. tabella 49) in tutti gli intervalli di fascia considerati (valori superiori allo 0,95). Spearman offre

invece una situazione simile a quanto visto nel caso precedente (v. tabella 50) poiché meno forte rispetto a Pearson.

Wilcoxon in questo caso non offre niente da segnalare (v. tabella 50) .

Caratteristiche morfo-sintattiche

Tra le caratteristiche morfo-sintattiche sono stati considerati i nomi propri e gli avverbi.

- **Nomi propri (SP)**

I nomi propri seguono lo stesso andamento in tutti gli intervalli di fascia

SP	pvalue_p	r_p
fascia 1/2	✅	0.26965353211509585
fascia 2/3	✅	0.678474811266384
fascia 3/4	✅	0.566967700000546
fascia 4/5	✅	0.5922209288962367
fascia 5/6	✅	0.5341717307696852
fascia 1/6	✅	0.36219140940580347

Tabella 51: tabella dei risultati statistici per SP nei testi narrativi.

considerati come si può vedere nei risultati per Spearman (v. tabella 52). Pearson presenta invece come unica differenza l'intervallo di fascia 1/2 (v. tabella 51), a causa dei numerosi vincoli di cui deve tener conto. I coefficienti di correlazione ottenuti sono positivi e moderati con casi tendenti alla forte correlazione come per esempio l'intervallo di fascia 2/3 o quello 5/6.

SP	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.6287372688954623	✓
fascia 2/3	✅	0.6942462806016555	❌
fascia 3/4	✅	0.6577518148395998	❌
fascia 4/5	✅	0.6784758371975167	❌
fascia 5/6	✅	0.5960504391200394	✓
fascia 1/6	✅	0.5396509681456018	❌

Tabella 52: tabella dei risultati statistici per SP nei testi narrativi.

Il p-value ottenuto da Wilcoxon indica come intervalli caratterizzanti quelli compresi tra la prima e la seconda fascia e tra la quinta e la sesta. Parliamo comunque di una significatività statistica che si assesta su valori maggiori di 0,01.

In merito dunque ai nomi propri, troviamo una correlazione positiva in tutti gli intervalli di fascia analizzati e una significatività statistica rispettivamente nelle parti iniziali e conclusive dei testi presi in esame.

- **Avverbi (B)**

Gli avverbi presentano una correlazione positiva e moderata in tuti gli intervalli di fascia analizzati, sia per Pearson sia per Spearman (v. tabella 53, tabella 54). Possiamo quindi parlare di un andamento tipico degli avverbi nei testi narrativi, mantenendo comunque il punto di vista degli intervalli di fascia. Tale correlazione non è tipica soltanto degli avverbi per quanto riguarda il suddetto corpus, ma la si ritrova anche nel caso degli aggettivi, delle congiunzioni, delle preposizioni e in altre caratteristiche di natura morfo-sintattica.

B	pvalue_p	r_p
fascia 1/2	✅	0.5178113380258623
fascia 2/3	✅	0.5498859217953753
fascia 3/4	✅	0.5587346165820039
fascia 4/5	✅	0.5164592164147332
fascia 5/6	✅	0.5861515674425269
fascia 1/6	✅	0.5023763073292433

Tabella 53: tabella dei risultati statistici per B nei testi narrativi.

B	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.5502318964483134	❌
fascia 2/3	✅	0.6109773150405587	✓
fascia 3/4	✅	0.6173087533531941	❌
fascia 4/5	✅	0.5459046365492007	❌
fascia 5/6	✅	0.6050122073918937	❌
fascia 1/6	✅	0.5466342608512808	✓

Tabella 54: tabella dei risultati statistici per B nei testi narrativi.

La scelta di inserire in quest'analisi il caso degli avverbi dipende dai risultati ottenuti per Wilcoxon, che vede gli intervalli di fascia 2/3 e 1/6 come caratterizzanti per il genere preso in esame (v. tabella 54).

Caratteristiche sintattiche

Per quanto riguarda le caratteristiche sintattiche sono stati considerati due casi: l'oggetto diretto e la media delle ampiezze degli alberi delle subordinate di primo grado.

- **Oggetto diretto (obj)**

L'"oggetto diretto" è la relazione fra la testa verbale e il suo oggetto diretto. Nei casi di Pearson e Spearman, per i testi narrativi, tale caratteristica sintattica presenta una tendenza positiva e moderata a co-variare in tutti gli intervalli di fascia considerati (v. tabella 55, tabella 56). La relazione sopra descritta presenta quindi un andamento costante tra i vari documenti analizzati.

obj	pvalue_p	r_p
fascia 1/2	✅	0.4287527120650885
fascia 2/3	✅	0.4455308911045386
fascia 3/4	✅	0.48109418507844265
fascia 4/5	✅	0.5933506153284483
fascia 5/6	✅	0.5542238677532054
fascia 1/6	✅	0.5002085942530616

Tabella 55: tabella dei risultati statistici per obj nei testi narrativi.

obj	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.45848049080760006	✅
fascia 2/3	✅	0.533596456507665	✔️
fascia 3/4	✅	0.5255327224557206	❌
fascia 4/5	✅	0.5795870849545319	❌
fascia 5/6	✅	0.5659200309827375	❌
fascia 1/6	✅	0.5209050795299006	✔️

Tabella 56: tabella dei risultati statistici per obj nei testi narrativi.

Per Wilcoxon troviamo invece una variazione estremamente significativa nell'intervallo tra la prima e la seconda fascia, e variazioni molto significative negli

intervalli i fascia 2/3 e 1/6 (v. tabella 56). In questo caso la relazione "oggetto diretto" risulta caratterizzante per il genere narrativo nei primi intervalli di fascia presenti nel testo e nella relazione tra la parte iniziale e la parte conclusiva dei documenti considerati.

- **Media delle ampiezze degli alberi sintattici delle subordinate di primo grado (mWeightSubMain)**

L'ampiezza di un albero sintattico corrisponde alla larghezza di quest'ultimo, proprio come l'altezza corrisponde alla profondità. Nel caso dei tesi narrativi i risultati dei test di Pearson di Spearman indicano una correlazione positiva e moderata in tutte gli intervalli di fascia analizzati. In Pearson manca l'intervallo 1/2 che trova però riscontro in Spearman. La feature considerata varia con lo stesso andamento tra i vari intervalli di fascia nei documenti di genere narrativo.

mWeightSubMain	pvalue_p	r_p
fascia 1/2	✅	0.24934920169899102
fascia 2/3	✅	0.37554144795572203
fascia 3/4	✅	0.527204274075358
fascia 4/5	✅	0.45160489732936526
fascia 5/6	✅	0.3516919675332232
fascia 1/6	✅	0.3711661926105981

Tabella 57: tabella dei risultati statistici per mWeightSubMain nei testi narrativi.

mWeightSubMain	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.34136019860505085	❌
fascia 2/3	✅	0.3416189903822392	✓
fascia 3/4	✅	0.39516417126844544	❌
fascia 4/5	✅	0.5994203958637107	❌
fascia 5/6	✅	0.5521861456301929	❌
fascia 1/6	✅	0.4256075627037985	✔

Tabella 58: tabella dei risultati statistici per mWeightSubMain nei testi narrativi.

In merito a Wilcoxon troviamo una significatività statistica nell'intervallo di fascia 2/3, corrispondente alla porzione di testo successiva all'antefatto o simili, e una

variazione moto significatività tra la prima la sesta fascia, corrispondente al rapporto fra l'antefatto e la conclusione.

4.1.4. Genere narrativo

Caratteristiche di base

- **Lunghezza media delle frasi (tokens)**

Nei testi appartenenti al genere "materiali didattici" la lunghezza media delle frasi presenta una correlazione positiva e moderata in tutti gli intervalli di fascia considerati. In questo caso Spearman non mette in luce differenze degne di nota rispetto a Pearson ma si limita a una minima variazione nei valori ottenuti (v. tabella 59, tabella 60).

Wilcoxon non individua intervalli di fascia significativi (v. tabella 60) e quindi caratterizzanti per il genere rispetto al parametro considerato.

Tokens	pvalue_p	r_p
fascia 1/2	✅	0.3222886840645066
fascia 2/3	✅	0.3855928550355543
fascia 3/4	✅	0.34483159406824093
fascia 4/5	✅	0.46368497852592433
fascia 5/6	✅	0.3572814196252291
fascia 1/6	✅	0.39377307522060107

Tabella 59: tabella dei risultati statistici per la lunghezza media delle frasi nei testi didattici.

Tokens	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.415184293577001	❌
fascia 2/3	✅	0.4698985853982828	❌
fascia 3/4	✅	0.40398737100496346	❌
fascia 4/5	✅	0.4520811881703526	❌
fascia 5/6	✅	0.4833649844452377	❌
fascia 1/6	✅	0.3940909822038274	❌

Tabella 60: tabella dei risultati statistici per la lunghezza media delle frasi nei testi didattici.

La lunghezza media delle frasi, negli intervalli di fascia considerati, varia quindi secondo lo stesso andamento senza presentare però casi di forte correlazione. In

merito alla significatività, nessuno degli intervalli considerati risultata caratterizzante per il genere.

- **Lunghezza media delle parole (mCxT)**

In merito alla lunghezza media delle parole vale più o meno quanto detto per la lunghezza media delle frasi. Troviamo infatti anche qui una correlazione positiva e moderata (v. tabella 61, tabella 62) in entrambi i casi statistici considerati (Pearson e Spearman) e l'assenza di significatività per Wilcoxon in tutti gli intervalli di fascia presi in esame (v. tabella 62).

mCxT	pvalue_p	r_p
fascia 1/2	✅	0.6391100848619113
fascia 2/3	✅	0.6265685821035505
fascia 3/4	✅	0.5633653867918177
fascia 4/5	✅	0.5675136264206737
fascia 5/6	✅	0.569467773642312
fascia 1/6	✅	0.4514825369114876

Tabella 61: tabella dei risultati statistici per mCxT nei testi didattici.

mCxT	pvalue_s	rho_s	pvalue_w
fascia 1/2	✅	0.6019664512919307	❌
fascia 2/3	✅	0.5717920701331208	❌
fascia 3/4	✅	0.5465235674046595	❌
fascia 4/5	✅	0.6063315910312574	❌
fascia 5/6	✅	0.603526677882582	❌
fascia 1/6	✅	0.5010350742769835	❌

Tabella 62: tabella dei risultati statistici per mCxT nei testi didattici.

Caratteristiche morfo-sintattiche

Come caratteristiche morfo-sintattiche sono state prese in esame gli aggettivi e i verbi.

- **Aggettivi (A)**

Nei testi di natura didattica gli aggettivi presentano una correlazione positiva e moderata nella seconda parte dei vari documenti analizzati, con alcune accezioni. Nello specifico Pearson individua negli intervalli di fascia 3/4, 5/6 e 1/6 un

andamento tipico, mentre Spearman, oltre agli intervalli già presentati per Pearson, aggiunge il caso che coinvolge la seconda e la terza fascia (v. tabella 63).

Il p-value di Wilcoxon non risulta, in nessuno degli intervalli considerati, sufficientemente piccolo perché si possa parlare di significatività statistica (v. tabella 64).

A	pvalue_p	r_p
fascia 1/2	❌	0.14433437205408886
fascia 2/3	✓	0.19585544082182244
fascia 3/4	✅	0.3517326430701736
fascia 4/5	✔	0.24914827609521895
fascia 5/6	✅	0.3899810834284357
fascia 1/6	✅	0.3432465022665968

Tabella 63: tabella dei risultati statistici per mCxT nei testi didattici.

Ciò che risulta quindi è che gli aggettivi nei testi didattici presentano un andamento tipico nelle fasce centrali di ogni documento e in quelle conclusive. Da segnalare è inoltre il simile andamento che ritroviamo nel confrontata le fasce iniziali e quelle conclusive.

A	pvalue_s	rho_s	pvalue_w
fascia 1/2	❌	0.10617940132190383	❌
fascia 2/3	✅	0.3126562553577867	❌
fascia 3/4	✅	0.41753818975848234	❌
fascia 4/5	✔	0.26220716549525624	❌
fascia 5/6	✅	0.31645076503172154	❌
fascia 1/6	✅	0.3773325090452757	❌

Tabella 64: tabella dei risultati statistici per A nei testi didattici.

- **Verbi(V)**

I verbi presenti nei testi di natura didattica non presentano un andamento tipico negli intervalli di fascia analizzati se non per un'unica eccezione, corrispondente all'intervallo 4/5 nell'analisi tramite Pearson (v. tabella 65).

Contrariamente a quanto però visto nelle feature considerate in precedenza per il presente genere, i verbi presentano una significatività statistica nell'intervallo di fascia 4/5 e una variazione molto significativa nell'intervallo di fascia 5/6 secondo Wilcoxon (v. tabella 66). In questo caso possiamo notare come i verbi siano caratterizzanti della parte conclusiva di ciascuno dei documenti considerati.

V	pvalue_p	r_p
fascia 1/2	❌	0.04493046444875847
fascia 2/3	✓	0.18567338718011842
fascia 3/4	✔	0.28359946010756804
fascia 4/5	✅	0.32132701775006006
fascia 5/6	✔	0.24569934719848063
fascia 1/6	✓	0.2284746008219825

Tabella 65: tabella dei risultati statistici per i verbi nei testi didattici.

V	pvalue_s	rho_s	pvalue_w
fascia 1/2	❌	0.0793928246004881	❌
fascia 2/3	✔	0.28913511830440713	❌
fascia 3/4	✅	0.2969160727504879	❌
fascia 4/5	✓	0.22832491832283539	✓
fascia 5/6	✔	0.23761202343154025	✔
fascia 1/6	✓	0.17938155513906523	❌

Tabella 66: tabella dei risultati statistici per i verbi nei testi didattici.

Caratteristiche sintattiche

Tra le caratteristiche sintattiche sono state selezionate le subordinate di grado superiore al primo e i complementi temporali.

- **Subordinate di grado superiore al primo (subMinor)**

Con "subordinate di grado superiore al primo" indico tutte le subordinate che non hanno come reggente la proposizione principale, ma che dipendono a loro volta da una subordinata. Nei testi appartenenti al genere "materiali didattici", la suddetta feature presenta un andamento tipico esclusivamente nell'intervallo di fascia 2/3 per Spearman (v. tabella 68).

subMinor	pvalue_p	r_p
fascia 1/2	❌	0.1347488857295599
fascia 2/3	✔️	0.2534782376120535
fascia 3/4	❌	0.09809390726243813
fascia 4/5	✓	0.23005087113373804
fascia 5/6	❌	0.1531135645224577
fascia 1/6	❌	0.11434069004338836

Tabella 67: tabella dei risultati statistici per subMinor nei testi didattici.

subMinor	pvalue_s	rho_s	pvalue_w
fascia 1/2	❌	0.07751776423865388	❌
fascia 2/3	✅	0.31665127095942186	❌
fascia 3/4	❌	0.10018664773243682	✓
fascia 4/5	✓	0.20373766233385096	❌
fascia 5/6	✓	0.1984930081485877	❌
fascia 1/6	❌	0.15342763327553716	✓

Tabella 68: tabella dei risultati statistici per subMinor nei testi didattici.

Per quanto concerne invece la significatività, negli intervalli 3/4 e 1/6 vi è una significatività statistica indicata dai valori assunti dal p-value di Wilcoxon (v. Tabella 68).

- **Complementi temporali (comp_temp)**

Un complemento temporale indica una relazione di tempo con una testa sintattica verbale. Nei documenti appartenenti al genere denominato “materiali didattici”, la suddetta feature presenta dei valori interessanti per l’analisi compiuta esclusivamente nel caso di Pearson (v. tabella 69).

comp_temp	pvalue_p	r_p
fascia 1/2	❌	0.07310893699203043
fascia 2/3	✅	0.5365710599139792
fascia 3/4	❌	-0.06387814477595104
fascia 4/5	✅	0.3997772236195221
fascia 5/6	✅	0.3229380457677914
fascia 1/6	❌	0.07389346610357073

Tabella 69: tabella dei risultati statistici per i comp_temp nei testi didattici.

Gli intervalli di fascia 2/3, 4/5, e 5/6 presentano un coefficiente di correlazione maggiore di 0,3 e quindi indicano come i complementi temporali abbiano un andamento tipico nei suddetti intervalli.

comp_temp	pvalue_s	rho_s	pvalue_w
fascia 1/2	✗	0.13552530829391915	✗
fascia 2/3	✓	0.2244585979646266	✗
fascia 3/4	✗	-0.016478148150358102	✗
fascia 4/5	✓	0.17923131056744046	✗
fascia 5/6	✓	0.2254685564786528	✗
fascia 1/6	✗	0.10588584703935625	✗

Tabella 70: tabella dei risultati statistici per i comp_temp nei testi didattici.

4.4. Quarto esperimento

Il quarto studio descritto nella presente relazione consiste nel determinare le feature più caratteristiche per ciascuno dei generi considerati. Per farlo si utilizza la deviazione standard, o scarto quadratico medio[7], che a seconda del valore assunto determina se una data feature è caratterizzante per il genere preso in esame. A valori di deviazione standard molto bassi corrisponde infatti una variazione minima della caratteristica considerata.

Il calcolo dello scarto quadratico medio viene eseguito tramite l'utilizzo di una funzione presente nella libreria *scipy*, a cui si accompagna inoltre il calcolo della media:

- `DataFrame.std()`: restituisce la deviazione standard dei valori per la colonna richiesta
- `DataFrame.mean()`: restituisce la media dei valori per la colonna richiesta

In questo caso il file in formato .csv utilizzato non è la tabella definitiva (ovvero quella contenente tutte le tabelle per singolo documento suddivise in 6 fasce) ma la tabella definitiva sottoposta al processo di normalizzazione. In questo modo tutti valori per singola feature sono compresi fra 0 e 1 e permettono di determinare quale feature sia più caratteristica rispetto alle altre e quale feature lo sia meno.

Ottenuta la tabella definitiva vengono calcolati i valori di media e di deviazione standard per le 89 feature in ciascuno dei generi considerati. Il tutto viene poi ordinato in modo crescente in base alla deviazione standard, così da poter identificare con precisione le caratteristiche con scarto quadratico medio minore.

Le media viene utilizzata per escludere tutti quei casi che sembrerebbero caratteristici, ma che lo risultano esclusivamente per la frequenza minima che hanno all'interno dei testi. Una feature con una media pari a 0.01 si manifesta in pochissimi casi all'interno del genere e i dati che la rappresentano sono quindi poco dispersi. Tutto ciò non può far sì che la feature venga considerata un tratto caratterizzante e quindi i casi con media inferiore allo 0,02 non sono stati considerati.

Per ciascuno genere ho selezionato le dieci feature con scarto quadratico medio minore.

---

[7] Lo scarto quadratico medio (o deviazione standard o scarto tipo) è un indice di dispersione statistico, vale a dire una stima della variabilità di una popolazione di dati o di una variabile casuale.

4.3.1. Genere giornalistico

Nella tabella 71 sono presenti le dieci feature più caratterizzanti per il genere giornalistico, ordinate in base alla deviazione standard.

features	media	deviazioneStandard
CPOS_A(%)	0.06	0.05
DEP_ROOT(%)	0.08	0.05
mChildren	0.93	0.05
mCxT	0.54	0.06
mDist	0.26	0.06
CPOS_N(%)	0.05	0.07
DEP_clit(%)	0.04	0.07
Tokens	0.16	0.08
maxDist	0.15	0.08
DEP_mod_temp(%)	0.03	0.08

Tabella 71: le 10 feature con deviazione standard più bassa nei testi giornalistici.

Caratteristiche di base

Tra le caratteristiche di base troviamo sia la lunghezza media della parole (mCxT) che la lunghezza media della frasi (Tokens) . Possiamo quindi affermare che le suddette caratteristiche siano in tutto e per tutto caratterizzanti per il genere giornalistico, e che mCxT e Tokens tendano a disperdersi intorno alla media con i valori rispettivi di 0,06 e 0,08.

Caratteristiche morfo-sintattiche

In merito alle caratteristiche morfo-sintattiche troviamo la percentuale di aggettivi (CPOS_A(%)), che è inoltre l'aspetto più caratteristico per il genere giornalistico, e la percentuale di numeri (CPOS_N(%)). Gli aggettivi e i numeri sono quindi due part of speech (POS) caratteristiche dei testi giornalistici.

Caratteristiche sintattiche

Le caratteristiche sintattiche più caratteristiche per i testi giornalistici sono invece la dipendenza radice (DEP_ROOT(%)), il numero medio di figli per token

(mChildren), la distanza media dei tokens dalla testa sintattica (mDist), la DEP_clit(%), la distanza massima dalla testa sintattica (maxDist) e la DEP_mod_temp(%). In questo ambito le feature caratteristiche per il genere analizzato sono principalmente quelle legate alla distanza dalla testa sintattica e al numero medio di figli per token, che con una deviazione pari a 0,06 risulta uno degli aspetti più caratteristici del genere giornalistico.

4.3.2. Genere scientifico

Le 10 feature, con deviazione standard minore e ordinate in modo crescente, selezionate per i testi scientifici sono indicate nella tabella 72.

features	media1	deviazioneStandard1
maxDist	0.09	0.05
POS_VM(%)	0.03	0.06
mCxT	0.56	0.07
POS_S(%)	0.26	0.07
mDist	0.21	0.07
DEP_modal(%)	0.04	0.07
mWeight	0.18	0.07
CPOS_S(%)	0.31	0.08
mWeightSubMinor	0.05	0.08
Tokens	0.18	0.09

Tabella 72: le 10 feature con deviazione standard più bassa nei testi scientifici.

Caratteristiche di base

Tra le caratteristiche di base, anche nel caso dei testi scientifici troviamo sia la lunghezza media della parole (mCxT) sia la lunghezza media delle frasi (Tokens). Come nel caso precedente, le caratteristiche di base risultano caratterizzanti per il genere e quindi sottolineano come i documenti di natura scientifica presentano la lunghezza dei periodi e la media di caratteri per parola tendenti al valori medio calcolato sull'intero corpus.

Caratteristiche morfo-sintattiche

In merito alle caratteristiche morfo-sintattiche troviamo la percentuale di verbi modali (POS_VM(%)), la percentuale di sostantivi comuni (POS_S(%)) e la

percentuale di sostantivi in generale (CPOS_S(%)). Possiamo quindi affermare che i sostantivi siano l'aspetto morfo-sintattico più caratteristico nei testi scientifici, come è lecito aspettarsi vista la presenza di numerosi termini tecnici e strettamente legati all'ambito di ricerca trattato.

Caratteristiche sintattiche

Feature legate alla distanza dalla testa sintattica, come la distanza media e la distanza massima, trovano ampio spazio per quanto riguarda la caratterizzazione del genere scientifico. Troviamo inoltre, tra le caratteristiche sintattiche con scarto quadratico medio inferiore, l'ampiezza dell'albero sintattico (mWeight), l'ampiezza dell'albero sinottico delle subordinate di grado superiore al primo (mWeightSubMinor) e la dipendenza DEP_modal(%).

4.3.3. Genere narrativo

Per il genere narrativo sono state selezionate le 10 feature con deviazione standard minima in ordine crescente (v. tabella 73).

features	media1	deviazioneStandard1
DEP_mod_temp(%)	0.03	0.06
DEP_sub(%)	0.04	0.06
CPOS_D(%)	0.03	0.07
POS_VM(%)	0.03	0.08
DEP_modal(%)	0.03	0.08
DEP_neg(%)	0.04	0.08
mHeightSubMinor	0.05	0.08
mWeightSubMinor	0.05	0.08
DEP_arg(%)	0.06	0.09
DEP_mod_rel(%)	0.05	0.09

Tabella 73: le 10 feature con deviazione standard più bassa nei testi narrativi.

Caratteristiche di base

Tra le caratteristiche di base, rispetto a quanto visto nei casi precedenti, non troviamo nessun caso che possa essere considerato caratteristico il genere narrativo. È quindi chiaro come la lunghezza delle frasi e la media dei caratteri per parola non

siano feature che permettano di determinare un genere giornalistico, poiché i dati a esse associati risultano fortemente dispersi.

Caratteristiche morfo-sintattiche

Tra le caratteristiche morfo-sintattiche troviamo la percentuale di determinanti (CPOS_D(%)) e la percentuale di verbi modali (POS_VM(%)). Soltanto due caratteristiche morfo-sintattiche risultano caratterizzanti per il genere narrativo.

Caratteristiche sintattiche

La maggior parte delle feature con deviazione standard minore nel genere narrativo sono quindi di natura sintattica. Tra esse troviamo diverse dipendenze come per esempio DEP_mod_temp(%) e DEP_sub(%) e la media dell'altezza e quella dell'ampiezza degli alberi delle subordinate di grado superiore al primo. Nei testi narrativi gli aspetti più caratteristici sono quindi quelli legati alle dipendenze sinottiche, ovvero alla costruzione dei periodi e all'utilizzo delle subordinate di grado superiore al primo, meno frequenti rispetto a quelle di primo grado e quindi caratterizzate da una dispersione minore.

4.3.4. Genere didattico

Nella tabella 74 sono elencate le 10 feature con deviazione standard minore nei materiali didattici.

features	media1	deviazioneStandard1
mChildren	0.94	0.05
mCxT	0.66	0.08
DEP_modal(%)	0.05	0.1
subMainPre(%)	0.04	0.1
CPOS_N(%)	0.06	0.11
POS_PD*(%)	0.06	0.11
mDist	0.46	0.11
LinkPost(%)	0.73	0.11
DEP_comp_temp(%)	0.03	0.11
DEP_punc(%)	0.25	0.11

Tabella 74: le 10 feature con deviazione standard più bassa nei testi didattici.

Caratteristiche di base

Tra le caratteristiche di base troviamo esclusivamente la m lunghezza media della parole, che presenta un valore di deviazione standard pari a 0,8. Tale feature risulta quindi caratteristica per i testi narrativi.

Caratteristiche morfo-sintattiche

In merito alle caratteristiche morfo-sintattiche troviamo invece la percentuale di numero (CPOS_N(%)) e la percentuale di pronomi dimostrativi (POS_PD(%)). Particolarmente interessante quest'ultimo dato in quanto parliamo di una sottocategoria morfo-sintattica che però presenta una dispersione dei dati nel genere didattico piuttosto minima.

Caratteristiche sintattiche

Come nel caso precedente, anche per i materiali didattici la maggior parte delle feature caratteristiche per il genere sono di natura sintattica. Troviamo diverse dipendenze, come DEP_modal(%), DEP_comp_temp (%) e DEP_punc(%), la media dei figli per token (mChildren), la distanza media dalla testa sintattica (mDist), la percentuale di token che seguono la testa sintattica (LinkPOST(%)) e la percentuale delle subordinate di primo grado che precedono principale.

CAPITOLO 5

Conclusioni

In questo elaborato è stato condotto uno studio sui fenomeni linguistici rispetto ai generi testuali. L'obbiettivo che è stato perseguito è quello di capire quali caratteristiche linguistiche siano caratterizzanti per un genere rispetto all'altro, non solo per quanto riguarda i documenti interi, ma anche in merito a singole porzioni, identificative di sezioni del testo, quali l'introduzione e la conclusione. Inoltre, ciascuna caratteristica linguistica può essere studiata anche all'interno del singolo corpus, di modo da determinare l'andamento e il grado di significatività che essa presenta per un determinato genere testuale.

Per prima cosa è stato presentato uno stato dell'arte sul monitoraggio linguistico, descrivendone l'origine, l'evoluzione e le possibilità di studio da esso offerte. Sono state poi descritte le caratteristiche linguistiche ad esso collegate, per poi illustrare il processo di estrazione delle prime. Inoltre, sono stati presi in esame le differenze di genere e di complessità di modo da evidenziare altre possibilità di analisi rispetto all'argomento trattato in questa relazione.

Successivamente sono stati descritti gli otto corpora utilizzati e le feature da essi estratte durante la fase di monitoraggio. Particolare attenzione è stata riservata a tutti i casi legati alla posizione di un token rispetto alla sua testa sintattica, e al fenomeno della subordinazione. A seguire, è stata descritta la creazione delle tabelle contenenti i risultati del monitoraggio, punto di partenza per tutte le analisi o studi successivi.

Le funzioni statistiche utilizzate sono state descritte e analizzate a partire dalla formula di base. Per ciascuna di esse è stata inoltre presentata la corrispondente funzione offerta dalla libreria S*cipy*, che ha permesso l'applicazione delle suddette funzioni sui dati estratti nel monitoraggio.

Sono stati eseguiti 4 tipi di analisi sui dati estratti. Il primo è un studio del grado di significatività dei fenomeni linguistici nel confronto fra generi testuali. Ciascuna delle 89 feature estratte è stata sottoposta al Wilcoxon rank sum test per ciascun confronto, in modo da determinare quanto una feature possa essere caratteristica per un confronto rispetto che ad un altro. I risultati hanno mostrato un elevato grado di significatività per quanto riguarda le caratteristiche morfo-sintattiche generiche, contrariamente a quanto è accaduto invece per le sottocategorie ad esse associate che

in alcuni casi non hanno presentato alcun tipo di variazione significativa, per esempio il caso dei pronomi possessivi. In merito all'analisi delle dipendenze, i risultati hanno mostrato un buon grado di significativa in tutti i confronti descritti, specialmente nei confronti fra testi didattici e testi narrativi e testi narrativi e testi scientifici. Per quanto riguarda lo studio della posizione di un token rispetto alla sua testa sintattica, i risultati ottenuti hanno evidenziato in ben 4 confronti un grado di significatività elevato per gran parte delle feature considerate, mentre nei casi Giornalistico vs Didattico e Didattico vs Scientifico non sono mancati, tra i risultati, l'assenza di significatività per una data feature. Infine, le feature inerenti all'albero sintattico, hanno descritto un alto grado di caratterizzazione per quanto riguarda i casi più generici, come per esempio l'ampiezza o l'altezza dell'albero sintattico, per poi risultare molto meno caratterizzanti in casi più particolari, come la percentuale di subordinate di grado superiore al primo che precedono la principale.

Il secondo studio è anch'esso basato sui fenomeni linguistici nei vari generi testuali ma in questo caso non si limita al testo ma considera le fasce in cui esso è stato diviso. Sono stati presentati dieci casi, corrispondenti a dieci feature tra quelle estratte, in cui il grado di significatività di ogni feature, nei singoli confronti, è stato analizzato fascia per fascia. L'obiettivo primario era determinare il grado di caratterizzazione di una feature in diverse porzioni di testo, sempre mantenendo l'aspetto fondamentale del confronto di genere. Per esempio, nel caso dei pronomi dimostrativi il primo studio ci mostrava una variazione estremamente significativa nel confronto fra testi giornalisti e testi scientifici, mentre lo studio per singole fasce ha mostrato una variazione estremamente significativa solo per la prima fascia, e quindi per la prima porzione di testo, mentre le altre fasce non hanno presentato alcun tipo di variazione significativa.

Il terzo studio utilizza il concetto di divisone in fasce visto nel precedente studio ma si limita al singolo genere. Ogni coppia di fasce, consecutive e non, genera un intervallo all'interno del testo. L'obbiettivo è comprendere se una data feature presenta un andamento simile nei vari intervalli di fascia considerati, e naturalmente se risulta in qualche modo caratteristica per il genere. Per ciascuno dei quattro corpora analizzati (gli otto di partenza sono stati accoppiati in base al genere) è stato descritto il comportamento di 6 feature, rispettivamente suddivise in base alla

caratteristica linguistica. I casi descritti per singolo genere sono stati selezionati a seconda di quanto fossero utili per il seguente elaborato.

Il quarto e ultimo studio è infine un'indagine sulle feature più caratteristiche per ciascuno dei generi considerati. Per ogni caratteristica linguistica presa in esame è stata calcolata la media e la deviazione standard così da determinare, per ciascun genere, le dieci feature con valori più stabili e quindi più caratterizzanti. Tra le possibili osservazioni in merito alla suddetta indagine è stato interessante notare come ogni genere si comporti in base alle feature. Per esempio i casi Giornalistico e Scientifico offrono un buon equilibrio tra caratteristiche di base, morfo-sintattiche e sintattiche, che però non viene rispettato dagli altri due generi, le cui feature più caratterizzanti sono quasi tutte legate all'ambito sintattico. Un'altra delle possibili considerazioni è quella di osservare come la percentuale di numeri e di sostantivi presenti nel testo siano rispettivamente caratteristiche per il genere giornalistico e il genere scientifico. Nei testi giornalistici infatti si utilizzano date, valori, indirizzi e quant'altro, mentre il secondo è caratterizzato dall'uso di termini tecnici o simili.

In conclusione, ciò che è stato descritto nel seguente elaborato offre un nuovo punto di vista nello studio dei fenomeni linguistici all'interno del testo in generi diversi. La divisione in fasce consente infatti di determinare se il grado di significatività di una data caratteristica linguistica sia globale o se si limiti solo a determinate fasce. Inoltre, in merito agli studi sul singolo genere, l'utilizzo degli intervalli di fascia comporta un'ulteriore livello di analisi e approfondimento, permettendo di studiare, oltre al grado di significatività, l'andamento di una determinata feature in un genere preso in esame.

Bibliografia

Brunato D.; Dell'Orletta F.; Pieri G. (2016). Studio sull'ordinamento dei costituenti nel confronto tra generi e complessità. In Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it), 5-6 December 2016, Napoli, Italy.

Brunato D.; Dell'Orletta (2017). On the order of words in Italian: a study on genre vs complexity. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), Pisa, Italy, September 18-20, 2017.

Dell'Orletta F.; Montemagni S.; Venturi G. (2013). Linguistic profiling of texts across textual gen- res and readability levels. an exploratory study on italian fictional prose. In Proceedings of Recent Advances in Natural Language Processing (RANLP2013), (Hissar, Bulgaria, Settembre 2013).

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

Montemagni S. (2013). Tecnologie linguistico-computazionali e monitoraggio della lingua ita- liana. Studi Italiani di Linguistica Teorica e Applicata (SILTA), (1), 145–172.

Sitografia

DueParole (2002). Due parole, mensile di facile lettura. http://www.dueparole.it/.

Wikipedia, su Coefficiente di correlazione per ranghi di Spearman, https://it.wikipedia.org/wiki/Coefficiente_di_correlazione_per_ranghi_di_Spearman

Wikipedia, su Indice di correlazione per ranghi di Spearman, https://it.wikipedia.org/wiki/Indice_di_correlazione_di_Pearson

Wikipedia, su Test di Wilcoxon-Mann-Whitney, https://it.wikipedia.org/wiki/Test_di_Wilcoxon-Mann-Whitney

Wikipedia, su p-value, https://it.wikipedia.org/wiki/Valore_p

Wikipedia su Python, https://it.wikipedia.org/wiki/Python

Wikipedia, su SciPy, https://it.wikipedia.org/wiki/SciPy