UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Studio dei fattori di complessità linguistica all'interno del testo in varietà di lingua diverse

Candidato: *Chiara Buongiovanni*

Relatore: *Felice Dell'Orletta*

Correlatori: *Alessandro Lenci*
Dominique Brunato

Anno Accademico 2017-2018

INDICE

CAPITOLO I

Introduzione

Negli ultimi decenni, l'uso di tecnologie del linguaggio basate su strumenti di annotazione automatica all'avanguardia, addestrati su grandi quantità di dati linguistici, ha permesso un notevole sviluppo degli studi di monitoraggio linguistico sotto diversi punti di vista, anche molto distanti l'uno dall'altro e tutti ugualmente interessanti.

All'interno di questa variegata realtà, numerosi e proficui sono gli studi che hanno affiancato l'utilizzo di tali tecnologie linguistico-computazionali al rinnovato interesse che negli ultimi anni si è riacceso intorno al concetto di complessità linguistica, il quale tutt'oggi non ha ancora trovato una sistemazione risolutiva in una definizione accolta da tutti gli studiosi che ad esso hanno indirizzato la loro attività di ricerca. In letteratura, ciò si è tradotto in diverse proposte, ognuna dei quali ha saputo declinare la nozione di complessità in altrettanti scenari applicativi: i più fruttuosi riguardano sicuramente i processi di semplificazione e il livello di leggibilità dei testi, l'apprendimento della lingua madre o di una seconda lingua, l'influenza della complessità linguistica sulle prestazioni degli strumenti di analisi automatico e, da ultimo, ma non per importanza, la ricostruzione del profilo linguistico dei testi attraverso il monitoraggio dei fattori che si fanno portatori di complessità linguistica nei vari livelli di descrizione linguistica del testo, in varietà di lingua o in domini testuali differenti.
Le indagini e le metodologie discusse in questo lavoro di tesi rientrano a pieno titolo in quest'ultimo filone di ricerca. In particolare, si sfrutteranno algoritmi di analisi linguistica per ricercare le tendenze linguistiche che caratterizzano quattro generi testuali, a loro volta distinti in due varietà linguistiche differenti per grado di complessità, il quale è stato definito in relazione al lettore di riferimento. Accanto a uno studio di questo tipo, si è sperimentato un nuovo approccio che guardi al documento testuale non nella sua interezza ma che, al contrario, tenga conto delle diverse "parti" in cui un generico testo è idealmente diviso. In altre parole, si è cercato di definire se e come variano i fenomeni linguistici all'interno di un testo, riservando particolare attenzione ad alcuni tratti sintattici ampiamente riconosciuti come indici di complessità.
Prima di entrare nel vivo della ricerca, si è voluto innanzitutto descrivere a grandi linee il terreno sul quale ci stiamo muovendo: nel secondo capitolo, dunque, verranno presentate le idee alla base del Trattamento Automatico del Linguaggio, chiave di accesso al

contenuto informativo dei documenti testuali. Più nel dettaglio, verrà introdotto il concetto di corpus, risorsa linguistica che ha contribuito allo sviluppo degli strumenti per l'elaborazione dell'informazione linguistica. Dunque, verranno descritti i passaggi alla base dell'annotazione linguistica automatica dei testi e le tecnologie più all'avanguardia che, ad oggi, permettono di effettuare questa operazione con il minor numero di errori.

Nel terzo capitolo, invece, verrà tratteggiato lo stato dell'arte sulla complessità linguistica: dopo un breve excursus sulla declinazione del concetto di complessità nella letteratura dal diciannovesimo secolo ad oggi, verranno distinti i diversi approcci attraverso i quali l'argomento è stato studiato, per poi focalizzare l'indagine sul monitoraggio della complessità nei diversi livelli di descrizione linguistica, con particolare attenzione al livello sintattico.

Il quarto capitolo verrà dedicato alle fasi preliminari di questo studio. La scelta di comparare due varietà linguistiche differenti per grado di complessità internamente a quattro generi testuali – didattico, giornalistico, narrativo e scientifico – ha reso necessaria la selezione di otto corpora differenti, ognuno dei quali sarà descritto nel primo paragrafo. Nel quarto capitolo, inoltre, verrà illustrata la fase di estrazione delle caratteristiche linguistiche dal testo, con riferimenti agli script implementati a tale fine, e il modo in cui sono state gestite le informazioni estratte. Infine, verranno presentati gli algoritmi statistici utilizzati per l'individuazione dei risultati più significativi in fase di analisi dei risultati: il *Wilcoxon rank-sum test*, il coefficiente di correlazione di Pearson e di Spearman.

Il quinto capitolo descrive nel dettaglio i risultati ottenuti: dopo aver fornito una panoramica dei dati generali di ciascun corpus nel primo paragrafo, nei paragrafi 2, 3, 4 e 5, invece, verranno riportati e commentati i risultati ottenuti da quattro diversi esperimenti condotti: i) sulla variabilità delle features linguistiche all'interno di ogni corpus, ii) sulla variazione dei valori delle features tra la varietà semplice e complessa all'interno di ogni genere testuale, iii) sulle variazioni dei valori delle features tra fasce consecutive all'interno di ogni corpus, iv) sulla variazione dei valori delle features tra fasce corrispondenti nei testi semplici e complessi all'interno di ogni genere.

CAPITOLO II

Il monitoraggio linguistico: principi, metodi e strumenti

In questo primo capitolo verrà affrontato il tema del monitoraggio linguistico applicato allo studio della lingua italiana, e dei fondamenti teorici che ne stanno alla base. Verranno presentati gli strumenti e le tecnologie linguistico-computazionali che, ad oggi, hanno permesso lo sviluppo e l'implementazione di queste metodologie, divenute fondamentali per lo studio della variazione linguistica sull'asse diastratico, diamesico, diafasico e diacronico.

2.1. Introduzione al Trattamento Automatico della Lingua per il monitoraggio linguistico

Sollecitate dalla crescente necessità di una gestione intelligente dell'informazione contenuta in basi documentali in linguaggio naturale, la Linguistica Computazionale e l'Ingegneria del Linguaggio hanno investito nello sviluppo del settore del Trattamento Automatico della Lingua (TAL), o Natural Language Processing (NLP).
Dopo quasi un trentennio di ricerca e innovazione, gli strumenti del TAL si sono dimostrati in grado di risolvere l'intricata rete di strutture e relazioni grammaticali e lessicali della comunicazione linguistica, custodi della conoscenza depositata nei documenti testuali, nonostante gli ostacoli posti dall'ambiguità intrinseca nel linguaggio umano.
Le nuove capacità del computer nello svolgere compiti di elaborazione e decodificazione del linguaggio hanno aperto la strada a sviluppi promettenti in molti domini dell'*Information Technology.* Ne sono esempi i contesti applicativi mirati alla ricerca di documenti in lingue diverse (*Crosslingual Information Retrivial*), alla gestione e organizzazione del materiale documentale, all'acquisizione dinamica di elementi di conoscenza su un certo dominio (*Text Mining*) e all'estrazione automatica di informazioni strutturate da documenti non strutturati o semi-strutturati (*Information Extraction*) (Calzolari e Lenci, 2004).
Accanto a questi contesti applicativi legati all'analisi del testo dal punto di vista del contenuto, più recentemente la ricerca in linguistica computazionale si è indirizzata verso

lo studio di aspetti legati alla forma linguistica. L'implementazione di tecnologie linguistico-computazionali sempre più affidabili ha permesso, infatti, un significativo incremento del numero dei parametri linguistici sui quali condurre studi statistici in grado di tener traccia dei processi di variazione linguistica sull'asse diacronico, diafasico, diamesico e diastratico.

In questo senso, la riflessione di Montemagni (2013) si pone come manifesto del salto qualitativo, avvenuto in seguito all'automatizzazione dei processi di analisi del testo, degli studi di monitoraggio linguistico, finalmente svincolati dalla lenta analisi manuale o semi-manuale del testo. Tale innovazione, congiuntamente alla disponibilità di enormi quantità di dati linguistici, rappresentativi di diverse varietà d'uso della lingua italiana, ha assicurato, da un lato, il raggiungimento di un livello di accuratezza e affidabilità sempre più elevato e, dall'altro, il coinvolgimento di una vasta gamma di parametri appartenenti a diversi livelli di descrizione linguistica (lessicale, morfologico e sintattico). In particolare, l'elemento che ha determinato un netto scarto rispetto agli studi antecedenti all'utilizzo del TAL è da individuare nel monitoraggio di particolari aspetti della struttura sintattica del testo, ricostruibili solo a partire dalle caratteristiche strutturali dell'albero sintattico.

2.2. Il corpus: l'habitat naturale del dato linguistico

Riprendendo il parallelismo proposto da Lenci (2005), il monitoraggio linguistico può essere visto come lo studio e l'attività di raccolta dati che uno scienziato naturalista (il linguista computazionale) svolge sul comportamento di un animale (il linguaggio umano), osservandolo nel suo habitat naturale (il testo). In questo caso, il dato viene definito *ecologico*, in quanto conserva tutta la sua naturalezza, essendo stato raccolto nel proprio ambiente, riducendo al minimo le perturbazioni e i condizionamenti causati dalla presenza dell'osservatore e dai metodi di registrazione.
Chiaramente, l'indagine svolta seguendo una metodologia di questo tipo necessita un grandissimo numero di osservazioni, molte delle quali irrilevanti o di disturbo. È per questo motivo che, in linguistica computazionale, a fronte dell'estrema ricchezza delle variabili da controllare, si è ritenuto opportuno combinare l'utilizzo di dati linguistici ecologici con dati linguistici *controllati*, più mirati ma meno spontanei, in quanto ottenuti somministrando ai parlanti test ad hoc elaborati appositamente da un linguista.

Il fatto che il dato linguistico più "autentico" sia quello estratto da documenti testuali è necessario, ma non sufficiente, a spiegare lo stretto rapporto che lega le nuove tecnologie linguistico-computazionali con la creazione e l'uso di corpora testuali, ovvero di collezioni di testi selezionati e organizzati secondo specifici criteri per scopi di analisi linguistica. Infatti, da sempre alla base dello studio empirico del linguaggio, la raccolta di testi nell'era informatica si è imposta come fonte di evidenza per l'analisi computazionale della lingua, poggiandosi sul rinnovato interesse per i metodi statistico-matematici e sullo sviluppo della tecnologia informatica, la quale ha permesso di immagazzinare quantità sempre crescenti di testi e di agevolare la loro esplorazione per scopi di ricerca di dati linguistici interessanti.

Vogliamo ora definire i criteri che guidano la creazione di un corpus, per definizione frutto di un'opera di selezione, e i parametri che differenziano una tipologia dall'altra in vista della presentazione dei corpora utilizzati nel presente elaborato:

- **l'estensione**: la cui unità di misura è il numero di parole unità (*tokens*) che esso contiene. Rappresenta uno degli aspetti cruciali per determinare la conformazione e l'usabilità del corpus;
- **il grado di generalità**: dipende dalla misura in cui i testi che compongono il corpus sono stati selezionati in maniera trasversale rispetto alle diverse varietà di una lingua. È massimo nei corpora *generali*, i cui testi spaziano tra diverse comunità d'uso, varietà e registri linguistici, e minimo nei corpora *specialistici* o *verticali*, che contengono testi appartenenti a uno specifico *sub-language* o a un particolare dominio tematico. I primi sono plurifunzionali, ovvero non destinati ad una specifica applicazione, e sono perlopiù utilizzati come fonte di risorse trasversali di riferimento per comporre il quadro descrittivo della lingua nel suo complesso; i secondi si prestano a particolari obiettivi di ricerca;
- **la modalità di produzione dei testi**: i corpora possono contenere testi prodotti originariamente in forma scritta (corpora *di lingua scritta*), in forma orale (corpora *di lingua parlata*) o una combinazione dei due (corpora *misti*);
- **la selezione in relazione all'asse temporale**: si distinguono corpora *sincronici*, i cui testi appartengono a uno stesso intervallo di tempo, e corpora *diacronici*, che comprendono testi appartenenti a periodi diversi;
- **la lingua**: in base a questo parametro si distinguono i corpora *monolingue* dai corpora *bilingue* o *multilingue*, a loro volta distinguibili in:
    - corpora *paralleli*, o paralleli *allineati*, nel caso in cui le unità linguistiche dei testi nella lingua originaria e quelle dei testi nella lingua in cui sono stati tradotti siano esplicitamente collegati;

    - corpora *comparabili*, che contengono testi originali in lingue diverse selezionati in base agli stessi criteri;
- **l'integrità dei testi**: i testi *interi* sono preferibili per preservare la naturalezza dei dati; la selezione di porzioni di testo di lunghezza uniforme, invece, può essere preferita per ragioni di bilanciamento;
- **la codifica digitale**: si intende il modo in cui sono rappresentati i testi digitali. Distinguiamo i corpora *codificati ad alto livello*, in cui i testi sono arricchiti da etichette che esplicitano informazione di tipo strutturale e compositivo, e corpora *annotati* nei quali si codificano informazioni riguardanti la struttura linguistica del testo a diversi livelli di rappresentazione.

A monte dei suoi studi, dunque, il linguista computazionale è chiamato ad effettuare una scelta tra uno o più corpora che possano fornirgli il tipo e la quantità di informazioni di suo interesse (Lenci *e altri*, 2005).
Nello specifico caso del monitoraggio linguistico, gli studi dovrebbero essere svolti su quello che John Sinclair (1991) definì *monitor corpus,* una raccolta *aperta* che, rimanendo fedele ai criteri stabiliti in fase progettuale, viene arricchita nel tempo di nuovi testi al fine di tener traccia dei processi di variazione linguistica, ma anche tra diversi generi testuali o tra varietà d'uso. La dinamicità di questo tipo di corpus viene ampiamente sfruttata per studiare l'evoluzione del linguaggio e, in particolar modo, si presta al monitoraggio delle dinamiche del lessico della lingua (Montemagni, 2013).

2.3. L'annotazione linguistica

Il contenuto informativo del dato testuale è racchiuso in complesse strutture linguistiche e testuali, organizzate in gerarchie multiple di tratti linguistici, per cui sono necessari livelli di conoscenza e competenza linguistica di cui il computer non dispone. Si rende così necessaria una codifica di alto livello, durante la quale alle porzioni di testo vengono associate esplicitamente struttura e funzione corrispondenti, solitamente tramite etichette (o *tag*) di marcatura.
Tale operazione di annotazione linguistica rende esplicita, interpretabile ed esplorabile dal computer – in altre parole *machine readable* – la struttura linguistica implicita nel testo. Quest'ultimo può essere esplorato su molteplici livelli di descrizione linguistica, per ognuno dei quali è stato predisposto un diverso livello di annotazione:

I. **annotazione morfo-sintattica**: assegnazione dell'informazione relativa alla categoria grammaticale a ogni parola (o *token*) nel contesto specifico in cui è

inserita (ad esempio nome, verbo, aggettivo, punteggiatura, ecc.), talvolta integrata da specificazioni morfologiche (ad esempio persona, genere, numero ecc.). A questo livello vengono disambiguate le omografie riguardanti parti del discorso ed è per questo spesso combinato all'annotazione per lemma, o *lemmatizzazione*, che alla forma flessa della parola associa la corrispondente forma canonica (non marcata);

II. **annotazione sintattica**: codifica di informazione relativa all'analisi sintattica delle frasi di un testo. Si distinguono due principali sistemi di rappresentazione sintattica:
    i. rappresentazioni a costituenti: identificazioni dei costituenti sintattici, detti sintagmi, e delle loro relazioni di incassamento gerarchico;
    ii. rappresentazioni a dipendenze funzionali: identificazione delle relazioni binarie di dipendenza tra le parole e delle rispettive funzioni sintattiche (soggetto, oggetto, modificatore, complemento…);

III. **annotazione semantica**: codifica del significato o contenuto semantico delle espressioni linguistiche di un testo. Generalmente riguarda la classificazione rispetto a categorie semantico concettuali predefinite e/o la marcatura del ruolo semantico, o tematico, che un certo costituente svolge;

IV. **annotazione pragmatica**: codifica di fenomeni riguardanti la funzione comunicativa di una particolare unità linguistica o le relazioni che coinvolgono la struttura del discorso o della macro-organizzazione linguistica del testo.

2.3.1. I moduli dell'Analisi Linguistica

Le analisi linguistiche svolte per esplicitare la struttura linguistica del testo si configurano su una linea di complessità crescente, in cui l'output di un modulo costituisce l'input del successivo:

- ***sentence splitting***: segmentazione del testo in frasi;
- ***tokenizzazione***: segmentazione delle frasi in *tokens* (parole ortografiche, numeri, sigle, segni di punteggiatura, nomi propri ed entità complesse);
- ***part-of-speech tagging***: processo di assegnazione della categoria morfosintattica di appartenenza e del lemma corrispondente a ciascun token;
- ***dependency parsing***: analisi della struttura sintattica della frase in termini di relazioni binarie di dipendenza.

Per svolgere compiti di annotazione linguistica in modo accurato, gli strumenti per l'analisi linguistica di testi e per l'acquisizione dinamica della conoscenza utilizzano algoritmi

di apprendimento automatico supervisionato, addestrati su corpora annotati con informazione morfo-sintattica e sintattica. In questo modo, l'annotazione linguistica viene trattata in termini di classificazione probabilistica: in altre parole, il sistema assegna alla parola in input l'annotazione più probabile, considerando i suoi tratti descrittivi e il contesto in cui è inserita.
L'accuratezza dell'annotazione si mantiene sempre più che sufficiente, sebbene decresca con l'aumentare della complessità del livello di descrizione linguistica.

2.3.2. Linguistic Annotation pipeline (*LinguA*)

Un utile strumento per effettuare e visualizzare le diverse operazioni di annotazione automatica del testo è *LinguA*[1], una pipeline di annotazioni linguistiche che combina algoritmi basati su regole (*rule-based*) con metodi di apprendimento automatico (*machine learning*).

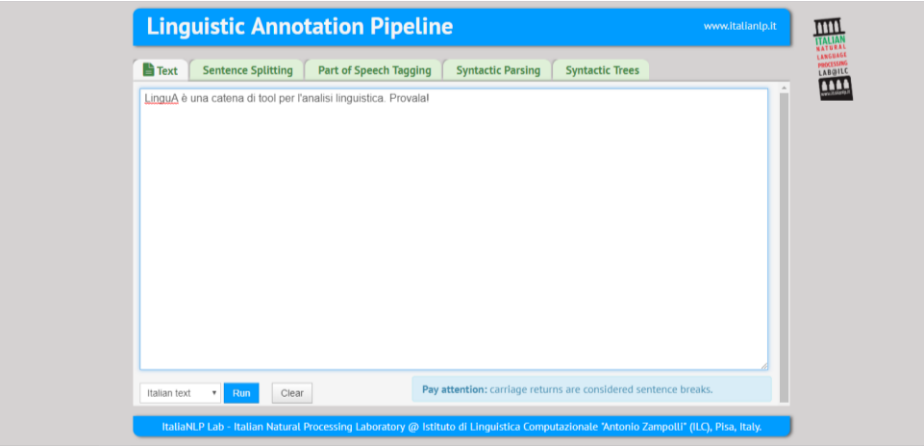

FIGURA 2.1: *Screenshot* del sito web di *LinguA*[2]

Questo software integra il *Part–of–Speech tagger* descritto in (Dell'Orletta, 2009), che registra un'accuratezza (calcolata come il rapporto tra il numero di token classificati correttamente e il numero totale di token analizzati) del 96,34% nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati; per quanto riguarda l'analisi sintattica, è stato utilizzato il *parser* a dipendenze DeSR (Attardi *e altri*, 2009), che raggiunge l'87,71% e l'83,38% in termini di *Unlabelled Attachment Score* (UAS) e *Labelled Attachment Score* (LAS) rispettivamente. Tali percentuali evidenziano come DeSR sia più affidabile nel ricostruire le relazioni di dipendenza che collegano le parole della frase (UAS), piuttosto che nell'identificare simultaneamente il tipo di dipendenza e la testa sintattica (LAS). In ogni caso, le sue prestazioni rappresentano lo stato dell'arte per

[1] `http://linguistic-annotation-tool.italianlp.it/`
[2] visitato il 21 marzo 2019.

la lingua italiana, come testimoniano i risultati delle campagne di valutazione di componenti per il trattamento automatico dell'italiano EVALITA[3], effettuate periodicamente.

LinguA permette di visualizzare le informazioni estratte dall'analisi incrementale del testo in una tabella (TABELLA 2.1) in cui le frasi sono separate da una riga vuota e ad ogni riga corrisponde un token, identificato con un ID.

				PART OF SPEECH TAGGING			SYNTACTIC PARSING	
	ID	**TOKEN**	**LEMMA**	**C-POS**	**F-POS**	**TRATTI**	**TESTA**	**TIPO DI DIPENDENZA**
1	1	LinguA	lingua	S	S	num=s | gen=f	2	subj
	2	è	essere	V	V	num=s | per=3 mod=i | ten=p	0	ROOT
	3	una	uno	R	RI	num=s | gen=f	4	det
	4	catena	catena	S	S	num=s | gen=f	2	pred
	5	di	di	E	E	_	4	comp
	6	tool	tool	S	S	num=s | gen=m	5	prep
	7	per	per	E	E	_	4	comp
	8	l'	il	R	RD	num=s | gen=n	9	det
	9	analisi	analisi	S	S	num=n | gen=f	7	prep
	10	linguistica	linguistico	A	A	num=s | gen=f	9	mod
	11	.	.	F	FS	_	2	punc
2	1	Prova-	provare	V	V	num=s | per=2 mod=m | ten=p	0	ROOT
	2	la	la	P	PC	num=s | per=3 | gen=f	1	obj
	3	!	!	F	FS	_	1	punc

TABELLA 2.1: Rappresentazione tabellare dell'annotazione linguistica del testo "LinguA è una catena di tool per l'analisi linguistica. Provala!" svolta da *LinguA*

Per ogni token, inoltre, vengono riportate le seguenti informazioni linguistiche: il lemma, due livelli di p*art-of-speech* (una più generale, la C-POS, e una più specifica, la F-POS), i tratti morfologici (numero, genere, persona, modo, tempo o superlativo), l'ID della testa sintattica da cui dipende e il tipo di dipendenza da questa, utilizzando il tagset morfo-sintattico e il tagset a dipendenze ISST-TANL[4].

LinguA, inoltre, è in grado di trasporre le informazioni linguistiche estratte in un albero a dipendenze sintattiche (FIGURA 2.2), in cui ogni arco rappresenta la dipendenza sintattica che lega la testa, da cui parte la freccia, al dipendente, su cui punta la freccia.

[3] `http://www.evalita.it/`

[4] L'inventario delle categorie morfo-sintattiche e delle dipendenze sintattiche utilizzate, con relativa descrizione, è consultabile alle pagine `http://www.italianlp.it/docs/ISST-TANL-POStagset .pdf` e `http://www.italianlp.it/docs/ISST-TANL-DEPtagset.pdf`.

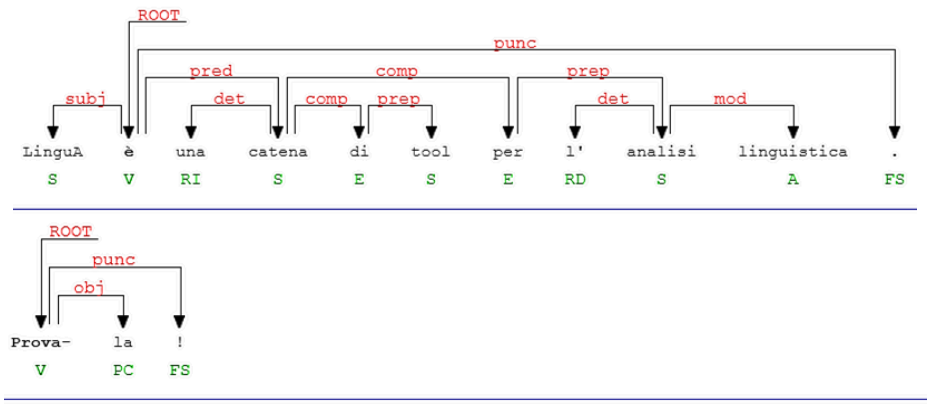

FIGURA 2.2: Rappresentazione grafica dell'annotazione linguistica in TABELLA 2.1

Questo strumento, infine, permette di effettuare il download i risultati dell'analisi in formato CoNLL (Nilsson *e altri*, 2007), di cui è riportato un esempio (FIGURA 2.3).

```
1       LinguA  lingua  S       S       num=s|gen=f     2       subj
2       è       essere  V       V       num=s|per=3|mod=i|ten=p 0       ROOT
3       una     uno     R       RI      num=s|gen=f     4       det
4       catena  catena  S       S       num=s|gen=f     2       pred
5       di      di      E       E       _       4       comp
6       tool    tool    S       S       num=s|gen=m     5       prep
7       per     per     E       E       _       4       comp
8       l'      il      R       RD      num=s|gen=n     9       det
9       analisi analisi S       S       num=n|gen=f     7       prep
10      linguistica     linguistico     A       A       num=s|gen=f     9       mod
11      .       .       F       FS      _       2       punc

1       Prova-  provare V       V       num=s|per=2|mod=m|ten=p 0       ROOT
2       la      la      P       PC      num=s|per=3|gen=f       1       obj
3       !       !       F       FS      _       1       punc
```

FIGURA 2.3: Esempio di download in formato CoNLL generato da *LinguA*

2.4. Le applicazioni del monitoraggio linguistico

L'elevato potere diagnostico delle analisi generate da questi strumenti ha incoraggiato i linguisti computazionali a utilizzare queste nuove tecnologie in diversi ambiti della ricerca. In particolare, numerosi sono stati gli studi condotti sull'acquisizione del linguaggio nel bambino o in pazienti con deficit. Si citano a tal proposito gli studi di Sagae e altri (2005) e Lu (2008) sul monitoraggio dello sviluppo della sintassi nel linguaggio infantile e gli studi condotti da Roarkn e altri (2007) per l'identificazione di deficit cognitivi attraverso misure di complessità sintattica.

Altre possibili applicazioni sono state individuate nell'ambito dello apprendimento della lingua e in particolare per misurare la leggibilità di testi per studenti di L1 e L2 (Heilman e altri, 2007; Collins-Thompson, 2005).

Inoltre, il monitoraggio linguistico condotto su vasti insiemi di documenti, appartenenti a diversi domini linguistici, ha permesso una sempre più dettagliata ricostruzione del profilo linguistico dei testi. In questi ultimi anni, infatti, numerosi sono stati i progetti che hanno sondato i limiti e le possibilità degli strumenti linguistico-computazionali in questo ambito. In questo senso, si sono mossi gli studi di Dell'Oglio, Brunato e Dell'Orletta (2018), sulla variazione del lessico e della sintassi tra generi testuali e varietà di lingua, e di Brunato e Dell'Orletta (2017), sull'ordine dei costituenti, nei quali, ognuno di questi fenomeni linguistici, è stato trattato come indice della complessità lessicale e sintattica del testo.

CAPITOLO III

Lo studio della complessità linguistica

Questo capitolo vuole offrire una panoramica generale sul tema della complessità linguistica, concetto complesso che ancora oggi è oggetto di dibattito e discussione negli studi di linguistica teorica e applicata. La natura articolata del tema ha reso necessario un approccio modulare, basato sui diversi livelli della lingua – fonetico, morfologico, semantico e pragmatico.

Alla fine del capitolo, verranno esposti gli spunti metodologici, suggeriti da Montemagni (2013), che hanno ispirato le misure di complessità linguistica condotte nel presente studio.

3.1. Cenni storici sulla nozione di complessità linguistica

Il concetto di complessità linguistica muove i suoi primi passi verso la fine del XIX secolo, all'interno del cosiddetto "razzismo scientifico", inteso come studio di tecniche e ipotesi a sostegno o giustificazione della fede nel razzismo (Wikipedia, 2019a), e delle categorizzazioni razziali. Questa ideologia si tradusse, negli studi di linguistica, nella volontà di stilare una sorta di classifica delle lingue sulla base della loro evoluzione e dunque della loro complessità.
Mettendo in relazione la complessità linguistica con la complessità del pensiero, infatti, alcuni studiosi tentarono di dimostrare che gli idiomi delle popolazioni sottomesse durante il colonialismo, considerate più primitive, fossero meno evoluti rispetto a quelli dei paesi dominanti. Di conseguenza, le lingue indoeuropee, tipicamente flessive, furono ritenute più adatte all'elaborazione di un pensiero complesso, rispetto a quelle non indoeuropee, in genere isolanti o agglutinanti (Fiorentino, 2009 e Gallissot *e altri*, 2001).
Con la conclusione del secondo conflitto, la ferma denuncia al razzismo scientifico e la rinnovata idea che tutti gli uomini fossero dotati delle stesse capacità cognitive fecero spazio alla teoria che le lingue fossero tutte ugualmente complesse. Ad oggi, anche questa idea è ritenuta obsoleta. È apparso chiaro, infatti, che il concetto di complessità non possa essere considerato in modo assoluto: una lingua, ad esempio, viene percepita più

o meno complessa a seconda della distanza dalla propria lingua madre; oppure, basti pensare che, in prospettiva inter-linguistica, la complessità possa essere relativizzata a livello fonologico, lessicale, sintattico, ecc.
È con questi presupposti che la teoria della complessità linguistica si è fatta di nuovo protagonista del dibattito sulla lingua degli ultimi anni, arrivando a ricoprire il ruolo di categoria esplicativa per interpretare il cambiamento linguistico, per descrivere l'evoluzione del linguaggio umano e per indagare fenomeni connessi all'acquisizione del linguaggio.

3.2. I due approcci al problema della complessità

La natura complessa del linguaggio ha suggerito ai linguisti due modalità di analisi della complessità linguistica: la complessità nel sistema linguistico, ovvero l'insieme dei suoi elementi e delle sue regole, e la complessità per l'utente.

3.2.1. La complessità nel sistema

La complessità può essere definita confrontando sistemi e strutture linguistiche prendendo in esame i criteri interni alle lingue.
Secondo la teoria di Cangelosi e Turner (2002), infatti, un sistema linguistico è caratterizzato da una serie di elementi in grado di interagire tra loro in maniera distribuita, autonoma e gerarchica: i livelli inferiori, come la fonetica, influenzano quelli superiori, come il livello sintattico o lessicale. Tale processo di interazione e auto-organizzazione porta alla formazione della sintassi e permette la comunicazione linguistica tra gruppi di individui. Il linguista, dunque, è chiamato a valutare tale complessità sulla base di criteri come il numero di regole necessarie a produrre un determinato output, il numero di eccezioni alle regole e la mancata trasparenza nella relazione forma-significato.
Come già in (von Humboldt, 1836), il concetto di complessità linguistica viene dunque traslato sul piano morfo-sintattico e associato a tratti linguistici determinanti.
In questo senso si erano mossi alcuni lavori su registri "semplificati", quali il *baby talk*, il *foreigner talk*, il *teacher talk*, come anche i *pidgin* e i *creoli*, idiomi originati dalla mescolanza di lingue di popolazioni differenti.
Su quest'ultimo ambito di ricerca si sono concentrati gli studi di Ferguson (1982) e McWhorter (2001), entrambi autori di una definizione di complessità di tipo comparatistico: per il primo, nel confronto tra due strutture linguistiche, quella che possiede tratti

maggiormente diffusi nelle lingue naturali rispetto all'altra risulterà più semplice e immediata durante il processo di acquisizione del linguaggio; per il secondo, un linguaggio complesso è quello che, a parità di necessità comunicative, contiene più distinzioni fonetiche, morfologiche, sintattiche e semantiche di un altro.
Kusters (2003), in un certo senso, adotta un punto di vista empirico che ben si concilia con la definizione di Ferguson, definendo la complessità linguistica in termini di difficoltà di apprendimento di un linguaggio da parte di un utente *outsider*, inteso come individuo che impara una nuova lingua in età adulta.
Da questo punto di vista, la complessità linguistica può dare voce anche a riflessioni sul piano sociale ed evoluzionistico: è stata infatti avanzata l'ipotesi che le trasformazioni della complessità di un sistema linguistico, in prospettiva filogenetica, facciano parte di schemi universali evolutivi capaci di descrivere il modo in cui le lingue cambino la loro complessità strutturale nel tempo.

3.2.2. La complessità per l'utente

La complessità per l'utente viene valutata in base all'efficienza comunicativa tra i parlanti. Il discorso si sposta sul piano psicolinguistico, chiamando in causa le strategie cognitive attivate dall'utente durante il processo di codifica e decodifica del linguaggio. In questo senso, si considera complesso ciò che richiede uno sforzo cognitivo e una elaborazione maggiore in fase di produzione o di comprensione.
I recenti studi di Hawkins si muovono in questa direzione. Egli sviluppa un quadro coerente in cui l'approccio cognitivista viene applicato alla sintassi dell'inglese, affrontando il concetto di complessità linguistica in termini di costi di *processing* dell'informazione (2004) e di efficienza comunicativa (2009). A tal proposito, individua tre operazioni in grado di assicurare una comunicazione "efficiente":

- **minimizzare i domini**. Al fine di alleggerire la memoria di lavoro dell'utente è necessario concentrare le sequenze connesse sintatticamente e semanticamente, rendendo i domini più brevi possibili;
- **massimizzare il processo *on-line***. L'utente elabora un elemento X più velocemente se le proprietà di X sono enunciate in modo sequenziale e possono essere assegnate a esso man mano che l'elemento viene processato: assegnazioni successive richiedono uno sforzo maggiore, con il rischio di aumentare il margine di errore;
- **minimizzare le forme**. Lo sforzo cognitivo richiesto per l'elaborazione delle forme linguistiche e delle loro proprietà può essere ridotto minimizzandone l'uso e sfruttando, invece, le informazioni extra-linguistiche già presenti nella comunicazione.

Hawkins, inoltre, rifiuta l'idea che la complessità linguistica sia da ricercare nel numero di unità strutturali e regole di una lingua e, anzi, sostiene che alla semplificazione di un livello linguistico corrisponda la complessità su altri livelli (2009).

3.3. La complessità nei vari livelli di descrizione linguistica

Ad ogni modo, effettuare una diagnosi onnicomprensiva di ogni lingua naturale per ottenere una classificazione precisa su una scala di complessità necessiterebbe di una metrica convenzionalmente concordata. Tale consapevolezza ha condotto McWhorter (2001) a proporre una metrica incentrata su fenomeni fonologici e morfologici. L'intuizione di McWhorter è che un'area della grammatica possa essere considerata più complessa rispetto alla stessa area in un'altra grammatica nella misura in cui la prima comprende maggiori distinzioni e regole della seconda.
Vengono riportati di seguito gli spunti proposti da McWhorter e altri per calcolare la complessità per ciascun livello della lingua.

La complessità fonologica. Il grado di complessità fonologica può essere definito in base alla presenza di fonemi marcati[5] nell'inventario fonemico di una lingua e al numero di toni in un sistema tonale, che necessitano il mantenimento di distinzioni rispettivamente intersegmentali e intertonali più sottili.

La complessità morfologica. La complessità morfologica è invece rintracciabile nella flessione. Essa, innanzitutto, è responsabile di processi morfofonologici con ripercussioni anche a livello fonetico. Inoltre, il fatto che non tutti i prodotti della flessione trovino corrispondenti nelle altre grammatiche, come la marca di genere e le declinazioni delle classi nominali, assenti nelle lingue isolanti, è spia di un livello maggiore di complessità.

A sostegno di questa idea citiamo i risultati ottenuti dagli studi di Kusters (2003) che, nel comparare la morfologia flessionale a quella derivazionale, ha registrato una minore complessità della seconda rispetto alla prima: la regola derivazionale di un linguaggio corrisponde solitamente a diverse regole lessicali di un'altra lingua, oltre a regolarizzare il lessico e a semplificare la formazione delle parole.

[5] Il termine è inteso in senso *cross-linguistico* e perciò utilizzato per indicare fonemi meno frequenti nelle lingue naturali.

La complessità semantica. Ad affrontare il discorso sulla complessità semantica è Berruto (1990). Secondo la sua teoria, infatti, la semplificazione di un livello della lingua può comportare un aumento della complessità negli altri livelli: ad esempio alla semplificazione dei tratti sintattici corrisponde un maggiore utilizzo della polisemia, e dunque l'aumento della complessità lessicale. Voghera (2001) seleziona tre tratti sintattici indici di maggiore complessità:

- **il significato astratto**, più complesso del significato concreto, in quanto non percepibile fisicamente;
- **la polisemia**, che facilita il lavoro di produzione del produttore ma che, al contrario, comporta uno sforzo maggiore per il ricevente;
- **il lessico funzionale**, composto da parole vuote quali congiunzioni, articoli, preposizioni, ecc., considerato più complesso del referenziale, costituito da parole piene, quali nomi, aggettivi, verbi, avverbi.

La complessità pragmatica. A costituire questo ulteriore livello della complessità, sono le informazioni desunte dal contesto all'interno di cui viene collocato il discorso. Tali conoscenze sono necessarie per la corretta comunicazione tra gli individui, in particolar modo qualora i fenomeni linguistici in questione risultassero incomprensibili se analizzati dal punto di vista strettamente linguistico, come nel caso dell'ironia e del sarcasmo.

3.3.1. La complessità sintattica

Particolare attenzione riserviamo alla complessità sintattica, sulla quale saranno basate le analisi discusse nei prossimi capitoli.
Per definire gli indici della complessità sintattica, si sceglie di far riferimento a quelli proposti da Berruto e Cerruti (2011):

- l'ordine lineare degli elementi di una frase, che permette la disambiguazione dei significati;
- la discontinuità, intesa come la possibilità che gli elementi legati semanticamente o sintatticamente non siano immediatamente adiacenti;
- le relazioni e le dipendenze fra gli elementi non contigui;
- il grado di incassatura fra gli elementi;
- la ricorsività, ovvero la possibilità di applicare lo stesso procedimento un numero illimitato di volte per ottenere strutture sempre nuove;
- le parti del discorso, che danno informazioni sulla sua struttura interna (per esempio, le congiunzioni coordinanti e subordinanti).

3.3.1.1. L’origine dell'ordine dei costituenti SVO

Secondo studi recenti, condotti da Gell-Mann e Ruhlen (2011), l’origine della maggior parte dei linguaggi umani è identificabile in un unico linguaggio antenato in cui l’ordine dei costituenti era affidato alla formula Soggetto-Oggetto-Verbo (SOV), e non al più canonico Soggetto-Verbo-Oggetto (SVO).

Gli studi di Gibson e altri (2013) hanno suggerito l’idea che l’ordine di tipo SVO si è affermato, in prospettiva interlinguistica, solo dopo un lento processo innescato dalla necessità di ridurre al minimo gli errori causati dal rumore presente nel segnale linguistico: effettivamente, nella teoria dell'informazione di Shannon, matematico statunitense, si pone l’attenzione sul forte impatto che il rumore può avere sulla corretta ricezione del messaggio da parte del destinatario e, conseguentemente, sull’efficacia della comunicazione.

L'ordine SVO è funzionale, ad esempio, in contesti linguistici che descrivono eventi semanticamente reversibili, in cui entrambi i referenti sono animati e potrebbero essere agenti dell'evento espresso dal predicato: si pensi alla frase “Francesca saluta Arianna”, di facile comprensione, e alla sua costruzione secondo l’ordine SOV: “Francesca Arianna saluta”, sicuramente meno immediata e possibile fonte di ambiguità.

3.3.1.2. L’ordine marcato e non marcato degli elementi

Negli studi condotti sull'ordine dei costituenti in una frase, si individuano due scuole di pensiero: se Hawking (1994) affida alla necessità di elaborare velocemente l’informazione l’impostazione dell’ordine dei costituenti, Diessel (2005) sostiene che l'ordine delle parole nella frase dipenda dalla struttura dell'informazione.

Nella lingua italiana, l’ordine tipico, dunque non marcato, è di tipo SVO. Tuttavia, al fine di focalizzare l'attenzione su una parte del discorso o su una informazione in particolare, non è raro che l'ordine degli elementi venga invertito. A tal proposito, Corpina (2009) definisce la marcatezza in base a tre livelli di descrizione linguistica:

- **marcatezza fonologica**: un enunciato è fonologicamente marcato se la melodia intonativa non è rappresentabile secondo una curva, ma presenta interruzioni, pause o picchi intonativi;
- **marcatezza sintattica**: un enunciato è sintatticamente marcato quando i costituenti non occupano le posizioni canoniche al fine di esprimere dei significati particolari; vi è una stretta correlazione tra marcatezza sintattica e fonologica, in quanto a un ordine inusuale degli elementi di un enunciato si accompagna spesso un'intonazione particolare;

- **marcatezza pragmatica**: una frase è marcata pragmaticamente quando si adatta ad un numero ridotto di contesti e situazioni linguistiche.

La lingua italiana concede agli elementi sintattici una certa libertà di movimento, a seconda delle esigenze comunicative. Di seguito, verranno approfonditi alcuni casi esemplari.

Soggetto. Il soggetto occupa solitamente una posizione preverbale. A differenza degli altri argomenti del verbo, inoltre, non è mai sottoposto a tematizzazione mediante dislocazione a sinistra: in italiano, infatti, non esistono forme di clitico soggetto che, oltretutto, non necessita di focalizzazione in quanto di norma svolge già il ruolo di tema (Treccani, 2011). È pur vero che l'italiano permette l'omissione del soggetto, forte indice della libertà di spostamento, e gli concede una certa flessibilità quando è necessario per essere evidenziato come elemento nuovo.
Ad esempio, la collocazione del soggetto in posizione postverbale si può riscontrare negli enunciati tetici, in cui è solo il rema a veicolare informazione nuova:

es: "È arrivato *Antonio*".

Inoltre, nelle frasi scisse, il soggetto viene rematizzato a inizio frase mediante l'utilizzo del verbo essere in funzione di copula e di tratti prosodici

es: "È *il soggetto* che precede il verbo".

Oggetto. A seconda dell'intenzione comunicativa, l'oggetto può essere dislocato a sinistra o a destra del verbo. Nel primo caso, normalmente l'oggetto viene preposto al verbo e ripreso con un pronome clitico:

es: "La cena, la prepara Arianna"

Nel caso della dislocazione a destra, invece, l'oggetto viene collocato dopo il verbo, con un ordine marcato rema-tema:

es: "Arianna non la prepara, la cena".

Modificatori del nome. Gli articoli, i determinanti, i numerali e i quantificatori sono invece costretti in strutture piuttosto rigide, che conservano l'ordine con testa a destra.

Aggettivo. Quando l'aggettivo ha funzione restrittiva, ovvero quando indica una qualità distintiva in grado di distinguere il nome al quale si riferisce da altri della stessa categoria, viene posto dopo il nome, dunque in posizione non marcata:

es: "Una casa *bella*".

Invece, quando l'aggettivo è posto prima del nome, dunque in posizione marcata, ha solitamente funzione descrittiva, ovvero indica una qualità del nome a cui è legato, di solito

accompagnato da una maggiore soggettività di giudizio in chi parla o scrive, una particolare enfasi emotiva o ricercatezza stilistica (Treccani, 2010a):

es: "Una *bella* casa".

Vi sono poi alcuni tipi di aggettivi che conoscono solo un ordine fisso. Gli aggettivi alterati ("un bimbo *piccolino*"), gli aggettivi che reggono un complemento ("una parete *piena di quadri*"), gli aggettivi derivati da un participio ("una macchina *incidentata*"), gli aggettivi che indicano il colore ("una palla *rossa*"), la forma ("una palla *rotonda*") e la nazionalità ("una pietanza *giapponese*") seguono il nome. Al contrario, in posizione prenominale troviamo gli aggettivi possessivi ("la *mia* coinquilina") – che vengono dislocati solo per motivi di focalizzazione ("quello è il piatto *mio*!") – e gli aggettivi usati in senso figurato ("un'*alta* carica").

Avverbi. La collocazione dell'avverbio, sebbene abbastanza libera, dipende dal tipo di elemento che accompagna: segue un verbo composto ("Bolt ha corso velocemente"), o si frappone tra l'ausiliare e il participio passato nel caso di avverbi di tempo (*ancora, appena, finalmente, già, mai, sempre, spesso, subito, talvolta*) e di giudizio (*certamente, forse, neanche, nemmeno, neppure, probabilmente, proprio, sicuramente*):

es: "Non ho *mai* tardato più di mezz'ora".

Si noti che talvolta l'interpretazione dell'enunciato può essere modificata a seconda della posizione dell'avverbio: "Ho *semplicemente* risposto" (*soltanto*, avverbio di tipo limitativo) non ha lo stesso significato di "Ho risposto *semplicemente*" (*in modo semplice*, valore modale).

La posizione degli avverbi focalizzatori (*solo, anche, proprio, soprattutto* ecc.), specializzati nel modificare il focus della frase, varia a seconda del significato che si vuole dare alla frase:

es:"Francesca ascolta *anche* musica rock" / "*Anche* Francesca ascolta musica rock".

Subordinate. Le subordinate costituiscono un caso di studio particolare. Si possono individuare tre linee di pensiero.

La teoria di natura più spiccatamente pragmatica sostiene che la dislocazione a sinistra della subordinata rispetto alla clausola principale abbia il solo scopo di riportare alla memoria informazioni già note al ricevente, in vista dell'introduzione di un nuovo argomento. La seconda teoria porta la firma di Hawkins (1994): la *Performance theory of order and constituency*, i cui principi sono stati già anticipati al PAR 3.2.2, spiega come la subordinata posta dopo la principale renda l'informazione più facilmente processabile. La terza linea di pensiero è proposta da Diessel (2005), il quale sostiene la stretta interazione tra *processing*, pragmatica e semantica nel determinare l'ordine di una frase.

Nell'ambito di questi studi, Fiorentino (2009) ha analizzato un corpus di italiano scritto elettronico composto da 2200 clausole avverbiali di modo finito (concessive, finali, temporali, causali e condizionali). I dati raccolti hanno, innanzitutto, registrato la presenza di subordinate sia posposte che preposte alla principale, con una netta superiorità delle prime sulle seconde, che rappresentano il 67% dei casi individuati.

Tipo di subordinata	% di subordinate posposte
Finale	97,7%
Condizionale	68,6%
Temporale	67,1%
Concessive	63,5%
Causali	58,2%

TABELLA 3.1: Schema riassuntivo della percentuale di subordinate posposte alla principale in base al tipo di subordinata individuate negli studi di Fiorentino (2009)

In TABELLA 3.1, è riportata la percentuale di subordinate posposte alla principale a seconda del tipo di subordinata in esame.
Come si può notare, le clausole finali sono le uniche, con buona approssimazione, a mantenere una posizione fissa, posposta alla reggente; le altre clausole, invece, mostrano una variazione della posizione a seconda delle congiunzioni utilizzate. Lo dimostra il fatto che le subordinate condizionali introdotte da *se* tendono a ricorrere prima della reggente, così come le causali introdotte da *siccome* e le temporali introdotte da *appena*. Al contrario, le proposizioni introdotte da *benché, qualora, sebbene, mentre* possono essere generalmente collocate sia prima, che dopo la principale.

3.4. "Misurare" la complessità linguistica

Come già ampiamente discusso nel CAP. 2 e specificatamente nel PAR 2.3, l'arricchimento del testo con informazioni linguistiche diventa il punto di partenza per l'identificazione di una vasta gamma di parametri rappresentativi di fenomeni dei diversi livelli della lingua, proficuamente sfruttati in compiti di monitoraggio linguistico.

Facendo riferimento agli studi di Montemagni per l'italiano (2013), è stato dimostrato come analizzando la distribuzione delle categorie morfo-sintattiche estratte automaticamente dal testo annotato sia possibile individuare tratti caratterizzanti dei diversi generi testuali, dello scritto e del parlato, così come di varietà semplificate del linguaggio. La possibilità di calcolare la distribuzione delle congiunzioni coordinanti e subordinanti ha permesso, ad esempio, di fornire una stima del rapporto tra costruzioni paratattiche e ipotattiche all'interno dei corpora selezionati e di discriminarli in base a una maggiore o

minore “leggerezza” sintattica. Inoltre, mettendo in relazione la frequenza di occorrenze di certe categorie grammaticali rispetto ad altre è stato possibile misurare la “densità lessicale”, intesa come la proporzione delle parole semanticamente piene rispetto al totale delle occorrenze (Dell’Orletta e Montemagni, 2012), e il rapporto tra diverse categorie morfo-sintattiche, ad esempio tra nomi e verbi. Analizzando la sequenza delle categorie morfo-sintattiche, invece, è possibile rintracciare la presenza di strutture particolari, ricalcando l’approccio seguito da Biber negli studi condotti sulla *register variation* (1988).

Tuttavia, un approccio basato solo sull’informazione morfosintattica non basta a ricavare indicazioni precise sulla struttura sintattica complessiva sottostante al testo in esame, soprattutto in relazione allo studio della complessità linguistica. Ad esempio, informazioni particolarmente utili, quali la categoria morfo-sintattica della testa da cui dipende una clausola e il suo livello di incassamento rispetto alla radice verbale della frase, sono rintracciabili solo nelle caratteristiche strutturali dell’albero sintattico.
In letteratura linguistica, linguistico-computazionale e psicolinguistica, infatti, i parametri atti al monitoraggio di tali caratteristiche rivestono un ruolo centrale nella valutazione della complessità di un testo: Yngve (1960), Frazier (1960) e Gibson (1998), ad esempio, la studiano in relazione alla profondità dell’albero sintattico associato a una frase, Lin (1996) e Gibson (1998) alla lunghezza delle relazioni di dipendenza, intesa come la distanza in parole tra testa e dipendente.

CAPITOLO IV

Risorse e strumenti per lo studio della complessità

Il seguente capitolo intende focalizzarsi sulle risorse testuali utilizzate in questo studio: i corpora. Per ogni collezione di testi verranno specificate informazioni quali la genesi, la tipologia, lo scopo della raccolta, la dimensione e il target di riferimento.

Verrà quindi descritto, pur senza entrare nel dettaglio, il lavoro di implementazione di funzioni in Python, grazie al quale è stato possibile estrarre dai corpora le caratteristiche linguistiche oggetto di indagine.

Inoltre, verrà illustrato il modo in cui i dati linguistici estratti sono stati adattati alle finalità dello studio e, infine, verranno presentati gli algoritmi statistici utilizzati per individuare le *features* significative.

4.1. I corpora analizzati

Il presente studio è stato svolto sull'analisi incrociata di otto corpora appartenenti a quattro generi testuali: proprio dei materiali didattici, giornalistico, narrativo e scientifico. All'interno di ciascun genere sono state selezionate due collezioni di testi appartenenti a due varietà linguistiche differenziate in base al destinatario a cui sono rivolti, ascrivibili perciò a due livelli di complessità linguistica opposti: uno difficile, o complesso, e uno facile, o semplificato.

Tutti i corpora sono stati arricchiti automaticamente con annotazione morfo–sintattica e sintattica utilizzando la catena di analisi linguistica *LinguA* (cfr. PAR 2.3.2).

4.1.1. I corpora dei materiali didattici

All'interno del genere didattico, sono stati scelti come rappresentativi due collezioni di testi inerenti all'ambito scolastico:

Materiali Didattici per la Scuola Superiore (*Sup*). Questo primo sub-corpus è stato creato al fine di rappresentare la varietà "complessa" del genere didattico. Si tratta

di una collezione di testi tratti da materiali didattici rivolti a studenti di scuola superiore. Consta di 70 documenti, di circa 48'000 tokens.

Materiali Didattici per la Scuola Elementare (*Elem*). Per rappresentare la varietà "semplice", si è scelto di affiancare un corpus di testi tratti da sussidiari per bambini frequentanti la scuola elementare. Il corpus in questione comprende 127 documenti, per un totale di circa 48'000 tokens.

4.1.2. I corpora giornalistici

Il genere giornalistico è rappresentato da due corpora monolingua, *La Repubblica*, rappresentativo della varietà complessa, e *Due Parole*, che si colloca al polo opposto della scala di complessità linguistica.

Repubblica (*Rep*). È un'ampia collezione di testi giornalistici tratti dall'omonimo quotidiano "La Repubblica": si tratta di un corpus monolingue, generale per dominio tematico, in quanto comprendente articoli di argomenti diversi, ma specialistico per tipologia testuale. Il corpus, sviluppato presso la SSLMIT dell'Università degli Studi di Bologna, includeva inizialmente le annate dal 1985 al 2000 e contava circa 175 milioni di tokens. Attualmente, però, è in via di ampliamento, arrivando a sfiorare i 400 milioni di tokens. La porzione di corpus oggetto del nostro studio è limitata agli articoli scritti tra gli anni 2000 e 2005, per un totale di circa 232.000 tokens.

Due Parole (*2Par*). Il corpus *Due Parole* comprende gli articoli del mensile "Due Parole"[6], giornale di facile lettura, fondato nel 1983 presso l'Università di Roma "La Sapienza". Il giornale è rivolto ad un pubblico adulto con un basso livello di scolarizzazione o con lievi disabilità intellettuali (Piemontese, 1996) ed è curato da linguisti specializzati in compiti di semplificazione dei testi. I testi sono scritti utilizzando in modo consapevole e sistematico criteri di scrittura *controllata*, seguendo criteri quali la brevità dei testi, la semplicità delle frasi, la scelta di un lessico comune e una curata organizzazione logico-concettuale dei testi. Il corpus qui analizzato raccoglie gli articoli scritti durante gli anni 2001-2006 e contiene circa 73.000 tokens.

[6] `http://www.dueparole.it/`

4.1.3. I corpora narrativi

Per quanto riguarda i testi narrativi, la scelta è ricaduta sulla risorsa descritta in (Brunato *e altri*, 2015), specificatamente sviluppata per gli studi di semplificazione automatica del testo in lingua italiana, creata dall'ItaliaNLP Lab dell'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa.
La risorsa è costituita da due sub-corpora, *Terence* e *Teacher*, rappresentative di due strategie di semplificazione differenti, rispettivamente "strutturale" e "intuitiva".
Entrambi i sub-corpora contengono due versioni dello stesso testo allineate a livello di frase, ovvero la versione originale del testo e la sua versione semplificata manualmente, destinata a categorie specifiche di lettori.

Terence. *Terence*, come *Teacher*, è un corpus allineato che comprende 32 racconti brevi e la rispettiva versione semplificata, rivolta a bambini e bambine udenti e non udenti di età compresa tra i 7 e gli 11 anni, affetti da disturbi di comprensione del testo. Tale collezione è stata creata nell'ambito del *Progetto Terence*, promosso dall'Unione Europea, negli anni che vanno dal 2007 al 2010.
La semplificazione adottata viene definita "strutturale" in quanto gli esperti che hanno svolto il lavoro di semplificazione e riscrittura dei testi hanno seguito una linea guida predefinita, attraversando tre livelli discendenti e sequenziali di difficoltà:

- coerenza globale: esplicitazione delle informazioni necessarie per la comprensione del significato generale;
- coerenza locale: miglioramento della comprensibilità tramite l'introduzione di elementi coesivi tra le frasi;
- lessico e grammatica: semplificazione dei vocaboli utilizzati, della sintassi e delle figure retoriche.

Per potersi concentrare solo sulle categorie lessicali, grammaticali e sintattiche, di cui si interessa la semplificazione automatica, si è scelto di prendere il livello di coerenza locale come l'equivalente del testo originario e il livello di semplificazione lessicale e grammaticale come equivalente del livello semplificato.

Teacher. Nel corpus *Teacher* sono raccolti 24 coppie di testi originali e semplificati provenienti da siti web educativi specializzati nella divulgazione di risorse gratuite per insegnanti, di vario genere testuale.
La strategia di semplificazione qui utilizzata è detta "intuitiva" in quanto ogni testo è stato semplificato da differenti insegnanti in maniera indipendente: in questo modo la semplificazione, svincolata da limitazioni di natura gerarchica e da regole predefinite,

risulta essere spalmata su diversi livelli linguistici. Il lavoro di semplificazione dei testi è stato indirizzato a studenti L2 (approssimabile al livello B2 per la lingua italiana).

Terence&Teacher, versione originale e versione semplifica. Per monitorare l'effetto della complessità linguistica all'interno di questo genere, sfruttando la caratteristica di allineamento dei due corpora paralleli appena presentati, sono stati creati due ulteriori raccolte: un corpus unico di testi narrativi "complessi" (*TT-orig*), contenente solo i testi originari di Terence e Teacher, e un corpus unico di testi narrativi "semplici" (*TT-sempl*), contenente solo le versioni semplificate (Brunato e Dell'Orletta, 2017).

4.1.4. I corpora scientifici

Per la prosa scientifica, sono state selezionate le seguenti collezioni di testi:

Articoli Scientifici (*ArtScient*)**.** Il corpus *Articoli Scientifici*, scelto come rappresentativo della varietà "complessa", comprende 84 documenti e conta complessivamente più di 470'000 tokens. Gli articoli, tratti da riviste scientifiche di settore, coprono vari argomenti e ambiti della ricerca, dai cambiamenti climatici a studi di linguistica.

Wikipedia: Ecologia e Ambiente (*Wiki*). La varietà "semplice" è rappresentata da un corpus di articoli di Wikipedia, per un totale di circa 200'000 tokens. I documenti di cui il corpus si compone sono stati estratti dal portale italiano "Ecologia e Ambiente".

4.2. L'estrazione delle *features* linguistiche di interesse

Attraverso l'implementazione di script in *Python*, sono state estratte dalle frasi dei diversi documenti di ogni corpus un ampio spettro di caratteristiche linguistiche, che sono state poi utilizzate nelle indagini e negli esperimenti descritti nel CAP. 5.
In seguito alle analisi eseguite dal programma, ogni frase risulta descritta da un vettore in cui sono stati inseriti i valori assunti dalle *features* estratte all'interno della frase stessa. I parametri oggetto di monitoraggio vengono di seguito raggruppati e descritti in base a tre categorie alle quali possono essere ricondotti i fenomeni di complessità che intercettano: di base, morfo-sintattiche, sintattiche. Per la visione delle caratteristiche complete si rimanda all'appendice A.

4.2.1. Le caratteristiche di base

In prima analisi, si è voluto studiare la composizione del documento in termini di caratteri e tokens. In particolare, tenendo conto della punteggiatura, sono stati calcolati i seguenti dati:

- lunghezza della frase, in termini di numero di tokens per frase;
- numero di caratteri per frase;
- lunghezza media dei tokens: numero medio di caratteri per tokens.

Inoltre, ad ogni frase è associato un indice, inteso come la posizione che essa occupa nel documento: la prima frase avrà indice 1, l'ultima indicherà il numero totale di frasi del documento.

4.2.2. Le caratteristiche morfo-sintattiche

Nelle caratteristiche morfo-sintattiche, rientrano le distribuzioni delle categorie morfo-sintattiche, o parti del discorso, generiche (*coarse-grained part-of-speech*) e delle categorie morfo-sintattiche più granulari (*fine-grained part-of-speech*) in percentuale, calcolate sul totale di tokens della frase. In particolare, sono state calcolate le distribuzioni degli aggettivi, degli avverbi, delle congiunzioni, delle preposizioni, della punteggiatura, dei numeri, dei pronomi, degli articoli, dei nomi e dei verbi.
Più nel dettaglio sono state indagate le distribuzioni dei nomi propri e comuni e dei verbi principali, ausiliari e modali, e ancora: la percentuale di pronomi dimostrativi, personali, indefiniti, possessivi, interrogativi, relativi e clitici sul totale di pronomi; di congiunzioni subordinanti e coordinanti sul totale di congiunzioni; di articoli determinativi e indeterminativi sul totale di articoli.

4.2.3. Le caratteristiche sintattiche

Il gruppo delle caratteristiche sintattiche è sicuramente il più numeroso e interessante per lo studio della complessità del testo.

La distanza testa-dipendente. Innanzitutto, è stata estrapolata la lunghezza media dei link sintattici, ovvero la media delle distanze tra testa e dipendente calcolata in numero di token (escludendo la punteggiatura), la massima distanza dalla testa sintattica registrata nella frase analizzata e la percentuale di parole che precedono o che seguono la propria testa sintattica.

Per poter ottenere informazioni di questo tipo viene innanzitutto calcolata la distanza del token dalla sua testa sintattica, sfruttando l'id ad esso associato (`tok.id`) e l'id della testa da cui dipende (`tok.head`), riportato nella colonna TESTA nella TABELLA 2.1.

```
distanzaTesta = tok.id - tokens[int(tok.head)-1].id
```

La distanza è di segno positivo quando il token segue la testa da cui dipende; viceversa, quando il token precede la testa da cui dipende, è di segno negativo. Questa considerazione ha permesso di "smistare" i tokens in base alla posizione occupata rispetto alla propria testa sintattica. Le distanze prese in valore assoluto sono state utilizzate per calcolare la distanza media e per individuare la massima distanza registrata nella frase.
Lo studio sul segno della distanza è stato utilizzato anche in altri contesti: una volta estratte le distribuzioni dei diversi tipi di relazioni di dipendenza (in percentuale sul totale di parole della frase), sono state calcolate la percentuale di soggetti, di oggetti e di avverbi in posizione preverbale e postverbale, come anche la percentuale di aggettivi prenominali e postnominali.

La struttura dell'albero sintattico. Sono state estratte diverse caratteristiche riguardanti la struttura dell'albero sintattico.
Per visualizzare i parametri estratti e le strategie utilizzate per estrapolarli dal dato testuale, viene proposta la rappresentazione a costituenti di un albero sintattico, di cui era stata già proposta la rappresentazione a dipendenze in FIGURA 2.2.

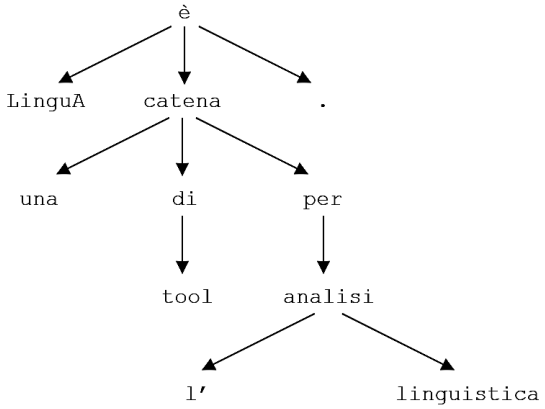

FIGURA 4.1: Rappresentazione a costituenti della frase
"LinguA è una catena di tool per l'analisi linguistica."

In primis, è stata calcolata l'altezza dell'albero sintattico, intesa come la massima distanza che intercorre tra la radice dell'albero e una foglia (parola del testo senza dipendenti), ed è espressa come numero di archi (relazioni di dipendenza) attraversati nel cammino radice-foglia, e la massima ampiezza dell'albero, intesa come il maggior numero di nodi, rappresentanti le parole del periodo, posizionati al medesimo livello.

Per quanto riguarda le dipendenze, invece, attraverso un controllo iterato sui figli di ogni nodo, è stato calcolato il numero medio di dipendenti per tokens, di dipendenti per teste verbali e di dipendenti per teste nominali, escludendo la punteggiatura.

Infine, è stato approfondito il fenomeno della subordinazione. In particolare, sono state estratte le seguenti caratteristiche linguistiche:

- il numero assoluto di subordinate per frase;
- la media delle altezze e delle ampiezze degli alberi associati alle subordinate della frase;
- la percentuale di subordinate di primo grado sul totale di subordinate;
- la media delle altezze e delle ampiezze degli alberi rappresentanti le subordinate di primo grado della frase;
- la percentuale di subordinate di primo grado che precedono e che seguono la principale da cui dipendono, adattando lo studio sul segno della distanza testa-dipendente descritto poc'anzi;
- la percentuale di subordinate di grado superiore al primo sul totale di subordinate, ripetendo su queste ultime tutte le analisi condotte su quelle di primo grado: media delle altezze, delle ampiezze, e calcolo della percentuale di subordinate preposte o posposte alla subordinata reggente.

Per svolgere uno studio di questo tipo, sono stati innanzitutto individuati i verbi, i nomi e i pronomi, ovvero le categorie morfo-sintattiche in grado di introdurre una subordinata. Un controllo sui figli individua quelli che intrattengono con la testa una relazione di dipendenza di tipo *argument*, *modifier*, *temporal modifier* o *relative modifier* e che, al contempo, siano preposizione, congiunzione subordinante o verbo. Con un controllo di questo tipo si riescono a selezionare i seguenti tipi di costrutti:

- V → E → V Es.: "Ha [detto] [di] [arrivare] presto."
- V → CS → V Es.: "Ha [detto] [che] [usciva]."
- V → V Es.: "[Desiderava dormire]."
- N → E → V Es.: "È giunto il [momento di andare]."
- V → CS → V Es.: "Hai [parlato] [mentre] [era] meglio tacere."
- V → E → V Es.: "[Per arrivare] in tempo, Gianni [uscì] presto."
- V → V Es.: "[Uscendo] dalla stanza, il ragazzo [salutò]."
- S → V Es.: "Non conosceva il [ragazzo] che la [chiamò]."
- P → V Es.: "Non si è mai accertato [chi] [volle] la sua incarcerazione."

Per estrarre parametri quali l'altezza o l'ampiezza dell'albero sintattico generato dalla subordinata, sono state utilizzate le stesse funzioni implementate per lo studio dell'albero

sintattico generato dalla frase, passando come parametro non la radice dell'albero, bensì la preposizione, la congiunzione subordinante o il verbo reggente la subordinata.
Per riconoscere e distinguere le subordinate di primo grado dalle subordinate di grado superiore, si è scelto di prendere in esame l'elemento reggente la subordinata: nel caso fosse inserito nella clausola principale, la subordinata veniva classificata come di primo grado. Al contrario, se l'elemento in questione è già inserito a sua volta in una subordinata, la subordinata da lui retta ha sicuramente grado di subordinazione maggiore di 1.

4.3. La preparazione degli esperimenti

L'operazione di estrazione dei parametri di monitoraggio linguistico (cfr. PAR. 4.2) ha condotto alla creazione di una tabella per ogni documento, in cui ogni riga corrisponde al vettore di caratteristiche linguistiche di una frase. La tabella, esemplificata in FIGURA 4.2, avrà, dunque, tante righe quante sono le frasi del documento, se consideriamo la prima come intestazione, e tante colonne quante sono le features estratte, nel nostro caso 88 (escludendo il parametro `Sentences`, indice della frase nel testo).

	A	B	C	D	E	F	G	H	I	J
1	Sentences	Tokens	Chars	mCxT	CPOS_A(%)	CPOS_B(%)	CPOS_C(%)	POS_CC*(%)	POS_CS*(%)	CPOS_D(%)
2	1	3	15	5	33,33	0	0	0	0	0
3	2	8	34	4,25	0	0	0	0	0	0
4	3	33	151	4,58	6,06	0	0	0	0	0
5	4	30	161	5,37	3,33	6,67	0	0	0	0
6	5	55	247	4,49	0	1,82	1,82	0	100	0
7	6	51	226	4,43	1,96	1,96	5,88	33,33	66,67	0
8	7	13	67	5,15	7,69	0	0	0	0	0
9	8	14	61	4,36	7,14	7,14	0	0	0	7,14
10	9	26	98	3,77	0	7,69	0	0	0	0
11	10	15	58	3,87	6,67	6,67	6,67	0	100	0

FIGURA 4.2: Struttura di un file Excel generato a partire da un documento in formato CoNLL lungo 10 frasi. Sono mostrati solo i primi 10 parametri estratti.

4.3.1. Il raggruppamento in fasce

L'ipotesi di ricerca iniziale aspira a mettere in luce l'esistenza di *features* il cui andamento nel testo avrebbe in qualche modo caratterizzato i documenti di una varietà linguistica particolare. Per poter ricercare questi andamenti tipici, è stato necessario rendere omogeneo il dato di partenza, la cui lunghezza sarebbe stata altrimenti enormemente variabile, avendo a che fare con documenti di 1 sola frase, fino a documenti di 644 frasi. Si è deciso quindi di suddividere ogni tabella in 6 fasce, in cui sono state "inglobate" un certo numero di righe, a seconda della lunghezza del documento stesso. Ogni fascia è stata popolata con le medie dei valori delle righe che ha inglobato.

Una scelta di questo tipo ha reso inutilizzabili i documenti contenenti meno di 6 frasi, che sono stati per questo esclusi: il numero di documenti "utili", a fronte del numero di documenti disponibili per corpus, è stato riportato in TABELLA 4.1. Per esempio, una tabella di 5 righe, generata dall'analisi di un documento di 5 frasi, infatti, avrebbe prodotto una nuova tabella di 6 righe, una per fascia, ognuna di queste popolata con i valori di una riga della precedente tabella, eccetto l'ultima, che sarebbe rimasta vuota.

GENERE	CORPUS	N. DOC.	N. DOC. UTILI
DIDATTICO	*Materiali didattici per la Scuola Elementare*	60	52
	Materiali didattici per la Scuola Superiore	70	69
GIORNALISTICO	*DueParole*	321	303
	Repubblica	318	304
NARRATIVO	*Terence&Teacher – versione semplificata*	56	54
	Terence&Teacher – versione originale	56	53
SCIENTIFICO	*Wikipedia: Ecologia e Ambiente*	293	249
	Articoli scientifici su argomenti specialistici	84	84

TABELLA 4.1: Schema dei corpora utilizzati nello studio in cui viene specificato il numero di documenti disponibili per corpus e il numero di documenti effettivamente analizzati.

La tabella in FIGURA 4.2, invece, poiché contiene un numero di righe maggiore di sei, ne genera una seconda, in cui ogni fascia è correttamente riempita:

	A	B	C	D	E	F	G	H	I	J
1	Sentences	Tokens	Chars	mCxT	CPOS_A(%)	CPOS_B(%)	CPOS_C(%)	POS_CC*(%)	POS_CS*(%)	CPOS_D(%)
2	1,5	5,5	24,5	4,63	16,66	0	0	0	0	0
3	3,5	31,5	156	4,97	4,7	3,33	0	0	0	0
4	5,5	53	236,5	4,46	0,98	1,89	3,85	16,66	83,34	0
5	7,5	13,5	64	4,76	7,42	3,57	0	0	0	3,57
6	9	26	98	3,77	0	7,69	0	0	0	0
7	10	15	58	3,87	6,67	6,67	6,67	0	100	0

FIGURA 4.3: Tabella generata dall'operazione di raggruppamento in fasce a partire dalla tabella in FIGURA 4.2.

Per ottenere questo risultato, è stato necessario innanzitutto definire il numero di righe da raggruppare in una fascia, dividendo il numero di righe della tabella per il numero delle fasce:

```
numFasce = 6
fascia = int(round(righe*1.0/numFasce)) # righe per fascia
```

Si procede, poi, alla creazione di due variabili:

```
x = 0 # limite inferiore
step = fascia # limite superiore
```

Queste verranno utilizzate per individuare nella tabella il *range* di righe su cui effettuare la media, utilizzando il metodo `iloc` fornito da *Pandas*[7]:

```
media = df.iloc[x:step].mean()
```

Inoltre, verranno utilizzate per avanzare nella tabella, insieme alle variabili:

```
fasceVuote = numFasce
righeRestanti = righe
```

La variabile `fasceVuote` tiene il conto delle fasce ancora da popolare: inizialmente, sono tutte vuote, dunque le si assegna il valore di `numFasce`. La variabile `righeRestanti` conta le righe della tabella d'origine che debbono ancora essere processate: le viene per questo assegnato, come primo valore, il numero delle righe totali della tabella.
Le due variabili `fasceVuote` e `righeRestanti` vengono sfruttate per valutare, di volta in volta, se dovesse essere necessario modificare il numero di righe per fascia definito in origine, al fine di popolare correttamente ognuna delle sei fasce.
Tale controllo è necessario in quanto il numero di frasi di un documento potrebbe non essere divisibile per il numero di fasce stabilito, generando un "resto", ovvero un certo numero di frasi che rimarrebbero escluse dall'operazione di raggruppamento.

Il programma, quindi, itera per il numero di fasce da popolare le seguenti operazioni:

I. controlla se il rapporto tra il numero di righe ancora da accorpare e il numero di fasce da riempire è uguale al numero di righe per fascia stabilito;

```
if math.ceil(righeRestanti*1.0/fasceVuote) == fascia:
```

II. se il controllo dà esito positivo: crea una riga popolata dalle medie dei valori delle righe inglobate e la inserisce nella nuova tabella; dunque incrementa il limite inferiore `x`, il limite superiore `step`, decrementa il numero di fasce ancora da riempire e sottrae dal numero di righe ancora da accorpare il numero di righe appena accorpate:

```
x = step
step = x + fascia
fasceVuote -= 1
righeRestanti = righe - x
```

[7] La libreria *Pandas* fornisce strutture e strumenti per l'analisi di dati in linguaggio *Python*. Tutta la documentazione in merito è fornita dal sito ufficiale `http://pandas.pydata.org/`.

III. se il controllo dà esito negativo: assegna alla variabile fascia il quoziente del rapporto tra il numero di righe ancora da accorpare e il numero di fasce ancora da riempire, sposta il limite superiore della "finestra" in base al nuovo valore assunto dalla variabile fascia. A questo punto, svolge le operazioni illustrate al punto II.

```
fascia = righeRestanti/fasceVuote
step = x + fascia
```

Questa operazione "tradurrà" le tabelle di dimensioni variabili prodotte durante il processo di estrazione delle *features* in altrettante tabelle di uguale dimensione: 6 righe x 89 colonne.

4.3.2. La normalizzazione dei dati

Per monitorare l'andamento dei parametri linguistici estratti, talvolta si è reso necessario prescindere dalla forma e dall'unità di misura del dato che avrebbe rischiato di fuorviare gli algoritmi statistici ai quali i dati sono stati sottoposti. Si è scelto, pertanto, di applicare ad ogni lista di *features* una funzione di normalizzazione, in modo da scalare ogni valore nell'intervallo [0,1]. Tale funzione divide i valori `x` di ogni colonna per il massimo di questi, `x.max()`:

```
lambda x: x/x.max()
```

Le tabelle generate da questa operazione sono esemplificate in FIGURA 4.4.

	A	B	C	D	E	F	G	H	I	J
1	Sentences	Tokens	Chars	mCxT	CPOS_A(%)	CPOS_B(%)	CPOS_C(%)	POS_CC*(%)	POS_CS*(%)	CPOS_D(%)
2	0,15	0,1	0,1	0,93	1	0	0	0	0	0
3	0,35	0,59	0,66	1	0,28	0,43	0	0	0	0
4	0,55	1	1	0,9	0,06	0,25	0,58	1	0,83	0
5	0,75	0,25	0,27	0,96	0,45	0,46	0	0	0	1
6	0,9	0,49	0,41	0,76	0	1	0	0	0	0
7	1	0,28	0,25	0,78	0,4	0,87	1	0	1	0

FIGURA 4.4: Tabella generata dall'operazione di normalizzazione effettuata sulla tabella in FIGURA 4.3.

4.4. Gli algoritmi statistici di correlazione

Prima di procedere con la descrizione degli esperimenti effettuati, è necessario introdurre le misure statistiche utilizzate per analizzare e valutare la significatività dei parametri estratti.

4.4.1. Wilcoxon rank-sum test

Il *Wilcoxon rank-sum test*, o della somma dei ranghi, o test di omogneità di *Mann-Whitney-Wilcoxon*, o *Mann-Whitney U test*, è un test statistico non parametrico per due campioni indipendenti di dimensioni differenti.

I metodi non parametrici sono adatti in quei casi per i quali non sia possibile utilizzare l'analisi della varianza. Le tecniche di analisi della varianza (incluso il *test t*) sono basate sull'assunto che le osservazioni siano tratte da popolazioni di valori normalmente distribuiti e con la stessa varianza, condizione non sempre soddisfatta, nemmeno in maniera approssimativa. Inoltre i metodi non parametrici sono adatti per l'analisi di variabili misurate su scale ordinali. Questo test non richiede nessuna ipotesi sulla distribuzione della popolazione, ed è adatto anche per l'analisi di dati misurati su scale ordinali (Crasta, 2007: 15).
Il test viene utilizzato per verificare se due campioni indipendenti provengono dalla stessa popolazione quando, per le variabili studiate, si raggiunge almeno un livello di misurazione ordinale.

Si suppone di avere un campione $X_1, \ldots, X_{n_P}$ di ampiezza n_P tratto dalla popolazione X e un campione $Y_1, \ldots, Y_{n_G}$ di ampiezza n_G tratto dalla popolazione Y, con $n_P \leq n_G$.
L'ipotesi nulla H_0 considera identica la distribuzione delle due popolazioni X e Y.
Si ordinano le $n = n_P + n_G$ osservazioni in ordine di grandezza crescente, e si associa a ciascuna di esse il proprio numero d'ordine, detto rango. Si sommano poi, separatamente, i numeri d'ordine associati alle osservazioni X, che definiamo T_X, e alle osservazioni Y, T_Y, con l'accortezza – nel caso di due o più osservazioni coincidenti – di assegnare un numero d'ordine pari alla media dei numeri d'ordine di tali osservazioni.
È chiaro che, se le distribuzioni X e Y coincidono, bisogna aspettarsi che le somme dei ranghi siano "vicine", in quanto i due campioni hanno la stessa numerosità.
La loro ulteriore somma è pari alla somma dei primi n numeri interi:

$$T_X + T_Y = \frac{n(n+1)}{2}$$

Se H_0 fosse vera si osserverà in media l'uguaglianza tra T_X e T_Y; se invece la somma dei ranghi di un gruppo è molto più grande o più piccola dell'altra, allora è legittimo pensare che i campioni non appartengano alla stessa popolazione.

Si procede, dunque, al calcolo della statistica W per il primo e per il secondo campione:

$$W_X = T_X - \frac{n_P(n_P+1)}{2} \qquad W_Y = T_Y - \frac{n_G(n_G+1)}{2}$$

Per rifiutare o accettare l'ipotesi nulla si fa fede al *p-value* del test.

Il *p-value* rappresenta la probabilità di sbagliare affermando che il test ha rilevato una differenza reale fra i campioni. Quindi, un basso valore del *p-value* indica che il test ha rilevato una differenza reale fra i campioni. Viceversa, un valore alto del *p-value* indica che il test non ha rilevato differenze significative fra i campioni. Generalmente:

- se $p\text{-}value \geq 0.05$, non c'è significatività statistica;
- se $p\text{-}value < 0.05$, c'è significatività statistica;
- se $p\text{-}value \leq 0.01$, la variazione è *molto* significativa;
- se $p\text{-}value \leq 0.001$, la variazione è *estremamente* significativa.

Il *p-value* può anche essere interpretato come il più piccolo livello di significatività α per il quale i dati osservati indicano che l'ipotesi nulla dovrebbe essere rifiutata.

4.4.2. Pearson Correlation

Il *coefficiente di correlazione di Pearson* è un indice che esprime in che misura c'è una relazione di linearità tra due variabili statistiche.

Date due variabili statistiche X e Y, l'indice di correlazione ρ_{XY} è dato dal rapporto tra la covarianza tra le due variabili e il prodotto tra le deviazioni standard delle due variabili:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

dove σ_{XY} è la covarianza tra le due variabili, mentre σ_X e σ_Y sono le deviazioni standard delle due variabili.

4.4.3. Spearman Correlation

Il coefficiente di correlazione di Spearman è una misura statistica non parametrica di correlazione. Viene definito come il coefficiente di correlazione di Pearson tra il rank delle variabili ed è utilizzato per calcolare la correlazione fra due variabili X e Y quando le stesse hanno una distribuzione che non risulta normale.

Assegnato un campione di dimensione n, i valori X_i e Y_i sono convertiti in rank, che denotiamo con $rg(X_i)$ e $rg(Y_i)$; definiamo coefficiente di correlazione di Spearman come il rapporto tra la covarianza del rank delle due variabili e il prodotto tra le deviazioni standard del rank di queste:

$$\rho_{rg(X)\,rg(Y)} = \frac{\sigma_{rg(X)\,rg(Y)}}{\sigma_{rg(X)}\,\sigma_{rg(Y)}}$$

Il valore dell'indice di entrambe le correlazioni è sempre compreso tra -1 e 1:

- se $\rho = 0$, le due variabili sono dette incorrelate, ovvero non esiste correlazione tra esse;
- se $\rho > 0$, esse sono dette direttamente o positivamente correlate;
- se $\rho < 0$, esse sono dette inversamente o negativamente correlate.

Il coefficiente assume valori tra -1 e 1, indicando nel segno e nel valore il tipo e la forza della correlazione. Si dice che si ha correlazione positiva perfetta quando $\rho = 1$, e correlazione negativa perfetta quando $\rho =$ -1 (Ross, 2014).

Definiamo come riferimento i seguenti *range* di valori:

- se $0 < \rho_{XY} < 0.3$, la correlazione è molto debole, non rilevante;
- se $0.3 \leq \rho_{XY} < 0.6$, la correlazione è moderata;
- se $0.6 \leq \rho_{XY} < 1$, la correlazione è molto forte.

La correlazione negativa è gestita in modo analogo: quanto più un valore è vicino all'estremo superiore 1 o a quello inferiore -1, tanto più la correlazione è forte.

CAPITOLO V

Gli esperimenti e l'analisi dei risultati

Le analisi descritte in questo capitolo sono state condotte al fine di monitorare l'andamento di fenomeni indicativi di complessità linguistica all'interno del testo nelle diverse varietà di lingua.

Dopo aver fornito una panoramica dei dati generali di ciascun corpus, si entrerà nel vivo degli studi effettuati. Sono stati svolti due tipi di esperimenti: il primo tipo di esperimento concentra la sua indagine entro i confini del corpus stesso e ha come obiettivo la ricostruzione del profilo linguistico dei testi ad esso appartenenti; la seconda tipologia, invece, effettua una serie di confronti tra la varietà semplice e complessa di ogni genere testuale al fine di rintracciare differenze e similarità tra questi. Entrambe le tipologie sondano il testo sia nella sua interezza che rispetto alla divisione in fasce.

Per ogni esperimento, verranno descritti il procedimento seguito nella sua realizzazione e i risultati ottenuti più interessanti, per i quali verranno avanzate delle possibili interpretazioni.

5.1. Una panoramica dei dati

Si vuole, innanzitutto, tratteggiare il profilo di base degli otto corpora in esame. Nella TABELLA 5.1, per ciascuna raccolta di testi, verranno indicate le seguenti grandezze:

- lunghezza media dei documenti, ovvero il numero medio di frasi per documento;
- numero medio di frasi per fascia, che altro non è, approssimativamente, che la lunghezza media dei documenti divisa per il numero di fasce. Il valore è comunque interessante in quanto evidenzia il variare della "portata" delle fasce tra i vari corpora;
- lunghezza media delle frasi, ovvero il numero medio di parole per frase;
- lunghezza media delle parole, ovvero il numero medio di caratteri per parola.

GENERE	CORPUS	MEDIA FRASI/DOC.	MEDIA FRASI/FASCIA	MEDIA TOK./FRASE	MEDIA CAR./TOKENS
DIDATTICO	*Elem*	21,0	3,5	20,5	4,3
	Sup	23,4	3,9	29,7	4,7
GIORNALISTICO	*2Par*	12,7	2,1	18,6	4,5
	Rep	30,5	5,1	24,9	4,6
NARRATIVO	*TT-sempl*	25,5	4,3	17,4	4,3
	TT-orig	25,3	4,2	19,3	4,3
SCIENTIFICO	*Wiki*	29,1	4,9	27,7	5,0
	ArtScient	215,6	35,9	26,1	5,0

TABELLA 5.1: Parametri del profilo di base degli otto corpora

Appare subito evidente che i parametri assumano valori tendenzialmente più alti in relazione ai corpora di varietà complessa. Questa divergenza si appiana per quanto riguarda la lunghezza delle parole, che si mantiene pressoché invariata nelle due varietà di lingua in ogni genere. È possibile notare un reale scarto tra i due livelli di complessità linguistica solo nel genere dei materiali didattici, giustificato dall'utilizzo di un linguaggio di dominio, caratterizzato da lessico specialistico tipicamente composto da parole più lunghe.

In generale, comunque, le statistiche rispecchiano la preferenza nei documenti complessi a utilizzare frasi più lunghe e con più parole rispetto ai corrispettivi semplici, anche rimanendo all'interno di uno stesso genere.
Caso particolare è quello relativo al genere scientifico: i testi di *Articoli Scientifici*, infatti, hanno frasi mediamente più corte rispetto a quelle dei testi di *Wikipedia*. Una plausibile spiegazione del dato può essere rintracciata nelle esigenze di chiarezza e rigore proprie del genere scientifico, le quali si traducono in frasi non più lunghe del necessario.

Rispetto ai corpora paralleli originale e semplificato di *Terence&Teacher*, si registrano valori molto vicini in ogni categoria analizzata, un risultato sicuramente da attribuire alle modalità strettamente controllate con cui la semplificazione è stata eseguita e al fatto che i corpora in questione siano paralleli e comparabili.

5.2. Esperimento n.1: sulla variabilità delle features linguistiche

I primi due esperimenti descritti di seguito non tengono conto della suddivisione in fasce, ma considerano il documento nella sua interezza. In particolare, il primo esperimento è

finalizzato alla valutazione del tasso di variabilità di ciascuna feature linguistica all'interno del corpus in esame. I parametri che attestano un valore vicino allo zero di deviazione standard saranno considerati come caratterizzanti; al contrario, varianze molto alte indicheranno una maggiore variabilità del parametro all'interno della varietà linguistica in esame e quindi non potranno essere considerati caratterizzanti di questa.

5.2.1. Il procedimento

Per ogni corpus, è stata creata una tabella che chiameremo "totale", concatenando le tabelle generate dall'operazione di estrazione dei parametri di monitoraggio linguistico per ogni documento, come descritto nel PAR. 4.3.
Ogni tabella "totale" è stata quindi sottoposta ad un processo di normalizzazione, necessario per rendere confrontabili i valori delle features estratte, di natura estremamente diversa: è stata applicata alla lista di valori di ogni *features*, una funzione di normalizzazione (cfr PAR. 4.3.2) al fine di scalare tutti i valori nell'intervallo [0,1], che si tratti di valori assoluti, di medie o di percentuali, lunghezze o distribuzioni. Il valore più alto di ogni *feature* è stato sostituito dal valore 1, il più basso dal valore 0, i valori intermedi dal rapporto tra il valore stesso e il valore massimo registrato per quella *feature*.

In seguito a questa operazione, per ogni *features* è stata calcolata la media aritmetica per fornire una previsione del valore più probabile che si ottiene in un gran numero di prove. Inoltre, per avere indicazione di quanto un generico valore della variabile possa differire dal suo valore medio e per misurare il grado di variabilità di una distribuzione, sono state calcolate la varianza e la deviazione standard, delle quali viene di seguito fornita la definizione.

Data una variabile casuale X, si definisce *scarto* la differenza fra un qualsiasi valore x_i della variabile e il valor medio μ.

$$(x_i - \mu\,) = x'$$

Poiché molte di queste quantità sono negative, viene considerato il quadrato per ottenere un valore indipendente dal segno: viene dunque introdotta la *varianza* σ^2.

$$\sum_i p_i\,(x_i - \mu\,)^2 = \sigma^2$$

Questo valore identifica la dispersione dei valori della variabile X attorno al valor medio. Se i valori x_i appaiono più volte nelle osservazioni con frequenze f_i diverse, si può dire che le probabilità p_i sono ottenute come $p_i = {}^{f_i}/_{n}$ e la formula della varianza espressa come:

$$\sum_i \frac{f_i\,(x_i - \mu\,)^2}{n} = \sigma^2$$

Poiché la varianza è una quantità di secondo grado, si preferisce, quindi, usare la *deviazione standard* o *scarto quadratico medio*: $\sigma = \sqrt{\sigma^2}$.

In generale, la varianza è uguale a zero quanto tutti i valori della variabile sono uguali e quindi non c'è variabilità nella distribuzione; è comunque sempre positiva. Quanto più i valori di X sono dispersi, tanto maggiore è la varianza; viceversa, essa è tanto minore quanto più i valori sono concentrati attorno al valore medio.

5.2.2. I risultati

Di seguito, verranno esposti i risultati dell'esperimento attraverso un confronto tra la varietà di lingua semplice e complessa, per ogni genere. Ogni set di dati è stato ordinato secondo il valore della deviazione standard, in ordine crescente. Per evidenziare le caratteristiche linguistiche che, all'interno del corpus, si sono mantenute relativamente stabili, si è scelto di mostrare le prime quindici *features*, escludendo dal conteggio quelle per cui la media ha assunto valori approssimabili allo zero, riportate in tabella con una barra. Si è scelto di non eliminare tali parametri in quanto la loro assenza, nel corpus, può essere ugualmente considerata come caratteristica del corpus stesso. Lo stesso procedimento è stato seguito per le *features* che variano di più.

Inoltre, accanto ad ogni parametro estratto dal corpus "semplice" è stato riportato il *rank*. Tale dato è stato sfruttato per un confronto diretto con il corpus "complesso" dello stesso genere.

Il genere didattico. Affrontiamo per prima l'analisi del genere dei materiali didattico, mettendo a confronto i risultati relativi al corpus *Materiali didattici per la Scuola Elementari* (*Elem*) e al corpus *Materiali didattici per la Scuola Superiore* (*Sup*).

Per entrambi i corpora in esame, la deviazione standard più bassa alta si stabilisce intorno allo 0,05, la più alta intorno allo 0,3. Inoltre, nel corpus di varietà semplice si può notare un numero maggiore di caratteristiche linguistiche con una deviazione standard maggiore di 0,2: questa osservazione suggerisce che nel corpus *Elem* le *features* varino in misura maggiore, indice di una minore omogeneità interna.

		Elem			*Sup*
rank	features	dev.stand.	rank	features	dev.stand.
1	mChildren	0,06	1	mChildren	0,05
~~2~~	~~POS_PP*(%)~~	~~0,08~~	~~2~~	~~POS_PP*(%)~~	~~0,05~~
~~3~~	~~subMinorPre(%)~~	~~0,08~~	5	mCxT	0,07
~~4~~	~~DEP_concat(%)~~	~~0,08~~	~~10~~	~~POS_SA(%)~~	~~0,08~~
5	mCxT	0,09	~~3~~	~~subMinorPre(%)~~	~~0,09~~
~~6~~	~~POS_PQ*(%)~~	~~0,09~~	27	DEP_modal(%)	0,09
7	LinkPost(%)	0,12	~~65~~	~~subMainPre(%)~~	~~0,09~~
8	mDist	0,12	~~6~~	~~POS_PQ*(%)~~	~~0,1~~
~~9~~	~~POS_PD*(%)~~	~~0,12~~	7	LinkPost(%)	0,1
~~10~~	~~POS_SA(%)~~	~~0,13~~	8	mDist	0,1
11	mWeight	0,13	17	DEP_punc(%)	0,1
12	mChildrenV	0,13	21	CPOS_F(%)	0,1
13	mChildrenS	0,13	~~28~~	~~DEP_mod_loc(%)~~	~~0,1~~
14	mHeight	0,13	11	mWeight	0,11
~~15~~	~~POS_VM(%)~~	~~0,13~~	12	mChildrenV	0,11
16	LinkPre(%)	0,13	29	DEP_ROOT(%)	0,11
17	DEP_punc(%)	0,14	52	DEP_pred(%)	0,11
18	CPOS_N(%)	0,14	66	DEP_comp_loc(%)	0,11
~~19~~	~~DEP_disj(%)~~	~~0,14~~	13	mChildrenS	0,12
20	CPOS_V(%)	0,14	14	mHeight	0,12
21	CPOS_F(%)	0,15	15	POS_VM(%)	0,12
22	CPOS_S(%)	0,15	16	LinkPre(%)	0,12
23	DEP_comp_temp(%)	0,15	18	CPOS_N(%)	0,12
…	…	…	…	…	…
75	DEP_arg(%)	0,23	70	CPOS_B(%)	0,19
76	POS_RI*(%)	0,23	76	POS_RI*(%)	0,19
77	POS_VA(%)	0,24	81	advPost(%)	0,19
78	POS_PI*(%)	0,25	60	POS_PR*(%)	0,2
79	subMinorPost(%)	0,26	26	POS_PE*(%)	0,22
80	POS_RD*(%)	0,27	82	POS_PC*(%)	0,22
81	advPost(%)	0,29	80	POS_RD*(%)	0,23
82	POS_PC*(%)	0,3	85	subMain(%)	0,25
83	advPre(%)	0,3	84	subjPre(%)	0,27
84	subjPre(%)	0,31	86	adjPost(%)	0,27
85	subMain(%)	0,32	89	subMainPost(%)	0,27
86	adjPost(%)	0,32	83	advPre(%)	0,28
87	POS_CC*(%)	0,32	87	POS_CC*(%)	0,28
88	objPost(%)	0,32	88	objPost(%)	0,28
89	subMainPost(%)	0,33	79	subMinorPost(%)	0,3

TABELLA 5.2: Ordinamento delle *features* monitorate rispetto al valore crescente della deviazione standard nei corpora *Elem* e *Sup*.

È interessante notare come, in entrambe le varietà linguistiche del genere didattico, siano le caratteristiche sintattiche a mantenersi particolarmente stabili: il numero di dipendenti per testa sintattica (`mChildren`), sia che questa rappresenti un token generico, un verbo (`mChildrenV`) o un sostantivo (`mChildrenS`), e ancora, la distanza lineare media tra testa e dipendente (`mDist`), la distribuzione in percentuale dei link sintattici verso entrambe le direzioni, come anche l'altezza e l'ampiezza degli alberi sintattici generati dalle frasi, sono parametri che in entrambi i corpora si attestano nelle prime 20 posizioni. Elementi di contrasto, invece, sono rappresentati dalle distribuzioni delle relazioni di dipendenza sintattica per quanto riguarda i complementi locativi e i complementi predicativi del soggetto e dell'oggetto, per i quali in ogni caso la media calcolata non supera lo 0,1.

Viene riscontrata una generale congruenza nel ranking anche tra i parametri che variano in modo maggiore, generalmente di carattere sintattico, legati alla subordinazione e all'ordine relativo dei costituenti. Ciò evidenzia il fatto che non esiste una marcata differenza tra i testi destinati alle scuole elementari e quelli destinati alle scuole superiori.

Il genere giornalistico. Di seguito verranno analizzati i risultati relativi al corpus *DueParole* (*2Par*) e al corpus *Repubblica* (*Rep*), rappresentanti rispettivamente della varietà semplice e complessa per il genere giornalistico.

In questo caso, per evidenziare casi particolari, è stato ritenuto opportuno riportare non solo il rank dei parametri del corpus complesso in relazione all'ordine assunto da questi nel corpus semplice ($rank_1$), ma anche l'opposto ($rank_2$), ovvero il rank dei parametri del corpus semplice in relazione all'ordine assunto da questi nel corpus complesso. Inoltre, il fatto che le medie siano state calcolate su circa 1800 fasce, in quanto sia *DueParole* che *Repubblica* raccolgono all'incirca 300 documenti ciascuno, ha comportato un numero elevato di parametri che, nel processo di normalizzazione, ha assunto un valore medio arrotondabile a 0. Poiché i parametri in questione sono per lo più gli stessi in entrambi i corpora e, in ogni caso, poco riscontrati, si è scelto di non riportarli nella tabella, ad eccezione di quelli con media approssimabile a zero in uno dei due corpora, per i quali è stata conservata la notazione della barra, e con media non trascurabile nell'altro.

			2Par				*Rep*
rank $_1$	features	d.s.$_1$	rank $_2$	rank $_2$	features	d.s.$_2$	rank $_1$
1	mChildren	0,03	7	1	CPOS_A(%)	0,04	45
2	DEP_ROOT(%)	0,06	8	~~2~~	~~DEP_comp_temp(%)~~	~~0,05~~	~~16~~
3	Tokens	0,06	19	3	mDist	0,06	5
4	Chars	0,06	34	4	mCxT	0,06	9
5	mDist	0,06	3	5	DEP_pred(%)	0,06	32
6	mHeight	0,07	26	~~6~~	~~CPOS_D(%)~~	~~0,06~~	~~39~~

~~7~~	~~DEP_clit(%)~~	~~0,08~~	~~20~~	7	mChildren	0,07	1
8	maxDist	0,08	21	8	DEP_ROOT(%)	0,07	2
9	mCxT	0,08	4	9	mChildrenS	0,07	21
~~10~~	~~mHeightSubMinor~~	~~0,1~~	~~49~~	10	DEP_conj(%)	0,07	26
11	mWeightSubMain	0,1	50	11	DEP_obj(%)	0,07	34
12	mWeightSubTOT	0,1	54	12	CPOS_C(%)	0,07	51
13	mHeightSubMain	0,1	51	13	CPOS_N(%)	0,07	57
14	mChildrenV	0,1	27	14	POS_V(%)	0,08	19
~~15~~	~~mWeightSubMinor~~	~~0,11~~	~~39~~	~~15~~	~~DEP_comp_loc(%)~~	~~0,08~~	~~24~~
16	DEP_comp_temp(%)	0,11	2	16	DEP_con(%)	0,08	25
17	DEP_sub(%)	0,11	28	~~17~~	~~DEP_modal(%)~~	~~0,08~~	~~30~~
18	subTOT	0,11	45	~~18~~	~~POS_VM(%)~~	~~0,08~~	~~31~~
…	…	…	…	…	…	…	…
58	POS_CS*(%)	0,2	59	58	subMinor(%)	0,18	50
59	advPost(%)	0,2	60	59	POS_CS*(%)	0,18	58
60	POS_RI*(%)	0,2	57	60	advPost(%)	0,18	59
61	adjPre(%)	0,21	61	61	adjPre(%)	0,19	61
62	subMinorPost(%)	0,24	65	62	POS_PC*(%)	0,21	63
63	POS_PC*(%)	0,25	62	63	POS_PR*(%)	0,21	64
64	POS_PR*(%)	0,27	63	64	POS_RD*(%)	0,25	65
65	POS_RD*(%)	0,3	64	65	subMinorPost(%)	0,26	62
66	subjPre(%)	0,35	69	66	subMain(%)	0,27	69
67	advPre(%)	0,36	67	67	advPre(%)	0,28	67
68	POS_CC*(%)	0,37	70	68	adjPost(%)	0,28	70
69	subMain(%)	0,37	66	69	subjPre(%)	0,29	66
70	adjPost(%)	0,38	68	70	POS_CC*(%)	0,29	68
71	objPost(%)	0,38	71	71	objPost(%)	0,3	71
72	subMainPost(%)	0,39	72	72	subMainPost(%)	0,31	72

TABELLA 5.3: Ordinamento delle *features* monitorate rispetto al valore crescente della deviazione standard nei corpora *2Par* e *Rep*

Sebbene le *features* che occupano l'ultima parte della tabella siano più o meno le stesse incontrate nel confronto tra i corpora dei materiali didattici, quanto si osserva nelle prime righe svela una situazione ben diversa: la maggior parte dei parametri che si mantengono stabili in un corpus, nell'altro registra variazioni di diverso grado. Caratteristiche di base come la lunghezza delle frasi in parole ortografiche (`Tokens`) e in caratteri (`Chars`), sebbene si mantengono stabili nella varietà semplice del genere, registrano variazioni importanti in quella complessa, classificandosi al 19esimo e al 34esimo posto, rispetto al terzo e al quarto posto della graduatoria di *2Par*.

Incongruenze simili si riscontrano anche nelle caratteristiche sintattiche e in particolare per quanto riguarda la media di dipendenti per verbi e la massima distanza registrata tra una testa e il suo dipendente, nonché nelle caratteristiche strutturali degli alberi sintat-

tici, come ad esempio per l'altezza media: i parametri appena elencati, infatti, si mantengono stabili in *2Par*, mentre sono relativamente variabili in *Rep*. Ciò evidenzia come, i testi di *DueParole* più che quelli di *Repubblica*, siano il risultato di una scrittura controllata, soprattutto rispetto a quei parametri che possono causare un aumento della complessità sintattica del testo.

Per decifrare correttamente il fatto che in *DueParole* molte caratteristiche riguardanti la subordinazione abbiano deviazioni standard basse, bisogna considerare il valore della loro media: questa infatti, sia per l'altezza che per l'ampiezza delle subordinate, sia per il numero di subordinate individuate in ogni frase (`subTOT`) che per la distribuzione delle proposizioni subordinate, si mantiene intorno allo 0,1, un valore molto basso che mette in luce la limitata portata di questi parametri e, in generale, del fenomeno della subordinazione, che al contrario si attesta molto variabile nel corpus *Repubblica*.

Il genere narrativo. In questo caso, i corpora esaminati sono il *Terence&Teacher* nella versione semplificata (*TT-sempl*) e in quella originale (*TT-orig*). Anche qui si è preferito non riportare nella tabella le caratteristiche con medie approssimabili allo 0.

		TT-sempl			*TT-orig*
rank	features	dev.stand.	rank	features	dev.stand.
1	mCxT	0,08	1	mCxT	0,07
2	CPOS_S(%)	0,09	3	mChildren	0,08
3	mChildren	0,09	5	mDist	0,08
~~4~~	~~DEP_mod_temp(%)~~	~~0,1~~	62	DEP_sub(%)	0,09
5	mDist	0,1	2	CPOS_S(%)	0,1
6	LinkPre(%)	0,11	28	mHeight	0,1
7	subMainPre(%)	0,12	~~31~~	~~CPOS_D(%)~~	~~0,1~~
8	DEP_neg(%)	0,12	9	DEP_arg(%)	0,11
9	DEP_arg(%)	0,12	11	DEP_mod(%)	0,11
10	DEP_ROOT(%)	0,12	~~29~~	~~POS_VM(%)~~	~~0,11~~
11	DEP_mod(%)	0,12	~~30~~	~~DEP_modal(%)~~	~~0,11~~
12	mChildrenS	0,12	46	CPOS_V(%)	0,11
13	POS_S(%)	0,12	6	LinkPre(%)	0,12
14	mChildrenV	0,12	14	mChildrenV	0,12
15	LinkPost(%)	0,12	15	LinkPost(%)	0,12
…	…	…	…	…	…
63	subMinorPost(%)	0,22	34	DEP_pred(%)	0,21
64	POS_CS*(%)	0,22	55	advPost(%)	0,21
65	POS_RI*(%)	0,22	61	adjPre(%)	0,21
66	Sentences	0,22	64	POS_CS*(%)	0,21
67	CPOS_P(%)	0,22	65	POS_RI*(%)	0,23
68	mHeightSubTOT	0,23	63	subMinorPost(%)	0,26
69	POS_PC*(%)	0,27	69	POS_PC*(%)	0,26

70	POS_CC*(%)	0,29	70	POS_CC*(%)	0,28
71	subMain(%)	0,29	71	subMain(%)	0,28
72	POS_RD*(%)	0,29	75	advPre(%)	0,28
73	adjPost(%)	0,3	72	POS_RD*(%)	0,29
74	subjPre(%)	0,3	74	subjPre(%)	0,29
75	advPre(%)	0,31	76	subMainPost(%)	0,3
76	subMainPost(%)	0,31	77	objPost(%)	0,3
77	objPost(%)	0,31	73	adjPost(%)	0,31

TABELLA 5.4: Ordinamento delle *features* monitorate rispetto al valore crescente della deviazione standard nei corpora *TT-sempl* e *TT-orig*

Il prospetto offerto da questo confronto può essere paragonato a quello del genere didattico. Anche in questo caso, infatti, i parametri più stabili sono quelli sintattici: dal numero medio di dipendenti per testa sintattica, alla distanza media testa-dipendente. Inoltre, il rank 6 del parametro `mHeight` relativo all'altezza media degli alberi sintattici svela che le frasi del corpus *TT-orig* generano alberi di altezza poco variabile.

Come nel confronto precedente, inoltre, si attestano nuovamente nelle ultime file i parametri che indagano l'ordine relativo degli elementi, in particolare di soggetto, oggetto, aggettivo e avverbio, e alcuni parametri descriventi le subordinate.

È interessante notare che, in questo confronto, la distribuzione della categoria morfosintattica dei sostantivi sia un parametro stabile in entrambi i corpora, mentre quella dei verbi subisca variazioni mediamente importanti nei testi semplici (rank 46), piuttosto che in quelli complessi (rank 9).

Il genere scientifico. L'ultima analisi di questo esperimento riguarda la varietà semplice e complessa del genere scientifico. Si tratta di un caso particolare in quanto il corpus *Wikipedia* (*Wiki*) registra ben 19 parametri con valori approssimabili allo 0, a differenza del corpus *Articoli Scientifici* (*ArtScient*), che ne attesta solo 3. Questo fatto è spia della marcata differenza che sussiste tra i due corpora, suggerendo che il secondo abbia una sintassi ben più ricca e diversificata del primo.

Wiki			*ArtScient*		
rank	Features	dev.stand.	rank	features	dev.stand.
~~1~~	~~POS_PQ*(%)~~	~~0,03~~	~~2~~	~~POS_PP*(%)~~	~~0,06~~
~~2~~	~~POS_PP*(%)~~	~~0,04~~	14	mCxT	0,07
~~3~~	~~subMinorPre(%)~~	~~0,04~~	~~3~~	~~subMinorPre(%)~~	~~0,08~~
~~4~~	~~DEP_mod_loc(%)~~	~~0,04~~	22	POS_S(%)	0,08
~~5~~	~~DEP_comp_loc(%)~~	~~0,05~~	60	CPOS_V(%)	0,08
~~6~~	~~POS_SA(%)~~	~~0,06~~	68	POS_V(%)	0,08
~~7~~	~~DEP_comp_ind(%)~~	~~0,06~~	23	CPOS_S(%)	0,09
~~8~~	~~DEP_concat(%)~~	~~0,06~~	27	DEP_neg(%)	0,09

~~9~~	~~DEP_comp_temp(%)~~	~~0,06~~	5	DEP_comp_loc(%)	0,1
10	maxDist	0,06	~~6~~	~~POS_SA(%)~~	~~0,1~~
~~11~~	~~POS_VM(%)~~	~~0,07~~	16	objPre(%)	0,1
12	mWeight	0,07	59	CPOS_C(%)	0,1
13	mDist	0,07	4	DEP_mod_loc(%)	0,11
14	mCxT	0,07	7	DEP_comp_ind(%)	0,11
~~15~~	~~DEP_mod_temp(%)~~	~~0,08~~	15	DEP_mod_temp(%)	0,11
~~16~~	~~objPre(%)~~	~~0,08~~	19	DEP_dis(%)	0,11
~~17~~	~~subMainPre(%)~~	~~0,08~~	45	DEP_mod(%)	0,11
~~18~~	~~POS_PE*(%)~~	~~0,08~~	63	POS_CS*(%)	0,11
~~19~~	~~DEP_dis(%)~~	~~0,08~~	17	subMainPre(%)	0,12
~~20~~	~~DEP_modal(%)~~	~~0,08~~	41	POS_SP(%)	0,12
21	mWeightSubMinor	0,08	42	CPOS_D(%)	0,12
22	POS_S(%)	0,08	1	POS_PQ*(%)	0,13
23	CPOS_S(%)	0,08	9	DEP_comp_temp(%)	0,13
24	Sentences	0,09	52	DEP_clit(%)	0,13
25	Tokens	0,09	66	DEP_punc(%)	0,13
26	Chars	0,09	67	CPOS_F(%)	0,13
~~27~~	~~DEP_neg(%)~~	~~0,1~~	74	CPOS_E(%)	0,13
28	DEP_subj_pass(%)	0,1	8	DEP_concat(%)	0,14
29	DEP_arg(%)	0,1	10	maxDist	0,14
30	DEP_mod_rel(%)	0,1	20	DEP_modal(%)	0,14
31	DEP_conj(%)	0,1	25	Tokens	0,14
32	mWeightSubMain	0,1	29	DEP_arg(%)	0,14
…	…	…	…	…	…
75	DEP_prep(%)	0,18	80	POS_PR*(%)	0,18
76	mChildren	0,18	36	POS_VA(%)	0,19
77	advPost(%)	0,19	65	adjPre(%)	0,19
78	POS_RI*(%)	0,19	75	DEP_prep(%)	0,19
79	POS_PC*(%)	0,22	76	mChildren	0,19
80	POS_PR*(%)	0,23	84	subMain(%)	0,19
81	subMinorPost(%)	0,26	53	mHeightSubMinor	0,2
82	POS_RD*(%)	0,29	85	advPre(%)	0,2
83	adjPost(%)	0,29	73	DEP_ROOT(%)	0,21
84	subMain(%)	0,3	82	POS_RD*(%)	0,21
85	advPre(%)	0,31	83	adjPost(%)	0,21
86	POS_CC*(%)	0,32	86	POS_CC*(%)	0,21
87	subMainPost(%)	0,33	87	subMainPost(%)	0,22
88	subjPre(%)	0,33	89	objPost(%)	0,22
89	objPost(%)	0,34	88	subjPre(%)	0,23

TABELLA 5.5: Ordinamento delle *features* monitorate rispetto al valore crescente della deviazione standard nei corpora *Wiki* e *ArtScient*

In questo genere, contrariamente a quanto riscontrato finora, il numero medio di dipendenti per testa sintattica (`mChildren`) viene relegato nelle posizioni più basse (rank 76),

sia per la varietà semplice che per quella complessa, insieme a parametri tipicamente variabili in ogni altro corpus analizzato.

Un ulteriore parametro che, a sorpresa, viene classificato tra quelli più variabili, è quello riguardante la distribuzione della ROOT, detta radice, o "testa della frase" (rank 73 in *Wiki* e rank 83 in *ArtScient*). Soprattutto in periodi lunghi, infatti, una frase può presentare più di una radice e non, come accade tipicamente, una per frase.

Prendendo in esame le *features* più caratterizzanti del corpus *Wiki*, appare subito evidente che siano tendenzialmente le stesse riscontrate nel corpus *2Par*, e in entrambe le varietà semplice e complessa dei generi didattico e narrativo. Possiamo, dunque, affermare che nei generi più semplici, destinati alla didattica o comunque ad un pubblico non adulto, ci sia la tendenza a controllare le caratteristiche più basilari, come ad esempio la lunghezza delle frasi e la distribuzione dei sostantivi.

Le due varietà registrano differenze interessanti rispetto alla variazione dei valori assunti dalle categorie morfosintattiche, i cui valori sono molto più stabili nella varietà complessa che in quella semplice: è il caso, ad esempio, dei verbi, attestati in quinta posizione in *ArtScient*, e solo alla sessantesima in *Wiki*, dei sostantivi, per i quali il rank 7 in *ArtScient* corrisponde al 23esimo in *Wiki*, e delle congiunzioni (rank 12 vs rank 59), soprattutto subordinanti (rank 18 vs rank 63).

5.3. Esperimento n.2: sulla variazione delle features tra la varietà semplice e complessa

Il secondo esperimento vuole individuare le caratteristiche linguistiche che variano in maniera più interessante tra le varietà linguistiche studiate. In questo caso, oltre allo studio della media aritmetica, della varianza e della deviazione standard, ci si è affidati al *Wilcoxon rank-sum test* per verificare la significatività dei confronti. Il test di Wilcoxon è stato applicato sfruttando la funzione messa a disposizione da *SciPy*, una libreria *open source* di algoritmi e strumenti matematici per *Python*.

5.3.1. Il procedimento

Per svolgere questo esperimento, sono state utilizzate le tabelle "totali" – la cui genesi è stata riassunta nel PAR. 5.2.1. In questo caso, a differenza dell'esperimento precedente, non è stato necessario normalizzare i dati, in quanto sono state confrontate liste di valori omogenei. Il confronto, infatti, è stato effettuato parametro per parametro: ogni vettore di *features* di un corpus è stato confrontato con il vettore corrispondente del corpus rappresentante la varietà linguistica di segno opposto.

A tale scopo, è stato implementato uno script in Python che ha permesso di applicare il *Wilcoxon rank-sum test* su colonne corrispondenti delle tabelle "totali" oggetto del confronto. Lo script, quindi, associa ad ogni *feature* il punteggio di significatività, detto *p-value*, calcolato dal test di Wilcoxon, la media aritmetica, la differenza tra le medie, le rispettive varianze e le deviazioni standard.
La differenza tra le medie dei valori è stata calcolata sottraendo dal valore del parametro estratto dal corpus "semplice", quello del corpus "complesso".

5.3.2. I risultati

Per poter discutere i risultati di questo esperimento, è necessario approfondire il tipo di contributo che un'informazione come il *p-value* apporta all'analisi dei dati. Si è già detto (cfr. PAR. 4.4.1) che il test di Wilcoxon viene utilizzato per verificare se due campioni indipendenti provengono dalla stessa popolazione: ciò significa una *feature* viene identificata come significativa se i valori che essa assume nei due corpora a confronto sono considerati appartenenti a popolazioni con diverse distribuzioni; viceversa, se il test considera i valori della *feature* come appartenenti a popolazioni con la stessa distribuzione,

la *feature* non sarà significativa. In questo caso, una *feature* significativa per il test di Wilcoxon può essere ascritta tra le caratteristiche che distinguono i testi della varietà linguistica "semplice" da quella "complessa".

Nei prospetti proposti, per offrire un confronto più immediato dei risultati ottenuti, si è scelto di assegnare alle *features* la cui variazione è stata classificata come statisticamente significativa il colore rosso, di tre tonalità differenti a seconda del grado di significatività individuato:

- *p-value* $< 0{,}05$, indice di significatività statistica;
- *p-value* $\leq 0{,}01$, indice di alta significatività;
- *p-value* $\leq 0{,}001$, indice di estrema significatività.

Inoltre, per tenere traccia della grandezza relativa delle *features*, è stato riportato il segno di ogni differenza tra le medie dei parametri: il segno + indica che il parametro monitorato nel corpus "semplice" ha in media dei valori più alti rispetto a quello monitorato nel corpus "complesso"; il segno -, al contrario, è indice del fatto che i valori estratti per quel parametro nel corpus rappresentante la varietà complessa sono mediamente maggiori rispetto a quelli estratti nel corpus appartenente allo stesso genere ma alla varietà semplice.

Nell'analisi dei risultati non sono stati presi in considerazione i valori assoluti (come il numero di tokens, il numero di caratteri e il numero di frasi) in quanto, facendo riferimento a fasce di dimensioni anche molto diverse, sarebbero risultati fuorvianti per il calcolo della varianza e conseguentemente della deviazione standard.

Le caratteristiche morfo-sintattiche. In primis, verranno discussi i risultati ottenuti dal monitoraggio delle caratteristiche morfo-sintattiche.
La tabella mostra chiaramente, ad un primo colpo d'occhio, come il test di Wilcoxon abbia individuato un numero molto alto di valori statisticamente significativi nel confronto tra i corpora appartenenti al genere giornalistico e al genere scientifico. Per quanto riguarda il confronto tra i corpora di carattere didattico, come quello tra corpora narrativi, il numero delle *features* morfo-sintattiche significative risulta essere nettamente inferiore.

Restringendo la finestra di analisi, si vuole innanzitutto analizzare i risultati ottenuti per il primo confronto. Nel genere dei materiali didattici, è risultata estremamente significativa la variazione degli aggettivi, dei determinanti, delle preposizioni, delle congiunzioni e, all'interno di queste, delle congiunzioni coordinanti sul totale di congiunzioni, i cui valori sono sempre superiori nella varietà complessa.

%	ELEM / SUP	2PAR / REP	TT-SEMPL / TT-ORIG	WIKI / ARTSCIENT
CPOS_A	-	-		
CPOS_B		-		
CPOS_C	-	+		
POS_CC*	-	-		
POS_CS*		-		-
CPOS_D	-	+		-
CPOS_E	-	-		-
CPOS_F		-		+
CPOS_N		+		-
CPOS_P		-	-	
POS_PD*	-	-		-
POS_PE*		-		-
POS_PI*		-	-	+
POS_PQ*		-		-
POS_PR*	-	-	-	-
POS_PC*		-	-	-
CPOS_R	+	+		+
POS_RD*		+		
POS_RI*		-		-
CPOS_S		+		+
POS_S	-	-		
POS_SA		-		-
POS_SP		+		+
CPOS_V	+	+		-
POS_VA	+	+	-	-
POS_VM	-	+		-
POS_V	+	+		-

TABELLA 5.6: Livelli di significatività delle caratteristiche morfo-sintattiche e segno della differenza tra il valore medio delle caratteristiche estratte dai testi semplici e dai testi complessi all'interno di ogni genere testuale.

Come osserva Montemagni (2013), il dato sulle congiunzioni, e in questo caso sull'assenza di significatività per quanto riguarda le congiunzioni subordinanti, può essere spiegato in relazione a fenomeni di "leggerezza sintattica". Allo stesso tempo, l'estrema significatività registrata per le congiunzioni coordinanti, indice di una sintassi ad organizzazione principalmente paratattica, può essere attribuito al fatto che queste, a differenza delle congiunzioni subordinanti, tipicamente associate esclusivamente a clausole subordinate, possono riguardare diverse categorie grammaticali, non solo i verbi. Per questo motivo, per considerazioni di maggiore precisione sulle strutture coordinate di tipo verbale, responsabili di costruzioni ipotattiche, si rimanda allo studio delle dipendenze sintattiche, condotto a breve.

Particolarmente interessante è il dato registrato per le distribuzioni dei verbi, per le quali si attesta una variazione estremamente significativa. La presenza di verbi risulta essere maggiore nel corpus *Elem (*come dimostra la differenza delle medie risulta essere di segno positivo), eccetto per la categoria dei verbi modali, la cui variazione è sì fortemente significativa, ma di segno negativo. Sempre facendo riferimento agli studi di Montemagni (2013), si può ricercare la giustificazione di questa particolarità nel fatto che i testi accademici, ai quali possono essere paragonati alcuni testi del corpus *Materiali didattici per la Scuola Superiore* considerati, si caratterizzano per un'alta densità informativa, con una maggiore distribuzione di nomi rispetto ai verbi – come testimoniato dalla variazione, più bassa ma comunque significativa, della categoria dei nomi comuni nella varietà complessa del genere didattico.

Per quanto riguarda il confronto tra i corpora *DueParole* e *Repubblica*, è evidente come tra le due varietà linguistiche vi siano differenze molto marcate. Di nuovo, come nel caso appena affrontato, spiccano le variazioni registrate per le distribuzioni dei verbi. Il segno generalmente positivo delle variazioni è indice di valori mediamente più alti registrati nel corpus *2Par*, i cui testi possono essere classificati come meno informativi e più legati alla struttura "narrativa". Tale ipotesi trova conferma nel fatto che le distribuzioni dei sostantivi attestino valori maggiori nel corpus *Repubblica*, generando una variazione rispetto alla varietà semplice dello stesso genere classificata come estremamente significative: i testi giornalistici, come i testi accademici, infatti, registrano un'alta densità informativa.

Il fatto che la categoria grammaticale dei nomi propri attesti un valore medio più alto nel corpus di varietà semplice trova giustificazione nel fatto che *DueParole* sia comunque un corpus di testi giornalistici in cui l'informazione è altamente condensata in poche righe per articolo (cfr. TABELLA 5.1). Le stesse considerazioni possono essere fatte per la variazione positiva e statisticamente significativa registrata per la categoria dei numerali.

Passando al confronto tra *Terence&Teacher semplificato* e *Terence&Teacher originale*, dunque interno al genere narrativo, troviamo pochissime categorie morfo-sintattiche segnalate come statisticamente significative. Tra queste, le più significative riguardano, come ci si aspetta, i pronomi, per i quali si attestano significatività più o meno alte anche a livello delle sottocategorie grammaticali. I pronomi, infatti, rappresentano un tratto caratterizzante del parlato, dunque presenti anche nei dialoghi che occorrono spesso nei testi narrativi.

Per il confronto interno al genere scientifico, come per quello giornalistico, sono numerose le categorie morfo-sintattiche segnalate come significative dal test di Wilcoxon. Interessante è la variazione tra congiunzioni subordinanti, indice della presenza di costruzioni ipotattiche all'interno del testo. Altre *features* in grado di distinguere i testi della varietà linguistica "semplice" da quella "complessa" appartengono alla categoria morfo-sintattica dei verbi, dei pronomi e dei sostantivi. Di questi ultimi, in particolare, si attesta una variazione positiva per quanto riguarda i nomi propri, e negativa per le abbreviazioni.

Le caratteristiche linguistiche sintattiche offrono spunti di riflessione ancora più interessanti.

	ELEM / SUP	2PAR / REP	TT-SEMPL / TT-ORIG	WIKI / ARTSCIENT
mDist	-	-	-	-
maxDist	-	-	-	-
LinkPre (%)	+	+		+
LinkPost (%)	-	-		-

TABELLA 5.7: Livelli di significatività di alcune caratteristiche sintattiche e segno della differenza tra il valore medio delle caratteristiche estratte dai testi semplici e dai testi complessi all'interno di ogni genere testuale.

Sia il dato relativo alla distanza lineare media del dipendente dalla sua testa sintattica che la massima distanza registrata tra testa e dipendente, costituiscono risultano estremamente significativi. Le variazioni hanno tutte un valore negativo: i testi complessi, com'era prevedibile, hanno relazioni di dipendenza più lunghe rispetto ai testi semplici.

Per quanto riguarda l'ordine relativo dei dipendenti rispetto alla propria testa sintattica, ad eccezione del genere narrativo, si segnala una significatività statistica molto marcata. Si noti, inoltre, che la variazione è sempre positiva in relazione alla media delle percentuali di tokens che precedono la propria testa sintattica: si può dunque pensare che si tratta di una *feature* caratterizzante la varietà linguistica semplice. Al contrario, la variazione del numero medio della percentuale di link a destra, generati da tokens che seguono la propria testa sintattica, è sempre negativa, dunque tale parametro può essere considerato come caratterizzante la varietà complessa.

La significatività delle variazioni delle caratteristiche sintattiche è illustrata nella seguente tabella. Come per le caratteristiche morfo-sintattiche, anche in questo caso il confronto interno al genere giornalistico e al genere scientifico individua un elevato numero di *features* la cui variazione è estremamente significativa. Un così elevato tasso di significatività contrasta, invece, con i risultati ottenuti dal confronto interno al genere dei

materiali didattici, e ancor più al genere narrativo: per entrambi, infatti, le poche variazioni significative raramente lo sono in modo forte.

DEP (%)	Elem / Sup	2Par / Rep	TTs / TTo	Wiki / ArtC	DEP (%)	Elem / Sup	2Par / Rep	TTs / TTo	Wiki / ArtC
arg		-		-	mod_loc		+		-
aux	+	+	-	-	mod_rel	-	+		
clit		-	-	-	mod_temp		-		+
comp	-	-		-	modal	-	+		+
comp_ind		-		-	neg	-	-		+
comp_loc		+		+	obj		-		-
comp_temp		+		+	pred		+		+
con	-	+			prep	-	-		-
concat		-		+	punc		-		+
conj	-				ROOT	+		+	-
det	+	+		+	sub		-		-
disj	-			+	subj	+	+	+	+
mod	-	-		-	subj_pass		-		+

Tabella 5.8: Livelli di significatività delle relazioni di dipendenza sintattica e segno della differenza tra il valore medio delle caratteristiche estratte dai testi semplici e dai testi complessi all'interno di ogni genere testuale.

Rispetto alla presenza di variazioni delle relazioni di dipendenza interne al genere dei materiali didattici, sono state attestate come estremamente significative le variazioni delle relazioni di dipendenza testa-complemento, testa-modificatore (aggettivale, avverbiale o frasale), testa verbale e verbo modale, testa preposizionale-complemento. Per ognuna di esse, la media dei valori registrati nei testi semplici è mediamente inferiore a quella calcolata sui valori estratti dai testi complessi. Al contrario, la distribuzione delle dipendenze articolo-nome, come era evidente già dal dato sulla distribuzione degli articoli e dei sostantivi in Tabella 5.6, è statisticamente significativa ma attesta valori mediamente più alti nella varietà semplice del genere didattico; lo stesso discorso vale per la distribuzione delle relazioni di dipendenza verbo-soggetto. In particolare, questo tipo di relazione è significativa in ogni confronto effettuato e sempre con segno positivo: ciò è chiaramente indice del fatto che nei testi semplici vi sia una più alta percentuale di relazioni sintattiche verbo-soggetto. Questo dato può trovare risposta nella volontà di rendere immediatamente intellegibili le informazioni nel testo, ricorrendo a strutture canoniche formate da soggetto, predicato e complemento.

Estremamente significative sono le variazioni che si riscontrano nel confronto tra *DuePa-role* e *Repubblica*. È interessante notare come, in questo caso, il segno delle differenze tra le medie dei valori varino molto da *feature* a *feature*. *2Par*, ad esempio, attesta valori

mediamente maggiori per quanto riguarda la distribuzione degli ausiliari e dei verbi modali, dei complementi di luogo, di tempo e dei complementi predicativi del soggetto o dell'oggetto, delle congiunzioni copulative, dei modificatori relativi e di quelli locativi. Di segno opposto risultano essere le variazioni statisticamente significative di parametri quali la distribuzione dei sintagmi argomentali non-soggetto, dei clitici, dei complementi preposizionali e indiretti con funzione di oggetto, dei modificatori – sia generici che legati al verbo da una relazione di tipo temporale. Estremamente significativi sono anche i dati ottenuti per quanto riguarda le relazioni tra testa verbale e complemento oggetto diretto e per le relazioni indicanti una frase subordinata. Infine, la variazione della relazione sintattica di soggetto, di cui si è già discusso, nel confronto tra *2Par* e *Rep* diventa significativa quando il soggetto è passivo.

Analogamente a quanto visto nel confronto appena analizzato, anche nel confronto tra il corpus *Wikipedia* e *Articoli Scientifici* si registra la quasi totalità di relazioni di dipendenza le cui variazioni risultano essere estremamente significative. Dai risultati, inoltre, si evince come molti di questi parametri siano maggiormente attestati nel corpus rappresentante la varietà semplice, piuttosto che nella varietà complessa. In particolare, la variazione, contraddistinta dal segno negativo, dei sintagmi argomentali non-soggetto, dei modificatori e dei modificatori locativi, delle clausole subordinate e delle relazioni tra teste preposizionali e complementi di natura frasale e non, sono una spia inequivocabile della maggiore "pesantezza" sintattica del corpus *Articoli Scientifici*, essendo quelli appena elencati parametri in qualche modo coinvolti nella costruzione di strutture subordinate. Parallelamente, si può notare una certa corrispondenza tra i parametri il cui valore medio risulta essere maggiore nella varietà semplice del genere scientifico (*Wiki)* e quelli della varietà semplice per il genere giornalistico (*2Par*): in particolare, si sta facendo riferimento alla distribuzione dei complementi e dei complementi indiretti con funzione di oggetto, i verbi modali, i complementi predicativi e la distribuzione dei soggetti, in questo anche di quelli passivi.

Si vuole ora analizzare i risultati ottenuti relativamente all'ordine sintattico facenti riferimento alla posizione lineare di un elemento rispetto alla "testa" da cui è retto in una rappresentazione sintattica a dipendenze.
Gli elementi considerati sono stati il soggetto, l'oggetto, l'avverbio, l'aggettivo e la clausola subordinata (di quest'ultima si tratterà nello specifico più avanti). Nelle considerazioni che seguiranno, si considera "canonica" la posizione rispetto alla matrice prevalente SVO dell'italiano (preposta o posposta alla testa a seconda dell'elemento indagato) e "marcata" la posizione opposta.

%	ELEM / SUP	2PAR / REP	TT-SEMPL / TT-ORIG	WIKI / ARTSCIENT
subjPre		+		
subjPost		-		-
objPre	-	-		-
objPost	-	+		-
adjPre		-		-
adjPost	-	+		
advPre		-		
advPost	-	-		-

TABELLA 5.9: Livelli di significatività dell'ordine relativo dei costituenti e segno della differenza tra il valore medio delle caratteristiche estratte dai testi semplici e dai testi complessi all'interno di ogni genere testuale.

In linea con i risultati analizzati in precedenza, anche in questo caso si nota come le varietà interne al genere giornalistico e quelle interne al genere scientifico siano ben distinte. Al contrario, il confronto tra *Terence&Teacher* semplificato e originale non produce alcun risultato significativo.

Nel confronto tra *Elem* e *Sup*, si attestano come significative le variazioni delle distribuzioni degli oggetti preverbali e postverbali, oltre che degli aggettivi postnominali e degli avverbi postverbali.

Per quanto riguarda il confronto tra *2Par* e *Rep*, invece, è stata riscontrata significatività statistica per ogni parametro. Sia per la posizione dell'aggettivo e dell'avverbio, che in italiano risulta variare in base alla funzione semantica svolta più che rispetto alla varietà linguistica in cui si manifesta (Cinque, 2010 citato in Pieri *e altri*, 2016), sia per la posizione del soggetto e dell'oggetto.
In particolare, il segno positivo della differenza tra il numero medio delle distribuzioni del soggetto preverbale e dell'oggetto postverbale nei due corpora dimostra come la varietà più semplice del genere giornalistico si attenga maggiormente all'ordine canonico SVO. La variazione del soggetto e dell'oggetto in posizione marcata, classificata come estremamente significativa, attesta valori maggiori nel corpus *Repubblica*.

Analogamente a quanto appena esposto, anche il confronto tra *Wiki* e *ArtScient* mostra un'estrema significatività statistica per quanto riguarda la variazione della distribuzione di soggetti e oggetti in posizione marcata, entrambi maggiormente attestati nella varietà complessa.

Analizzando i dati relativi alle caratteristiche strutturali dell'albero sintattico, si evidenzia un andamento chiaro e unilaterale: per tutti i confronti le variazioni statisticamente

significative sono "a favore" della varietà complessa, com'è testimoniato dalla quasi totalità dei segni negativi. È chiaro, dunque, che gli alberi sintattici abbiano altezze e ampiezze maggiori nelle varietà complesse di ogni genere testuale, nonché un numero mediamente maggiore di figli per tokens. Le frasi dei testi complessi, com'è facile immaginare, quindi, non solo risultano essere più lunghe, ma anche più "profonde", ovvero caratterizzate da una organizzazione del periodo di tipo ipotattico.

	ELEM / SUP	2PAR / REP	TT-SEMPL / TT-ORIG	WIKI / ARTSCIENT
mHeight	-	-	-	-
mWeight	-	-	-	-
mChildren	-		-	-
mChildrenS	-		-	-
mChildrenV		+		

TABELLA 5.10: Livelli di significatività delle caratteristiche strutturali dell'albero e segno della differenza tra il valore medio delle caratteristiche estratte dai testi semplici e dai testi complessi all'interno di ogni genere testuale.

A corroborare la tesi appena esposta, vi è sicuramente il seguente prospetto che riporta le variazioni statisticamente significative riguardanti le subordinate.

La subordinazione appare una discriminante estremamente significativa nella distinzione tra testi semplici e complessi, registrando valori in ogni caso mediamente più alti nei corpora rappresentanti la varietà difficile.

	ELEM / SUP	2PAR / REP	TT-SEMPL / TT-ORIG	WIKI / ARTSCIENT
subTOT	-	-	-	-
mHeightSubTOT	-	-	-	-
mWeightSubTOT	-	-	-	-
subMain(%)		-		-
mHeightSubMain	-	-	-	-
mWeightSubMain	-	-	-	-
subMainPre(%)		-		-
subMainPost(%)	-	-	-	-
subMinor(%)	-	-	-	-
mHeightSubMinor	-	-	-	-
mWeightSubMinor	-	-	-	-
subMinorPre(%)		-		+
subMinorPost(%)	-	-	-	-

TABELLA 5.11: Livelli di significatività dei parametri riguardanti la subordinazione e segno della differenza tra il valore medio delle caratteristiche estratte dai testi semplici e dai testi complessi all'interno di ogni genere testuale.

Un dato da sottolineare è l'assenza di significatività nel confronto interno al genere didattico e al genere narrativo per quanto riguarda la distribuzione delle subordinate preposte alla principale, sia di primo grado che di grado superiore al primo. I testi didattici e i testi narrativi risultano essere complessivamente più "semplici" di quelli giornalistici e scientifici, in quanto i primi sono destinati alla didattica, i secondi ad un pubblico di bambini. In questi generi, dunque, la preferenza ricade generalmente sulla posizione della subordinata posposta alla principale, in accordo con quanto predetto dai modelli di processing secondo cui un ordinamento di questo tipo comporta un impegno cognitivo minore nell'utente, per il quale le relazioni sintattiche risultano più facilmente riconoscibili (Hawkings, 1994).

5.4. Esperimento n.3: sulla variazione delle features tra fasce consecutive

Il terzo esperimento ha come scopo quello di individuare, nello studio interno ad ogni documento, variazioni linguistiche significative nel passaggio da una fascia all'altra. Inoltre, si vuole monitorare l'andamento delle *features* all'interno di ogni fascia, per rintracciare un eventuale andamento tipico.
Questo tipo di informazione è stata ottenuta calcolando la correlazione dei valori assunti dalle *features* linguistiche nelle frasi di una fascia con quelli assunti nella fascia successiva utilizzando il coefficiente di correlazione di Spearman e il coefficiente di correlazione di Pearson. Il giudizio sulla significatività della variazione calcolata tra una fascia e l'altra è stato affidato al *p-value* del test di Wilcoxon.

5.4.1. Il procedimento

A partire dal raggruppamento dei documenti in fasce, che ha generato tabelle (esemplificata in FIGURA 4.3) con un numero di colonne pari al numero di *feature* oggetto di monitoraggio e con un numero di righe pari a 6, ovvero il numero delle fasce, tramite l'ausilio di uno script in Python, è stato effettuato un ciclo su ogni file CSV, che ha permesso di generare una tabella per ogni *feature*, avente in colonna le sei fasce e in riga i valori che il parametro ha assunto nelle differenti fasce in ognuno dei documenti analizzati (FIGURA 5.1).

	A	B	C	D	E	F
1	fascia 1	fascia 2	fascia 3	fascia 4	fascia 5	fascia 6
2	15,5	19,5	40	19	22,5	35,67
3	34,6	23	23,4	33	30,6	28,5
4	15,33	11,67	14,33	13	12,5	16
5	7,5	11,5	7,5	8,5	8	7
6	9,8	8,2	8,2	17	10,83	20
7	26	21	19	31	12,33	26,67
8	15	13,1	14,6	11,3	9	13,22
9	22	15	33	24	20	32
10	...					

FIGURA 5.1: Esempio di tabella "totale" di una *feature*, in cui si riportano i valori assunti da questa in ogni fascia per ogni documento del corpus

Solo a questo punto è stato possibile applicare le funzioni statistiche sui dati, tramite l'ausilio delle funzioni della libreria SciPy: il *Wilcoxon rank-sum test*, il coefficiente di correlazione di Spearman e il coefficiente di correlazione di Pearson. A tali funzioni è stato fornito come primo parametro la lista dei valori della fascia *n* e come secondo parametro la lista dei valori della fascia *n+1*.

Inoltre, si è scelto di indagare anche la variazione e l'andamento della prima e dell'ultima fascia, supponendo di riscontrare, in questo caso specifico, una spiccata significatività per due motivi: i) si tratta di parti del testo molto distanti; ii) sia l'introduzione che la conclusione sono generalmente fortemente caratterizzanti di un certo genere testuale.

Nel file CSV che ha raccolto l'output delle analisi, per confronto tra fasce, sono state messe a confronto la media dei valori che la *feature* ha assunto nelle fasce coinvolte nel confronto, la loro differenza, la varianza e la deviazione standard.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		pvalue_p	r_p	pvalue_s	r_s	pvalue_w	media_A	media_B	mediaDIFF	varA	varB	devA	devB
2	fascia1-2	1,03E-05	0,56999	1,14E-05	0,56772	0,03406	4,43	4,22	0,21	0,17	0,23	0,41	0,48
3	fascia1-6	7,10E-05	0,52244	4,00E-05	0,53737	0,32947	4,43	4,38	0,05	0,17	0,19	0,41	0,43
4	fascia2-3	9,98E-07	0,61904	1,23E-06	0,61491	1	4,22	4,21	0,01	0,23	0,31	0,48	0,56
5	fascia3-4	8,88E-05	0,51645	1,37E-05	0,56339	0,07276	4,21	4,4	-0,19	0,31	0,4	0,56	0,63
6	fascia4-5	0,00037	0,47564	0,00053	0,4643	0,34919	4,4	4,36	0,04	0,4	0,19	0,63	0,43
7	fascia5-6	4,95E-06	0,5863	9,30E-06	0,57231	0,97925	4,36	4,38	-0,02	0,19	0,19	0,43	0,43

FIGURA 5.2: Esempio di tabella generata per l'esperimento n.3. In colonna si hanno in ordine: *p-value* e *rho* dell'indice di correlazione di Pearson, *p-value* e *rho* dell'indice di correlazione di Spearman, *p-value* di Wilcoxon, media della fascia *n*, media della fascia *n+1*, differenza tra le medie, varianza e deviazione standard delle fasce *n* e *n+1*.

5.4.2. I risultati

Procediamo, anche in questo caso, suddividendo la vasta gamma di *features* estratte per tipologia, analizzando i risultati ottenuti corpus per corpus. Scegliamo, inoltre, di segnalare con tre colori differenti il coefficiente che ha individuato una correlazione nell'andamento delle fasce:


per il coefficiente di correlazione di Pearson;


per il coefficiente di correlazione di Spearman;


quando sia il coefficiente di Pearson che di Spearman registrano un valore abbastanza alto del ρ.

I segni di spunta, invece, indicano che la caratteristica linguistica è significativa per discriminare le fasce coinvolte nell'intervallo studiato secondo il test di Wilcoxon. Anche in questo esperimento, inoltre, si è voluto monitorare il segno della differenza tra le medie dei valori che le *features* hanno registrato nelle fasce consecutive.

Materiali Didattici per la Scuola Elementare. Com'è possibile osservare dalla TABELLA 5.12, le caratteristiche di base del corpus *Elem* seguono andamenti fortemente correlati in ogni fascia: esiste, perciò, una funzione in grado di trasformare i dati della prima fascia in quelli della seconda.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	+	+	+	+	- ✓	-
Chars	+	+	-	+	- ✓	-
mCxT	+ ✓	0	-	0	0	0

TABELLA 5.12: Andamenti e significatività delle caratteristiche di base per il corpus *Elem*.

È evidente, inoltre, l'andamento decrescente della lunghezza delle frasi che, però, nell'ultima fascia, subisce un repentino incremento, che risulta essere statisticamente significativo (GRAFICO 5.1).

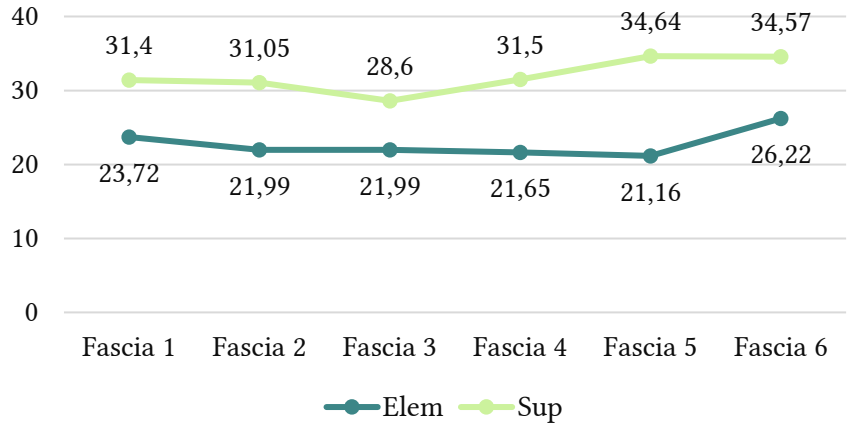

GRAFICO 5.1: Variazione della lunghezza media delle frasi in tokens nelle diverse fasce nei corpora *Elem* e *Sup*

Per quanto riguarda l'ambito morfo-sintattico, gli andamenti più interessanti riguardano la distribuzione dei nomi, e in particolare dei nomi comuni e di quelli propri. Per questi ultimi, si segnala una variazione statisticamente significativa nei passaggi tra la prima e la seconda fascia e tra la seconda e la terza.

Importanti sono anche i risultati ottenuti per quanto riguarda i verbi principali, il cui andamento correla tra la terza e la quarta fascia, tra la quarta e la quinta come anche tra la quinta e la sesta. In queste fasce, inoltre, la variazione dei valori registrati è statisticamente significativa.

%	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
CPOS_S	+ ✓	-	-	+	-	+
POS_S	+	+	+	+	+	+
POS_SP	+ ✓	+ ✓	-	-	+	+ ✓
POS_V	-	-	+ ✓	-	+ ✓	-/0

TABELLA 5.13: Andamenti e significatività delle distribuzioni delle *part-of-speech* di nomi e verbi in *Elem*

Nello studio delle caratteristiche sintattiche più generali, quali la distanza media (mDist) delle relazioni tra testa e dipendente e le distribuzioni delle relazioni con testa a destra (LinkPost(%)) o a sinistra (LinkPre(%)), gli andamenti assunti nel passaggio da una fascia all'altra risultano essere in varia misura correlati, ma mai statisticamente significativi.

Questo dato conferma quanto visto nei risultati dell'esperimento n.1: questi parametri, infatti, risultano essere i meno variabili nel corpus, e quindi si può supporre che si mantengano relativamente stabili nel passaggio da una fascia all'altra. La stessa considerazione può essere fatta per i risultati ottenuti nello studio delle caratteristiche strutturali dell'albero, quali l'altezza media degli alberi sintattici (rank 14 per l'esperimento n.1), la loro ampiezza media (rank 11), e il numero di figli per nodo (rank 1).

Per quanto riguarda l'ordine sintattico degli elementi, il test di Wilcoxon individua una variazione statisticamente significativa nel confronto tra la prima e l'ultima fascia per la maggior parte delle *features* monitorate. Gli andamenti, tuttavia, non risultano essere particolarmente correlati e dunque, si può supporre, che i valori delle distribuzioni varino in modo per lo più casuale. È interessante, invece, il fatto che le distribuzioni del soggetto e dell'oggetto in posizione canonica si mantengano costanti nel documento, al contrario di quanto accade per le rispettive posizioni marcate, le quali subiscono un incremento nella parte centrale del testo (GRAFICO 5.2).

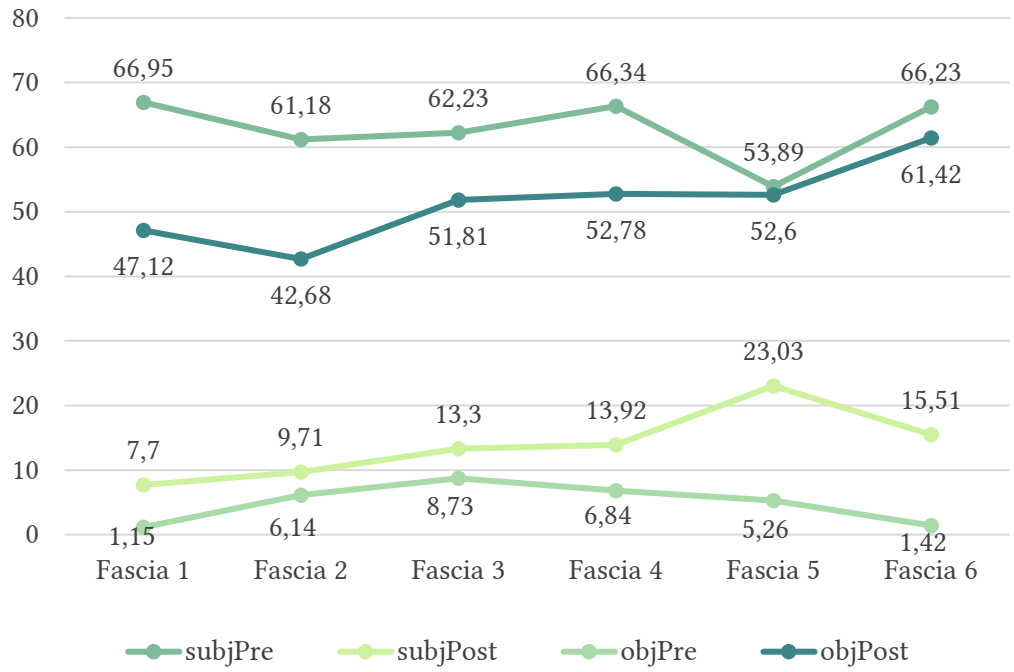

GRAFICO 5.2: Variazione della distribuzione (percentuale) dell'ordine relativo dei costituenti in *Elem.*

Significativi sono i risultati ottenuti per quanto riguarda la subordinazione. Infatti, si può notare un andamento tendenzialmente crescente sia dell'utilizzo delle subordinate, sia dell'ampiezza dell'albero sintattico generato da queste. Inoltre, si noti che il differente utilizzo delle subordinate di grado superiore al primo nella terza e nella quarta fascia sia discriminante per distinguere la prima dalla seconda.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
subTOT	0	0	-	0	-	-
mHeightSubTOT	+	-	+	-	-	- ✓
mWeightSubTOT	+	- ✓	+ ✓	-	-	- ✓
subMain(%)	-	- ✓	+ ✓	-	- ✓	-
mHeightSubMain	+	-	+	-	-	- ✓
mWeightSubMain	+	- ✓	+	-	-	- ✓
subMinor(%)	- /	+	- ✓	-	+	-
mHeightSubMinor	0	-	- ✓	+	+	-
mWeightSubMinor	-	+	- ✓	+	-	-

TABELLA 5.14: Andamenti e significatività della subordinazione nel corpus *Elem*.

Materiali Didattici per la Scuola Superiore. A differenza di quanto accade nel corpus *Elem*, la variazione della lunghezza delle frasi tra una fascia e l'altra nel corpus *Sup* non produce alcuna significatività statistica, sebbene si denoti un certo andamento tipico nel confronto tra la seconda fascia e la terza, e tra la quarta e la quinta.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	+/0	+	-	-	0	-
Chars	0	+	-	-	0	-
mCxT	-/0	+/0	-/0	+/0	+/0	+/0

TABELLA 5.15: Andamenti e significatività delle caratteristiche di base per il corpus *Sup*

Per questo corpus non sono state individuate variazioni interessanti dal punto di vista statistico; le poche correlazioni individuate tra le fasce sono relative agli andamenti delle distribuzioni di alcune caratteristiche linguistiche non interessanti (punteggiatura, preposizioni, complementi, ecc.).

Maggiormente degni di nota sono i parametri relativi alla sintassi e in particolare lunghezza del link sintattico più lungo `maxDist`, l'ampiezza media dell'albero sintattico generato dalle frasi `mWeight` e la media di figli per tokens `mChildren`. Per ognuno di questi parametri, infatti, la variazione che si verifica tra la terza e la quarta fascia risulta essere significativa. Si noti che, proprio nel confronto in cui vi è significatività, l'andamento non risulta più essere tipico.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
maxDist	0	+	- ✓	-	0	-
mDist	0	0	0	0	0	0
mHeight	0	+	-	-	+	-
mWeight	-/0	+	- ✓	-/0	0	-
mChildren	0	0	0 ✓	0	0	0

TABELLA 5.16: Andamenti e significatività di alcune caratteristiche sintattiche in *Sup*

DueParole. Nell'analisi dei risultati ottenuti per il corpus 2Par, si nota innanzitutto una generale assenza di correlazione tra le fasce.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	+ ✓	–	+	–	0	+
Chars	+ ✓	– ✓	+	–	–/0	–
mCxT	0	0	0	0	0	0 ✓

TABELLA 5.17: Andamenti e significatività delle caratteristiche di base per il corpus *2Par*

L'unica osservazione interessante riguarda l'andamento della lunghezza delle frasi in tokens: essa riconosce una variazione statisticamente significativa nel passaggio dalla prima alla seconda fascia, che viene registrata anche nel corpus della varietà opposta, *Repubblica*, ma di segno opposto. Come appare evidente nel grafico proposto, infatti, la lunghezza delle frasi della prima fascia è generalmente maggiore di quelle della seconda.

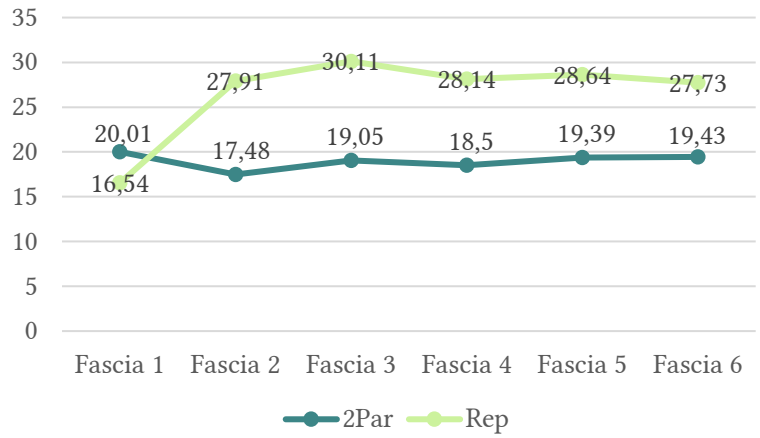

GRAFICO 5.3: Variazione della lunghezza media delle frasi in tokens nelle diverse fasce nei corpora *2Par* e *Rep*

Tale significatività statistica interesserà la maggior parte delle *features* monitorate. Ciò appare evidente già nei risultati ottenuti per quanto riguarda le distribuzioni delle principali categorie morfo-sintattiche.

%	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
CPOS_A	+ ✓	0	+	-	-/0	+
CPOS_B	- ✓	-	+/0	0	-/0	- ✓
CPOS_C	- ✓	-	+/0	-	+	- ✓
CPOS_S	+ ✓	+ ✓	-	+	-	+ ✓
POS_S	+ ✓	+	-	+ ✓	- ✓	+ ✓
POS_SP	+ ✓	-/0	+/0	-	+	+ ✓
CPOS_V	- ✓	- ✓	-	+	+	- ✓
POS_V	- ✓	-/0	0	-	+	- ✓

TABELLA 5.18.: Andamenti e significatività delle distribuzioni delle principali *part-of-speech* per il corpus *2Par.*

Nei risultati ottenuti dal monitoraggio delle *features* sintattiche di base non si manifestano andamenti tipici. La distanza media testa-dipendente subisce una variazione statisticamente significativa nel confronto tra la prima fascia, che registra valori maggiori, e la seconda, per poi mantenersi per lo più costante all'interno del testo – con deviazioni standard generalmente vicine allo zero.

Può essere interessante soffermarsi sulle variazioni dell'altezza e sull'ampiezza media degli alberi sintattici generati dalle frasi. Nello specifico, le variazioni registrate dai valori dell'altezza tra la prima e l'ultima fascia, così come tra la quarta e la quinta, sono statisticamente significative; per l'ampiezza, invece, solo tra la prima e la seconda. Nel grafico seguente, sono state messi a confronto gli andamenti delle grandezze in questione, dopo aver normalizzato i rispettivi valori.

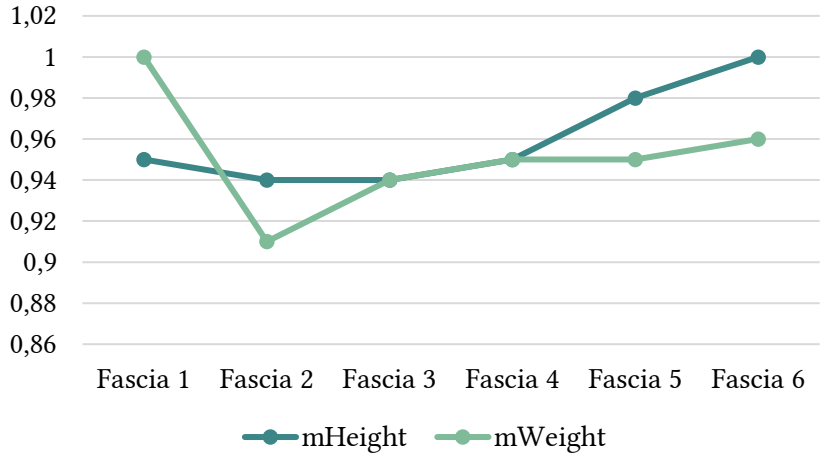

GRAFICO 5.4: Andamento dei valori delle altezze e delle altezze degli alberi sintattici tra le fasce.

L'utilizzo delle subordinate in *2Par* è ovviamente molto esiguo: in media, ogni frase ha meno di una subordinata. A livello statistico, le uniche significatività sono riconosciute nel confronto tra la prima e l'ultima fascia, in tutti i parametri mirati al monitoraggio del fenomeno della subordinazione. Nell'ultima fascia, infatti, si registra una media di subordinate per frase lievemente maggiore.

Repubblica. Nell'analisi dei risultati ottenuti dall'applicazione delle funzioni statistiche sui dati estratti dal corpus Rep, si evidenzia una tendenza in un certo senso complementare a quella analizzata nella varietà semplice del genere giornalistico.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	- ✓	-	+	-	+	- ✓
Chars	- ✓	-	+	-/0	+	- ✓
mCxT	0	0	0	0	0	0

TABELLA 5.19: Andamenti e significatività delle caratteristiche di base in *Rep.*

Nel grafico 5.5, si evidenzia il netto incremento del numero di tokens per frase che si verifica nel passaggio dalla prima alla seconda fascia: si tratta di un dato fortemente caratterizzante del passaggio tra la prima e la seconda fascia, come conferma il *p-value* calcolato dal test di Wilcoxon. Questo andamento, come si vedrà anche nel corpus *Wikipedia*, è determinato dalla presenza, nelle prime righe del documento, del titolo dell'articolo in questione, di lunghezza ovviamente ridotta e con una sintassi molto particolare, propria dei soli titoli giornalistici che tendono alla brevità e all'evidenza tramite l'omissione di alcuni elementi dell'enunciato. La percentuale di sostantivi sul totale di tokens risulterà dunque maggiorata, a differenza di quella dei verbi.

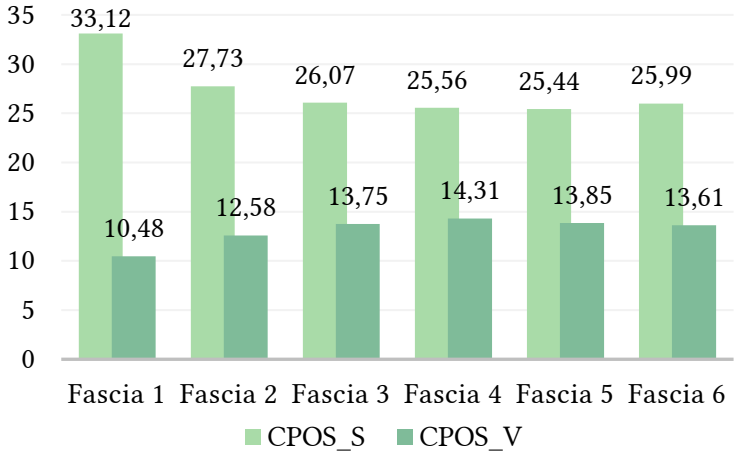

GRAFICO 5.5: Confronto tra la variazione delle distribuzioni (percentuali) di nomi e verbi tra le fasce

In particolare, l'andamento della percentuale di sostantivi correla la prima e la seconda fascia e la seconda e la terza; invece, la percentuale di verbi manifesta un andamento tipico nel confronto tra la quarta e la quinta fascia e tra la quinta e la sesta.

%	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
CPOS_S	+ ✓	+ ✓			-	+ ✓
POS_S	+ ✓		0	+ ✓	- ✓	+ ✓
POS_SP	+ ✓	+ ✓		-	0	+ ✓
CPOS_V	- ✓	- ✓	-	+	+	- ✓
POS_VA	- ✓	- ✓	0	+/0	0	- ✓
POS_VM	0 ✓	-/0 ✓	0	0	0	- ✓
POS_V	- ✓	-	-	+	+/0	- ✓

TABELLA 5.20: Andamenti e significatività delle distribuzioni delle principali *part-of-speech* in *Rep.*

Oltre alla generale significatività delle variazioni calcolata tra la prima e la seconda fascia e tra la prima e l'ultima, è evidente come anche la seconda e la terza fascia siano interessante da scarti particolarmente significativi in relazione alle distribuzioni delle caratteristiche morfo-sintattiche.

Altro dato interessante è quello riguardante la distanza media e la distanza massima testa-dipendente: in entrambi i casi, oltre alle ormai canoniche significatività nel confronto Fascia 1/Fascia 2 e Fascia 1/Fascia 6, si segnala nel confronto Fascia 3/Fascia 4. Nella fascia 3, infatti, si attesta un numero leggermente più alto di tokens per frase, che si traduce in relazioni sintattiche più lunghe.

I diversi tipi di relazioni sintattiche non offrono spunti di riflessione ulteriori rispetto a quelli già esposti, se non per quanto riguarda la distribuzione dei modificatori relativi e temporali, utilizzati in questo studio per individuare le subordinate.

DEP (%)	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
mod_rel	- ✓	- ✓	+	+	+ ✓	- ✓
mod_temp	- ✓	+	+ ✓	-	+	- ✓

TABELLA 5.21: Andamenti e significatività delle relazioni di dipendenza relative ai modificatori relativi e temporali in *Rep*

La variazione dei modificatori relativi è significativa anche tra la seconda e la terza fascia, nonché tra la quinta e la sesta, quella dei temporali, invece, tra la terza e la quarta. Guardando al segno della variazione, è evidente che la fascia 3 registri una distribuzione maggiore di relazioni di questo tipo. Questo dato è in qualche modo confermato dall'andamento del numero di subordinate per frase che, sebbene non registri una variazione statisticamente significativa, conferma il dato appena analizzato.

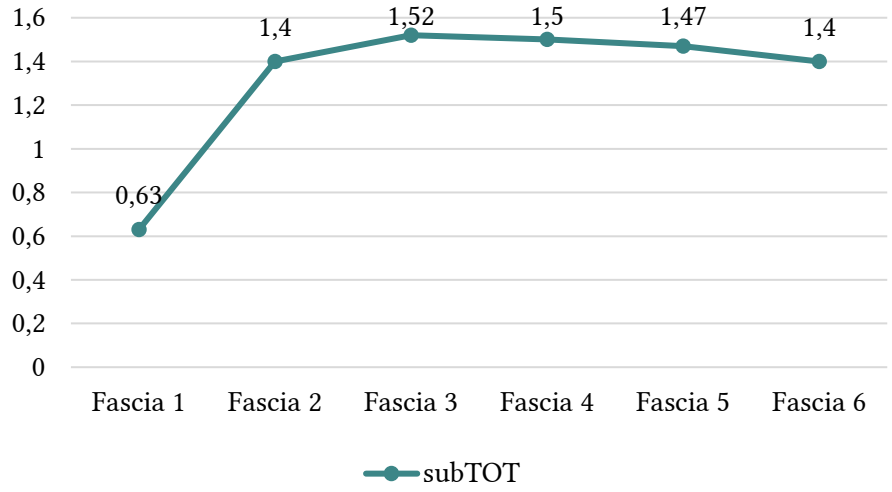

GRAFICO 5.6: Variazione del numero di subordinate per frase tra le fasce

Terence&Teacher semplificato. Per quanto concerne lo studio del corpus narrativo *TT-sempl*, i risultati ottenuti sono stati generalmente poco rilevanti, anche in caratteristiche di base quali la lunghezza delle frasi in tokens o in caratteri, generalmente significative negli altri corpora.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6

Tokens	-	-	+	0	-	- ✓
Chars	-	-	+	-/0	-	- ✓
mCxT	0	0	0	0	0	0

TABELLA 5.22: Andamenti e significatività delle caratteristiche di base in *TT-sempl*

Scorrendo velocemente i risultati relativi alle distribuzioni delle *part-of-speech*, si incontra significatività statistica esclusivamente per quanto riguarda la distribuzione dei nomi e dei verbi, in particolare tra la fascia 1 e la fascia 2. Inoltre, si attesta una correlazione abbastanza forte nell'andamento della distribuzione dei nomi propri nelle fasce centrali e conclusive del testo.

%	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
CPOS_S	+ ✓	+	-	+	-	+ ✓
POS_SP	+	-	+/0	+	-	+
CPOS_V	- ✓	+	-	-	+	-

TABELLA 5.23: Andamenti e significatività della distribuzione delle *pos* in *TT-sempl*

Il GRAFICO 5.7 mette in evidenza come la particolare distribuzione percentuale di sostantivi e verbi nella prima fascia, valutata come statisticamente significativa dal test di Wilcoxon, assuma poi un andamento più regolare nelle fasce successive.

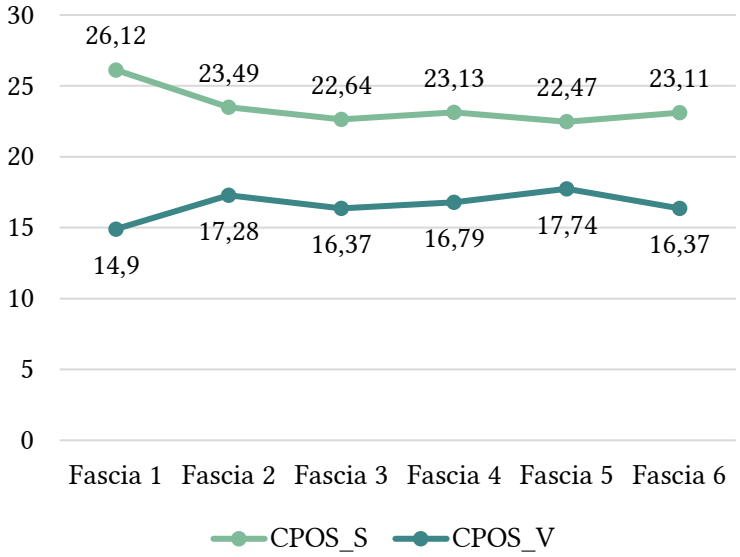

GRAFICO 5.7: Variazione della distribuzione di sostantivi e verbi tra le fasce.

Se non si registrano particolari andamenti o variazioni per quanto riguarda la distanza media o per la massima distanza di un token dalla sua testa sintattica, i risultati ottenuti dallo studio delle caratteristiche sintattiche dell'albero denotano una maggiore correlazione.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
mHeight	0	-	-	-	-	-

mChildren	0	0	0	0	0	0 ✓
mChildrenS	0	0	0	0	0	0

TABELLA: Andamenti e significatività delle caratteristiche dell'albero sintattico in *TT-sempl*

Dalla tabella, si evince che l'altezza media degli alberi sintattici generati dalle frasi conosce un andamento tipico per quanto riguarda la parte iniziale e la parte finale dei testi. Inoltre, per quanto riguarda il numero di figli per token, sebbene si mantenga mediamente stabile in tutte le fasce, si registrano andamenti simili tra la terza e la quarta fascia e tra la quinta e la sesta.

Per quanto riguarda la subordinazione, i risultati sono in linea con quanto finora esposto: significatività importanti sono riscontrate solo nel confronto tra la prima e la sesta fascia, mentre andamenti tipici delle grandezze calcolate sulle subordinate sono riscontrati per lo più nelle ultime fasce.

Terence&Teacher originale. I risultati ottenuti dallo studio condotto sul corpus *TT-orig* ha prodotto risultati ancor più diluiti. La variazione della lunghezza delle frasi, anche in questo caso, non ha prodotto alcuna significatività statistica, eccetto che nel confronto tra la prima e l'ultima fascia: dunque si può dire che la lunghezza delle frasi in tokens è un dato discriminante per distinguere la prima fascia, caratterizzata da frasi più corte, dall'ultima, popolata da frasi relativamente più lunghe. L'andamento ha invece assunto tipicità solo nelle fasce più interne dei documenti.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	-	0	+	-	+	- ✓
Chars	-	+	+	-	+	-
mCxT	0	0	0	0	0	0

TABELLA 5.24: Andamenti e significatività delle caratteristiche di base in *TT-orig*

Nello studio delle caratteristiche morfo-sintattiche, è stata attestata significatività statistica esclusivamente nel passaggio dalla prima alla seconda fascia per quanto riguarda la distribuzione dei sostantivi e dei verbi principali. Inoltre, l'andamento della distribuzione dei sostantivi ha permesso di correlare le due fasce iniziali e quelle intermedie.

%	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
CPOS_A	+	+	-	+	-	+
CPOS_S	+ ✓	+	-	-/0	-	+ ✓
POS_SP	+	0	+/0	+	-	+
POS_V	- ✓	+/0	-	+	+	-

TABELLA 5.25: Andamenti e significatività delle pos in *TT-orig*

La Tabella ricalca più o meno fedelmente quella ottenuta per il corpus *TT-sempl*.

Per quanto riguarda le caratteristiche sintattiche, invece, non si registrano andamenti degli di nota. La distanza media si mantiene pressoché costante nel passaggio da una fascia all'altra, mentre l'andamento dei valori da essa assunti all'interno di ogni fascia è stato valutato come tipico solo tra la terza e la quarta fascia.

In effetti, generalmente, le correlazioni più forti e frequenti si registrano proprio nel confronto tra le fasce interne dei documenti, quindi tra la seconda e la terza, la terza e la quarta, e in alcuni casi anche negli ultimi due intervalli di fasce. Ciò è evidente nel seguente prospetto, in cui sono riassunte le *features* sintattiche che hanno registrato i risultati più significativi o rappresentativi:

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
mHeight	-	-	+	-	+	-
mWeight	-	+/0	0	-/0	+/0	-
mChildren	0	0	0	0	0	0 ✓
subTOT	- ✓	0	0	0	+	- ✓

TABELLA 5.26: Andamenti e significatività delle *features* sintattiche in *TT-orig*

Degna di nota è la variazione statisticamente significativa registrata tra il numero medio di subordinate per frase nella prima e nella seconda fascia. In generale, infatti, si registrano valori tendenzialmente più alti nella parte centrale del testo, con un decremento in quella finale (GRAFICO 5.8).

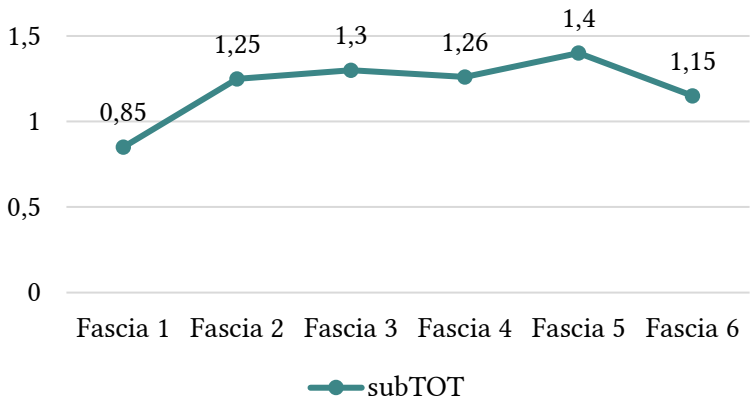

GRAFICO 5.8: Variazione del numero di subordinate per frase tra le fasce di *TT-orig*

Wikipedia. L'analisi dei risultati ottenuti dall'analisi del corpus *Wiki* ha messo in luce un tipico andamento generale, già riscontrato in entrambi i corpora giornalistici: una estrema significatività delle variazioni registrate tra la prima e la seconda fascia, e conseguentemente tra la prima e la sesta sia nelle caratteristiche di base, mostrate in Tabella, che nella quasi totalità delle altre *features* linguistiche monitorate.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	- ✓	+	+/0	+	-	- ✓
Chars	- ✓	+	+	+	-	- ✓
mCxT	+/0 ✓	+/0	0	0	0	0 ✓

TABELLA 5.27: Andamenti e significatività delle caratteristiche di base in *Wiki.*

Ciò va sicuramente indagato nella "conformazione" della prima fascia, che appare tanto diversa dalle successive.

La lunghezza media delle frasi, innanzitutto, subisce un incremento repentino nel passaggio alla seconda fascia, una variazione che non si verifica nella varietà complessa dello stesso genere e che quindi non può essere attribuita ad una caratteristica del genere testuale scientifico.

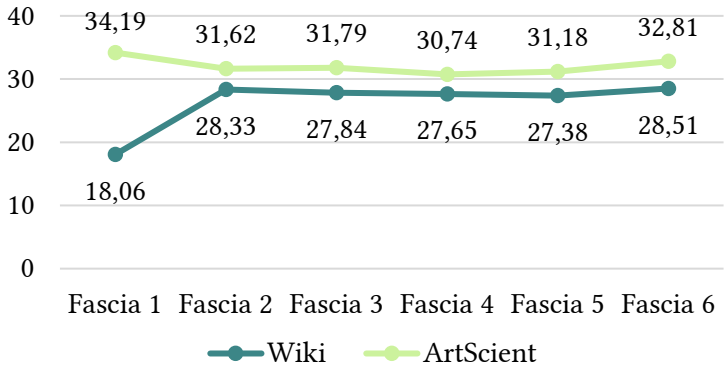

GRAFICO 5.9: Variazione della lunghezza media delle frasi in *Wiki* e *ArtScient*

Questa particolarità si riflette chiaramente in tutti i confronti effettuati, com'è evidente dal prospetto delle distribuzioni delle *part-of-speech*:

%	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
CPOS_A	- ✓	+	-	+	0	+ ✓
CPOS_B	- ✓	-	0	0	-	- ✓
POS_CC*	- ✓	+	+	+	-	- ✓
POS_CS*	- ✓	- ✓	+	+	+	- ✓
CPOS_D	- ✓	-	+/0	-/0	-/0	- ✓
CPOS_E	- ✓	-/0	-	+	-	- ✓
CPOS_P	- ✓	+	+	+	-	- ✓
CPOS_S	+ ✓	+	-	-	+	+ ✓
POS_S	-	+	- ✓	+	+	-
POS_SP	+ ✓	-	+	-	+/0	+ ✓
CPOS_V	- ✓	-	0	+	-	- ✓

TABELLA 5.28: Andamenti e significatività delle caratteristiche di base in *Wiki*

La giustificazione di una tale differenza tra i valori assunti dalle *features* nella prima fascia rispetto a quelli assunti dalle stesse nella seconda e, in generale, nelle altre fasce,

è da ricercare nel fatto che i documenti di Wikipedia, nella prima frase, riportano il titolo dell'articolo. Ciò comporta in primis un numero medio di tokens per frase nella prima fascia relativamente minore, essendo i titoli tipicamente costituiti da poche parole; in secundis, una distribuzione di sostantivi maggiore rispetto a qualsiasi altra categoria morfo-sintattica, essendo i titoli principalmente costituiti da sostantivi, articoli e preposizioni. Inoltre, solitamente, gli articoli di Wikipedia presentano nella parte iniziale una sorta di introduzione o di sunto dell'argomento trattato, di alto contenuto informativo, il che si traduce in uno scarto ancora più accentuato rispetto a quello riscontrato nelle altre fasce tra la percentuale di sostantivi e di verbi.

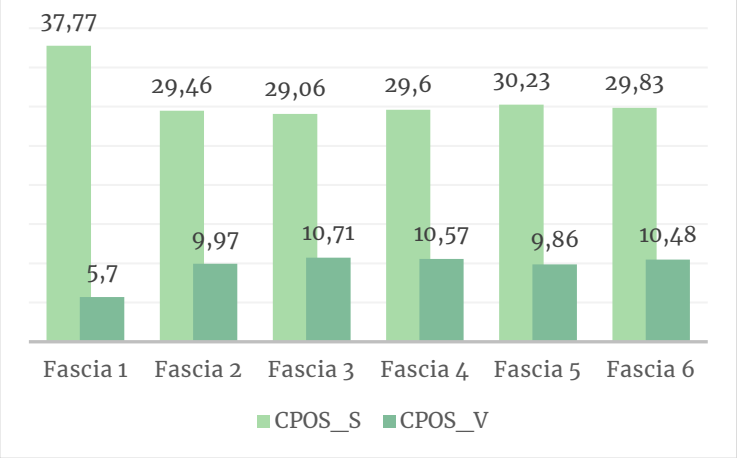

GRAFICO 5.10: Confronto tra la variazione delle distribuzioni (percentuali) di nomi e verbi tra le fasce

Nello studio di tutte le *features* riguardanti la sintassi si riscontra una situazione analoga: andamenti interni per lo più casuali, con rarissime correlazioni, e una significatività statistica nel confronto della prima fascia con la seconda e con la sesta.

Articoli Scientifici. L'analisi del corpus *ArtScient* ha prodotto risultati inediti: ogni parametro monitorato ha prodotto andamenti interni alle fasce fortemente correlati sia per il coefficiente di correlazione di Spearman che per quello di Pearson. Le caratteristiche di base sono spia di questa situazione generale:

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
Tokens	+	-/0	+	-	-	+
Chars	+	+	+	-	-	+
mCxT	0	0	0 ✓	0	0	0

TABELLA 5.29: Andamenti e significatività delle caratteristiche di base in *ArtScient*.

I coefficienti di correlazione relativi all'andamento delle distribuzioni delle categorie morfo-sintattiche superano abbondantemente la soglia minima dello 0,3 in ogni confronto. Questo dato, unitamente al fatto che il test di Wilcoxon non abbia individuato

alcuna significatività statistica, suggerisce che non esista una reale variazione tra i fenomeni linguistici delle diverse fasce, ma che, al contrario, questi si somiglino in modo molto importante.

In generale, quanto detto vale anche per le distribuzioni delle relazioni di dipendenza sintattica. D'altronde, spesso l'unico confronto a non registrare una correlazione tra gli andamenti delle *features* nelle fasce coinvolte è proprio quello che negli altri corpora aveva mostrato i risultati più significativi: il confronto tra la prima e l'ultima fascia.

Le distribuzioni delle caratteristiche sintattiche si comportano in modo analogo: ogni fascia risulta essere correlata con la successiva. Ciò conferma che per i test statistici, le fasce in questioni provengano dalla "stessa popolazione", ovvero non vi sono fenomeni in grado di discriminare l'una dall'altra.

Per quanto riguarda l'ordine relativo dei soggetti, degli oggetti, degli aggettivi e degli avverbi, gli andamenti conservano la loro tipicità quando gli elementi ricorrono in posizione canonica (preverbale per il soggetto, postverbale per l'oggetto, postnominale per l'aggettivo e preverbale per l'avverbio). Al contrario, in posizione marcata, i valori delle *features* variano in maniera più causale. Anche in questo caso, comunque, non è stata registrata alcuna significatività statistica per le variazioni subite dalle distribuzioni in questione tra le fasce.

	Fascia 1-2	Fascia 2-3	Fascia 3-4	Fascia 4-5	Fascia 5-6	Fascia 1-6
subjPre						
subjPost						
objPre						
objPost						
adjPre					✓	
adjPost						
advPre						
advPost						

TABELLA 5.30

Le subordinate, invece, conservano un andamento tipico in tutte le fasce, con l'unica eccezione della fascia centrale, a livello della quale si registra un repentino decremento del totale di subordinate. Nel confronto tra la terza e la quarta fascia, infatti, anche l'altezza e l'ampiezza media delle subordinate subisce un cambio di rotta, interrompendo quell'andamento tipico che invece contraddistingue i confronti tra tutte le altre fasce. È bene segnalare, inoltre, che generalmente gli andamenti si mantengono tipici quando la subordinata, sia di primo che di secondo grado, occupa la sua posizione canonica, ovvero posposta alla clausola dalla quale dipende.

5.5. Esperimento n.4: sulla variazione dei valori delle *features* tra le fasce dei testi semplici e complessi all'interno di ogni genere.

Il quarto esperimento è stato condotto mettendo a confronto le fasce dei testi appartenenti ai corpora rappresentanti la varietà semplice con le fasce corrispondenti nei testi della varietà complessa al fine di individuare variazioni statisticamente significative. Il test statistico utilizzato è stato il *Wilcoxon rank-sum test*.

5.5.1. Il procedimento

Gli studi sono stati condotti internamente ad ogni genere testuale, a partire dalle tabelle esemplificate in FIGURA 5.1, la cui genesi è stata descritta al PAR. 5.4.1. Nello specifico, è stato effettuato il test di Wilcoxon tra la fascia *n*, corrispondente alla *n-esima* colonna della tabella, del corpus *X*, e la corrispettiva fascia del corpus *Y*, dove *X* e *Y* sono corpora appartenenti allo stesso genere ma di complessità linguistica opposta.

Da questo confronto, sono state generate un numero di tabelle pari al numero di *features*, in cui ad ogni fascia è stato associato il *p-value*, la media dei valori registrati per quella *feature* in ogni fascia di entrambi i corpora coinvolti nel confronto, la differenza tra le medie, le variazioni e le deviazioni standard (FIGURA 5.3).

	A	B	C	D	E	F	G	H	I
1	features	pvalue	media1	media2	mediaDIFF	var1	var2	dev1	dev2
2	fascia 1	0,08492	0,76	0,93	-0,17	1,43	1,56	1,2	1,25
3	fascia 2	0,37311	1,41	0,79	0,62	3,91	1,09	1,98	1,04
4	fascia 3	0,11121	0,53	0,95	-0,42	0,65	1,66	0,81	1,29
5	fascia 4	0,47052	0,87	0,94	-0,07	1,97	1,75	1,4	1,32
6	fascia 5	0,7013	1,02	0,9	0,12	1,83	0,8	1,35	0,9
7	fascia 6	0,20489	1,04	1,14	-0,1	3,34	1,88	1,83	1,37

FIGURA 5.3: Esempio di prospetto finalizzato al confronto
tra la variazione dei valori nelle fasce corrispondenti in ogni corpus

5.5.2. I risultati

Per interpretare correttamente i risultati ottenuti da questo esperimento, è necessario fare qualche premessa.

Innanzitutto, bisogna tener conto del fatto che il test di Wilcoxon venga tipicamente utilizzato per verificare se due campioni indipendenti provengono dalla stessa popolazione. In questo caso, i due campioni, ovvero le liste dei valori attestati nelle fasce del corpus semplice e in quelle del corpus complesso, sicuramente appartengono a popolazioni diverse, ovvero a varietà appartenenti allo stesso genere testuale ma complementari dal punto di vista della complessità linguistica. Dunque, questo studio è interessante non tanto per individuare le variazioni in grado di discriminare un campione dall'altro, quanto più per rintracciare, nel confronto tra i corpora appartenenti alle differenti varietà linguistiche, le porzioni di testo all'interno delle quali non si verifichi una variazione staticamente significativa delle *features* monitorate. In quei casi, il fatto che non vi sia una variazione è sicuramente attribuibile ad una caratteristica del genere testuale, piuttosto che della varietà linguistica.

Una seconda precisazione è necessaria per giustificare la scelta di non considerare, in questo studio, le distribuzioni percentuali delle *part-of-speech* e delle relazioni di dipendenza. Per poter confrontare due percentuali, infatti, è necessario innanzitutto che queste appartengano alla stessa quantità *base*, rappresentante il 100% del totale, o, nel caso limite, a *basi* diverse aventi però lo stesso numero di partizioni. Nelle analisi condotte sui corpora, tuttavia, le distribuzioni delle caratteristiche morfo-sintattiche o delle relazioni di dipendenza non hanno mai presentato regolarità di questo tipo. Per chiarire quanto detto, viene proposto un semplice esempio: siano X e X_1 le quantità base descriventi la totalità dei fenomeni linguistici attestati in due corpora differenti, pari a n nel primo corpus e a m nel secondo, con $n > m$. Sia in X che in X_1, un certo fenomeno si manifesta con la stessa frequenza x; tuttavia, poiché in X si registra un numero maggiore di fenomeni linguistici, la percentuale che tale fenomeno assumerà sul totale risulterà essere minore della percentuale calcolata su X_1, in quanto dovrà dividere la quantità *base* con un numero maggiore di partizioni, rappresentante ognuna un fenomeno individuato. Nel confronto tra le due percentuali, dunque, il fenomeno verrebbe erroneamente considerato più frequente in X_1, in cui la sua percentuale risulta maggiore.

Ultimo appunto riguarda la notazione utilizzata per indicare il segno della differenza tra le medie dei valori di ogni fascia: si è scelto, infatti, di segnalare esclusivamente i casi particolari, ovvero quelli di segno positivo, dove il valore più alto si riscontra nella varietà semplice piuttosto che in quella complessa.

Il genere didattico. Uno sguardo d'insieme alle caratteristiche di base evidenzia immediatamente il fatto che la variazione della lunghezza delle frasi in tokens o in caratteri tra le fasce del corpus *Elem* e quelle di *Sup* sia sempre estremamente significativa.

Elem / Sup	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
Tokens	✓	✓	✓	✓	✓	✓
Chars	✓	✓	✓	✓	✓	✓
mCxT	✓	✓	✓	✓	✓	✓

TABELLA 5.31: Livelli di significatività delle caratteristiche di base nel confronto tra i corpora *Elem* e *Sup*

Una tale evidenza è resa manifesta dal GRAFICO 5.1, in cui è messo a confronto l'andamento della lunghezza delle frasi dei testi semplici con quello nei testi complessi del genere didattico.

Interessante è anche il confronto riguardante la distanza media tra un token e la sua testa sintattica (`mDist`), la quale mette in luce l'andamento altalenante che il parametro assume nei testi di *Elem*.

Elem / Sup	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
maxDist	✓	✓	✓	✓	✓	✓
mDist		✓		✓		✓

TABELLA 5.32: Livelli di significatività delle caratteristiche riguardanti la lunghezza dei link sintattici nel confronto tra i corpora *Elem* e *Sup*

Per rendere conto di questa particolarità, viene proposto il GRAFICO 5.11, in cui è evidente come la variazione tra i due valori assuma significatività statistica nei punti in cui lo scarto tra le due medie diventa più importante.

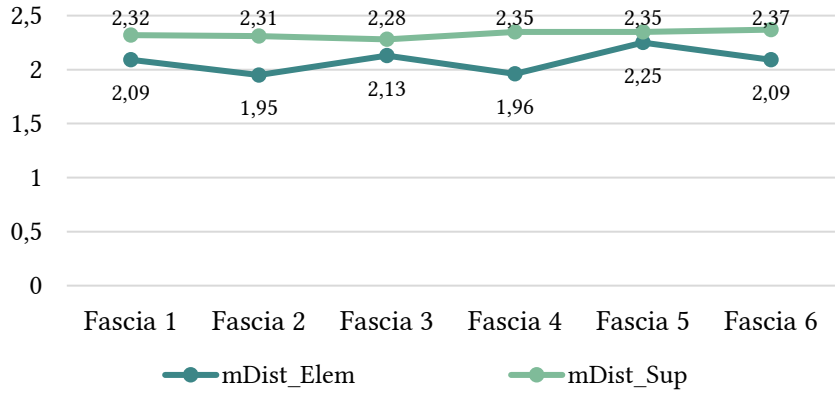

GRAFICO 5.11: Lunghezza media (in parole) fra testa e dipendente nel confronto tra i corpora *Elem* e *Sup*

Non si evidenziano particolari interessanti per quanto riguarda l'altezza degli alberi sintattici generati dalle frasi, generalmente maggiore in ogni fascia dei testi complessi. La variazione delle ampiezze dei suddetti alberi, invece, perde significatività nelle fasce conclusive del testo. La variazione del numero di figli per tokens, invece, perde di significatività statistica nella fascia centrale e nell'ultima.

Elem / Sup	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
mHeight	✓	✓	✓	✓	✓	✓
mWeight	✓	✓	✓	✓		
mChildren	✓	✓		✓	✓	
mChildrenS		✓	✓	✓	✓	✓
mChildrenV				+	✓	+

TABELLA 5.33: Livelli di significatività delle caratteristiche sintattiche nel confronto tra i corpora *Elem* e *Sup*

Generalmente poco significative sono, al contrario, le variazioni registrate rispetto ai fenomeni relativi all'ordine sintattico degli elementi. In particolare non si registra alcuna significatività per quanto riguarda i soggetti, sia in posizione preverbale che postverbale, e per gli oggetti in posizione canonica. La variazione della distribuzione degli oggetti in posizione marcata, con anteposizione dell'oggetto alla testa verbale, invece, risulta essere significativa nel confronto tra le prime e le ultime fasce dei due corpora in esame. Generalmente non significative sono, invece, le distribuzioni dell'ordine degli avverbi, al contrario di quanto accade per gli aggettivi, come si evince dalla TABELLA 5.34.

Elem / Sup	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
objPre(%)	✓	+	+	+		✓
objPost(%)						
adjPre(%)	+		✓		+	+
adjPost(%)		✓		✓	✓	
advPre(%)				+	+	+
advPost(%)					✓	

TABELLA 5.34: Livelli di significatività dell'ordine relativo dei costituenti nel confronto tra i corpora *Elem* e *Sup*

La subordinazione ricalca la generale differenza di complessità sintattica tra i due corpora in esame, riscontrando variazioni per lo più significative nella quasi totalità dei parametri monitorati (TABELLA 5.35). Nello specifico, sono soprattutto le prime due fasce a registrare le differenze più significative, che pure non mancano nella parte conclusiva dei testi. Generalmente, dunque, la subordinazione risulta essere un fenomeno maggiormente attestato nella varietà complessa del genere didattico.
Anche quanto avviene nella terza, quarta e quinta fascia è degno di nota: la variazione del numero di subordinate non viene valutata come discriminante per distinguere le fasce di uno o dell'altro corpus, anche se appare evidente che le subordinate dei testi complessi registrino valori generalmente più alti sia in termini di altezza che di ampiezza degli alberi sintattici da esse generati.

Inoltre, il fatto che, nella terza fascia, la percentuale di subordinate di primo grado (`subMain(%)`) registri valori mediamente maggiori nei testi semplici piuttosto che in quelli

complessi è prova della presenza di un numero maggiore di subordinate di secondo grado o di grado superiore nella varietà complessa del genere, indice di una più profonda organizzazione sintattica della parte centrale dei testi.

Elem / Sup	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
subTOT	✓	✓				✓
mHeightSubTOT	✓	✓		✓	✓	✓
mWeightSubTOT	✓	✓		✓	✓	
subMain(%)		✓	+ ✓	✓	✓	+
mHeightSubMain	✓	✓		✓		✓
mWeightSubMain	✓	✓		✓	✓	✓
subMinor(%)	✓	✓	✓			✓
mHeightSubMinor	✓	✓	✓			✓
mWeightSubMinor	✓	✓	✓			✓

TABELLA 5.35: Livelli di significatività dei parametri riguardanti la subordinazione nel confronto tra i corpora *Elem* e *Sup*

Il genere giornalistico. Nei risultati ottenuti per il genere giornalistico, si manifestano chiaramente le particolarità già riscontrate nel terzo esperimento per *DueParole* e *Repubblica*. In entrambi i corpora, infatti, i valori assunti dalle *features* nella prima fascia si erano resi responsabili di andamenti anomali rispetto a quelli registrati nel confronto tra le altre fasce. In particolare, la lunghezza media delle frasi nella prima fascia di *Repubblica* risultava essere sensibilmente inferiore non solo a quello delle altre fasce ma anche al valore calcolato nella prima fascia del corpus *DueParole*. Quest'ultimo, al contrario, si era attestato come il maggiore tra i valori delle altre fasce, caratterizzate da frasi mediamente più corte (GRAFICO 5.4). In generale, questa particolare situazione era stata riscontrata per la maggior parte dei fenomeni linguistici.
In questo studio, quanto appena detto si traduce in variazioni estremamente significative tra i valori assunti dalle *features* nella prima fascia dei testi di *DueParole* e in quella dei testi di *Repubblica*. Inoltre, il segno positivo delle differenze tra le loro medie ha messo in luce come i valori estratti dalle diverse *features* nella prima fascia di *2Par* siano mediamente maggiori di quelli estratti dalle stesse *features* nella prima fascia di *Rep*.

In tutte le altre fasce, invece, la situazione si uniforma ai risultati ottenuti nel confronto tra gli altri corpora oggetto di studio: variazioni per lo più statisticamente significative, con valori medi generalmente più alti nei testi complessi.

2Par/Rep	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
Tokens	+ ✓	✓	✓	✓	✓	✓
Chars	+ ✓	✓	✓	✓	✓	✓
mCxT	✓	✓		✓	+	+
maxDist	+ ✓	✓	✓	✓	✓	✓

mDist	+	✓		✓		✓		✓		✓		✓
mHeight	+	✓		✓		✓		✓		✓		✓
mWeight	+	✓		✓		✓		✓		✓		✓
mChildren	+	✓	0	✓	0	✓	0	✓	0		0	

TABELLA 5.35: Livelli di significatività di alcune *features* di interesse nel confronto tra i corpora *2Par* e *Rep*

In una situazione tanto regolare, è interessante il dato registrato per il numero medio di figli per token. Innanzitutto, la differenza dei valori registrati in entrambi i corpora, ad eccezione di quella calcolata tra le rispettive prime fasce, è approssimabile a 0: ciò significa che i valori sono pressoché identici, nonostante i testi appartengano a due complessità linguistiche differenti. Ancora più interessante, è il fatto che il *p-value* non raggiunga la soglia minima per discriminare la fascia 5 e la fascia 6 del corpus *2Par* dalle corrispondenti in *Rep*.

Il prospetto dedicato all'ordine dei costituenti (TABELLA 5.36) mostra con chiarezza quello che era già apparso chiaro nel confronto *2Par/Rep* effettuato nel secondo esperimento: gli elementi monitorati nel corpus semplice si manifestano perlopiù in posizione canonica, quella preverbale per il soggetto e quella postverbale per l'oggetto; al contrario, l'ordine marcato che il soggetto posposto e l'oggetto preposto al verbo diventa cifra stilistica della varietà complessa. Unica nota di contrasto riguarda la distribuzione del soggetto posposto al verbo nella prima fascia, che registra un valore comunque più alto nel corpus *DueParole*.

2Par/Rep	Fascia 1		Fascia 2		Fascia 3		Fascia 4		Fascia 5		Fascia 6	
subjPre(%)	+	✓	+	✓	+	✓	+	✓	+	✓	+	✓
subjPost(%)	+	✓		✓		✓		✓		✓		✓
objPre(%)		✓		✓		✓		✓		✓		✓
objPost(%)	+	✓	+		+	✓	+	✓	+	✓	+	✓

TABELLA 5.36: Livelli di significatività dell'ordine relativo dei costituenti nel confronto tra i corpora *2Par* e *Rep*

L'utilizzo delle subordinate è chiaramente prerogativa della varietà complessa del genere giornalistico, come è evidente dalle variazioni estremamente significative ottenute confrontando le fasce dei testi giornalistici semplici e quelle dei testi complessi. Per quanto riguarda il fenomeno della subordinazione relativo alla fascia 1, invece, la situazione è ben diversa: la variazione del numero di subordinate non produce alcuna significatività nel confronto tra *DueParole* e *Repubblica*. Fanno eccezione i parametri riguardanti le subordinate di grado superiore al primo: la quasi totale assenza di queste strutture nel corpus *DueParole* genera variazioni di segno negativo statisticamente significative.

2Par/Rep	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
subTOT		✓	✓	✓	✓	✓

mHeightSubTOT	+	✓	✓	✓	✓	✓
mWeightSubTOT	+	✓	✓	✓	✓	✓
subMain(%)	+ ✓	✓	✓	✓		
mHeightSubMain	+	✓	✓	✓	✓	✓
mWeightSubMain	+	✓	✓	✓	✓	✓
subMinor(%)	✓	✓	✓	✓	✓	✓
mHeightSubMinor	✓	✓	✓	✓	✓	✓
mWeightSubMinor	✓	✓	✓	✓	✓	✓

TABELLA 5.37: Livelli di significatività dei parametri riguardanti la subordinazione nel confronto tra i corpora *2Par* e *Rep*

Il genere narrativo. Anche in questo esperimento, le analisi condotte sui corpora appartenenti al genere narrativo non hanno condotto a risultati particolarmente significativi. Ciò è chiaramente indice dell'elevata similarità dei testi che compongono l'uno e l'altro corpus.

Gli unici parametri che hanno assunto variazioni degne di nota appartengono principalmente al livello sintattico. In particolare, sono significative le differenze tra i valori assunti dalle altezze e dalle ampiezze degli alberi sintattici delle frasi nella quarta fascia e quinta fascia.

TT-sempl/TT-orig	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
mHeight				✓	✓	
mWeight					✓	

TABELLA 5.38: Livelli di significatività delle caratteristiche strutturali degli alberi sintattici nel confronto tra i corpora *TT-sempl e TT-orig*

I risultati ottenuti dal monitoraggio delle subordinate, invece, risultano maggiormente interessanti. Com'è evidente dalla TABELLA 5.39, infatti, la subordinazione diviene caratterizzante a livello della seconda fascia e della quarta, nel confronto dei valori registrati per le suddette fasce in *TT-sempl* e in *TT-orig*.

TT-sempl/TT-orig	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
subTOT		✓		✓		
mHeightSubTOT				✓		
mWeightSubTOT		✓		✓		
subMain(%)						
mHeightSubMain		✓		✓		
mWeightSubMain		✓		✓		
subMinor(%)		✓		✓	✓	
mHeightSubMinor		✓				
mWeightSubMinor		✓		✓		

TABELLA 5.39: Livelli di significatività dei parametri riguardanti la subordinazione nel confronto tra i corpora *TT-sempl e TT-orig*

Il genere scientifico. Già nel secondo esperimento si era resa evidente la marcata differenza tra le due collezioni di testi scelte per rappresentare la varietà semplice e complessa del genere scientifico. Si vuole ora individuare in quali sezioni del testo questa discrepanza risulti essere più o meno marcata.
Innanzitutto, la significatività delle variazioni del numero di tokens per frase assume un andamento altalenante: è attestata nella prima, nella terza e nelle ultime due fasce, mentre nella seconda e nella quarta i due corpora non registrano differenze significative.

Wiki/ArtScient	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
Tokens	✓		✓		✓	✓
Chars	✓	✓	✓	✓	✓	✓
mCxT		✓		+		

TABELLA 5.40: Livelli di significatività delle caratteristiche di base nel confronto tra i corpora *Wiki* e *ArtScient*

Per quanto riguarda le caratteristiche sintattiche di base, si riscontrano variazioni particolari.

Wiki/ArtScient	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
maxDist	✓	✓	✓	✓		✓
mDist	✓	+	+	+	+	✓

TABELLA 5.41: Livelli di significatività delle caratteristiche riguardanti la lunghezza dei link sintattici nel confronto tra i corpora *Wiki* e *ArtScient*

Specificatamente, la distanza lineare in tokens tra testa e dipendente risulta essere un parametro discriminante esclusivamente per la prima fascia e per l'ultima, a livello delle quali si registrano le uniche variazioni di segno negativo: in ogni altra fascia la distanza media risulta essere maggiore, seppur lievemente, nei testi semplici.

Le caratteristiche strutturali degli alberi sintattici delle frasi forniscono ulteriori osservazioni interessanti.

Wiki/ArtScient	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
mHeight	✓	✓	✓		✓	✓
mWeight	✓				+	+
mChildren	✓	+	+	+	+	0
mChildrenS	✓	0	0	0	0	0
mChildrenV	✓	+	+	+	+	+

TABELLA 5.42: Livelli di significatività delle caratteristiche sintattiche nel confronto tra i corpora *Wiki* e *ArtScient*

Innanzitutto, si nota che nelle parti centrali e conclusive dei testi dei due corpora non vi sono reali differenze. Infatti, l'unico parametro per il quale il test di Wilcoxon individua una significatività statistica al di fuori della prima fascia è la media delle altezze degli alberi sintattici generati dalle frasi del testo. In particolare, la variazione di questo valore

tra le fasce corrispondenti nei due corpora risulta essere sempre significativa, eccetto che nella quarta fascia, a livello della quale si riscontra una identità quasi perfetta dei valori registrati nella varietà complessa e in quella semplice del genere scientifico, come si evidenzia nel GRAFICO 5.12.

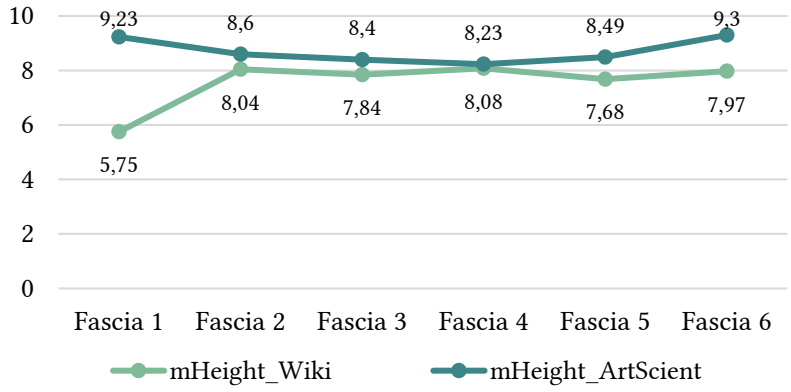

GRAFICO 5.12: Altezza media degli alberi sintattici nel confronto tra i corpora *Wiki* e *ArtScient*

Infine, il fatto che le frasi della quarta fascia dei testi complessi in un certo senso "riducano" la loro complessità, avvicinandosi a quelle della fascia corrispondente dei testi semplici, è testimoniato anche dall'utilizzo delle subordinate. Generalmente, infatti, i valori dei parametri monitorati relativi al fenomeno linguistico della subordinazione subiscono variazioni significative, nella direzione attesa, ovvero con valori più elevati nella varietà complessa. Eccezione importante avviene a livello della quarta fascia, in cui si perde la significatività statistica del confronto tra i valori attestati nei due corpora per la quasi totalità dei parametri.

	Fascia 1	Fascia 2	Fascia 3	Fascia 4	Fascia 5	Fascia 6
subTOT	✓	✓	✓		✓	✓
mHeightSubTOT	✓	✓	✓		✓	
mWeightSubTOT	✓		✓		✓	
subMain(%)	✓	+	+		✓	+
mHeightSubMain	✓	✓	✓		✓	✓
mWeightSubMain	✓	✓	✓		✓	
subMinor(%)	✓	✓	✓		✓	✓
mHeightSubMinor	✓	✓	✓		✓	✓
mWeightSubMinor	✓	✓	✓	✓	✓	✓

TABELLA 5.43: Livelli di significatività dei parametri riguardanti la subordinazione nel confronto tra i corpora *Wiki* e *ArtScient*

Conclusioni

In questo elaborato è stato effettuato, internamente a quattro generi testuali differenti (didattico, giornalistico, narrativo e scientifico), uno studio comparativo sulla distribuzione dei fattori di complessità linguistica nei testi appartenenti a due varietà linguistiche distinte per grado di complessità, definito in relazione al lettore di riferimento.

Per poter realizzare uno studio di questo tipo, sono stati necessari diversi passaggi preliminari. Innanzitutto, i corpora scelti sono stati annotati in maniera automatica fino al livello di analisi sintattica a dipendenze. Dalle frasi dei documenti di ogni corpus sono state estratte, tramite l'implementazione di *script* ad hoc, 88 caratteristiche linguistiche oggetto di monitoraggio, appartenenti a tre categorie generali: di base, morfo-sintattiche e sintattiche. I vettori di *features* risultanti sono stati accorpati in sei fasce, corrispondenti a porzioni consecutive di testo, in modo da poter confrontare documenti con un numero diverso di frasi. Ogni fascia è stata popolata dalle medie dei valori che ha racchiuso.

Le analisi sono state condotte su due livelli: uno più generale, considerando il testo nella sua complessità, l'altro più specifico, incentrato sulle parti in cui il testo è stato suddiviso. Il primo livello è stato indagato al fine di individuare i fenomeni linguistici più stabili e caratterizzanti della varietà linguistica semplice o complessa di ogni genere e, al contempo, quelle che variano maggiormente tra la varietà semplice e complessa dello stesso genere. Il secondo livello è stato indagato al fine di individuare gli andamenti dei fenomeni monitorati all'interno delle fasce adiacenti e il grado di correlazione di questi. Inoltre, si è voluto confrontare questi andamenti nella varietà semplice e in quella complessa di ciascun genere e valutarne la significatività statistica.

A monte degli studi, inoltre, si era ipotizzato che tra la prima e l'ultima fascia si sarebbero sicuramente individuate delle variazioni statisticamente significative, essendo l'introduzione e la conclusione parti del testo molto caratterizzati, così si è scelto di approfondire il confronto tra la prima e la sesta fascia.

I risultati sono stati in varia misura interessanti, a seconda del genere testuale dal quale sono stati ottenuti, delle fasce coinvolte nel confronto o del tipo stesso di confronto effettuato. Ma procediamo con ordine.

Si può sicuramente affermare che i corpora che hanno prodotto risultati generalmente più interessanti sono stati appartenenti al genere giornalistico e al genere scientifico. Le varietà semplice e complessa, in questi casi, sono risultate essere estremamente distanti, sia nel confronto condotto sul testo nella sua totalità, sia per quanto riguarda lo studio sulle fasce. I corpora appartenenti al genere didattico, e in misura ancora maggiore per il genere narrativo, invece, hanno restituito un quadro generale ben diverso, mostrando come il genere risulti meno influenzato dalla variazione interna rispetto alla complessità.

Già nel primo esperimento, che ha condotto le sue indagini sul livello più generale, i corpora del genere didattico e quelli del genere narrativo hanno riscontrato una quasi totale coincidenza dei parametri più stabili, indipendentemente dal livello di complessità considerato. Generalmente, questi parametri riguardavano aspetti sintattici della frase e caratteristiche strutturali degli alberi sintattici. Al contrario, il confronto interno al genere giornalistico ha evidenziato come i parametri più stabili del corpus *DueParole* fossero quelli considerati più variabili in *Repubblica*. Questi parametri riguardavano generalmente caratteristiche linguistiche di base, come la lunghezza delle frasi, e sintattiche, come l'altezza media degli alberi sintattici o ancor di più i parametri preposti al monitoraggio della subordinazione. La situazione si è poi presentata inversa per *Repubblica*, i cui parametri più caratterizzanti sono risultati tra quelli più altalenanti in *DueParole*.
Situazione analoga, se non ancor più estrema, è stata riscontrata nel confronto interno al genere scientifico, nel quale è risultata evidente la maggior complessità delle strutture sintattiche della varietà complessa in confronto a quella ben più basilare e scarna della varietà semplice.

Ciò che il primo esperimento aveva anticipato, è stato poi in qualche misura confermato dalla seconda analisi, che ha avuto il compito di individuare i parametri in grado di discriminare la varietà semplice da quella complessa per ciascun genere. Se per i corpora appartenenti al genere giornalistico e al genere scientifico la variazione della quasi totalità dei parametri monitorati è risultata significativa nel confronto interno, la distinzione tra i testi semplici e complessi del genere didattico e, di nuovo, ancor più del genere narrativo, è stata molto meno accentuata. Sicuramente, in entrambi i casi, non si può dire che le distribuzioni delle caratteristiche morfo-sintattiche, delle relazioni di dipendenza e dell'ordine canonico e marcato dei costituenti, abbiano giocato un ruolo fondamentale nei tentativi di discriminazione tra le due

varietà di lingua. Molto più interessante è, invece, ciò che accade in relazione a caratteristiche puramente sintattiche, quali la distanza media e la distanza massima registrata tra testa e dipendente, l'altezza e l'ampiezza media degli alberi sintattici delle frasi, la media di dipendenti per tokens e, con assoluta evidenza, l'utilizzo della subordinazione, tutta a vantaggio della varietà complessa.

Una volta appurata l'esistenza di *features* più caratterizzanti della varietà complessa piuttosto che in quella semplice, si è voluto indagare il modo in cui queste varino all'interno del testo, attraverso la suddivisione dei testi in parti consecutive.
Due sono stati gli esperimenti dedicati a questo studio: il primo, come già anticipato, ha voluto evidenziare quali siano gli andamenti tipici delle *features* tra una fascia e la successiva all'interno dello stesso corpus, utilizzando l'opera congiunta del test di Wilcoxon e dell'assegnazione dei coefficienti di correlazione di Pearson e di Spearman; il secondo, invece, ha voluto mettere a confronto i valori attestati nelle fasce dei testi semplici con quelli attestati nelle fasce corrispondenti dei testi complessi, al fine di individuare variazioni statisticamente significative.

Il terzo esperimento ha registrato tra le fasce dei testi del genere didattico andamenti perlopiù altalenanti. Le fasce dei testi semplici hanno riscontrato correlazioni più frequenti, in particolare per quanto riguarda le caratteristiche di base, e variazioni di maggiore interesse. Ad esempio, l'utilizzo della subordinazione nella varietà semplice del genere didattico genera variazioni statisticamente significative tra la seconda e la terza fascia e andamenti correlati nelle fasce centrali e in quelle conclusive. Anche le distribuzioni dell'ordine dei costituenti offrono spunti di riflessione di rilievo, in quanto la percentuale di soggetti e oggetti in posizioni marcate registrano un incremento nella parte centrale del testo. In generale, grazie ai risultati del quarto esperimento, è risultato evidente che ogni fascia dell'uno e dell'altro corpus generi variazioni significative, in particolare in relazione ai fenomeni sintattico.

Relativamente poco significativi sono stati i risultati riscontrati in entrambi gli esperimenti per i corpora del genere narrativo, unico dominio per il quale si è dovuto ricorrere al confronto tra la prima e la sesta fascia per rintracciare una qualche significatività. I risultati del quarto esperimento, invece, svelano l'utilizzo della subordinazione come uno dei fattori di complessità in grado di discriminare la seconda e la quarta fascia delle due varietà linguistiche del genere.

Nell'analisi di *DueParole* e *Repubblica*, invece, sono stati individuati risultati ben più definiti. Innanzitutto, entrambi i corpora hanno registrato andamenti anomali a livello della prima fascia, dato tranquillamente attribuibile ad una caratteristica del genere testuale. Il fatto curioso riscontrato è che questa "anomalia" abbia andamenti inversi: se nella prima fascia *DueParole* ha registrato una lunghezza media superiore, e in generale di qualsiasi altro parametro linguistico, in Repubblica, allo stesso livello, sono state riscontrata frasi più brevi e una sintassi meno complessa, come lo dimostrano i risultati ottenuti nel quarto esperimento per quanto riguarda la subordinazione: l'utilizzo delle subordinate in *Repubblica*, infatti, è discriminante nel confronto con ogni fascia di *DueParole*, eccetto che nella prima. Questo fatto può essere attribuito alla presenza, nella prima parte dei testi di Repubblica, dei titoli degli articoli giornalistici, i quali sono caratterizzati da una brevità tutta tipica del genere giornalistico, che invece mancano in *DueParole*.

Una situazione simile è stata riscontrata nello studio delle fasce del corpus *Wikipedia:* la prima fascia, infatti, si è distinta per valori generalmente inferiori di ogni parametro linguistico rispetto a quelli registrati nelle restanti fasce in cui è stato diviso il testo, a cominciare dalla lunghezza delle frasi fino all'ultima delle *features* sintattiche. La giustificazione di una tale differenza è stata rintracciata di nuovo nella presenza, in primis, del titolo dell'articolo e, in secundis, di una sorta di introduzione o di sunto dell'argomento trattato, ad alto contenuto informativo, responsabile oltretutto del massimo scarto registrato tra la percentuale di sostantivi e di verbi, tra quelli calcolati nelle altre fasce.
I documenti di *Articoli Scientifici*, al contrario, non mostrano alcuna variazione al loro interno: le fasce che li costituiscono sono generalmente correlate, con andamenti dei valori delle *features* in grado di generare tipicità. Tra i due corpora, come già detto in precedenza, risulta evidente la differenza di complessità linguistica. Il quarto esperimento, infatti, mostra in modo chiaro quasi ogni fascia assuma valori più importanti nei testi complessi, con rare eccezioni.

In sintesi, possiamo affermare che un monitoraggio linguistico che tenga conto non solo della presenza dei fenomeni linguistici, ma anche della loro distribuzione in posizioni diverse del testo, permetta di guardare al dato testuale da una nuova angolazione, avvicinandoci così ad una ricostruzione sempre più profonda del profilo linguistico dei testi.

APPENDICE A

Lista delle caratteristiche linguistiche

In questa appendice sono elencate le caratteristiche linguistiche estratte dalle frasi dei vari documenti, utilizzate per il monitoraggio della complessità delle frasi, con relativa descrizione. Le caratteristiche sono suddivise in 3 categorie: caratteristiche di base, morfo-sintattiche e sintattiche.

Caratteristiche di base

`Sentences`	indice della frase
`Tokens`	numero di tokens per frase (compresa la punteggiatura)
`Chars`	numero di caratteri per frase (compresa la punteggiatura)
`mCxT`	numero medio di caratteri per token

Caratteristiche morfo-sintattiche

`CPOS_#(%)`	distribuzione delle parti del discorso generiche
`POS_#(%)`	distribuzione delle parti del discorso più granulari
`POS_#*(%)`	distribuzione di POS_# sul totale di CPOS_#

Caratteristiche sintattiche

`mDist`	lunghezza media dei link sintattici (distanza tra testa e dipendente calcolata in numero di token, esclusa la punteggiatura)
`maxDist`	lunghezza del link sintattico più lungo, escludendo la punteggiatura
`LinkPre(%)`	percentuale di tokens che precedono la testa sintattica (link a sinistra)
`LinkPost(%)`	percentuale di tokens che seguono la testa sintattica (link a destra)

• relative alle relazioni di dipendenza:

`DEP_#(%)`	distribuzione delle relazioni sintattiche

• relative all'ordine dei costituenti:

`subjPre(%)`	percentuale di soggetti preverbali

subjPost(%)	percentuale di soggetti preverbali
objPre(%)	percentuale di oggetti preverbali
objPost(%)	percentuale di oggetti postverbali
adjPre(%)	percentuale di aggettivi prenominali
adjPost(%)	percentuale di aggettivi postnominali
advPre(%)	percentuale di avverbi preverbali
advPost(%)	percentuale di avverbi postverbali

▪ relative alle caratteristiche dell'albero sintattico:

mHeight	profondità, o altezza, media degli alberi sintattici
mWeight	ampiezza media degli alberi sintattico
mChildren	numero medio di figli per token
mChildrenS	numero medio di figli per sostantivo
mChildrenV	numero medio di figli per verbo

▪ relative alla subordinazione:

subTOT	numero di proposizioni subordinate
mHeightSubTOT	profondità media degli alberi sintattici delle subordinate
mWeightSubTOT	ampiezza media degli alberi delle subordinate
subMain(%)	percentuale di subordinate di primo grado sul totale delle subordinate
mHeightSubMain	profondità media degli alberi delle subordinate di primo grado
mWeightSubMain	ampiezza media degli alberi delle subordinate di primo grado
subMainPre(%)	percentuale di subordinate di primo grado che precedono la principale
subMainPost(%)	percentuale di subordinate di primo grado che seguono la principale
subMinor(%)	percentuale di subordinate di grado superiore al primo (II grado, III grado, ecc.) sul totale di subordinate
mHeightSubMinor	profondità media degli alberi delle subordinate di grado superiore al primo
mWeightSubMinor	ampiezza media degli alberi delle subordinate di grado superiore al primo
subMinorPre(%)	percentuale di subordinate di grado superiore al primo che precedono la subordinata reggente
subMinorPost(%)	percentuale di subordinate di grado superiore al primo che seguono la subordinata reggente

Bibliografia

Attardi, Giuseppe; Felice Dell'Orletta; Maria Simi; Joseph Turian. 2009. *Accurate dependency parsing with a stacked multilayer perceptron*. In "Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009", Reggio Emilia, Italia, Dicembre 2009.

Berruto, Gaetano. 1990. *Italiano regionale, commutazione di codice e enunciati mistilingui*. In "L'italiano regionale. Atti del XVIII Congresso della SLI", Michele Cortellazzo, Alberto Mioni. Roma: Bulzoni, pp. 103-127, 1990.

Berruto, Gaetano; Massimo Cerruti. 2011. *La linguistica. Un corso introduttivo*. UTET Universitá, Novara, Maggio 2011.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge & New York, Cambridge University Press.

Brunato, Dominique; Felice 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)", pp. 25-31, Pisa, Italia, September 18-20, 2017

Brunato, Dominique; Felice Dell'Orletta; Giulia Venturi; Simonetta Montemagni. 2015. *Design and annotation of the first italian corpus for text simplification*. In "Proceedings of LAW IX - The 9th Linguistic Annotation Workshop". Denver, Colorado, Giugno 2015.

Calzolari, Nicoletta; Alessandro Lenci. 2004. *Linguistica computazionale - strumenti e risorse per il trattamento automatico della lingua*. Mondo Digitale, 2, pp. 56–69.

Cangelosi, Angelo; Huck Turner. 2002. *L'emergere del linguaggio*. In "Scienze della Mente". A cura di Borghi A. M., Iachini T., pp. 227–244. Il Mulino, Bologna.

Corpina, Barbara. 2009. *Topic e Focus in Hdi. Strategie a confronto e analisi di testi*, 2009.

Crasta, Graziano. *Laboratorio di statistica*. 2007.

Dell'Orletta, Felice. 2009. *Ensemble system for part-of-speech tagging*. In "Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009", Reggio Emilia, Italia, Dicembre 2009.

Dell'Orletta, Felice; Simonetta Montemagni. 2012. *Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico*. In "S. Ferreri (a cura di), *Linguistica Educativa. Atti del XLIV Congresso Internazionale di Studi della SLI*", Roma, Bulzoni Editore, pp. 343-359.

Diessel, Holger. 2005. *Competing motivations for the ordering of main and adverbial clauses*. In "Linguistics", 43(3), pp. 449–470. 2005.

Ferguson, Charles. 1982. *Simplified registers and linguistic theory*. In "Exceptional Language and Linguistics". pp. 49–66. Academic Press, New York.

Fiorentino, Giuliana. 2009. *Complessità linguistica e variazione sintattica*. In "Studi Italiani di Linguistica Teorica e Applicata", Anno XXXVIII, Numero 2, pp. 281-312

Frazier, Lyn. 1985. *Syntactic complexity*. In D.R. Dowty, L. Karttunen e A.M. Zwicky (a cura di), "Natural Language Parsing", Cambridge University Press, Cambridge, UK.

Gallissot, René; Mondher Kilani; Annamaria Rivera. 2001. *L'imbroglio etnico in quattordici parole-chiave*. Nuova Biblioteca Dedalo: Antropo-logiche. Dedalo.

Gell-Mann, Murray; Merritt Ruhlen. 2011. *The origin and evolution of word order*. In "Proceedings of the National Academy of Sciences of the United States of America", pp. 17290-5, volume 108(42), October 2011.

Gibson, Edward. 1998. *Linguistic complexity: Locality of syntactic dependencies*. "Cognition", 68(1), pp. 1-76.

Gibson, Edward; Steven T. Piantadosi; Kimberly Brink;. Leon Bergen; Eunice Lim; Rebecca Saxe. 2013. *A noisy-channel account of crosslinguistic word-order variation*. Psychological Science, 24(7), pp. 1079–88.

Hawkins J. A. (1994). *A Performance Theory of Order and Constituency*. In "Cambridge Studies in Linguistics", 73. Cambridge: Cambridge University Press.

Hawkings, John. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John. 2009. *An efficiency theory of complexity and related phenomena*. In "Exceptional Language and Linguistics", Numero 13, pp. 252-268

Humboldt, Wilhelm von. 1836. *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts*. Berlin: Druckerei der königlichen Akademie der Wissenschaften.

Kusters, Wouter. 2003. *Linguistic Complexity The Influence of Social Change on Verbal Inflection*. In "LOT Dissertation Series". Numero 77. LOT, Utrecht.

Lenci, Alessandro; Simonetta Montemagni; Vito Pirrelli. 2005. *Testo e computer*. Roma, Carocci.

Lin, Dekan. 1996. *On the structural complexity of natural language sentences*. In "*Proceedings of COLING 1996"*, pp. 729–733.

McWhorter J. H. 2001. *The world's simplest grammars are creole grammars*. In "Linguistic Typology", Numero 5, pp. 125–166.

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

Nilsson, Jens; Sebastian Riedel; Deniz Yuret. 2007. *The CoNLL 2007 shared task on dependency parsing*. In "Proceedings of the CoNLL shared task session of EMNLP-CoNLL". sn. 2007, pp. 915–932.

Piemontese, Maria Emanuela. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnodid.

Pieri, Giulia; Dominique Brunato; Felice Dell'Orletta. 2016. *Studio sull'Ordine dei Costituenti nel Confronto tra Generi e Complessità (Analysis of Constituents Order Across Textual Genres and Complexity). CLiC-it/EVALITA* (2016).

Ross, Sheldon M. 2014. *Introduzione alla statistica*. Maggioli Editore.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, UK.

Treccani, 2011. voce *Ordine degli elementi*. http://www.treccani.it/enciclopedia/ordine-degli-elementi_(Enciclopedia_dell'Italiano)/ Ultima visita: 13/04/2019.

Treccani, 2010a. voce *Aggettivi*. http://www.treccani.it/enciclopedia/aggettivi_(altro)/. Ultima visita: 13/04/2019.

Wikipedia, 2019a. voce *Razzismo Scientifico*. https://it.wikipedia.org/wiki/Razzismo_scientifico. Ultima visita: 08/04/2019.

Yngve, Victor H.A. 1960. *A model and an hypothesis for language structure*. In: "Proceedings of the American Philosophical Society", pp. 444-466.