UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Lessico e sintassi: studio della variazione tra generi e complessità

Candidato: *Pietro dell'Oglio*

Relatori: *Dott.ssa Dominique Brunato,*
Dott. Felice Dell'Orletta

Correlatore: *Dott.ssa Simonetta Montemagni*

Anno Accademico 2016-2017

Alla mia famiglia

Sommario

CAPITOLO 1

Introduzione

Il concetto di complessità linguistica è stato oggetto di opinioni e riflessioni contrastanti. I principali punti di vista sono tre: quello matematico, quello psicolinguistico e quello empirico. Dal punto di vista linguistico, McWhorter (2001) afferma che una lingua può considerarsi complessa o più semplice sulla base delle caratteristiche fonetiche, morfologiche, sintattiche e semantiche non necessarie ai fini dell'efficacia comunicativa; da questo punto di vista, una semplificazione in uno solo di questi livelli potrebbe portare un aumento della complessità in altri.

Questo elaborato si focalizzerà sul lessico della lingua italiana e come esso possa variare tra i generi e i diversi livelli di complessità linguistica, e come correla la complessità lessicale alla complessità sintattica.

Per definizione, il lessico (o vocabolario) di una lingua è l'insieme di tutte le sue parole o locuzioni. Sottoinsiemi della lingua italiana sono il vocabolario considerato di base comprensivo di circa 7000 lessemi (*Chiari*, *De Mauro*, 2014), a sua volta distinto in fondamentale, di alto uso e di alta disponibilità, e il vocabolario comune. Insieme, vocabolario di base e comune rappresentano il *vocabolario corrente* della lingua italiana. Naturalmente vanno distinti altrettanti lessici settoriali e regionali.

L'obiettivo ultimo dell'analisi che sarà descritta in questo elaborato sarà quello di capire se e come varia la distribuzione del lessico (anche in relazione al VdB) e come questa sia influenzata dalle strutture sintattiche, all'interno di generi e varietà di lingua diversi.

L'analisi si propone di rispondere ad alcune domande. Quanto varia la complessità lessicale al variare dei generi? Come varia il lessico al variare della complessità del testo all'interno dello stesso genere? Ci si aspetta di trovare variazioni significative al variare della complessità; invece è molto meno chiaro prevedere possibili evidenze se consideriamo le variazioni di genere. Ancora, un ulteriore interrogativo che cercherà una risposta è relativo all'esistenza di una correlazione tra complessità lessicale e sin-

tattica: dato un testo di un certo livello di complessità, parole semplici e difficili occorrono in strutture linguistiche simili o diverse? E se questa correlazione esiste, come varia al variare dei generi e del livello di complessità del testo?

L'analisi è stata condotta rispetto a quattro generi e, per ognuno, due livelli di complessità, uno semplice e uno difficile.

Nel capitolo 2 sarà dato uno stato dell'arte sulla complessità linguistica, verranno introdotte le principali problematiche che comporta, la differenza tra lingua scritta e parlata e quella di complessità del sistema e per l'utente; nel capitolo 3 saranno descritti gli otto corpora utilizzati per l'analisi, divisi per quattro generi testuali differenti e due livelli di complessità ciascuno: rappresentanti del genere giornalistico sono *2Parole*, per quanto riguarda il livello di complessità basso, e *Repubblica*, per il livello di complessità alto; per il genere scientifico sono presenti *Wikipedia* e *Articoli Scientifici*, il primo considerato semplice e il secondo complesso; il genere dei materiali didattici (o scolastico) è rappresentato da un corpus di testi di *Scuola Elementare* e uno di *Scuola Superiore*; infine i corpora narrativi sono *Terence* e *Teacher*, composti da testi per l'infanzia e presenti sia nella loro versione originale che in quella semplificata. Sempre nel capitolo 3 saranno descritti il *Nuovo Vocabolario di Base della Lingua Italiana* (NVdB) di Tullio De Mauro e il corpus di riferimento itWaC; sarà, inoltre, introdotto il Wilcoxon rank-sum test, un test statistico non parametrico per due campioni indipendenti che, implementato in uno script, è stato utilizzato per calcolare il p-value, a sua volta utile a stabilire il livello di significatività delle variazioni riscontrate nei due esperimenti descritti all'interno dei due capitoli successivi. Il capitolo 4, infatti, è relativo all'analisi dei dati rispetto al lessico nel confronto tra generi e complessità, il capitolo 5, invece, è dedicato all'analisi dei dati estratti dallo studio sul lessico che cooccorre con alcuni fenomeni linguistici di tipo sintattico. Nel capitolo 6 saranno discussi i principali risultati dello studio, alla luce delle domande di ricerca iniziali, e riassunte le conclusioni riscontrate.

CAPITOLO 2

La complessità linguistica

In questo capitolo viene discusso il tema della complessità linguistica nella letteratura moderna. Verrà introdotto il concetto di complessità linguistica e spiegati alcuni problemi che lo riguardano (2.1), in particolare sarà discusso il problema della lingua scritta e parlata (2.1.1) ed evidenziata la differenza che intercorre tra la complessità dei sistemi linguistici e la complessità percepita dall'utente (2.1.2); infine, saranno riassunte le metriche della complessità nei differenti livelli della lingua (2.2).

2.1 – La complessità linguistica

Il dibattito sulla complessità prende avvio all'interno della matematica, della teoria dell'informazione e delle scienze computazionali. Nella letteratura moderna, ha iniziato ad apparire all'interno delle scienze sociali; in particolare nell'ambito della riflessione linguistica. Negli ultimi anni sono stati organizzati alcuni convegni dedicati, dimostrando come sia un tema sentito e molto discusso da linguisti anche appartenenti a orientamenti diversi (*Approaches to complexity in language*, Helsinki, 2005; *Language complexity as an evolving variable*, Max Planck, Lepizig, 2007; *TX The genesis of syntactic complexity*, Rice University, Houston, 2008).

I punti di vista più diffusi sono tre: quello matematico della teoria dell'informazione applicato allo studio del cambiamento linguistico; quello psicolinguistico, che si basa sul concetto di *processing*[1]; quello empirico, che si basa sulle complessità o ritardi dell'acquisizione di alcune strutture linguistiche dal punto di vista di un utente outsider (*Fiorentino*, 2009).

[1] Il costo di processamento dell'informazione da parte del parlante/ricevente.

La linguista Giuliana Fiorentino, nel suo breve saggio "*Complessità linguistica e variazione sintattica*", ripercorre brevemente come si è evoluta la nozione di *complessità* all'interno della storia del pensiero linguistico. All'inizio del XIX secolo essa viene ricollegata al concetto di *complessità di pensiero*. In questo panorama, le lingue indoeuropee venivano considerate più utili per esprimere il pensiero complesso e le lingue flessive venivano considerate più complesse di quelle isolanti e agglutinanti.
Più recentemente il dibattito ha preso una piega tutta nuova. Si sono susseguite idee contrastanti, come ad esempio la teoria delle pari complessità delle lingue, una teoria che è stata spesso obiettata e McWhorter (2001) dà spunto al dibattito a partire dai suoi studi di creolistica e da alcuni articoli sulle lingue creole, che lui stesso ha definito come *lingue dotate delle grammatiche più semplici*. In sintesi, per McWhorter una lingua può considerarsi complessa o più semplice sulla base delle caratteristiche fonetiche, morfologiche, sintattiche e semantiche che esulano dalla necessità comunicativa. Ferguson (1982) ha invece spostato l'indagine nel campo delle varietà e dei registri, e quindi verso fenomeni quali il *baby talk*, il *foreigner talk*, il *teacher talk* e i *pidgin* e i *creoli*.
Berruto (1990) dà una definizione lucida ed esauriente del concetto di complessità linguistica che si basa sull'impegno cognitivo che vede come protagonisti la lingua e l'utente (si veda sezione 2.1.2 per la complessità dal punto di vista dell'utente). La definizione è la seguente:

> "Per semplificazione linguistica, proponiamo di intendere il processo secondo cui a un elemento, forma o struttura X di una certa lingua o varietà di lingua si sostituisce/contrappone/paragona un corrispondente elemento, forma o struttura Y della stessa lingua o varietà di lingua o di un'altra lingua o varietà di lingua, tale che Y sia di più immediata processabilità, cioè più facile, più agevole, meno complesso, meno faticoso, meno impegnativo cognitivamente ecc. a qualche livello per l'utente."

In Italia il concetto di complessità è stato indagato e dibattuto nell'ambito degli studi sull'italiano popolare e parlato, spesso in senso contrastivo rispetto allo scritto, considerando le differenze di struttura e sintassi come un processo di semplificazione.

2.1.1 – Lingua scritta e lingua parlata

Tullio De Mauro (1970) spiega come la semplificazione della lingua scritta standard sia avvenuta per influenza del parlato; anche Berruto (1983) affronta il tema della semplificazione, ma lo fa a proposito di studi sull'italiano popolare.

Voghera (2001) afferma come un'opinione errata nel campo della complessità linguistica è quella che si riferisce alla lingua parlata come una varietà semplificata della lingua scritta. Parlato e scritto presentano differenze sostanziali nel modo e nei tempi in cui la comunicazione viene a operarsi. Infatti, sostiene, non è possibile valutare *ceteris paribus* la semplicità o la complessità di un atto linguistico nella sua struttura sintattica; al contrario, vanno sempre considerati i vincoli che vengono imposti dalla modalità di trasmissione usata, in particolare, appunto, se consideriamo la lingua parlata e scritta. Per ognuna di queste modalità vanno infatti presi in considerazione i meccanismi di produzione e ricezione, sempre diversi, di tipo fonico-uditivo per il parlato e di tipo grafico-visivo per lo scritto. Per le due modalità la pianificabilità consentita è completamente differente: nel caso della lingua scritta si riscontra spesso una distanza notevole dal parlante al ricevente dell'atto linguistico, mentre generalmente la distanza che intercorre tra i due nel caso del parlato è molto più breve.

Ridurre il materiale sintagmatico, porta Voghera (2001) come esempio, non equivale sempre ad avere una semplificazione, perché in questo modo si producono testi molto densi dal punto di vista lessicale e poco ridondanti che, nel caso del parlato, porterebbero a un innalzamento della complessità. La lingua parlata, infatti, proprio per la sua natura spesso dialogica, necessita di una rapida fruibilità, quindi fenomeni come la polisemia e la ridondanza sono spesso indice di semplificazione, al contrario di grandi nominalizzazioni e monosemia. La non ridondanza, in particolare, spesso è motivo di semplificazione per il produttore dell'atto linguistico, mentre per il ricevente può avere l'effetto opposto.

Voghera (2001) conclude che il problema principale nell'utilizzo della nozione di semplificazione è quello di considerare la lingua un'entità amodale, ovvero un insieme di relazioni che possono essere descritte a prescindere da quale sia la modalità di trasmissione utilizzata.

Fiorentino (2009), nonostante le posizioni critiche di Voghera a riguardo, afferma che in effetti è possibile considerare il parlato come più semplice dello scritto, almeno per

quanto riguarda alcune scelte sintattiche e lessicali (ad esempio, l'uso di frasi più brevi e lessico meno specialistico). Il tutto con le dovute cautele e precisazioni del caso.

2.1.2 – Complessità del sistema e complessità per l'utente

Il dibattito sulla complessità portato avanti dai linguisti è stato sviluppato in due modi: quello sulla complessità nel sistema e quello sulla complessità per l'utente.

La complessità nel sistema è definita confrontando i sistemi e strutture linguistiche sulla base di alcuni criteri interni alle lingue. Si tratta di definizioni di complessità che si basano sulla descrizione fatta dal linguista.

Nel caso della complessità per l'utente essa dipende da strategie cognitive messe in atto dal produttore, in fase di produzione del linguaggio, e dal ricevente, in fase di ricezione dello stesso. Si tratta di un approccio che valuta come più complesso qualcosa che richiede un tempo elevato per poter essere fruito e compreso, quindi uno sforzo cognitivo più alto.

Entrambi questi modi di definire la complessità sono nozioni generalmente considerate relative, cioè non assolute e dipendenti dalla marcatezza di un tratto linguistico in particolare; ciononostante alcuni linguisti tendono ad assolutizzare queste definizioni (*Fiorentino*, 2009).

2.2 – Una metrica della complessità

Se si ragiona sul concetto di complessità linguistica in termini assoluti sorge la necessità di definire una lingua come semplice o complessa sulla base di una metrica valida. Fiorentino (2009) porta come esempio la metrica proposta da McWhorter (1998, 2001), nella quale si considerano determinati fenomeni fonologici, morfologici e sintattici. Egli propone alcuni esempi per ciascun livello. Per quanto concerne la fonologia, la complessità si basa sulla marcatezza dei membri dell'inventario fonologico di una lingua; per quanto riguarda la morfologia abbiamo che quella flessiva è generalmente considerata più complessa rispetto a quella isolante; per la sintassi il livello di complessità aumenta di pari passo con il numero di regole.

Una metrica della complessità che possa invece essere applicata a tutte le strutture sintattiche è stata tentata più volte. In particolare Ferguson (1982) ha elencato alcuni tratti sintattici che sono dotati di più gradi di complessità (la paratassi rispetto all'ipotassi, la presenza o l'assenza di parole funzionali).

La complessità dal punto di vista del livello della semantica è strettamente connessa con la sintassi. Infatti la semplificazione a un certo livello della lingua potrebbe portare complessità in un altro livello della stessa (*Berruto*, 1990). Secondo Voghera (2001) il significato astratto, non conoscibile in altri modi se non attraverso la nostra mente, è più complesso del significato concreto; la polisemia, cioè il poter dare più significati a una stessa parola è indice di complessità maggiore rispetto alla monosemia[2]; generalmente il lessico funzionale (costituito dalle parole "vuote") è considerato più complesso del lessico referenziale.

Per calcolare la complessità morfologica dei testi è stato ideato un indice di complessità morfologica o ICM (*Pallotti,* 2015), basato su una formula semplice e immediata. È un indice che può essere calcolato in automatico attraverso un programma online che effettua il campionamento di un grande numero di campioni casuali, in modo tale da raggiungere stime affidabili della complessità media del testo esaminato. Non è, inoltre, dipendente dalla lunghezza del testo; infatti può essere applicato a testi anche molto brevi.

[2] Come abbiamo visto in 2.1.1 questo non è sempre vero.

CAPITOLO 3

Metodi d'indagine, strumenti e corpora utilizzati

Questo terzo capitolo è dedicato agli strumenti e i corpora utilizzati per portare avanti le analisi e i confronti che verranno indagati nei capitoli quarto e quinto. In particolare, a seguire saranno descritti gli otto corpora utilizzati per l'analisi, suddivisi per genere (giornalistico, scientifico, scolastico e narrativo) e complessità (facile o difficile), it-Wac, il Nuovo Vocabolario di Base per la lingua italiana di Tullio de Mauro, gli script utilizzati per estrarre i dati utili alle analisi e, ultimo ma non meno importante, il Wilcoxon rank-sum test.

3.1 – I Corpora analizzati

Un corpus è una collezione di testi e rappresenta una fonte di dati linguistici "ecologici" (*Testo e computer*, Lenci e altri, 2016), ovvero prodotti del linguaggio raccolti e sistematicamente organizzati così che soddisfino specifici criteri.

Generalmente, i corpora sono la fonte più importante di dati in linguistica computazionale, perché il computer offre la capienza utile a conservare un gran numero di dati testuali e a interrogarli in maniera intelligente. I corpora elettronici sono un caso specifico di corpus, perché quest'ultimo termine si riferisce in maniera molto più generale a una collezione di testi di grandi dimensioni; oggi, però, il termine sembra essere utilizzato per descrivere quasi esclusivamente grandi collezioni di testi digitalizzate, in un formato machine-readable (Nesselhauf, 2005).

Gli usi principali che si fanno dei corpora possono essere la progettazione di strumenti intelligenti, dotati di conoscenze linguistiche specialistiche o meno, oppure, come è il caso che si vuole indagare all'interno di questa trattazione, analizzare e descrivere fenomeni linguistici.

Considerato che i corpora sono il risultato di un'operazione di selezione e di organizzazione, essi possiedono una natura più o meno specifica e sono condizionati nelle

possibilità di utilizzo. Esistono dei parametri rilevanti per la classificazione dei corpora, e sono i seguenti:

- Generalità: un corpus può essere generale o specialistico. Nel primo caso si tratta di una collezione di testi che ha come finalità quella di proporsi come risorsa di riferimento per la descrizione e/o la rappresentazione di un linguaggio (*vedi 3.2 per itWac);* un corpus specialistico, invece, intende rappresentare e descrivere un dominio ristretto del linguaggio.
- Modalità: un corpus può riferirsi a informazioni di scritto, di parlato trascritto o miste.
- Cronologia: si intende un corpus sincronico, i cui testi appartengono a un particolare lasso di tempo nella storia, o un corpus diacronico, che descrive uno o più mutamenti linguistici nel tempo.
- Lingua: il corpus può essere descritto come monolingue o plurilingue. Nel secondo caso i corpora possono essere distinti in paralleli, con il testo rappresentato in traduzione "parallela" con una o più altre lingue, e comparabili, con testi originali appartenenti a due o più lingue relative allo stesso argomento o dominio, ma non in traduzione.
- Integrità dei testi: specifica l'appartenenza a un corpus di testi integri o meno.
- Codifica dei testi: nello specifico, i testi possono essere **codificati** o **annotati**. I testi codificati presentano delle etichette che forniscono informazioni aggiuntive del testo di tipo strutturale o composizionale, mentre i testi annotati esplicitano informazioni di tipo linguistico.

Tutti i corpora utilizzati per questa analisi sono stati annotati morfo-sintatticamente dal POS tagger descritto in *Dell'Orletta* (2009) citato in *Dell'Orletta e altri*(*2013)* e l'analisi delle dipendenze è stata effettuata con il DeSR parser usando Support Vector Machine come algoritmo di apprendimento (*Attardi*, 2006).

3.1.1 – I corpora giornalistici

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

- **La Repubblica** è un corpus che comprende 321 documenti per un totale di 232.908 token. Si tratta degli articoli dell'omonimo quotidiano pubblicati tra il

2000 e il 2005. La versione originale del corpus, invece, sviluppato dall'Università degli Studi di Bologna, includeva tutti gli articoli pubblicati tra il 1985 e il 2000.

- **Due Parole** è un corpus che comprende 322 documenti per un totale di 73.314 token. Il corpus prende il nome dall'omonimo giornale di facile lettura, nato nel 1989 dall'iniziativa di un gruppo di ricerca dell'Università di Roma "La Sapienza" e che proseguì con la fondazione, nel 1998, dell'associazione "*Parlar chiaro. Associazione per la semplificazione della comunicazione di interesse pubblico*". È un giornale rivolto a lettori con scarso livello di alfabetizzazione o difficoltà linguistiche lievi, che hanno bisogno di testi che danno informazione in modo leggibile e comprensibile. Gli articoli di Due Parole sono resi da linguisti o figure esperte, con criteri di scrittura controllata. «I criteri principali della scrittura controllata sono: la brevità dei testi, la semplicità delle frasi, la scelta di parole più comuni della lingua italiana e perciò note alla quasi totalità dei parlanti» (www.dueparole.it).

3.1.2 – I corpora scientifici

I corpora di genere scientifico utilizzati per l'analisi sono *Articoli Scientifici* e *Wikipedia: Ecologia e Ambiente*.

- **Articoli Scientifici** è un corpus che comprende 84 documenti per un totale di 471.969 token. I documenti sono tratti da riviste scientifiche specialistiche e riguardano differenti argomenti, ad esempio cambiamenti climatici e linguistica (*Dell'Orletta e altri, 2013*).
- **Wikipedia** è un corpus che comprende 293 documenti per un totale di 205.071 token. I documenti sono tratti dal portale italiano dell'omonima enciclopedia online sul tema di "Ecologia e Ambiente".

3.1.3 – I corpora dei materiali didattici (scolastici)

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

- **Materiali Didattici per la Scuola Elementare** è un corpus che comprende 127 documenti per un totale di 48.036 token. Si tratta di testi scolastici che

hanno come target primario gli studenti di scuola elementare. Generalmente sono testi per propria natura molto semplici.

- **Materiali Didattici per la Scuola Superiore** è un corpus che comprende 70 documenti per un totale di 48,103 token. Anche in questo caso si tratta di testi scolastici, ma a differenza di quelli presenti nel corpus per la Scuola Elementare, in questo caso il target principale sono studenti di scuola superiore, quindi hanno un grado di complessità lievemente superiore.

3.1.4 – I corpora narrativi

I corpora di genere narrativo utilizzati per l'analisi sono *Terence* e *Teacher*, con i documenti presi sia nella loro versione originale che semplificata.

- **Terence** è un corpus che comprende 32 documenti, corrispondenti a racconti brevi per bambini, con le rispettive versioni semplificate manualmente. Il nome del corpus deriva dal Progetto Terence, un progetto ideato e progettato nel 2007 e nato nel 2010, che ha come obiettivo quello di potenziare la comprensione del testo scritto in bambini tra i sette e i dieci anni con deficit o con particolari difficoltà di comprensione.
- **Teacher** è un corpus che comprende 24 documenti provenienti da siti web educativi dedicati a risorse gratuite per insegnati ed esperti di educazione. Il corpus comprende sia le versioni originali di questi testi che quelle semplificate, facenti parte di vari generi. La semplificazione dei testi è stata effettuata da esperti, indipendentemente l'uno dall'altro, e quindi si basa su diversi livelli linguistici, privata da gerarchie o da regole.

3.1.5 – Anatomia di un corpus

L'immagine 3.1 dà una panoramica generale ma molto rappresentativa di come sono stati annotati gli otto corpora utilizzati per gli esperimenti descritti di seguito. Come è chiaramente visibile nell'immagine che riporta un estratto dell'annotazione in formato CoNLL, le prime due colonne a partire da sinistra danno rispettivamente un Id (prima colonna) alla forma annotata, o token (seconda colonna). L'Id è un dato crescente e si azzera ogni qualvolta si conclude una frase o un periodo. La terza colonna mostra il lemma del token corrispondente (ad esempio "Le" diventa "il", "tecnologie" diventa "tecnologia"). Le tre successive colonne danno informazioni di tipo morfo-sintattico

che riguardano il token della stessa riga: la prima di queste tre colonne riporta la categoria grammaticale ("R" per "articolo", "V" per "verbo", ecc.), la seconda ne riporta una specificazione maggiore (ad esempio, se un articolo è determinativo o indeterminativo); mentre sotto la colonna successiva si danno informazioni aggiuntive come il genere o il numero. Le ultime due colonne riguardano l'annotazione sintattica a dipendenze: viene specificata la testa del token, ovvero a quale altro token quello corrispondente è immediatamente dipendente, e nell'ultima colonna è specificato il tipo di tale dipendenza (ad esempio il soggetto può essere dipendente dal verbo, o un articolo dal soggetto). Il tipo di relazione ROOT indica che si tratta della radice dell'albero sintattico.

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

Immagine 3.1: esempio di una rappresentazione di parte di un testo annotato linguisticamente (immagine presa da *Montemagni*, 2013).

3.2 – Un corpus di riferimento per la lingua italiana: itWaC

Un corpus di riferimento (o reference corpus) è una collezione di documenti generale, trasversale rispetto alle varie tipologie di generi testuali e alle differenti varietà di una lingua. È tendenzialmente un corpus plurifunzionale atto a rappresentare una lingua e a proporsi come punto di riferimento della stessa. In genere è di grandi dimensioni e può accadere che sia organizzato a sua volta in sotto-corpora, ciascuno specializzato per singola varietà di lingua o genere testuale.

ItWaC è un corpus di riferimento che contiene più di un miliardo di token. È generalmente considerato come la più grande risorsa pubblica documentata per la lingua italiana.

Insieme a ukWaC, per la lingua inglese, e deWaC, per il tedesco, itWaC è stato costruito attraverso web crawling, ovvero utilizzando il web come fonte primaria, tra il 2005 e il 2007 come parte del progetto WaCky (Web as Corpus kool ynitiative), un consorzio informale di esperti interessati all'esplorazione del web come fonte di dati linguistici (*Baroni e altri*, 2008). Tutti e tre i corpora contengono annotazione linguistica di base, si tratta di part-of-speech tagging e lemmatizzazione. In particolare, itWaC è stato taggato per part-of-speech con il TreeTagger, mentre la lemmatizzazione è stata effettuata utilizzando il lessico *Morph-it!*.

ItWaC, ukWac e deWaC non sono le prime risorse ad essere state costruite attraverso il metodo del web crawling, però sono uniche in quanto offrono un compromesso tra una grande ampiezza e la creazione di annotazioni utili a scopi linguistici (*Baroni e altri,* 2018).

3.3 – Il Nuovo Vocabolario di Base della Lingua Italiana

Il Nuovo Vocabolario di Base (NVdB) per la Lingua Italiana è una risorsa linguistica che descrive le parole più usate e conosciute dalla maggioranza della popolazione italiana che abbia un'istruzione media inferiore (*Chiari*, *De Mauro*, 2014).

La prima versione del Vocabolario di Base della Lingua Italiana apparve per la prima volta all'interno di *Guida all'uso delle parole* (De Mauro, 1980) e ha ricevuto modifiche di piccole dimensioni nel corso del tempo. L'ultima versione risale al 2007 ed era strutturata in tre categorie: Vocabolario Fondamentale (FO), Vocabolario di Alto Uso (AU) e Vocabolario di Alta Disponibilità (AD). La prima categoria comprende i lemmi

considerati più frequenti in assoluto nella lingua italiana (appartamento, commercio, fiore, ecc.); la seconda categoria si riferisce a lemmi meno frequenti rispetto a quelli appartenenti al Vocabolario Fondamentale, ma comunque abbastanza frequenti da risultare significativi (acciaio, concerto, fase, ecc.); l'ultima categoria, invece, riguarda quei lemmi che, pur non avendo una grande frequenza nei testi scritti, sono noti alla maggior parte della popolazione italiana perché hanno una specifica relazione con la concretezza della vita ordinaria (abbaiare, ago, forchetta, ecc.).

A partire da questa edizione del Vocabolario di Base fino al Nuovo Vocabolario di Base, la frequenza d'uso dell'italiano è cambiata radicalmente. Se prima questa lingua era usata solo dal 50% della popolazione italiana, oggi la stima è salita fino al 95% (*Chiari*, *De Mauro*, 2014).

In riferimento a chi lo ha creato, il NVdB della Lingua Italiana è una risorsa linguistica atta a perseguire tre finalità differenti e, allo stesso tempo, parallele: uno scopo linguistico, in senso teorico e descrittivo; uno scopo educativo, per lo sviluppo di applicazioni utili all'insegnamento e/o all'apprendimento; uno scopo regolativo, utile ad esempio per lo sviluppo di particolari linee guida per la scrittura professionale.

Il NVdB contiene informazioni statistiche e non statistiche degli elementi del lessico. Le categorie del Vocabolario Fondamentale e del Vocabolario di Alto Uso sono state costruite sulla base dell'analisi di un particolare corpus sull'italiano contemporaneo, organizzato in sotto-corpora di grandezza simile ma appartenenti a generi e mezzi di fruizione diversi (ad esempio giornali, letteratura o intrattenimento).

L'organizzazione interna del NVdB è gerarchica, ogni lemma è in genere associato a una o più istanze grammaticali, in modo da evitare possibili e plausibili ambiguità. In definitiva, si compone di circa 7400 lessemi: circa 2000 fanno parte del Vocabolario Fondamentale; circa 3000 fanno parte del Vocabolario di Alto Uso; circa 2400 appartengono al Vocabolario di Alta Disponibilità. Molti dei termini che sono entrati a far parte del Vocabolario Fondamentale, in questa edizione, prima appartenevano al Vocabolario di Alto Uso (*Chiari*, *De Mauro*, 2014).

3.4 – Metodi d'indagine

In questa sezione saranno oggetto d'indagine gli algoritmi, e quindi gli script utilizzati per raccogliere i dati analizzati nel capitolo quarto e quinto. Si tratta di due esperimenti,

entrambi con oggetto il lessico. Il primo (*vedi 3.4.3*) riguarda lo studio del lessico che compone i testi e le variazioni interne al genere e di genere. In questo caso gli script da analizzare sono quattro: con il primo sono stati estratti i dati che riguardano le frequenze, nei corpora oggetto di analisi (*vedi 3.1*), dei lemmi appartenenti alle tre categorie del NVdB della Lingua Italiana (*vedi 3.3*); con il secondo script sono stati estratti i dati riguardanti le classi di frequenza medie, nei corpora in esame, delle forme e dei lemmi presenti in itWaC (*vedi 3.2*); con il terzo script sono stati estratti i dati riguardanti le classi di frequenza medie, nei corpora analizzati, delle forme e dei lemmi presenti in itWaC tenendo conto delle rispettive part of speech; con l'ultimo script sono stati estratti i dati riguardanti le classi di frequenza medie delle forme e dei lemmi presenti in itWaC divisi per part of speech.

Il secondo esperimento (*vedi 3.4.4*) riguarda lo studio di alcuni fenomeni sintattici che co-occorrono con sostantivi semplici e difficili. In questo caso lo script utilizzato è servito per navigare l'albero sintattico per ogni corpus.

In entrambi gli esperimenti, per dare confronti tra i risultati e ottenere determinate evidenze, è stato utilizzato un algoritmo di correlazione, il Wilcoxon rank-sum test (vedi 3.4.1), fornito dall'Istituto di Linguistica Computazionale A. Zampolli del CNR di Pisa, che è stato utile per il calcolo del p-value, necessario a determinare la significatività dei vari confronti (*vedi 3.4.2*).

3.4.1 – Wilcoxon rank-sum test

Il Wilcoxon rank-sum test (anche noto come test di Mann-Whitney-Wilcoxon o Mann-Whitney U test) è un test statistico non parametrico per due campioni indipendenti alternativo al t-test (o Wilcoxon signed-rank test) per campioni dipendenti (*Chris Wild*, 1997). Verifica la presenza di valori che provengono da una distribuzione continua e, quindi, se i due campioni provengono da una stessa popolazione.

Il test richiede il calcolo della statistica comunemente chiamata U, la quale ha una distribuzione nota sotto l'ipotesi nulla[3]. Una statistica equivalente è quella della somma dei ranghi.

[3]In genere l'ipotesi nulla è indicata con H_0. In seguito si decide se accettarla o meno, e in questo secondo caso l'ipotesi alternativa è generalmente indicata con H_1.

Tutte le osservazioni vanno ordinate in una serie di rango, indifferentemente da quale campione provengano. I ranghi provenienti dal primo campione vengono sommati. La somma di tutti i ranghi vale:

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

"N" è il numero delle osservazioni in totale.
La statistica U, invece, è data dalla seguente formula:

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

dove n_1 è la dimensione del campione e R_1 è la somma dei ranghi, in entrambi i casi si tratta sempre del primo campione.
Vale lo stesso per il secondo campione:

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

Sommando U_1 e U_2, abbiamo:

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

Quindi:

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

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

Sappiamo inoltre che $N = n_1 + n_2$, quindi:

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

Semplificando:

$$U_1 + U_2 = n_1n_2$$

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

3.4.2 – P-value e significatività

Il p-value di un test statistico indica la probabilità di ottenere un risultato uguale o più estremo rispetto a quello osservato. Spesso ci si riferisce a questo valore come al "livello di significatività osservato".

Supponiamo di avere due campioni A e B contenenti n_a e n_b osservazioni rispettivamente. Vogliamo capire se la distribuzione di X valori sia la stessa in A e in B. Per i test unilaterali l'ipotesi H_0 vale che $A = B$ (ipotesi nulla). Le altre due possibilità sono che $H_1: A > B$ (A è spostato alla destra di B) o $H_1: A < B$ (A è spostato alla sinistra di B). Contro l'ipotesi nulla, se $A > B$ è vera, allora il $p - value = pr(X \geq R_1)$ dove X è una variabile aleatoria ed R_1 la somma dei ranghi del primo campione. Se, invece, $A < B$ è vera, allora il $p - value = pr(X \leq R_1)$.

Per i test bilaterali, l'ipotesi nulla vale $H_0: A = B$ e la sua alternativa è $H_1: A \neq B$. In quest'ultimo caso viene calcolata la probabilità che i valori finiscano spostati più verso destra o più verso sinistra e raddoppiato il valore precedentemente calcolato per i test unilaterali. Quindi, $p - value = 2pr(X \geq R_1)$ o $p - value = 2pr(X \leq R_1)$.

Se il p-value è molto piccolo, allora vuol dire che la significatività è alta. Generalmente, se il p-value calcolato è pari o superiore a 0.05 non c'è significatività, se è inferiore a 0.05 c'è significatività statistica. Se il valore scende al di sotto di 0.01 si tratta di una variazione **molto significativa**, mentre è inferiore a 0.001, allora si tratta di un caso **estremamente significativo**.

3.4.3 – Il primo esperimento

Questo primo esperimento è servito a uno studio del lessico che compone testi di generi e complessità differenti. Entrando nel particolare, è stata indagata la variazione di frequenza del Vocabolario Fondamentale, di Alto Uso e di Alta Disponibilità (*vedi 3.3*) all'interno degli otto corpora esaminati (*vedi 3.1*). In seguito, il centro dell'indagine si è spostato dal NVdB al corpus di riferimento itWaC (*vedi 3.2*): sono state calcolate le classi di frequenza medie dei termini presenti in itWaC per ogni documento di ognuno dei corpora in esame. Inizialmente lo si è fatto senza tenere conto delle part of speech, poi tenendone conto e, infine, si è data un'analisi per singola part of speech.

3.4.3.1 – Frequenze e Nuovo Vocabolario di Base

Il primo passo è stato quello di costruire la funzione principale che prendesse in input un corpus e il NVdB. Sono state create le liste per ognuna delle tre categorie (Vocabolario Fondamentale, di Alto Uso e di Alta Disponibilità) e del totale delle stesse e sono state inizializzate a zero le rispettive frequenze.

```
def main(file1, file2):
        diz = codecs.open(file1, "r", "utf-8")
        corpus = codecs.open(file2, "r", "utf-8")
        dizionario = []
        dizionarioFond = []
        dizionarioAltUso = []
        dizionarioAltDisp = []
        dizionarioTot = []
        DizionarioDB=caricaDizionario(diz)
        totDoc=0
        currentDoc = []
        frequenzaF = 0
        frequenzaAU = 0
        frequenzaAD = 0
        frequenzaTot = 0
        totDoc = 0
        n = 1
```

È stato caricato il NVdB all'interno del programma come struttura dati (dizionario).

```
def caricaDizionario(diz):
        Diz={}
        for line in diz:
                ls=line.strip().split(" ")
                if not(ls[0] in Diz):
                        Diz[ls[0]]=ls[1]
        return Diz
```

Per ogni documento del corpus è stata calcolata la frequenza di ognuna delle tre categorie presenti all'interno del NVdB e di tutte e tre insieme. Finché il documento non è concluso, il programma continua a contare:

```
ls = line.split("\t")
totDoc+=1
forma=ls[1]
lemma=ls[2]
if forma in DizionarioDB:
        if DizionarioDB[forma]=="LF":
                frequenzaF=frequenzaF+1
        elif DizionarioDB[forma]=="AU":
                frequenzaAU=frequenzaAU+1
        elif DizionarioDB[forma]=="AD":
                frequenzaAD=frequenzaAD+1
        frequenzaTot=frequenzaTot+1
elif lemma in DizionarioDB:
        if DizionarioDB[lemma]=="LF":
                frequenzaF=frequenzaF+1
        elif DizionarioDB[lemma]=="AU":
                frequenzaAU=frequenzaAU+1
        elif DizionarioDB[lemma]=="AD":
                frequenzaAD=frequenzaAD+1
        frequenzaTot=frequenzaTot+1
```

Quando il documento è concluso, le variabili delle frequenze vengono azzerate e il procedimento riprende.

3.4.3.2 – Classi di frequenza medie con itWaC

Lo script presenta tre funzioni. Quella principale carica in input il corpus e un file contenente le entrate di itWaC e la loro frequenza. Restituisce come output le classi di frequenza medie per documento delle forme e dei lemmi di itWaC all'interno del corpus considerato.

La classe di frequenza per ogni lemma è stata calcolata utilizzando la seguente funzione:

$$\left\lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \right\rfloor$$

Dove MFL è il lemma più frequente all'interno del corpus (itWaC) e CL è la frequenza del lemma o della forma considerata all'interno di uno degli otto corpora in esame.
La seguente funzione genera due dizionari contenenti le forme e i lemmi di itWaC:

```
def caricaItwac(itwac):
        ItwacForme={}
        ItwacLemmi={}
        var = 0
        for line in itwac:
                if line == "\n":
                        continue
                if line[0:2]=="FO":
                        var=1
                        continue
                if line[0:2]=="LE":
                        var=2
                        continue
                if var==1:
                        ls=line.strip().split("\t")
                        ItwacForme[ls[0]]=ls[1]
                        continue
                if var==2:
                        ls=line.strip().split("\t")
                        ItwacLemmi[ls[0]]=ls[1]
                        continue
        return ItwacForme, ItwacLemmi
```

La seguente funzione calcola la forma e il lemma più frequenti in itWaC:

```
def FormaLemmaPiuFrequente(ItwacForme, ItwacLemmi):
        maxLemma=0
        maxForma=0
        for forma in ItwacForme:
                if forma <> "(" and forma <> ")" and
forma <> "'" and forma <> "'" and forma <> "_" and
forma <> "-" and forma <> "," and forma <> ";" and
forma <> "." and forma <> "?" and forma <> "!":
                        if int(ItwacForme[forma])>maxForma:
                                maxForma=int(ItwacForme[forma])
        for lemma in ItwacLemmi:
                if lemma <> "(" and lemma <> ")" and
lemma <> "'" and lemma <> "'" and lemma <> "_" and
lemma <> "-" and lemma <> "," and lemma <> ";" and
lemma <> "." and lemma <> "?" and lemma <> "!":
                        if int(ItwacLemmi[lemma])>maxLemma:
                                maxLemma=int(ItwacLemmi[lemma])
        return maxLemma, maxForma
```

3.4.3.3 – Classi di frequenza medie considerando le part of speech con it-WaC

Lo script è sostanzialmente identico a quello descritto in 3.4.3.2, con la differenza che in questo caso tra le forme uguali e i lemmi uguali viene distinta l'appartenenza a una specifica part of speech.

3.4.3.4 – Classi di frequenza media per part of speech con itWaC

In questo caso lo script calcola le classi di frequenza medie per documento per ognuna delle part of speech considerate. Quindi, inizialmente nella funzione principale vengono inizializzate tante variabili quante sono le part of speech:

```
def main(file1, file2):
        itwac=codecs.open(file1, "r", "utf-8")
        corpus=codecs.open(file2, "r", "utf-8")
        ItwacForme, ItwacLemmi = caricaItwac(itwac)
        maxLemma, maxForma = FormaLemmaPiuFrequente(ItwacForme, ItwacLemmi)
        totDocForme=0
        totDocLemmi=0
        SommaClasseDiFreqAForme=0
        SommaClasseDiFreqBForme=0
        SommaClasseDiFreqCForme=0
        SommaClasseDiFreqDForme=0
        SommaClasseDiFreqEForme=0
        SommaClasseDiFreqIForme=0
        SommaClasseDiFreqNForme=0
        SommaClasseDiFreqPForme=0
        SommaClasseDiFreqRForme=0
        SommaClasseDiFreqSForme=0
        SommaClasseDiFreqTForme=0
        SommaClasseDiFreqVForme=0
        SommaClasseDiFreqXForme=0
        SommaClasseDiFreqALemmi=0
        SommaClasseDiFreqBLemmi=0
        SommaClasseDiFreqCLemmi=0
        SommaClasseDiFreqDLemmi=0
        SommaClasseDiFreqELemmi=0
        SommaClasseDiFreqILemmi=0
        SommaClasseDiFreqNLemmi=0
        SommaClasseDiFreqPLemmi=0
        SommaClasseDiFreqRLemmi=0
        SommaClasseDiFreqSLemmi=0
        SommaClasseDiFreqTLemmi=0
        SommaClasseDiFreqVLemmi=0
        SommaClasseDiFreqXLemmi=0
```

In riferimento al tagset morfo-sintattico ISST-TANL: “A” si riferisce agli aggettivi, “B” agli avverbi, “C” alle congiunzioni, “D” ai determinanti, “E” alle preposizioni, “F” ai segni di punteggiatura, “I” alle interiezioni, “N” ai numeri, “P” ai pronomi, “R” agli articoli, “S” ai nomi, “T” ai predeterminanti, “V” ai verbi e “X” ai residui, ad esempio formule, parole non classificate, simboli alfabetici e via di questo passo.
Il resto dello script, pur tenendo conto della differenza specificata, è concettualmente molto simile ai precedenti già analizzati.

3.4.3.5 – Confronti e calcolo del p-value

Per determinare la significatività delle variazioni esplicitate dai dati che saranno analizzati nel quarto capitolo, attraverso uno script del Wilcoxon rank-sum test messo a disposizione dall’Istituto di Linguistica Computazionale A. Zampolli del CNR di Pisa, sono stati fatti i seguenti confronti per il calcolo del p-value:

- Corpora complesso contro corpora semplice dello stesso genere;

- Corpora difficili di un particolare genere contro corpora difficili di un altro genere;
- Corpora facili di un particolare genere contro corpora facili di un altro genere.

3.4.4 – Il secondo esperimento

Questo esperimento ha richiesto un'analisi che comparasse il lessico nominale e determinati fenomeni linguistici, in particolare sintattici: si tratta della distanza delle occorrenze di un sostantivo dalla sua testa in termini di numero di parole, il numero dei suoi fratelli e dei suoi dipendenti e la sua distanza dalla radice. Il motivo della scelta è dettato dal fatto che questi fenomeni sono rilevanti dal punto di vista sintattico ma anche da quello lessicale. Le quattro coppie di corpora appartengono ad altrettanti generi, ogni coppia presenta un corpus rappresentativo di una varietà complessa e uno semplice per quel genere (*vedi 3.1*). Per ognuno di questi corpora sono stati selezionati i dieci sostantivi più semplici e alcuni dei sostantivi più difficili. Il criterio per la selezione è stato il seguente: maggiore è risultata la frequenza di un sostantivo all'interno di itWaC, maggiore è stata considerata la semplicità e la diffusione di quella particolare parola. Per quanto riguarda i sostantivi difficili si è scelto di estrarne l'ultimo quarto dalla lista di tutti i sostantivi, per un dato corpora, ordinati per complessità. Questa scelta è stata motivata dal fatto che se i dieci sostantivi più semplici presentano un numero di occorrenze sufficiente all'analisi, non avviene lo stesso per i dieci difficili; infatti questi ultimi sono in genere i termini meno frequenti.

In seguito sono state estratte alcune informazioni sintattiche per ogni singola occorrenza di ogni sostantivo in ognuno di questi gruppi all'interno del corpus di appartenenza. In particolare, per ogni sostantivo è stato ricavato il tipo di dipendenza all'interno della frase, la distanza dalla sua testa in termini di numero di parole, il numero dei suoi dipendenti, il numero dei suoi fratelli e la distanza dalla radice. In più, per ognuno dei sostantivi sono state calcolate le distribuzioni di specifiche dipendenze (ad esempio, quante volte un dato sostantivo è considerato soggetto o oggetto di una frase). Per costruire l'albero sintattico sono stati utilizzati degli script in python forniti dall'Istituto di Linguistica Computazionale A. Zampolli del CNR di Pisa. Diversamente è avvenuto per quanto riguarda la navigazione del suddetto. Lo script che è stato scritto per l'operazione prende in input le parole facili o difficili in esame e il loro corpus di appartenenza.

La seguente funzione trasforma il file contenente i sostantivi (in questo caso facili) in una struttura dati a dizionario utile per navigarli:

```
def caricaParole(ListaParoleBuone):
        listaFacili={}
        var=0
        for line in ListaParoleBuone:
                ls=line.strip().split("\t")
                listaFacili[ls[0]]=" "
                continue
        return listaFacili
```

Per estrarre il numero di figli per ogni occorrenza dei dieci sostantivi nel corpus è stato banalmente considerato il termine come testa ed estratte le parole a lui dipendenti (ad esempio un soggetto può avere come dipendenti un articolo o un aggettivo).
I fratelli di un sostantivo sono tutti i termini che hanno il suo stesso livello di dipendenza, dipendenti da uno stesso "padre"; quindi il dato riguardante i fratelli del sostantivo esaminato ha richiesto un'elaborazione solo lievemente più complessa. Data la parola corrente, sono stati contati tutti i dipendenti del suo elemento padre; dal dato risultante è stata sottratta un'unità.

```
numeroFratelli=len(tokens[tok.head-1].children)-1
```

La distanza dalla testa dell'occorrenza del sostantivo in termini di numero di parole è estratta grazie alla sottrazione tra l'id del termine e l'id della sua testa:

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

La distanza dalla radice è, invece, stata calcolata grazie alla seguente funzione ricorsiva:

```
def calcolaDistRadice(identificatore, testa, tokens):
        if testa==0:
                return 0
        return 1 + calcolaDistRadice(testa, tokens[int(testa)-1].head, tokens)
```

La funzione prende come parametri l'Id del token, l'Id della sua testa e la funzione che costruisce l'albero sintattico a dipendenze. Finché la testa non è uguale a zero, e quindi finché non viene trovata la radice (perché l'Id della testa del token radice è considerato

0), la funzione somma uno e chiama se stessa con i seguenti parametri: l'Id della sua testa come Id del token, la testa del token che prima era la testa del token considerato e la funzione che costruisce l'albero sintattico. La funzione continua a girare in questo modo finché non riconosce la radice. Come risultato genera la distanza dal token considerato alla radice.

3.4.4.1 – Confronti e calcolo del p-value

Anche in questo caso sono stati fatti alcuni confronti per il calcolo del p-value. Atti a determinare la significatività delle variazioni esplicitate dai dati che saranno analizzati nel quinto capitolo:

- I sostantivi più difficili contro i sostantivi più facili dello stesso corpus;
- I sostantivi più facili contro i sostantivi più facili di due corpora dello stesso genere (corpus semplice e corpus complesso);
- I sostantivi più difficili contro i sostantivi più difficili di due corpora dello stesso genere (corpus semplice e corpus complesso);
- I sostantivi più difficili di un corpus complesso contro i sostantivi più difficili di un altro corpus complesso;
- I sostantivi più facili di un corpus complesso contro i sostantivi più facili di un altro corpus complesso;
- I sostantivi più difficili di un corpus semplice contro i sostantivi più difficili di un altro corpus semplice;
- I sostantivi più facili di un corpus semplice contro i sostantivi più facili di un altro corpus semplice;
- I sostantivi facili contro i sostantivi complessi considerando tutti i quattro corpora semplici come un unico corpus;
- I sostantivi facili contro i sostantivi complessi considerando tutti i quattro corpora difficili come un unico corpus;
- I sostantivi facili contro i sostantivi complessi considerando tutti e otto i corpora come un unico corpus;
- Tutti i sostantivi di un corpus semplice contro tutti i sostantivi di un corpus complesso;
- Tutti i sostantivi di un corpus semplice contro tutti i sostantivi di un altro corpus semplice;

- Tutti i sostantivi di un corpus complesso contro tutti i sostantivi di un altro corpus complessi.

CAPITOLO 4

Analisi dei dati: studio sul lessico

In questo capitolo e nel seguente saranno analizzati i dati statistici estratti attraverso i due esperimenti descritti nel capitolo 3. Questo lavoro si inserisce negli studi di monitoraggio linguistico fatti da Dell'Orletta e altri (2013) e Montemagni (2013).
Di seguito saranno oggetto d'indagine i dati relativi al lessico. In particolare, fine ultimo di questa esposizione sarà lo studio delle variazioni lessicali nel confronto tra generi e complessità. Strumento dell'indagine sono stati gli otto corpora descritti nel terzo capitolo di questo elaborato. Qui di seguito verranno esposti, analizzati e messi a confronto i dati relativi alle frequenze delle uscite e dei lemmi appartenenti alle tre categorie del Nuovo Vocabolario di Base di Tullio De Mauro all'interno degli otto corpora oggetto d'analisi (4.1), successivamente quelli relativi alle variazioni di frequenza, negli otto corpora, dei termini appartenenti al corpus di riferimento itWaC (4.2).
Per concludere sarà data una panoramica dei dati in funzione di conclusione del capitolo (4.3).

4.1 – Lessico e Nuovo Vocabolario di Base

In questa sezione saranno analizzati i dati ed esposte le evidenze relative alle frequenze delle forme e dei lemmi appartenenti alle categorie del vocabolario fondamentale, di alto uso e di alta disponibilità del Nuovo Vocabolario di Base (NVdB) della Lingua Italiana di Tullio De Mauro all'interno degli otto corpora analizzati.
L'obiettivo è quello di capire se ci sono variazioni e se queste variazioni possano essere considerate significative per lo studio di corpora di diverso genere e complessità.
Nel grafico 4.1 è presente una vista d'insieme della percentuale media delle frequenze di ognuna delle tre categorie del NVdB all'interno di ciascuno dei corpora analizzati. Nel grafico, l'asse delle ordinate rappresenta la percentuale e l'asse delle ascisse le tre categorie del NVdB e una panoramica di tutte e tre le categorie insieme (totale). Ogni colore utilizzato all'interno del grafico si riferisce a uno degli otto corpora.

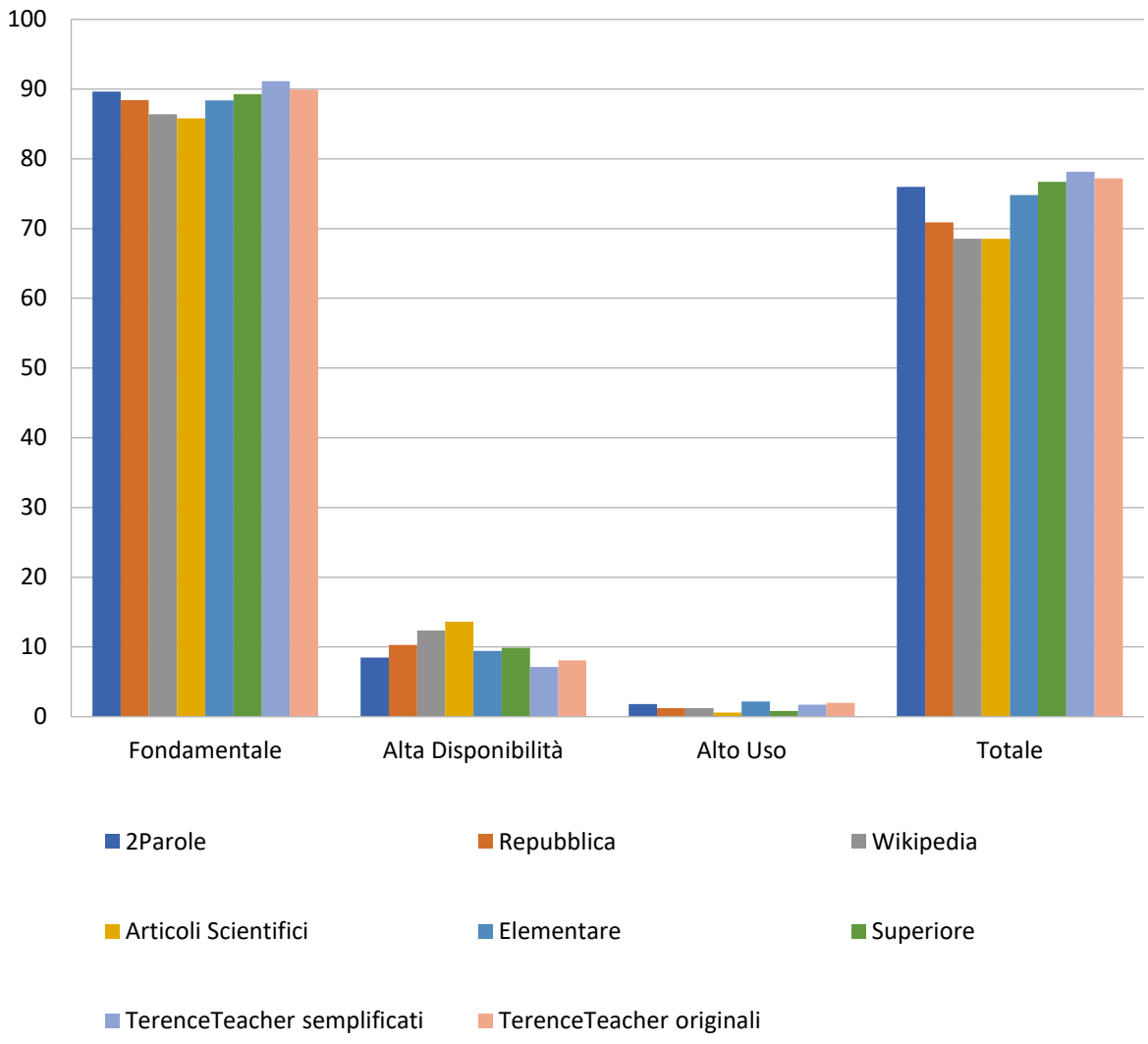

Grafico 4.1: percentuale della media delle frequenze del vocabolario fondamentale, di alto uso e di alta disponibilità negli otto corpora.

Il primo dato che salta all'occhio è il grande divario percentuale che c'è tra il vocabolario fondamentale e le altre due categorie del NVdB. Se ne deduce che generalmente in tutti e otto i corpora esiste una grossa preponderanza di termini facenti parte del vocabolario fondamentale, che sono considerati i lemmi più frequenti in assoluto nella lingua italiana e indipendentemente dal genere; quindi siamo di fronte a un'evidenza che ci saremmo aspettati. Inoltre, tenendo conto di tutte e tre le categorie del NVdB notiamo come esse in media rappresentino circa il 75% di quasi tutti i corpora. Le eccezioni più evidenti sono tre: in R*epubblica* siamo poco sopra il 70%, mentre per quanto riguarda *Wikipedia* e il corpus di *Articoli Scientifici* addirittura scendiamo al di sotto e ci aggiriamo intorno al 68%. Entrambi questi ultimi due corpora appartengono

allo stesso genere e sono rappresentanti di un livello di complessità alto (Articoli Scientifici) e basso (Wikipedia).
Una chiarezza sui dati veri e propri, in una forma meno gradevole all'occhio ma più precisa, ce la dà la tabella 4.1.

	Lessico fondamentale	Alta disponibilità	Alto uso	Totale
2Parole	0,896	0,085	0,018	0,761
Repubblica	0,884	0,103	0,012	0,709
Wikipedia	0,863	0,123	0,012	0,685
Articoli scientifici	0,858	0,136	0,005	0,685
Scuola elementare	0,883	0,094	0,021	0,748
Scuola superiore	0,892	0,099	0,008	0,767
TerenceTeacher semplificato	0,911	0,071	0,017	0,781
TerenceTeacher originale	0,899	0,080	0,02	0,772

Tabella 4.1: frequenza relativa media del vocabolario fondamentale, di alto uso e di alta disponibilità negli otto corpora.

La tabella 4.1 mostra le frequenze relative medie dei lemmi del NVdB negli otto corpora. Le righe mostrano le variazioni tra le categorie all'interno del singolo corpus, mentre le colonne mostrano le variazioni di una categoria all'interno dei corpora.
Per chiarire se le variazioni di frequenza tra i vari corpora (e quindi tra i generi e i due livelli di complessità) sono significative, e quindi la loro differenza non sia casuale, è stato usato un algoritmo di correlazione, quello del Wilcoxon rank-sum test (*vedi capitolo 3*). Attraverso di esso è stato calcolato il p-value per determinati confronti.
Nella tabella 4.2 possiamo notare come, nel confronto tra i due corpora giornalistici, le variazioni di complessità sono tutte estremamente significative, anche se con uno sguardo alla tabella 4.1 e al grafico 4.1 è chiaro che non sussistano grosse differenze per quanto riguarda il vocabolario fondamentale, di alto uso e di alta disponibilità. Al contrario, invece, se consideriamo tutte e tre le categorie insieme; infatti, come abbiamo visto, i lemmi del NVdB rappresentano circa il 70% di *Repubblica*, ben oltre il 75% di *2Parole*.

	Lessico Fondamentale	Alta disponibilità	Alto uso	Totale
2Parole vs Repubblica	☑	☑	☑	☑
Wikipedia vs Articoli Scientifici	☒	✓	✓	☒
Scuola elementare vs Scuola superiore	☒	☒	☑	☑
Terence-Teacher originale vs semplificato	✓	✓	☒	☒

Tabella 4.2: calcolo del p-value nel confronto tra corpora di stesso genere e complessità differente[4].

Notiamo invece variazioni molto significative nel confronto tra i due corpora scientifici riguardo al vocabolario di Alta Disponibilità e di Alto uso e nessuna significatività per il vocabolario fondamentale e tutte e tre le categorie prese insieme. Le categorie di lessico Fondamentale e di Alta Disponibilità nei due corpora scolastici non presentano alcuna significatività nelle loro variazioni; al contrario, estremamente significativa è la variazione del vocabolario di Alto Uso. Nei due corpora narrativi, invece, possiamo notare molta significatività in alcune categorie, ma nella vista totale non vi è nulla di importante.

Se osserviamo le tabelle 4.3 e 4.4, possiamo notare come passando al confronto dei risultati dati dai corpora di genere diverso facenti parte dello stesso livello di complessità, le variazioni tra i campioni considerati non sono quasi mai casuali, a parte qualche lieve eccezione.

[4]Da qui fino alla fine dell'elaborato, per le tabelle che mostrano la significatività dei confronti vale che: ☑ estremamente significativo; ✓ molto significativo; × poco significativo; ☒ nessuna significatività.

Da sottolineare, per quanto riguarda i corpora complessi, il fatto che il confronto tra *Terence e Teacher originali* e i testi di *Scuola Superiore* metta alla luce una non significatività tra le variazioni di frequenza delle tre categorie del NVdB in totale nei due corpora.

Un simile scenario è riscontrabile nel confronto tra *2Parole* e i testi di *Scuola Elementare* o tra *2Parole* e *Terence e Teacher* semplificati, se passiamo alla tabella che si riferisce ai corpora semplici. Questa evidenza è verosimile in quanto i corpora citati presentano tutti testi molto semplici (anche rispetto agli altri corpora in questa analisi considerati facili).

	Lessico fondamentale	Alta disponibilità	Alto uso	Totale
Articoli scientifici vs Scuola superiore	☑	☑	✗	☑
Repubblica vs Articoli scientifici	☑	☑	☑	☑
Repubblica vs Scuola superiore	✓	☒	☑	☑
Repubblica vs TerenceTeacher originale	☑	☑	☑	☑
TerenceTeacher originale vs Articoli scientifici	☑	☑	☑	☑
TerenceTeacher originale vs Scuola superiore	☒	☑	☑	☒

Tabella 4.3: calcolo del p-value nel confronto tra corpora complessi e di diverso genere.

	Lessico fondamentale	Alta disponibilità	Alto uso	Totale
2Parole vs	☑	✓	☑	☒

Scuola elementare				
2Parole vs TerenceTeacher semplificati	☒	✗	☑	☒
2Parole vs Wikipedia	☑	☑	☒	☑
TerenceTeacher semplificati vs scuola elementare	☑	☑	☒	☑
TerenceTeacher semplificati vs Wikipedia	☑	☑	☑	☑
Wikipedia vs Scuola elementare	☑	☑	☑	☑

Tabella 4.4: calcolo del p-value nel confronto tra corpora semplici e di diverso genere.

4.2 – Lessico, classi di frequenza e itWaC

In questa sezione saranno oggetto di analisi i dati relativi alle classi di frequenza delle forme e dei lemmi di itWaC all'interno degli otto corpora descritti nel terzo capitolo. Anche in questo caso il fine ultimo è quello di studiare similarità e differenze dovute al genere e alla complessità. L'analisi sarà divisa in due parti. Nella prima (4.2.1) verranno descritti i dati riguardanti le classi di frequenza medie per le forme e per i lemmi relativi a tutti e otto i corpora; nella seconda (4.2.2) a essere oggetto d'analisi saranno le classi di frequenza medie per le forme e per i lemmi per ognuna delle part of speech.

4.2.1 – Prima parte

Nel grafico 4.2 vediamo una panoramica media delle classi di frequenza delle forme e dei lemmi di itWaC all'interno degli otto corpora. Sull'asse delle ascisse abbiamo le quattro varianti possibili: le classi di frequenza (C. d. F.) delle forme, dei lemmi, delle forme tenendo conto delle diverse part of speech e dei lemmi tenendo conto delle diverse part of speech.

La prima dinamica generale che salta all'occhio è che, se per i lemmi il dato è molto simile, per quanto riguarda le forme abbiamo un'evidente diminuzione numerica se non escludiamo le part of speech. Inoltre, le differenze relative al livello di complessità sono minime: la media delle classi di frequenza delle coppie di corpora facili-complessi è pressoché simile, al contrario le differenze nei vari generi si notano chiaramente, ad esempio i testi scientifici e i testi narrativi di *Terence* e *Teacher* sono parecchio distanti: quasi 7,5 i primi, sotto il 7 i secondi. Non in misura così importante, ma lo stesso vale anche per i corpora giornalistici e scolastici.

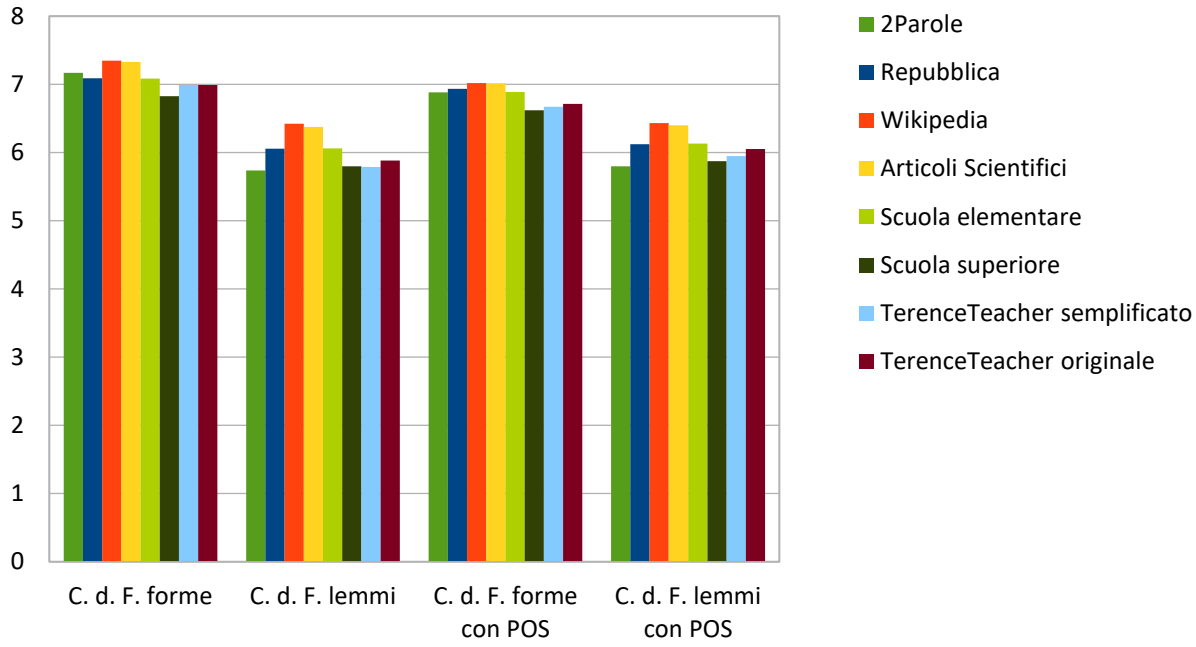

Grafico 4.2: media delle classi di frequenza delle forme e dei lemmi di itWaC all'interno degli otto corpora, senza tenere conto delle part of speech (POS) e tenendone conto.

Anche in questo caso sono stati fatti alcuni confronti tra generi e complessità per determinare se le variazioni tra i campioni esaminati sono o meno significative.
La tabella 4.5 mostra i confronti dei corpora dello stesso genere nel caso in cui non si tenga conto delle part of speech, ed evidenzia come, a livello di complessità, non ci sia significatività se facciamo eccezione per il confronto tra la media delle classi di frequenza delle forme e dei lemmi dei corpus di testi di *Scuola Elementare* e del corpus

di testi di *Scuola Superiore.* Se ne deduce che, avendo come riferimento itWaC, il lessico varia significativamente al variare della complessità in testi di tipo scolastico, e si riscontra una variazione dei lemmi anche nei due corpora di genere giornalistico.

	Media classi di frequenza forme	Media classi di frequenza lemmi
2Parole vs Repubblica	✗	✓
Wikipedia vs Articoli Scientifici	✗	✗
Scuola elementare vs Scuola superiore	✓	✓
TerenceTeacher originale vs semplificato	✗	✗

Tabella 4.5: calcolo del p-value nel confronto tra corpora dello stesso genere e differente complessità senza tenere conto delle part of speech.

Nelle tabelle 4.6 e 4.7, invece, notiamo che, se teniamo conto delle variazioni tra genere all'interno della stessa complessità, la situazione si allarga. Per i corpora complessi abbiamo mediamente un livello di significatività estremamente elevato, che cala leggermente nel caso dei confronti tra *Repubblica* e i testi di *Terence* e *Teacher* originali, e che scompare del tutto se questi ultimi sono messi a confronto con i testi di *scuola superiore*.

	Media classi di frequenza forme	Media classi di frequenza lemmi
Articoli scientifici vs Scuola superiore	✓	✓
Repubblica vs Articoli scientifici	✓	✓

Repubblica vs Scuola superiore	☑	☑
Repubblica vs TerenceTeacher originale	✗	✓
TerenceTeacher originale vs Articoli scientifici	☑	☑
TerenceTeacher originale vs Scuola superiore	☒	☒

Tabella 4.6: calcolo del p-value nel confronto tra corpora complessi di diverso genere senza tenere conto delle part of speech.

Per i corpora semplici abbiamo una situazione più varia. Più della metà dei confronti danno come risultato un valore estremamente significativo: variano significativamente, infatti, le classi di frequenza estratte da *Wikipedia* nel confronto con *2Parole*, i testi di *Terence* e *Teacher* semplificati e quelli di *scuola elementare*.

	Media classi di frequenza forme	Media classi di frequenza lemmi
2Parole vs Scuola elementare	☒	☑
2Parole vs TerenceTeacher semplificati	✓	☒
2Parole vs Wikipedia	☑	☑
TerenceTeacher semplificati vs scuola elementare	☒	✓

TerenceTeacher semplificati vs Wikipedia	☑	☑
Wikipedia vs Scuola elementare	☑	☑

Tabella 4.7: calcolo del p-value nel confronto tra corpora semplici di diverso genere senza tenere conto delle part of speech.

Se passiamo ora al caso in cui si tengano conto del part of speech, notiamo come le evidenze siano tutte molto simili a quelle analizzate fino ad ora.

	Media classi di frequenza forme	Media classi di frequenza lemmi
2Parole vs Repubblica	☒	☑
Wikipedia vs Articoli Scientifici	☒	☒
Scuola elementare vs Scuola superiore	☑	☑
TerenceTeacher originale vs semplificato	☒	☒

Tabella 4.8: calcolo del p-value nel confronto tra corpora dello stesso genere e differente complessità tenendo conto delle part of speech.

La tabella 4.8 relativa al confronto tra corpora dello stesso genere e di differente complessità, infatti, non ci mostra grosse differenze rispetto alla tabella 4.5; anzi, i due casi, quello in cui si tiene conto delle part of speech e quello in cui si ignorano, sono del tutto analoghi.

Passando al confronto tra generi per i corpora difficili (tabella 4.9), invece, si possono riscontrare alcune differenze rispetto al caso in cui non abbiamo tenuto conto delle part of speech; ma generalmente quasi tutte le variazioni sono estremamente significative.

	Media classi di frequenza forme	Media classi di frequenza lemmi
Articoli scientifici vs Scuola superiore	☑	☑
Repubblica vs Articoli scientifici	☒	☑
Repubblica vs Scuola superiore	☑	☑
Repubblica vs TerenceTeacher originale	☑	☒
TerenceTeacher originale vs Articoli scientifici	☑	☑
TerenceTeacher originale vs Scuola superiore	☒	✓

Tabella 4.9: calcolo del p-value nel confronto tra corpora complessi di diverso genere tenendo conto delle part of speech.

Per i corpora semplici la situazione è quella chiaramente riportata nella tabella 4.10. Notiamo come le variazioni delle classi di frequenza in *2Parole* e nei testi di *scuola elementare* sono molto significative. C'è significatività anche nelle variazioni tra *2Parole* e *Wikipedia* e nelle variazioni tra *Wikipedia* e i testi di *Terence* e *Teacher* semplificati.

	Media classi di frequenza forme	Media classi di frequenza lemmi
2Parole vs Scuola elementare	☒	☑
2Parole vs	✓	✓

TerenceTeacher semplificati		
2Parole vs Wikipedia	✗	☑
TerenceTeacher semplificati vs scuola elementare	✓	☒
TerenceTeacher semplificati vs Wikipedia	☑	☑
Wikipedia vs Scuola elementare	☒	☑

Tabella 4.10: calcolo del p-value nel confronto tra corpora semplici di diverso genere tenendo conto delle part of speech.

4.2.2 – Seconda parte

In questa sezione saranno esposti i dati dell'ultima parte di questo esperimento, trattando di un caso un po' particolare. Nel grafico 4.3 è data la media delle classi di frequenza delle forme di itWaC presenti negli otto corpora e calcolati per singola part of speech.

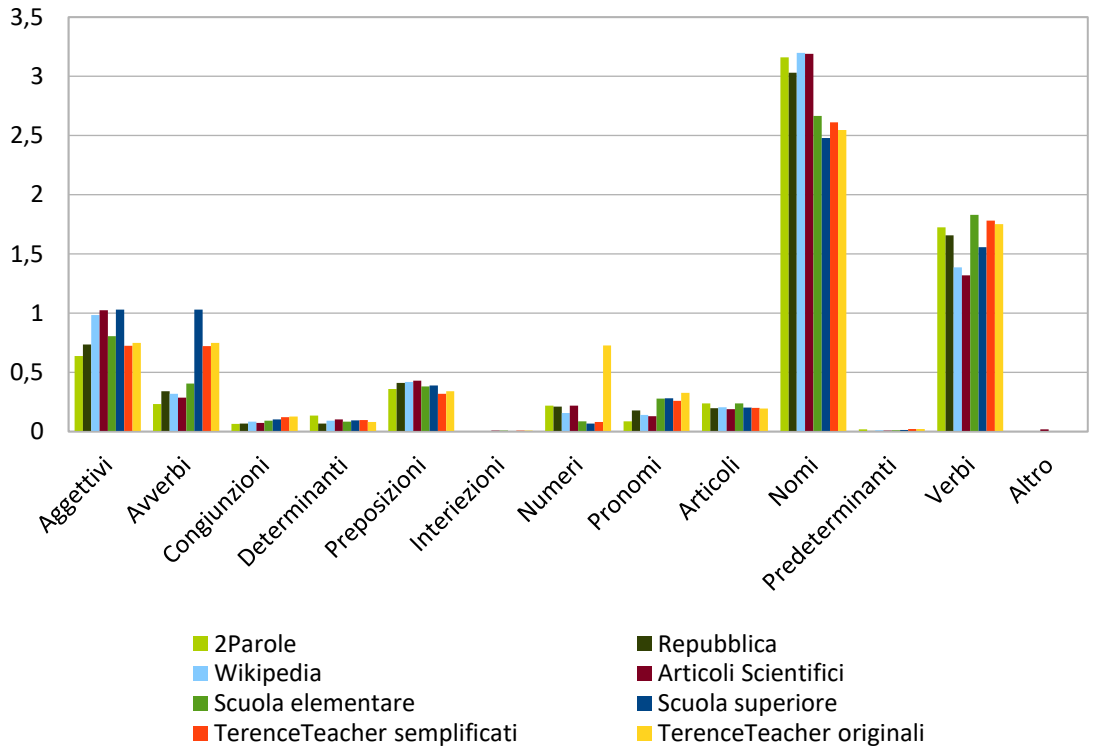

Grafico 4.3: classi di frequenza medie delle forme di itWaC all'interno degli otto corpora per singola part of speech.

È possibile notare come i nomi e i verbi siano le parti del discorso più frequentemente usate in tutti e otto i corpora, con particolare riguardo dei nomi. Se ne deduce che, al di là di qualsiasi differenza di genere e di complessità, in tutti i testi c'è una netta preponderanza dei nomi, prima, e dei verbi, poi. Entrando nel dettaglio, variazioni di complessità possiamo notarne osservando gli avverbi: c'è un grosso stacco tra i testi di *Scuola Elementare* e i testi di *Scuola Superiore*; però attraverso il calcolo del p-value nella tabella 4.11 notiamo chiaramente che la variazione non è significativa. Stessa evidenza e una conclusione molto simile per quanto riguarda gli avverbi messi a confronto tra i testi di *Terence* e *Teacher* originali e semplificati. In questo caso, in realtà, la significatività c'è, ma è poca.

Forme	Articoli scientifici vs Wikipedia	Repubblica vs 2Parole	Scuola superiore vs Scuola elementare	TerenceTeacher originale vs semplificato

Aggettivi	✗	✓	✓	✗
Avverbi	✗	✓	✗	~
Congiunzioni	✗	✓	✗	✗
Determinanti	✓	✓	~	✗
Preposizioni	✗	✓	✗	✗
Interiezioni	✓	~	✗	✗
Numeri	✓	✗	✗	✗
Pronomi	✗	✓	✗	~
Articoli	✗	✓	✓	✗
Nomi	✗	~	~	✗
Predeterminanti	✓	✗	~	✗
Verbi	✗	✗	✓	✗
Altro	✓	✗	✗	✗

Tabella 4.11: calcolo del p-value nel confronto tra corpora dello stesso genere per singola part of speech (POS).

Passando dalle variazioni di complessità a quelle di genere (per la significatività fare riferimento alle tabelle 4.12 e 4.13) possiamo notare come ci sia un divario tra i testi giornalistici e scientifici e i testi scolastici e narrativi nell'uso dei nomi; in questo caso la significatività è positiva per tutti i possibili confronti, ad eccezione di quello tra *2Parole* e *Wikipedia*. Vi è una forte variazione di frequenza nei verbi, tra i corpora scientifici e i corpora scolastici. Anche in questo caso possiamo considerare la variazione di frequenza non casuale, e quindi significativa. Altre variazioni di genere le riscontriamo tra i verbi e i numeri.

C'è una variazione significativa tra gli avverbi nei testi di *scuola superiore* e nei testi narrativi se messi a confronto con i testi giornalistici e scientifici. In genere non è una sorpresa riscontrare una discreta concentrazione di avverbi all'interno di testi di tipo narrativo, infatti anche in questo caso abbiamo che queste variazioni sono tutte significative.

Infine, come è già stato sottolineato, i testi narrativi di *Terence* e *Teacher* originali sembrano avere un'alta concentrazione di numeri, non solo in riferimento agli stessi

testi ma semplificati, ma anche all'interno del confronto con testi di altro genere: giornalistico, scientifico e scolastico. A differenza del confronto eseguito con la controparte semplificata dei testi narrativi, in questo caso abbiamo che, eccezion fatta se confrontati con i testi scolastici, le variazioni sono tutte statisticamente significative.

Forme	Articoli scientifici vs Scuola superiore	Repubblica vs Articoli scientifici	Repubblica vs Scuola superiore	Repubblica vs Terence Teacher originale	Terence Teacher originale vs Articoli scientifici	Terence Teacher originale vs Scuola superiore
Aggettivi	✗	✓	✓	✗	✓	✓
Avverbi	✓	✓	✓	✓	✓	✓
Congiunzioni	✓	✗	✓	✓	✓	✗
Determinanti	✗	✓	✓	✗	✓	✗
Preposizioni	✓	✓	✗	✓	✓	✓
Interiezioni	✗	✓	✗	✗	✗	✗
Numeri	✓	✗	✓	✓	✓	✗
Pronomi	✓	✓	✓	✓	✓	✗
Articoli	✗	✗	✗	✗	✗	✗
Nomi	✓	✓	✓	✓	✓	✗
Predeterminanti	✓	✓	✓	✓	✓	✗
Verbi	✓	✓	✗	✗	✓	✓
Altro	✓	✓	✗	✗	✓	✗

Tabella 4.12: calcolo del p-value nel confronto tra corpora complessi e di diverso genere per singola part of speech (POS).

Forme	2Parole vs Scuola elementare	2Parole vs TerenceTeacher semplificati	2Parole vs Wikipedia	TerenceTeacher semplificati vs scuola elementare	TerenceTeacher semplificati vs Wikipedia	Wikipedia vs Scuola elementare
Aggettivi	✓	✗	✓	✗	✓	✓

Avverbi	☑	☑	☑	☒	☑	☑
Congiunzioni	☑	☑	☑	✓	☑	☒
Determinanti	☑	✓	☑	☒	☒	☒
Preposizioni	☒	✓	☑	✓	☑	✓
Interiezioni	✓	✗	✓	☒	☒	☒
Numeri	☑	☑	☑	☒	✓	✓
Pronomi	☑	☑	☑	☒	☑	☑
Articoli	☒	✗	☑	☑	☒	☑
Nomi	☑	☑	☒	☒	☑	☑
Predetermi-nanti	☒	✗	☒	✓	☑	☒
Verbi	☒	☒	☑	☒	☑	☑
Altro	☒	☒	✗	☒	☒	☒

Tabella 4.13: calcolo del p-value nel confronto tra corpora semplici e di diverso genere per singola part of speech (POS).

Passiamo ora a fare la stessa analisi ma considerando i lemmi invece delle semplici forme. Il grafico a colonne 4.4 dà un'idea generale delle distribuzioni. Notiamo che il quadro generale è pressoché identico a quello delle forme con dei valori numerici generalmente poco più bassi.

Entrando nel particolare possiamo notare come per quanto riguarda gli aggettivi abbiamo variazioni anche molto diffuse. Se consideriamo i testi scolastici nelle due varianti di complessità, abbiamo che la media delle classi di frequenza degli aggettivi nei testi di *scuola superiore* è superiore rispetto a quella dei testi di *scuola elementare*, e questa variazione, come si può notare facilmente nelle tabella 4.13 è estremamente significativa. Anche in questo caso, tra le altre cose, notiamo come i testi di *Terence* e *Teacher* semplificati sembrano presentare un valore medio delle classi di frequenza di "numeri" molto elevato nel confronto sia tra generi che complessità.

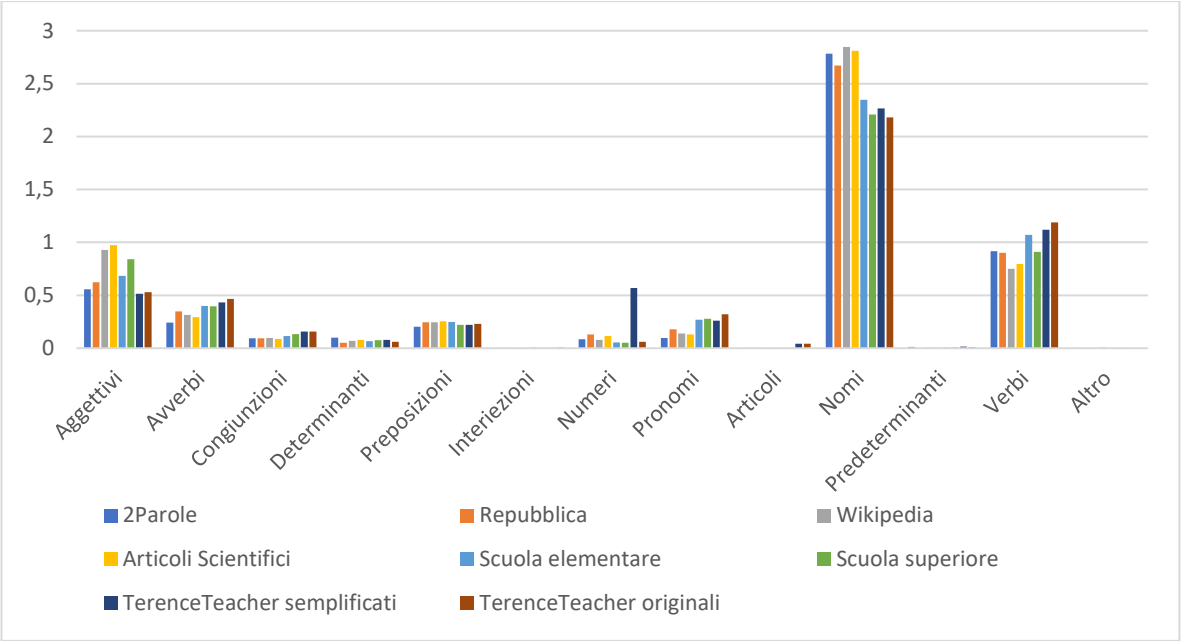

Grafico 4.4: classi di frequenza medie dei lemmi di itWaC all'interno degli otto corpora per singola part of speech.

Ed è facile notare come questo valore sia casuale: se consideriamo la variazione in riferimento all'altro corpus dello stesso genere e a corpora della stessa complessità ma di diverso genere (tabella 4.16), la significatività è sempre nulla. Rimanendo nei confronti per i due corpora di uno stesso genere e differente complessità, abbiamo variazioni visibili nei verbi e nei nomi, ma con poca significatività.

Lemmi	Articoli scientifici vs Wikipedia	Repubblica vs 2Parole	Scuola superiore vs Scuola elementare	TerenceTeacher originale vs semplificato
Aggettivi	✗	☑	☑	☒
Avverbi	☒	☑	☒	☒
Congiunzioni	☒	☒	☒	☒
Determinanti	☑	☑	✗	☒
Preposizioni	☒	☑	✓	☒
Interiezioni	☒	✗	☒	☒
Numeri	☑	☑	☒	☒
Pronomi	☒	☑	☒	✗

Articoli	☑	✗	☒	☒
Nomi	☒	✗	☒	☒
Predeterminanti	✓	☒	✗	☒
Verbi	✓	☒	☑	☒
Altro	☑	☒	☒	☒

Tabella 4.14: calcolo del p-value nel confronto tra corpora dello stesso genere per singola part of speech (POS).

Ancora una volta, se passiamo a confrontare le variazioni riscontrate nei generi allo stesso livello di complessità, abbiamo una situazione molto varia se consideriamo gli aggettivi, i verbi e i nomi. Nel caso dei corpora complessi (tabella 4.15), vediamo come tutte le variazioni che riguardano gli aggettivi sono estremamente significative; lo stesso nei corpora semplici (tabella 4.16), fatta eccezione per la – in effetti molto piccola – variazione tra *2Parole* e i testi di *Terence* e *Teacher* semplificati. I nomi nei corpora complessi variano quasi sempre significativamente mentre in due casi dei corpora semplici abbiamo una significatività nulla: nel confronto tra *2Parole* e *Wikipedia*, e nel confronto tra i testi di *Terence* e *Teacher* semplificati e i testi di *scuola elementare*. Infine, guardando i verbi, sia per i corpora complessi che per quelli semplici abbiamo anche in questo caso una significatività generalmente molto alta (un solo caso per gli uni e per gli altri in cui non ce ne sia affatto).

Lemmi	Articoli scientifici vs Scuola superiore	Repubblica vs Articoli scientifici	Repubblica vs Scuola superiore	Repubblica vs TerenceTeacher originale	TerenceTeacher originale vs Articoli scientifici	TerenceTeacher originale vs Scuola superiore
Aggettivi	☑	☑	☑	☑	☑	☑
Avverbi	☑	☑	✓	☑	☑	☑
Congiunzioni	☑	☒	☑	☑	☑	✗
Determinanti	☒	☑	☑	☒	☑	✓
Preposizioni	☑	✓	✓	☒	✓	☒
Interiezioni	☒	✗	☒	☒	☒	☒

Numeri	☑	☒	☑	☑	☑	☒
Pronomi	☑	☑	☑	☑	☑	☒
Articoli	☑	☑	☒	☑	☑	☑
Nomi	☑	✗	☑	☑	☑	☒
Predetermi-nanti	✗	✓	☑	☑	✓	☒
Verbi	☑	☑	☒	☑	☑	☑
Altro	☑	☑	☒	☒	☑	☒

Tabella 4.15: calcolo del p-value nel confronto tra corpora complessi per singola part of speech (POS).

Lemmi	2Parole vs Scuola elementare	2Parole vs TerenceTeacher semplificati	2Parole vs Wikipedia	TerenceTeacher semplificati vs scuola elementare	TerenceTeacher semplificati vs Wikipedia	Wikipedia vs Scuola elementare
Aggettivi	☑	☒	☑	☑	☑	☑
Avverbi	☑	☑	☑	☒	☑	☑
Congiunzioni	☑	☑	✗	☑	☑	✓
Determinanti	☑	✗	☑	☒	☒	☒
Preposizioni	☑	✗	☑	✗	✓	☒
Interiezioni	✓	☒	✓	☒	☒	☒
Numeri	☒	☒	✗	☒	☒	☒
Pronomi	☑	☑	☑	☒	☑	☑
Articoli	☒	☑	☑	☑	☑	☒
Nomi	☑	☑	☒	☒	☑	☑
Predetermi-nanti	☒	✗	☒	✓	☑	☒
Verbi	☑	☑	☑	☒	☑	☑
Altro	☒	☒	✗	☒	☒	☒

Tabella 4.16: calcolo del p-value nel confronto tra corpora semplici per singola part of speech (POS).

4.3 – Studio sul lessico: brevi considerazioni

In questo quarto capitolo sono stati analizzati i dati statistici riguardanti il primo esperimento descritto in questo elaborato, atto a uno studio sul lessico che si propone come confronto fra generi e complessità. Sono stati utilizzati corpora di quattro generi differenti per due livelli di complessità: il genere giornalistico ha come rappresentanti 2Parole e *Repubblica*, il genere scientifico ha come rappresentanti *Wikipedia* e *Articoli Scientifici*, il genere scolastico presenta testi di *Scuola Elementare* e testi di *Scuola Superiore*, infine il genere narrativo è rappresentato dai testi di *Terence* e *Teacher*, sia originali che semplificati.

L'esperimento è stato diviso in due fasi: nella prima si è analizzato come variano le tre categorie del Nuovo Vocabolario di Base di Tullio de Mauro (vocabolario Fondamentale, vocabolario di Alto Uso e vocabolario di Alta Disponibilità) al variare dei generi e delle complessità. Quel che ne è venuto fuori è che il vocabolario di Alto Uso riflette leggere variazioni sia di genere che di complessità, al contrario del vocabolario di Alta Disponibilità, che subisce variazioni anche significative al variare dei generi. In generale, però, la visione d'insieme è come potevamo aspettarcela, con il vocabolario Fondamentale che ricopre quasi la totalità delle forme e dei lemmi all'interno di ciascuno degli otto corpora.

La seconda fase dell'esperimento è a sua volta divisa in altrettante parti, una delle quali è stata dedicata a indagare le variazioni di frequenza delle forme e dei lemmi di itWaC all'interno degli otto corpora in esame sia tenendo conto delle part of speech che ignorandole; l'ultima fase invece è rivolta allo stesso studio, ma per ognuna delle part of speech che compongono il testo (ad eccezione dei segni di punteggiatura). Ne è risultato che, come è possibile aspettarsi, nomi e verbi sono le forme più diffuse per qualsiasi genere e livello di complessità. In particolare è stato notato come varino significativamente, entrambi, al variare dei generi ma non del livello di complessità. Ad esempio, ci sono forti variazioni significative al livello della complessità tra gli aggettivi dei testi di *scuola superiore* e *scuola elementare*, mentre per quanto riguarda gli avverbi si tratta di variazioni casuali e non significative. Queste conclusioni sono state fatte sulla base del calcolo del p-value con il Wilcoxon rank-sum test.

CAPITOLO 5

Analisi dei dati: lessico e sintassi

In questo capitolo saranno analizzati i dati relativi allo studio di alcuni fenomeni sintattici che co-occorrono con tutti i sostantivi degli otto corpora esaminati (5.1), in particolare sarà data una panoramica generale delle tipologie di dipendenza (5.1.1) e delle informazioni sintattiche relative alla distanza dalla testa di un particolare sostantivo, al numero dei suoi dipendenti, al numero di fratelli e alla sua distanza dalla radice (5.1.2). Un'indagine simile verrà esposta in 5.2, questa volta relativamente alle tipologie di dipendenze dei sostantivi più facili e più difficili all'interno di ognuno dei corpora (5.2.1) e dei fenomeni sintattici di quelli più facili e difficili (5.2.2).
In funzione di conclusione del capitolo sarà data una discussione panoramica dei dati analizzati (5.3).

5.1 – Tutti i sostantivi

5.1.1 – Tipologie di dipendenza

Per gli otto corpora utilizzati sono stati estratti tutti i sostantivi e ordinati per complessità a seconda della frequenza con la quale apparivano in itWaC e nel corpus che si voleva utilizzare. Per ogni occorrenza di queste parole all'interno del corpus sono state estratte alcune informazioni, tra le quali la relazione di dipendenza che il token ha con la sua testa; in seguito, per ognuna di esse è stato contato il numero totale di occorrenze e quante, tra queste, rappresentassero il soggetto della loro testa (subj), quante l'oggetto (obj), quante il modificatore (mod), quante, ancora, dipendevano da una preposizione (prep) e quante erano la radice della frase (ROOT).
In *2Parole*, il corpus giornalistico rappresentante di un livello di complessità basso, occorrono 22889 sostantivi e nel grafico 5.1 è possibile contestualizzare questo numero: solo 3077 di questi sono soggetti rispetto alla propria testa (13%), un dato solo di poco inferiore a quello riguardante gli oggetti (14%). Ben 4892 sostantivi occorrono come modificatori (21%), mentre quasi il doppio sono quelli che hanno con la loro testa una relazione di tipo preposizionale e che dipendono quindi da una preposizione (38%).

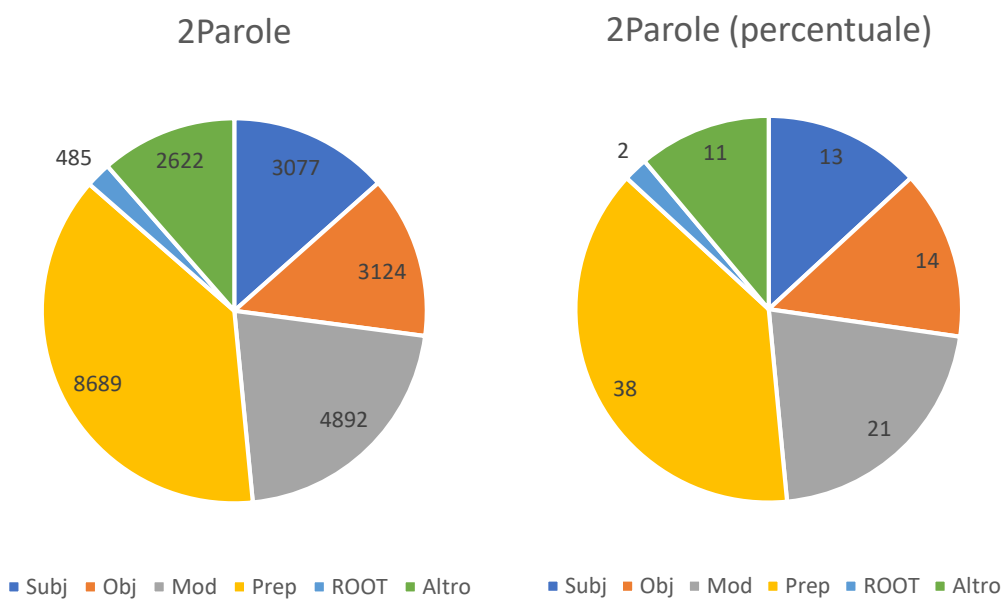

Grafico 5.1: numero di occorrenze (a sinistra) e relativa percentuale (a destra) dei sostantivi all'interno di 2Parole *divisi per tipo di dipendenza.*

Se passiamo a osservare gli stessi dati relativi al corpus giornalistico più complesso, *Repubblica* (grafico 5.2), salta immediatamente all'occhio che anche se numericamente i valori sono tutti più alti (questo perché il corpus è più grosso) la situazione è pressoché identica: il numero di occorrenze più grosso è relativo ai sostantivi retti da una preposizione (39%), numeri simili si hanno per i soggetti e gli oggetti, uno particolarmente elevato per i modificatori (25%), mentre sono poche le occorrenze di sostantivi come radice della frase (4%).

La situazione non varia di molto se cambiamo genere. Nel grafico 5.3 abbiamo i numeri che si riferiscono al corpus scientifico rappresentante del livello di complessità basso: *Wikipedia.*

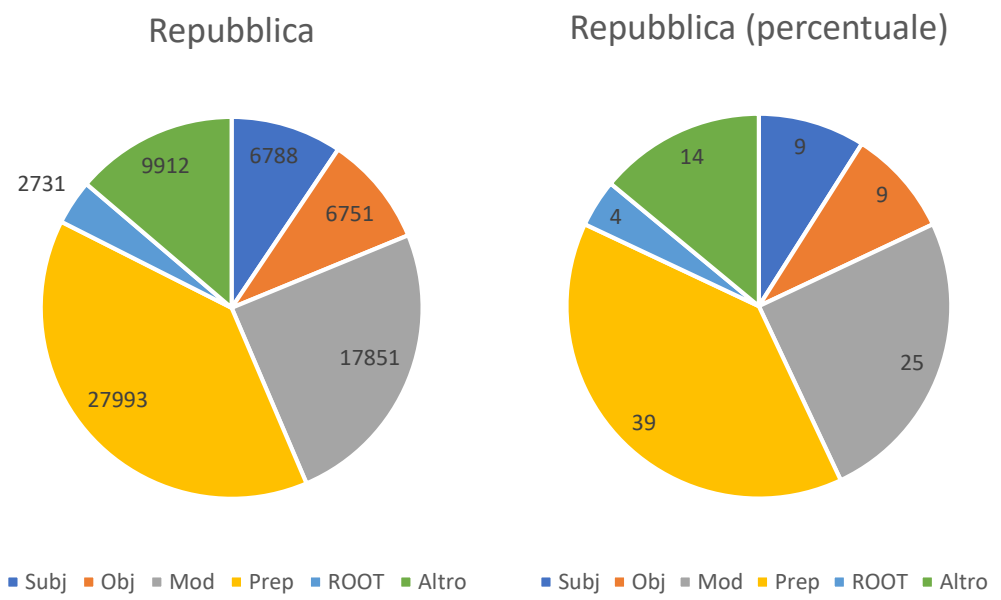

Grafico 5.2: numero di occorrenze (a sinistra) e relativa percentuale (a destra) dei sostantivi all'interno di Repubblica *divisi per tipo di dipendenza.*

Nel caso di *Wikipedia* abbiamo numeri che si avvicinano molto a quelli già visti per *Repubblica* (entrambi i corpora hanno un numero di occorrenze di sostantivi molto simile). Invece, una leggera diversificazione possiamo notarla se poniamo la nostra attenzione al corpus scientifico complesso. Il grafico 5.4 ci mostra come per il corpus di *Articoli Scientifici* il numero di occorrenze maggiore, ancora una volta, riguarda i sostantivi che dipendono da una preposizione (33%); ma c'è un numero particolarmente elevato di sostantivi che hanno un tipo di relazione di dipendenza diverso da quelli evidenziati nel grafico (31%). Inoltre, il numero di sostantivi che sono radice della frase in cui occorrono è particolare: 9568 (5%) contro i 12201(6%) sostantivi che occorrono come soggetto e i 11591 (6%) che occorrono come oggetto, una differenza sottile che evidenzia come all'interno di questo corpus il numero di sostantivi che possiamo considerare come radice è molto vicino a quello dei sostantivi che fungono da soggetto o oggetto.

I dati ritornano a essere coerenti a quelli considerati in precedenza se cambiamo ancora una volta genere e passiamo ai corpora di tipo scolastico. Per quanto riguarda i testi di *Scuola Elementare*, rappresentanti del grado di complessità basso, facciamo riferimento al grafico 5.5, mentre per i testi di *Scuola Superiore*, rappresentanti del grado di complessità più elevato, ci riferiremo al grafico 5.6.

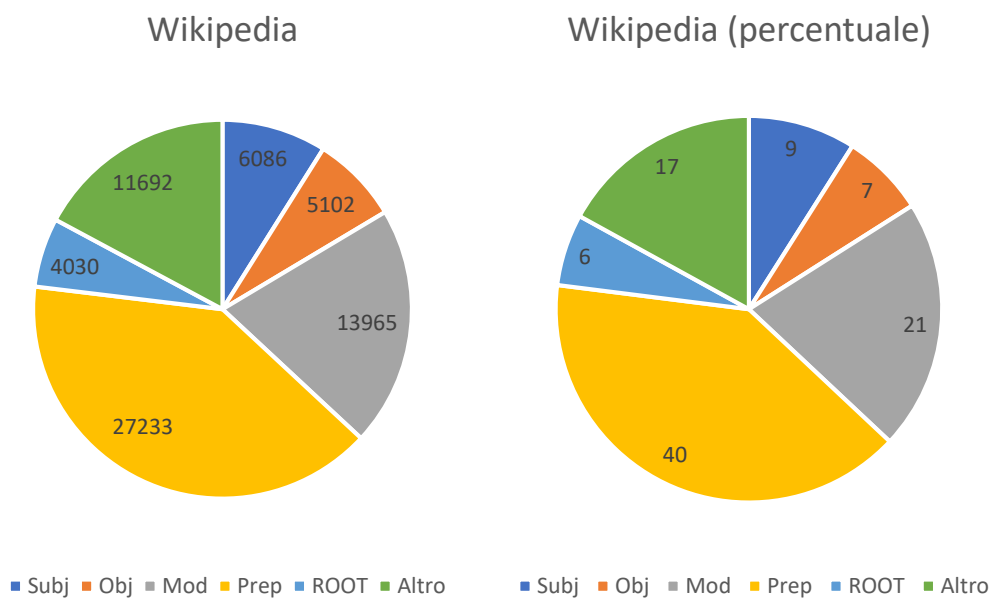

Grafico 5.3: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi all'interno di Wikipedia *divisi per tipo di dipendenza.*

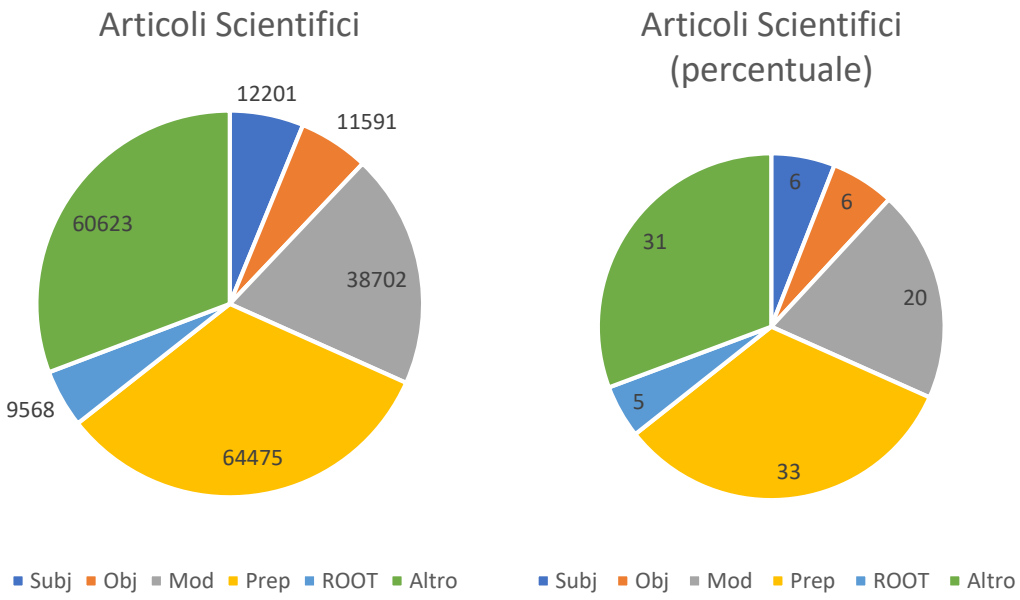

Grafico 5.4: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi all'interno di Articoli Scientifici *divisi per tipo di dipendenza.*

Il corpus di testi di *Scuola Elementare* è una collezione molto piccola, e questo è il motivo per cui i numeri relativi alle sue occorrenze sono bassi. All'interno di questi testi sono state estratte 5144 occorrenze di sostantivi e quasi la metà di esse si riferiscono a relazioni di dipendenza di tipo preposizionale (44%); 838 sostantivi occorrono

come soggetto (16%), 688 come oggetto (13%), 678 come modificatori della propria testa (13%) e appena 122 occorrenze sono radice (2%).

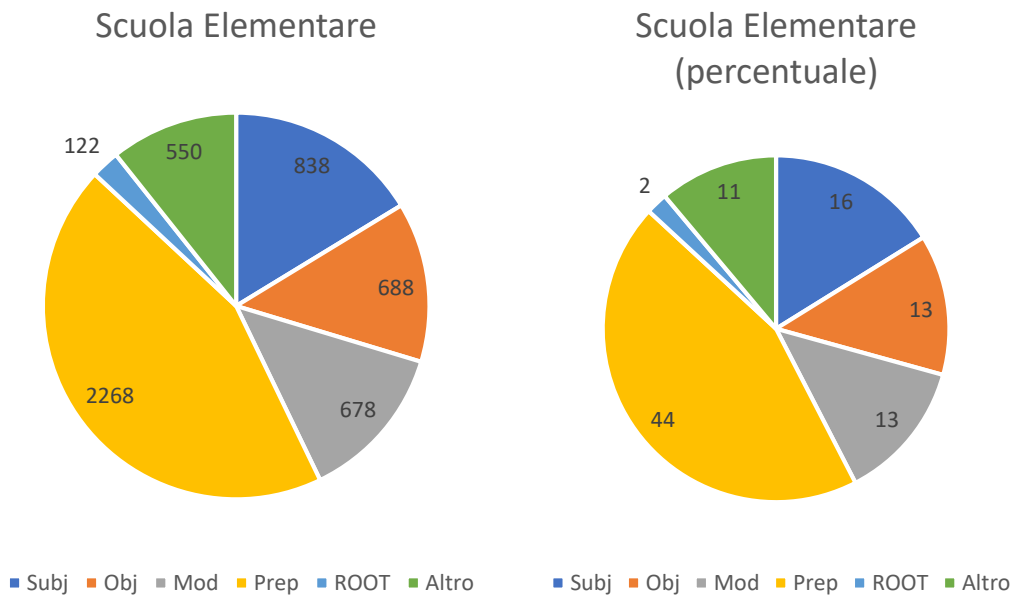

Grafico 5.5: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi all'interno di testi di scuola elementare *divisi per tipo di dipendenza.*

Nel caso dei testi di *Scuola Superiore* abbiamo numeri più alti. Si tratta infatti di una collezione leggermente più corposa. Anche qui abbiamo un numero risicato di occorrenze di sostantivi che sono radice della frase, e rispetto ai soggetti (1391) e agli oggetti (1396) il numero di occorrenze che sono modificatori della loro testa è più del doppio (19%), un'evidenza che ricorda quella di entrambi i corpora scientifici e di *Repubblica*. È un dato che sembra verificarsi all'interno di testi più complessi (ed è chiaro se non consideriamo, per un attimo, *Wikipedia* un corpus semplice e la collezione di testi di *Terence* e *Teacher* originali un corpus complesso).

Le due collezioni di testi di *Terence* e *Teacher*, sia originali che semplificati, infine, presentano numeri quasi analoghi se confrontati tra di loro: il grafico 5.7 presenta i numeri relativi ai semplificati, mentre il grafico 5.8 quelli relativi agli originali. I primi hanno un numero di sostantivi che occorrono come soggetto e come oggetto leggermente maggiori rispetto ai secondi: 1022 soggetti per i semplificati (14%) contro i 970 dei testi originali (12%) e 879 oggetti per i semplificati (12%) contro gli 868 degli originali (11%).

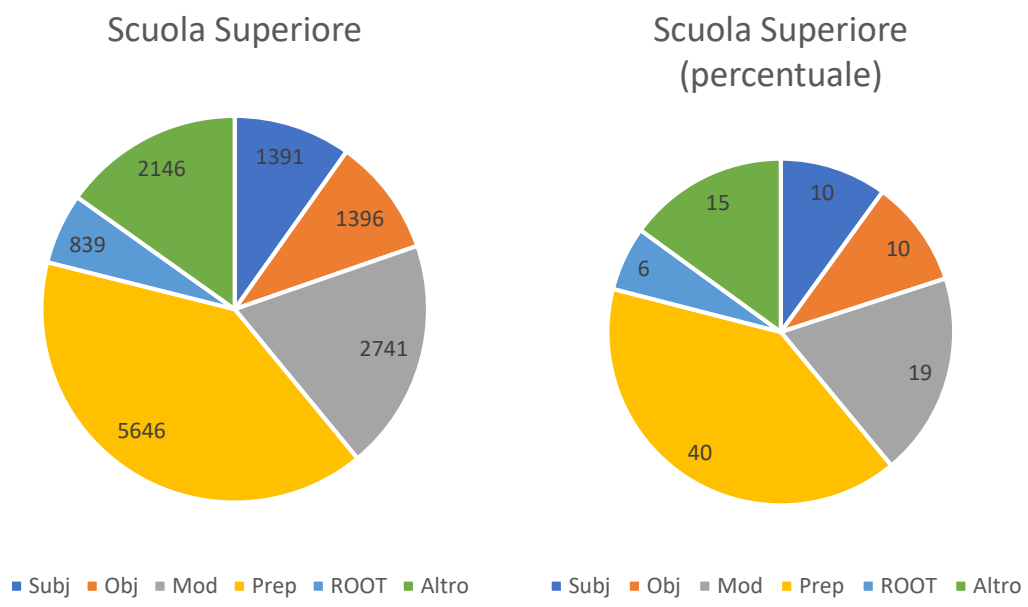

Grafico 5.6: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi all'interno di testi di scuola superiore *divisi per tipo di dipendenza.*

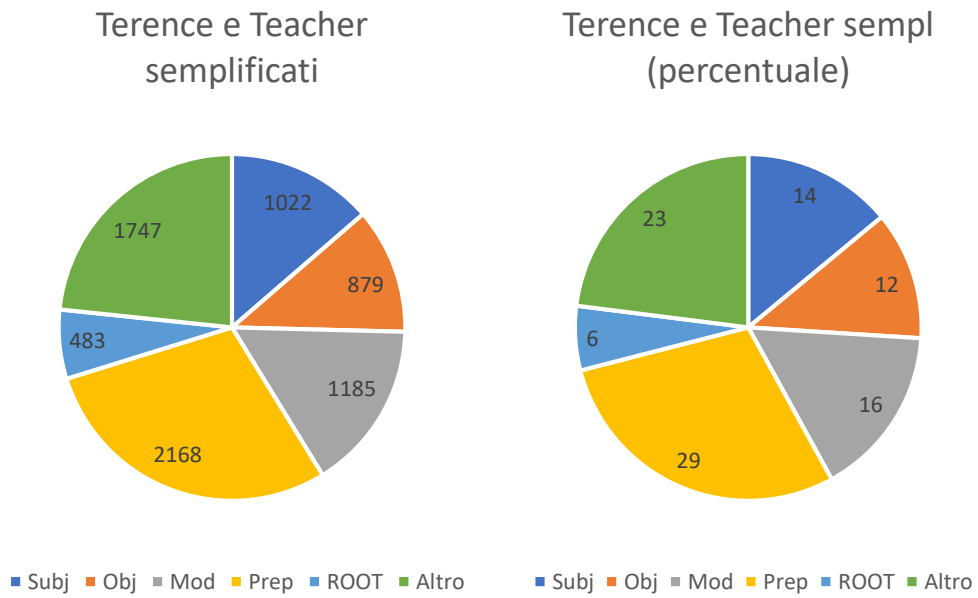

Grafico 5.7: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi all'interno di testi di Terence e Teacher semplificati *divisi per tipo di dipendenza.*

Per quanto riguarda i modificatori, entrambi i corpus presentano un numero di occorrenze molto simile (16%), così come per i sostantivi che occorrono come radice (6%); invece per i sostantivi che dipendono da una preposizione, i primi offrono un dato leggermente minore (29% contro 32%).

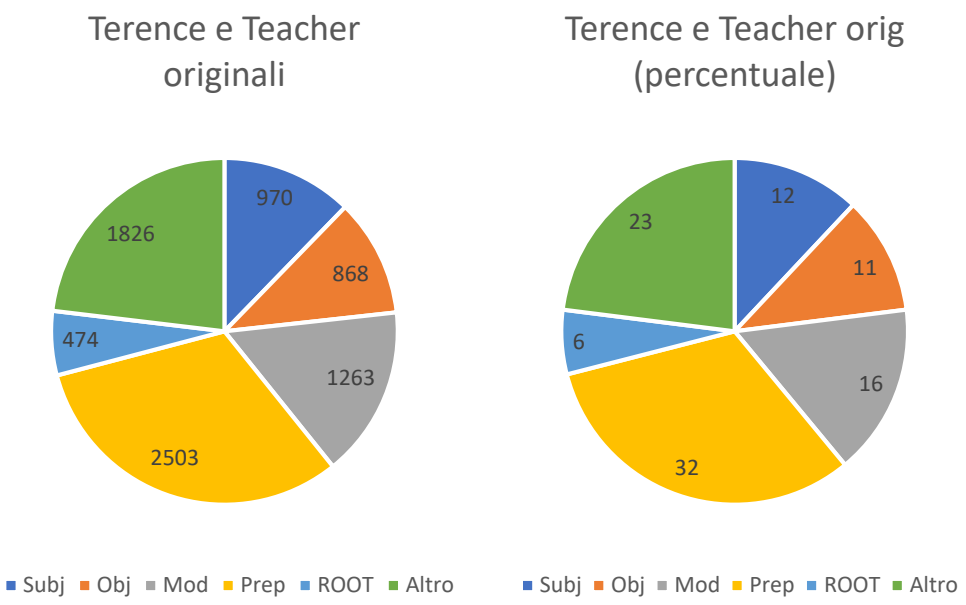

Grafico 5.8: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi all'interno di testi di Terence e Teacher originali *divisi per tipo di dipendenza.*

5.1.2 – Fenomeni sintattici

Per ogni occorrenza dei sostantivi di ogni corpus sono state estratte informazioni relative alla distanza del singolo sostantivo dalla sua testa (in termini di numero di parole), il numero dei suoi figli (quanti token sono suoi dipendenti), il numero di fratelli (quanti token dipendono dalla sua stessa testa) e la sua distanza dalla radice. Sono fenomeni rilevanti sia dal punto di vista sintattico che da quello lessicale. Come abbiamo visto in 5.1.1, alcune occorrenze dei sostantivi sono risultate essere effettivamente la radice della frase entro cui compaiono: in questo caso la loro distanza dalla radice sarà uguale a zero. La tabella 5.1 dà una panoramica di dati relativi alla media delle informazioni estratte per ogni corpus. Vediamo come per i due corpora giornalistici non ci sono grosse variazioni per quanto riguarda la distanza dalla testa, il numero di figli e il numero di fratelli, ma c'è un grosso divario se consideriamo la distanza dalla radice: *2Parole* presenta una distanza dalla radice media di 3,017 contro la media di 4,234 vista in *Repubblica*. La variazione, oltre che interessante, è estremamente significativa, come possiamo notare dalla tabella 5.2.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole	2,188	1,154	1,646	3,017
Repubblica	2,203	1,159	1,546	4,234
Wikipedia	2,55	1,350	1,627	4,311
Articoli scientifici	2,198	1,230	1,427	4,806
Scuola elementare	2,160	1,278	1,524	3,4
Scuola superiore	2,578	1,417	1,642	4,275
TerenceTeacher semplificato	2,109	1,255	1,734	2,987
TerenceTeacher originale	2,209	1,266	1,707	3,247

Tabella 5.1: media delle informazioni estratte per ogni corpus sulla base di tutti i sostantivi di ciascuno dei corpus.

Estremamente significative sono anche le variazioni che riguardano i due corpora scientifici, anche se lievi. In particolare, tutte le medie, eccezion fatta quella relativa alla distanza dalla radice, sono più alte se consideriamo *Wikipedia*, la collezione di genere scientifico rappresentante di una complessità bassa, contro *Repubblica*, il corpus giornalistico complesso. L'evidenza non stupisce in quanto *Wikipedia* rispetto agli altri corpora semplici utilizzati per quest'indagine è comunque molto più complesso.
Ancora estremamente significative sono le variazioni che riguardano la distanza dalla radice nel confronto tra i due corpora scolastici e anche in quello tra i due corpora narrativi. Per questi ultimi due confronti non ci sono altri dati significativi, o comunque non si tratta di grosse variazioni (ad esempio, il numero di figli nel confronto tra i testi di *Scuola Elementare* e di *Scuola Superiore*).
Se passiamo ad analizzare i dati dal punto di vista di un confronto per generi, invece che per complessità, dobbiamo fare riferimento alle tabelle 5.3 e 5.4. La prima evidenzia la significatività delle variazioni nei confronti tra corpora semplici di diverso genere, mentre la seconda riguarda i corpora difficili di diverso genere. In linea di massima la distribuzione di fenomeni sintattici rispetto al lessico di *2Parole* varia significativamente rispetto agli altri tre corpora semplici, mentre invece abbiamo significatività nulla se consideriamo il numero di figli e fratelli nel confronto tra *Wikipedia* e i testi di *Scuola Elementare*. È nulla anche nel caso della distanza dalla testa nel confronto tra i testi di *Terence* e *Teacher* semplificati e quelli di *Scuola Elementare*.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole vs Repubblica	☑	☑	☑	☑
Wikipedia vs Articoli scientifici	☑	☑	☑	☑
Scuola elementare vs Scuola superiore	☒	✓	☒	☑
TerenceTeacher semplificato vs originale	☒	☒	☒	☑

Tabella 5.2: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte da tutti i sostantivi nei corpora facili e difficili dello stesso genere

In tutti gli altri casi riscontriamo mediamente un livello di significatività molto alto, e quindi possiamo concludere che, considerando questi quattro corpora facenti parte di uno stesso livello di complessità, quello semplice, essi variano significativamente al variare del genere di appartenenza. Spesso, alcune variazioni sono anche importanti (ad esempio, la distanza dalla radice dei sostantivi in *2Parole* rispetto a quella in *Wikipedia* o in testi di *Scuola Elementare* o nei testi di *Terence* e *Teacher* semplificati). L'altra tabella, quella relativa alla significatività delle variazioni nei confronti tra corpora complessi e di diverso genere ci dà una panoramica non troppo diversa.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole vs Scuola elementare	☑	☑	✗	☑
2Parole vs TerenceTeacher semplificati	✓	☑	☑	☑
2Parole vs Wikipedia	☑	☑	☑	☑

TerenceTeacher semplificati vs scuola elementare	✗	✗	✓	✓
TerenceTeacher semplificati vs Wikipedia	✓	✓	✓	✓
Wikipedia vs Scuola elementare	✓	✗	✗	✓

Tabella 5.3: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte da tutti i sostantivi nei corpora semplici di diverso genere.

Quasi tutte le variazioni sono estremamente significative, tranne per il numero di fratelli nel confronto tra *Repubblica* e i testi di *Scuola Superiore*, che esprime molta significatività, e la distanza dalla radice nello stesso confronto, con significatività nulla. Hanno significatività nulla anche le variazioni relative al numero di figli nel confronto tra i testi di *Terence* e *Teacher* originali e *Articoli Scientifici* e la distanza dalla testa del sostantivo nel confronto tra i testi di *Terence* e *Teacher* originali e quelli di *Scuola Superiore*. Nel complesso, però, le variazioni non danno differenze molto grandi come succede a volte con i corpora semplici. È evidente come i valori mediamente più alti appartengano al corpus di testi di *Scuola Superiore* e ad *Articoli Scientifici*.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
Articoli scientifici vs Scuola superiore	✓	✓	✓	✓
Repubblica vs Articoli scientifici	✓	✓	✓	✓
Repubblica vs Scuola superiore	✓	✓	✓	✗
Repubblica vs	✓	✓	✓	✓

TerenceTeacher originale				
TerenceTeacher originale vs Articoli scientifici	☑	☒	☑	☑
TerenceTeacher originale vs Scuola superiore	☒	☑	☑	☑

Tabella 5.4: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte da tutti i sostantivi nei corpora complessi di diverso genere.

5.2 – Sostantivi facili VS sostantivi difficili

Dopo aver considerato tutti i sostantivi di ognuno dei corpora sono stati estratti, a partire da questi, quelli più facili e più difficili, avendo come corpus di riferimento itWaC (si veda il capitolo 3); inoltre sono state calcolate le stesse informazioni già esaminate in 5.1 per ogni occorrenza, quindi il tipo di dipendenza, la distanza del sostantivo dalla testa, il suo numero di figli e di fratelli e la sua distanza dalla radice.

5.2.1 – Tipologie di dipendenza

Secondo un meccanismo simile a quanto descritto in 5.1.1, per ognuno dei sostantivi facili e difficili di tutti e otto i corpora sono state contate le occorrenze all'interno della collezione e quante di esse sono soggetto della propria testa (subj), quante l'oggetto (obj), quante il modificatore (mod), quante dipendono da una preposizione (prep) e quante sono la radice della frase (ROOT).

Per quanto riguarda i corpora giornalistici, nei grafici 5.9 e 5.10 sono mostrate le occorrenze dei sostantivi più facili e più difficili in *2Parole* divise per tipo di dipendenza; invece, i grafici 5.11 e 5.12 sono relativi a *Repubblica*.

Per i sostantivi più facili, possiamo notare una distribuzione simile nei due corpora se confrontiamo quelli del primo con quelli del secondo: in entrambi i casi abbiamo che il numero più alto di occorrenze è relativo a quei sostantivi che dipendono da una preposizione. Per *2Parole* possiamo notare un maggior numero di soggetti rispetto agli oggetti, il contrario di quanto è evidente nel corpus giornalistico complesso; invece,

tutti e due hanno un piccolo numero di sostantivi che fungono da modificatori della loro testa e un numero ancora più piccolo di sostantivi radice.

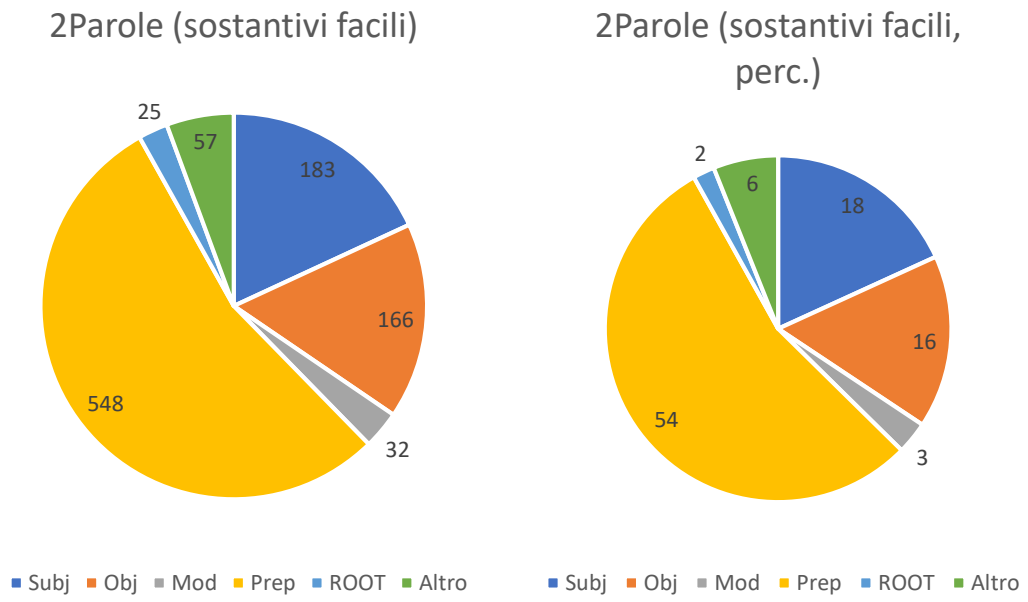

Grafico 5.9: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno di 2Parole *divisi per tipo di dipendenza.*

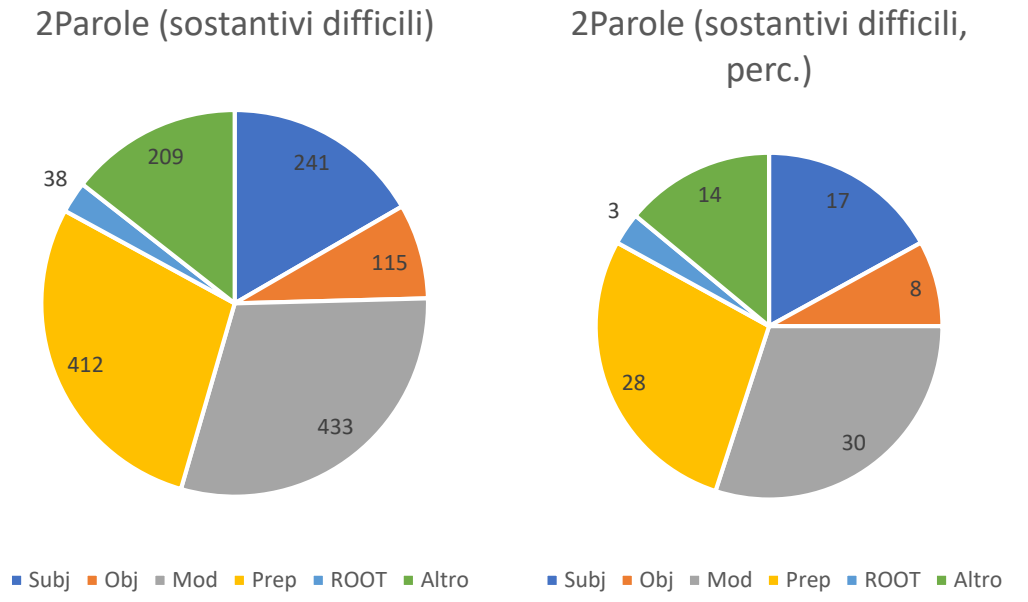

Grafico 5.10: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno di 2Parole *divisi per tipo di dipendenza.*

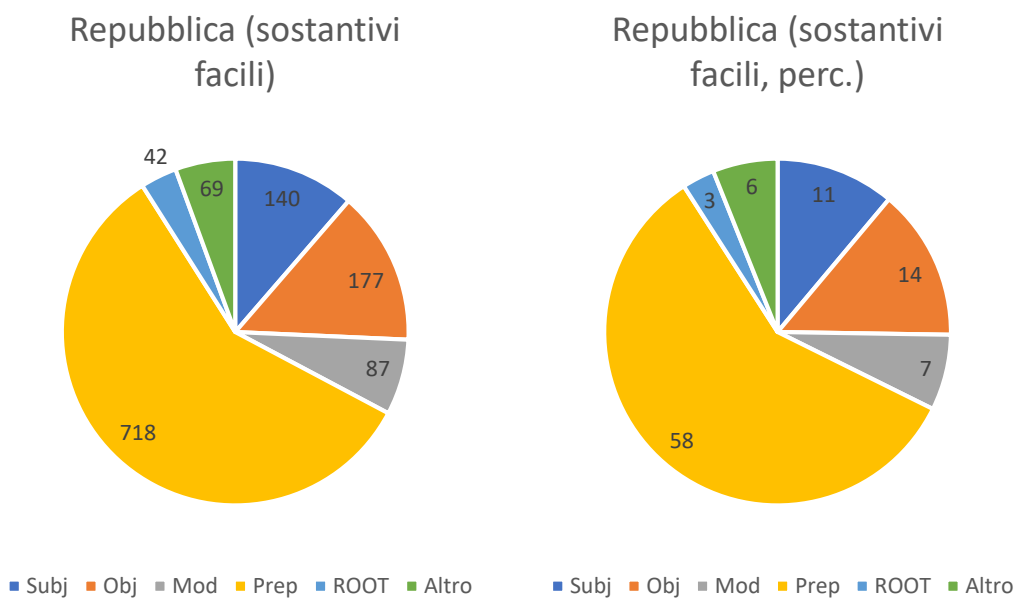

Grafico 5.11: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno di Repubblica *divisi per tipo di dipendenza.*

Se guardiamo, invece, ai sostantivi più complessi dei due corpora, è evidente come la distribuzione cambi. Nel caso di *2Parole*, il numero di occorrenze dei sostantivi che sono modificatori della propria testa è 433 (30%), leggermente più alto rispetto a quello dei sostantivi retti da una preposizione (28%); una situazione simile per numeri più alti la ritroviamo in *Repubblica*. Ciò che se ne potrebbe dedurre è che nel confronto tra due corpora dello stesso genere e di complessità differente c'è similarità nella distribuzione delle dipendenze sintattiche, in particolare relative ai sostantivi.

Prima di giungere a delle conclusioni bisogna però valutare anche gli altri casi in esame. Il grafico 5.13 dà una panoramica del numero di occorrenze dei sostantivi più facili, divisi per tipo di dipendenza, in *Wikipedia*. È evidente come anche in questo caso la distribuzione è in linea con quanto già visto in precedenza: la maggior parte delle occorrenze è relativa ai sostantivi che hanno con la propria testa una relazione di dipendenza di tipo preposizionale (60%); soggetti e oggetti hanno valori molto simili tra di loro (con una lieve preponderanza dei primi), ci sono pochi modificatori e ancor meno radici. La stessa (pressoché identica) situazione la riscontriamo nei sostantivi facili del corpus scientifico complesso: *Articoli Scientifici* (il grafico di riferimento è il 5.15). Anche qui abbiamo un grande numero di sostantivi retti da una preposizione, soggetti e oggetti con valori simili e poco altro.

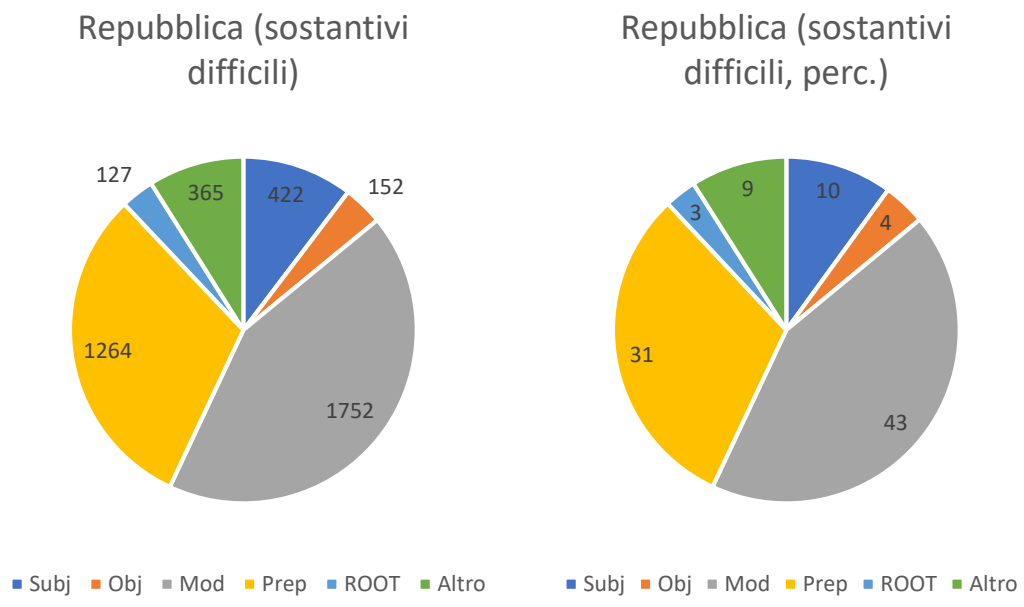

Grafico 5.12: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno di Repubblica *divisi per tipo di dipendenza.*

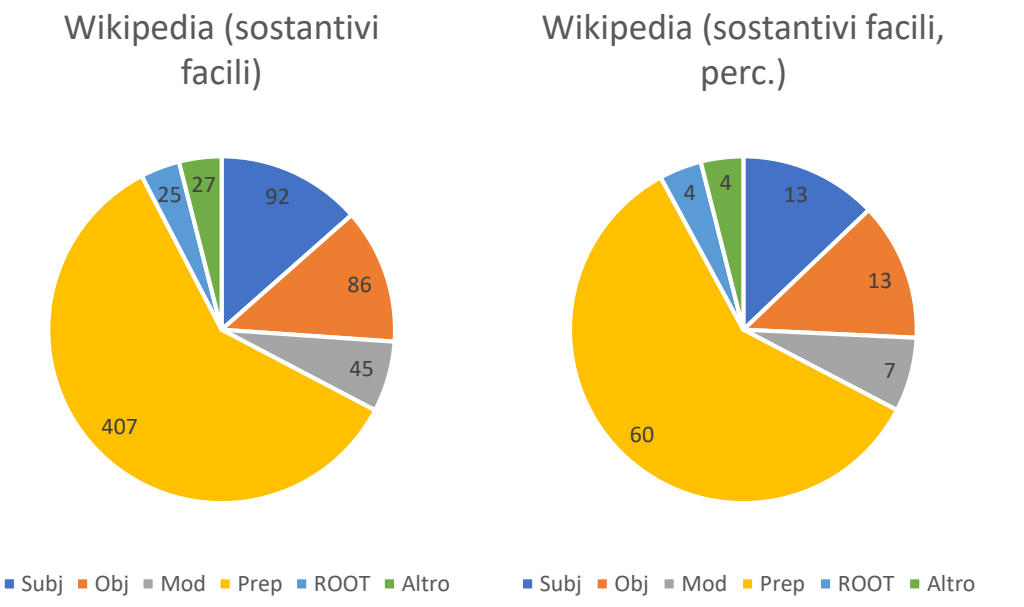

Grafico 5.13: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno di Wikipedia *divisi per tipo di dipendenza.*

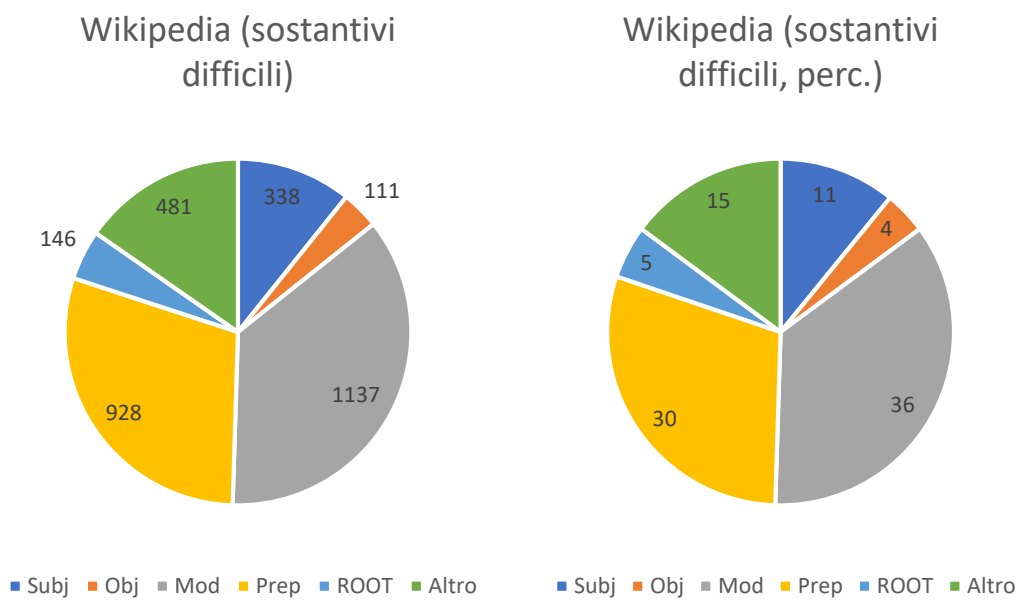

Grafico 5.14: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno di Wikipedia *divisi per tipo di dipendenza.*

Passando ai sostantivi più difficili in *Wikipedia* e in *Articoli Scientifici*, abbiamo andamenti simili a quanto visto con quelli dei due corpora giornalistici. Come possiamo notare nel grafico 5.14, nel corpus scientifico semplice il maggior numero di dipendenze che riguardano i sostantivi complessi sono di tipo modificazione, il 36% contro il 30% dei sostantivi retti da una preposizione.

È simile l'andamento dei sostantivi difficili in *Articoli Scientifici*. È evidente osservando il grafico 5.16. Anche qui, il numero di occorrenze maggiore è quello relativo ai modificatori, quasi il doppio rispetto ai sostantivi con una relazione di dipendenza di tipo preposizionale. È particolare, inoltre, che il numero di sostantivi che dipendono come radice sia più alto del numero dei soggetti e di quello degli oggetti.

Iniziamo a notare qualche lieve differenza per i corpora di genere scolastico, principalmente se guardiamo ai sostantivi difficili: notiamo infatti che sia per quanto riguarda la collezione di testi di *Scuola Elementare* (per i sostantivi difficili si veda il grafico 5.18) che per la collezione di testi di *Scuola Superiore* (si veda il grafico 5.20), rispetto ai casi analizzati in precedenza la distribuzione porta alcune variazioni. Il numero dei modificatori, in questo caso, è più piccolo di quello dei sostantivi che reggono una preposizione. E il numero di radici è molto inferiore al numero di soggetti e oggetti.

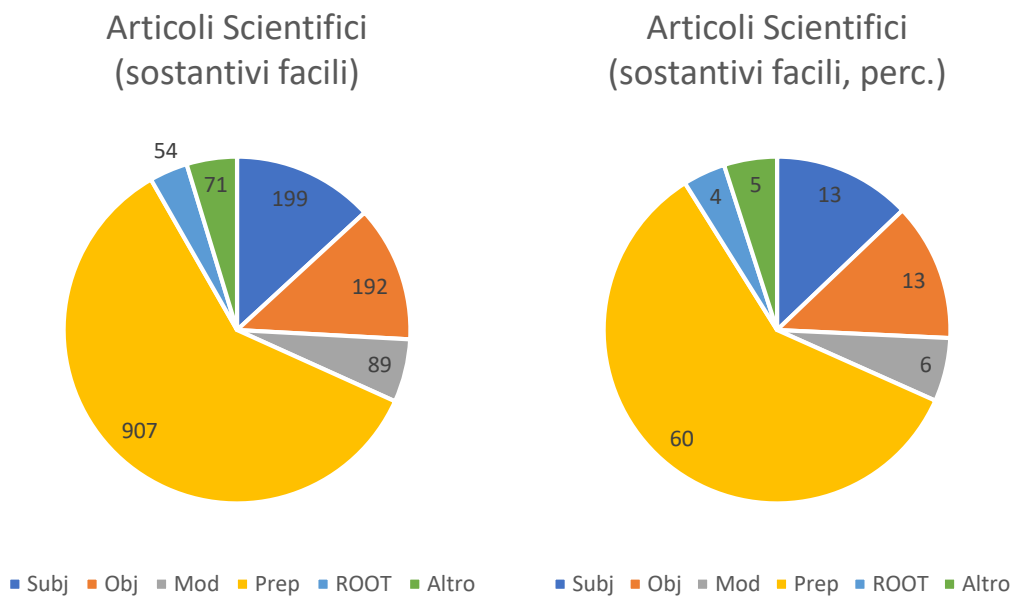

Grafico 5.15: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno di Articoli Scientifici *divisi per tipo di dipendenza.*

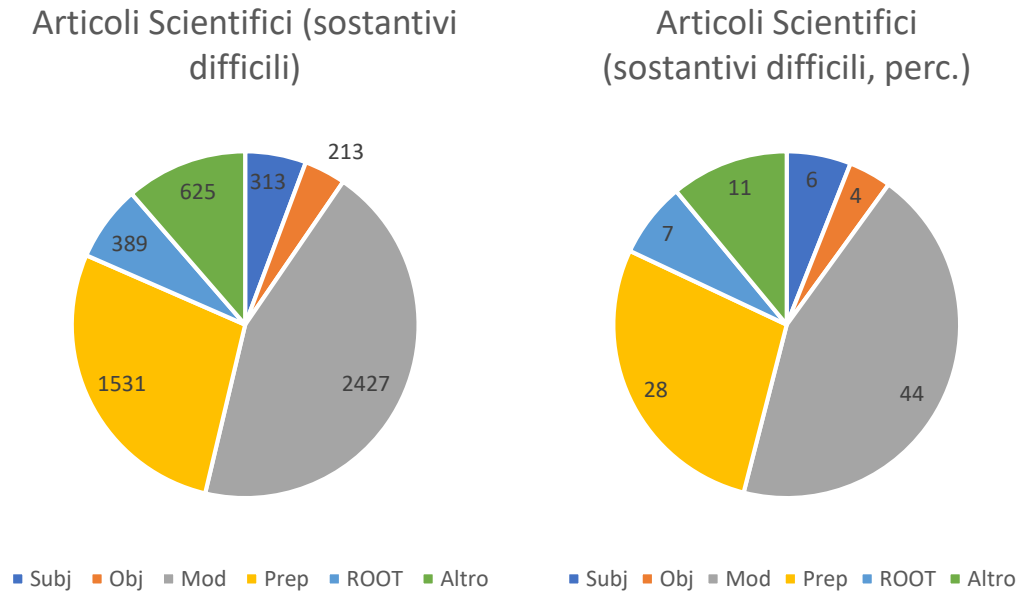

Grafico 5.16: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno di Articoli Scientifici *divisi per tipo di dipendenza.*

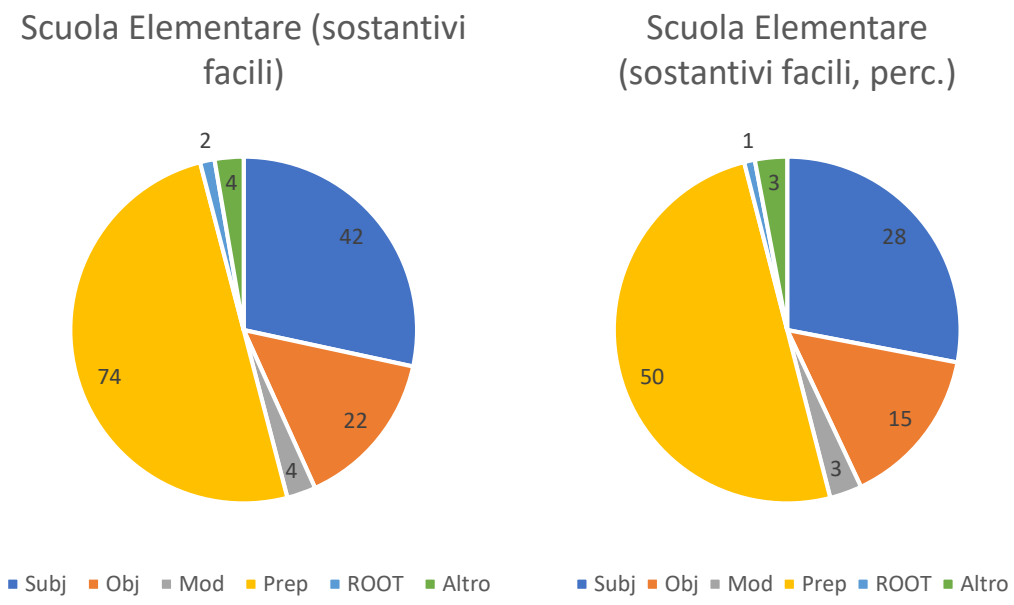

Grafico 5.17: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno dei testi di Scuola Elementare *divisi per tipo di dipendenza.*

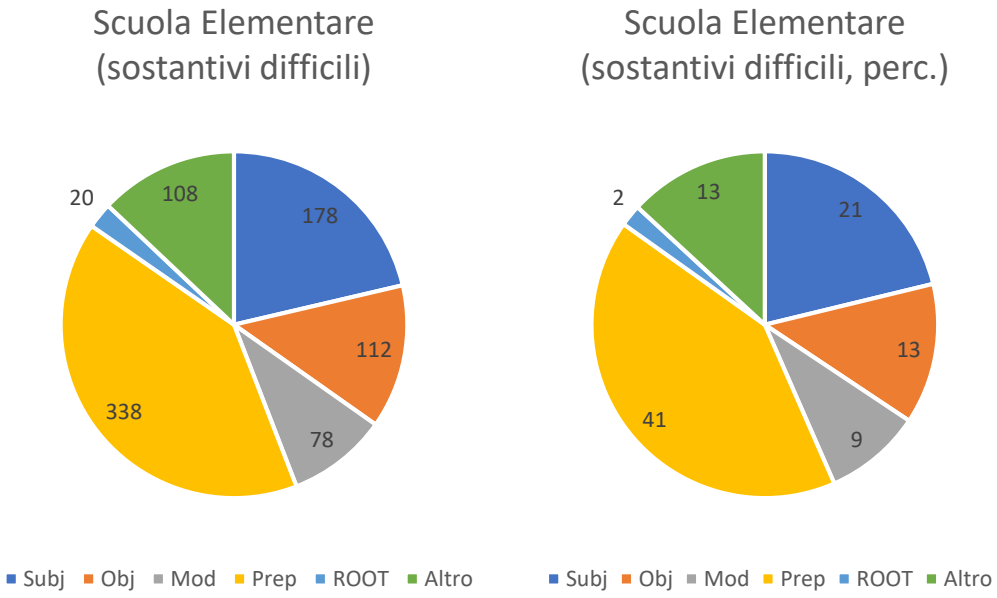

Grafico 5.18: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno dei testi di Scuola Elementare *divisi per tipo di dipendenza.*

Queste variazioni le riscontriamo anche nei sostantivi complessi dell'ultima coppia di corpora, le due collezioni di testi narrativi per l'infanzia originali e semplificati. Anche qui, come è possibile notare mettendo a confronto il grafico 5.22 e il grafico 5.24, le

occorrenze dei sostantivi (difficili) che reggono una preposizione sono in numero maggiore rispetto alle occorrenze dei modificatori; ancora una volta, il numero di radici è molto piccolo rispetto ai soggetti e agli oggetti.

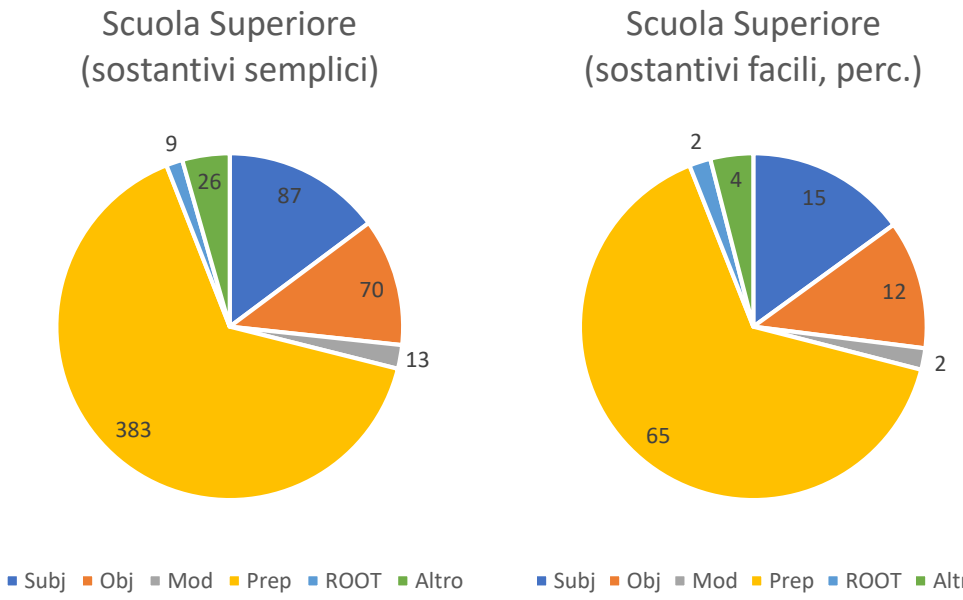

Grafico 5.19: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno dei testi di Scuola Superiore *divisi per tipo di dipendenza.*

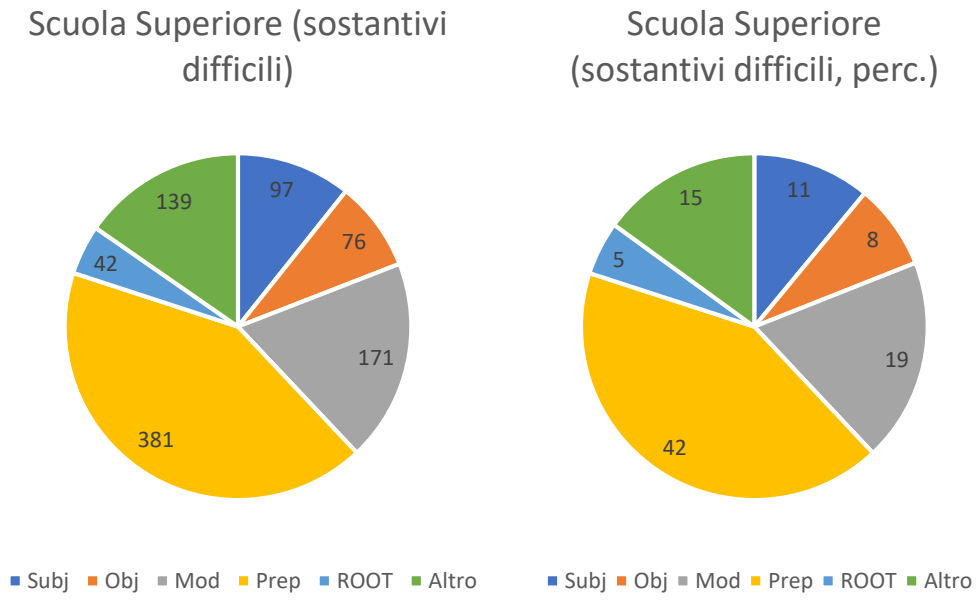

Grafico 5.20: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno dei testi di Scuola Superiore *divisi per tipo di dipendenza.*

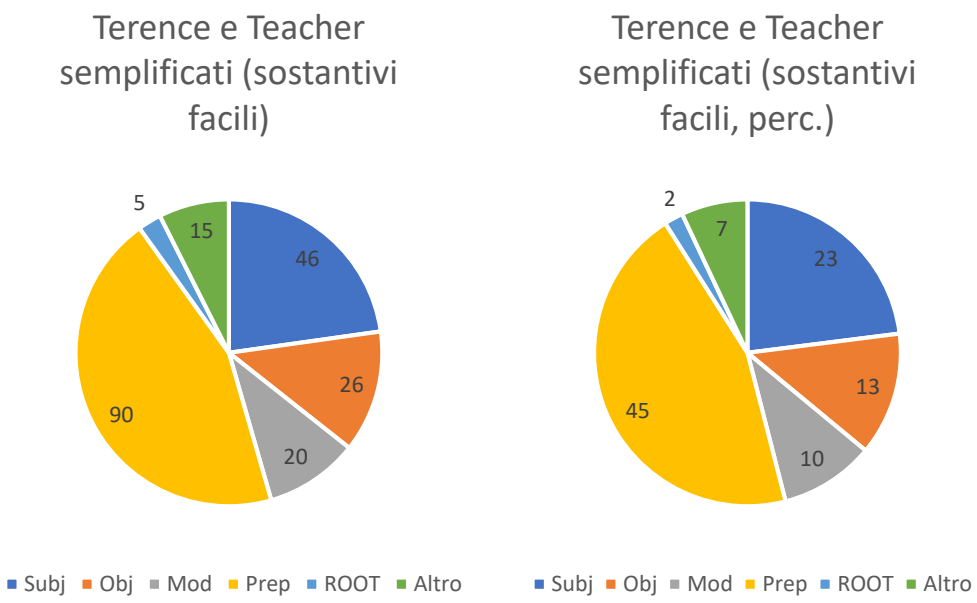

Grafico 5.21: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno dei testi di Terence e Teacher semplificati *divisi per tipo di dipendenza.*

Considerando i sostantivi semplici, sia per i corpora di tipo scolastico che per quelli narrativi per l'infanzia, la situazione, invece, è coerente con tutti gli altri casi analizzati fino a ora. Nel grafico 5.17 abbiamo la distribuzione dei sostantivi facili per il corpus di testi di *Scuola Elementare* e nel grafico 5.19 quella dei sostantivi facili per il corpus di testi di *Scuola Superiore*. I valori sono molto piccoli perché entrambi i corpora non sono di grosse dimensioni. Invece, per quanto riguarda le distribuzioni dei sostantivi facili all'interno della collezione di testi di *Terence* e *Teacher* semplificati bisogna fare riferimento al grafico 5.21, mentre per i testi di *Terence* e *Teacher* originali al grafico 5.23.

Sembra, dunque, che nel confronto dei due corpora dello stesso genere e di complessità differente ci sia effettivamente similarità per quanto riguarda la distribuzione delle occorrenze dei sostantivi facili e difficili. Nei confronti di genere, invece, se consideriamo i sostantivi facili la conclusione è la stessa, mentre per i sostantivi difficili abbiamo alcune variazioni che non ci permettono di considerare uniforme la distribuzione del lessico.

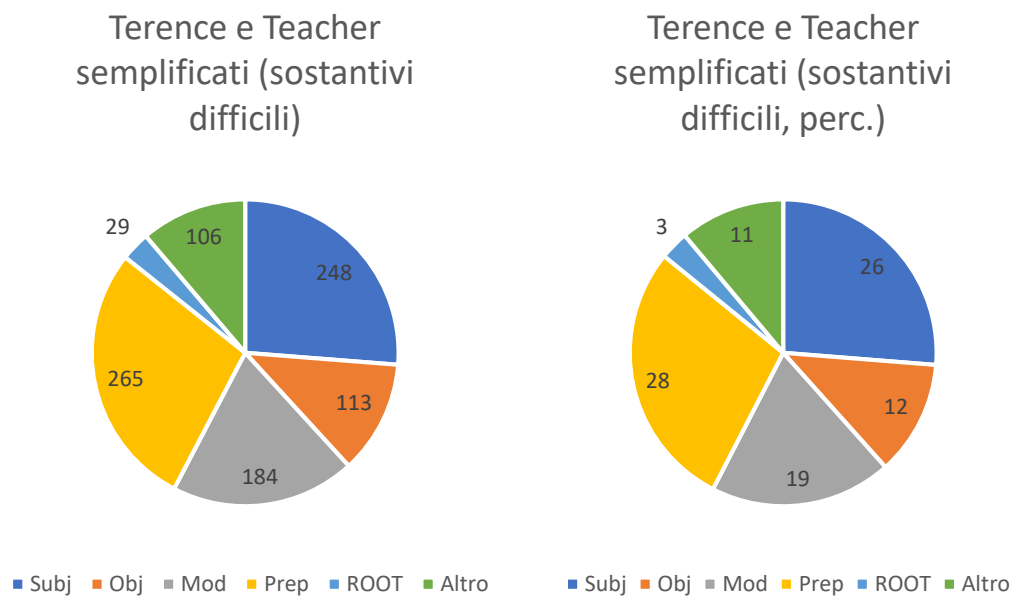

Grafico 5.22: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno dei testi di Terence e Teacher semplificati *divisi per tipo di dipendenza.*

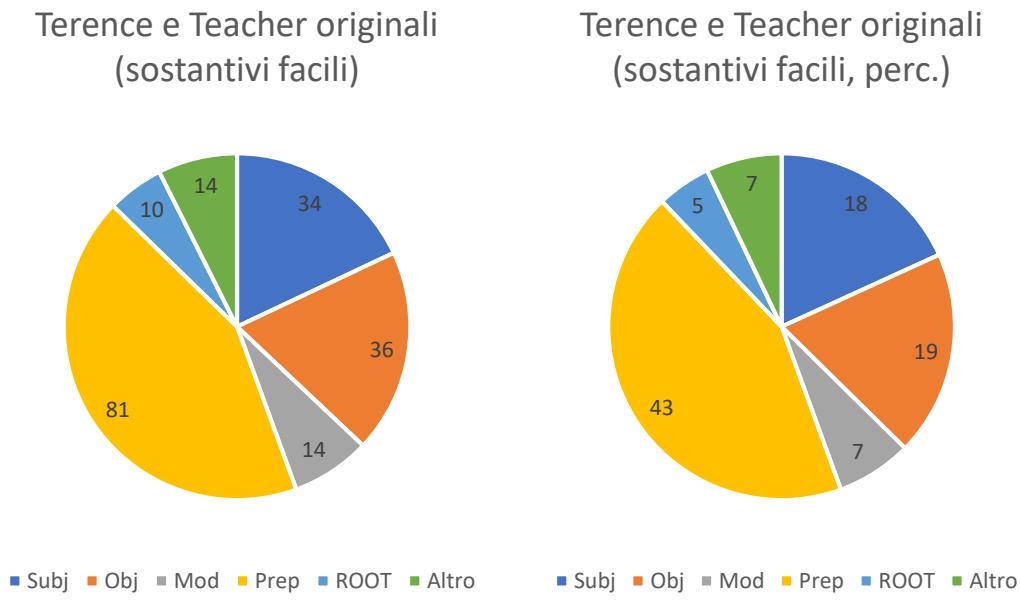

Grafico 5.23: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più facili all'interno dei testi di Terence e Teacher originali *divisi per tipo di dipendenza.*

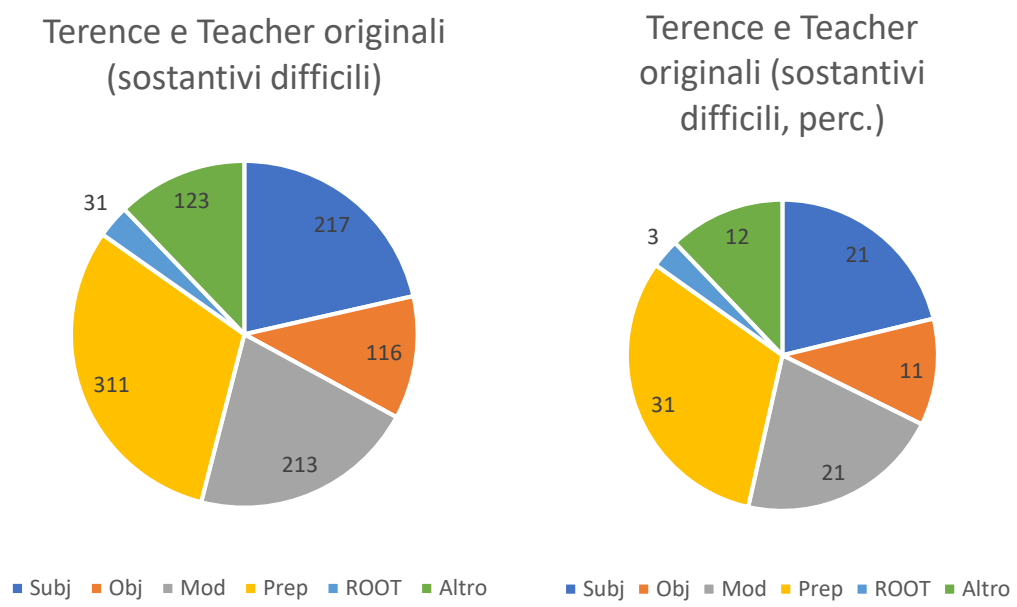

Grafico 5.24: numero di occorrenze (sinistra) e relativa percentuale (destra) dei sostantivi più difficili all'interno dei testi di Terence e Teacher originali *divisi per tipo di dipendenza.*

5.2.2 – Fenomeni sintattici

Come fatto per tutti i sostantivi (si veda 5.1.2), anche in questo caso per ogni occorrenza di quelli facili e difficili di ogni corpus sono state estratte le informazioni che riguardano la distanza del sostantivo dalla sua testa in termini di numero di parole, il numero dei suoi figli e dei suoi fratelli e la sua distanza dalla radice. Nelle tabelle 5.5 e 5.9 abbiamo rispettivamente i dati medi relativi ai sostantivi più difficili e quelli relativi ai sostantivi più facili per tutti e otto i corpora. Consideriamo prima i sostantivi difficili. Notiamo come nei due corpora giornalistici ci sono alcune variazioni, ad esempio, passando da *2Parole* a *Repubblica*, la media della distanza dalla testa del sostantivo difficile occorrente sale da 2,213 a 2,346, cala leggermente nel caso del numero di figli e di fratelli, mentre invece aumenta considerevolmente quando guardiamo la distanza dalla radice. Questi confronti, come possiamo notare nella tabella 5.6, sono tutti significativi, in particolar modo sono *estremamente* significative le variazioni che riguardano il numero di fratelli e la distanza dalla radice. Nei corpora scientifici abbiamo minime variazioni per quanto riguarda la distanza dalla testa (2,825 in *Wikipedia* e 2,496 in *Articoli Scientifici*), il numero di figli e di fratelli (entrambi in

discesa se passiamo dal corpus semplice al corpus complesso) e la distanza dalla radice. In questo caso, se facciamo eccezione per la variazione poco significativa relativa al numero di figli, tutte le altre sono estremamente significative.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole	2,213	0,945	1,79	2,896
Repubblica	2,346	0,886	1,582	4,382
Wikipedia	2,825	1,254	1,954	4,301
Articoli scientifici	2,496	1,14	1,598	4,98
Scuola elementare	2,357	1,259	1,592	3,432
Scuola superiore	2,706	1,268	1,593	4,496
TerenceTeacher semplificato	2,072	1,001	2,017	2,757
TerenceTeacher originale	2,318	1,107	1,902	2,988

Tabella 5.5: media delle informazioni estratte per ogni corpus sulla base dei sostantivi più complessi di ciascuno dei corpora.

Le variazioni che riscontriamo nei due corpora scolastici hanno quasi tutte significatività nulla, tranne per quella relativa alla distanza dalla radice: c'è una crescita estremamente significativa se passiamo dai testi di *Scuola Elementare* (3,432) a quelli di *Scuola Superiore* (4,496). Un discorso simile va fatto per i testi di *Terence* e *Teacher*. Non c'è significatività se consideriamo la distanza dalla testa e il numero di figli e di fratelli; c'è molta significatività invece nella leggera variazione della distanza dalla radice tra i testi semplificati e quelli originali.

È interessante anche operare un confronto tra le medie di questi fenomeni sintattici nei corpora di diverso genere. Partiamo dai corpora semplici. Per la significatività facciamo riferimento alla tabella 5.7 e notiamo come tutte le variazioni relative alla distanza dalla testa del sostantivo non sono significative. Per quanto riguarda il numero di figli notiamo come il valore più basso sia relativo a *2Parole* e abbiamo che due dei tre confronti che riguardano questo corpus sono estremamente significativi (nel particolare, quello con i testi di *Scuola Elementare* e con *Wikipedia*). La media del numero di fratelli varia molto nel confronto tra i testi di *Terence* e *Teacher* semplificati e quelli di *Scuola Elementare*, ed è un valore estremamente significativo; come lo sono anche tutte le variazioni tra i corpora semplici relative alla distanza dalla radice.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole vs Repubblica	✗	✓	☑	☑
Wikipedia vs Articoli scientifici	☑	✗	☑	☑
Scuola elementare vs Scuola superiore	☒	☒	☒	☑
TerenceTeacher semplificato vs originale	☒	☒	☒	✓

Tabella 5.6: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più difficili nei corpora facili e difficili dello stesso genere.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole vs Scuola elementare	☒	☑	☒	☑
2Parole vs TerenceTeacher semplificati	☒	☒	☑	☑
2Parole vs Wikipedia	☒	☑	☒	☑
TerenceTeacher semplificati vs scuola elementare	☒	☑	☑	☑
TerenceTeacher semplificati vs Wikipedia	☒	✗	☑	☑
Wikipedia vs Scuola elementare	☒	✗	✗	☑

Tabella 5.7: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più difficili nei corpora facili di diverso genere.

Passiamo invece a considerare i confronti dei valori relativi ai sostantivi più difficili tra i corpora rappresentanti di un livello di complessità più alto e di genere diverso (tabella 5.8). Per quanto riguarda le medie relative alla distanza dalla testa, i valori sono simili per tutti i corpora complessi, con un picco nella collezione di testi di *Scuola Superiore*. Si tratta di variazioni quasi sempre estremamente significative tranne nel caso del confronto tra *Repubblica* e *Articoli Scientifici* e tra i testi di *Terence* e *Teacher* originali e quelli di *Scuola Superiore*. Per le medie relative al numero di figli abbiamo variazioni estremamente significative nel confronto tra *Repubblica* e *Articoli Scientifici* (si passa da 0,886 a 1,14), in quello tra *Repubblica* e i testi di *Scuola Superiore* (da 0,886 a 1,268) e tra *Repubblica* e i testi di *Terence* e *Teacher* originali (da 0,886 a 1,107); una variazione molto significativa la riscontriamo nel confronto tra *Articoli Scientifici* e i testi di *Scuola Superiore*. Gli altri confronti mostrano significatività nulla.

Se invece guardiamo ai dati relativi al numero di fratelli, abbiamo che le uniche variazioni significative riguardano i confronti tra i testi di *Terence* e *Teacher* con *Repubblica*, con *Articoli Scientifici* e con la collezione di *Scuola Superiore*. Per quanto riguarda i dati relativi al fenomeno della distanza dalla radice abbiamo un unico caso in cui la significatività è nulla e riguarda il confronto operato tra *Repubblica* e i testi di *Scuola Superiore*.

Ne possiamo dedurre che le strutture in cui compaiono i sostantivi più difficili nei corpora complessi variano significativamente ma con qualche eccezione in tutti i fenomeni sintattici analizzati; se, invece, passiamo ai corpora semplici, sistematicamente non c'è significatività nelle variazioni delle distanze medie dei sostantivi dalla testa e, al contrario, tutte le variazioni che riguardano la distanza dalla radice sono estremamente significative.

Ora passiamo a considerare la tabella 5.9 che dà i valori riguardanti le medie delle informazioni estratte per ogni corpus sulla base dei sostantivi più facili di ciascuna collezione. Se mettiamo a confronto le medie della distanza dalla testa nei due corpora giornalistici notiamo una piccola crescita nel passaggio da *2Parole* a *Repubblica* e in maniera simile cresce il valore medio riguardante il numero dei figli; invece decresce il valore medio del numero di fratelli e torna a crescere (in modo più importante) quello

relativo alla distanza dalla radice. In accordo con la tabella 5.10 vediamo che quest'ultima variazione è estremamente significativa, quella relativa al numero di figli è molto significativa, mentre le altre due hanno significatività nulla.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
Articoli scientifici vs Scuola superiore	☑	✓	☒	✓
Repubblica vs Articoli scientifici	☒	☑	☒	☑
Repubblica vs Scuola superiore	☑	☑	☒	☒
Repubblica vs TerenceTeacher originale	☑	☑	☑	☑
TerenceTeacher originale vs Articoli scientifici	☑	☒	☑	☑
TerenceTeacher originale vs Scuola superiore	☒	☒	☑	☑

Tabella 5.8: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più difficili nei corpora difficili di diverso genere.

Nel caso dei corpora scientifici, l'unica variazione (estremamente) significativa, quindi non casuale, è quella relativa al numero di figli, che da *Wikipedia* ad *Articoli Scientifici* cala quasi impercettibilmente da 1,231 a 1,217. Nei testi scolastici abbiamo poca significatività nel caso del numero di fratelli, si passa dalla media di 1,514 per i testi di *Scuola Elementare* a quella di 1,177 per la collezione di testi di *Scuola Superiore*, e una variazione estremamente significativa per quanto riguarda la distanza dalla

radice (una media di 3,608 i primi contro quella di 4,995 dei secondi). Nel caso dei corpora narrativi per l'infanzia, nessuna delle variazioni è significativa.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole	2,04	1,260	1,211	3,437
Repubblica	2,271	1,38	1,156	4,436
Wikipedia	2,32	1,24	1,231	4,767
Articoli scientifici	2,418	1,521	1,217	4,944
Scuola elementare	2,095	1,108	1,514	3,608
Scuola superiore	2,359	1,27	1,177	4,995
TerenceTeacher semplificato	2,183	1,52	1,525	2,965
TerenceTeacher originale	2,254	1,662	1,705	3,59

Tabella 5.9: media delle informazioni estratte per ogni corpus sulla base dei sostantivi più facili di ciascuno dei corpora.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole vs Repubblica	☒	✓	☒	☑
Wikipedia vs Articoli scientifici	☒	☑	☒	☒
Scuola elementare vs Scuola superiore	☒	☒	✗	☑
TerenceTeacher semplificato vs originale	☒	☒	☒	☒

Tabella 5.10: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più facili nei corpora facili e difficili dello stesso genere.

Passiamo ora a operare il confronto anche tra le medie di questi fenomeni sintattici relativi ai sostantivi più facili tra i corpora di diverso genere. Nella tabella 5.11 è mostrato il livello di significatività delle variazioni nei confronti di genere per i corpora

semplici. Notiamo come nel confronto tra *2Parole* e i testi di *Scuola Elementare* non c'è mai significatività, al contrario se confrontiamo il corpus giornalistico con quello con testi di *Terence* e *Teacher* semplificati. Ritorniamo a una situazione di (quasi) totale non significatività se consideriamo le variazioni tra *2Parole* e *Wikipedia*: per questo confronto c'è un'unica variazione estremamente significativa ed è quella relativa alla distanza dalla radice, con una crescita che va da una media di 3,437 per il primo corpus a una media di 4,767 per il secondo. C'è una situazione nella media nel caso del confronto tra i testi di *Terence* e *Teacher* semplificati e quelli di *Scuola Elementare*; infatti per quanto riguarda la distanza dalla testa e il numero di fratelli non riscontriamo alcuna significatività, nel caso del numero di figli c'è una variazione nella media dei due valori estremamente significativa e possiamo ritenere molto significativa la variazione nel caso della distanza dalla radice.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
2Parole vs Scuola elementare	☒	☒	☒	☒
2Parole vs TerenceTeacher semplificati	☑	☑	☑	✓
2Parole vs Wikipedia	☒	☒	☒	☑
TerenceTeacher semplificati vs scuola elementare	☒	☑	☒	✓
TerenceTeacher semplificati vs Wikipedia	☑	☑	☑	☑
Wikipedia vs Scuola elementare	☒	☒	✗	✓

Tabella 5.11: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più facili nei corpora semplici di diverso genere.

Tutte le variazioni nel confronto tra i testi di *Terence* e *Teacher* semplificati e *Wikipedia* sono estremamente significative. Infine, la media della distanza dalla testa e del numero di figli nel confronto tra *Wikipedia* e i testi di *Scuola Elementare* sono variazioni casuali; poca significatività si riscontra nella variazione del numero medio di fratelli e molta significatività in quello della distanza dalla radice.
Nella tabella 5.12 sono esemplificati i livelli di complessità per i sostantivi più facili nel confronto tra i corpora di genere diverso e rappresentanti di un livello di complessità più alto.

	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
Articoli scientifici vs Scuola superiore	☒	☑	☒	☒
Repubblica vs Articoli scientifici	☒	✓	☒	☑
Repubblica vs Scuola superiore	☒	☒	☒	☑
Repubblica vs TerenceTeacher originale	✓	✓	☑	☑
TerenceTeacher originale vs Articoli scientifici	☑	☒	☑	☑
TerenceTeacher originale vs Scuola superiore	✓	☑	☑	☑

Tabella 5.12: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più facili nei corpora complessi di diverso genere.

Notiamo subito come la situazione sia piuttosto variegata anche in questo caso. Risulta evidente come tutti i confronti relativi alla distanza dalla radice siano estremamente

significativi, eccezion fatta per quello tra *Articoli Scientifici* e i testi di *Scuola Superiore*, che ha mediamente poca (se non nulla) significatività. I casi più felici, in questo senso, sono quelli relativi al confronto tra *Repubblica* e i testi di *Terence* e *Teacher* originali, con molta significatività se consideriamo i valori medi della distanza del sostantivo dalla testa e il numero di figli e variazioni estremamente significative, invece, se guardiamo al numero di fratelli e alla distanza dalla radice. Nel confronto tra i testi di *Terence* e *Teacher* originali e *Articoli Scientifici*, tutte le variazioni sono estremamente significative tranne quella relativa al numero di figli che può essere considerata casuale e, infine, nel confronto tra i testi narrativi originali e quelli di *Scuola Superiore* abbiamo livelli di significatività mediamente molto alti per tutte le variazioni.
Da tutto ciò si potrebbe dedurre che le strutture in cui si distribuiscono i sostantivi più facili nei corpora dello stesso genere e di complessità differente non variano significativamente. Abbiamo casi più vari se passiamo al confronto tra corpora di generi differenti e stessa complessità. Un caso che sembra sistematico è quello della distanza dalla radice: il suo valore varia significativamente al variare dei generi in entrambi i livelli di complessità, tranne quando il confronto è operato tra testi di tipo giornalistico e testi di tipo scolastico.

Tutti facili vs tutti difficili	Distanza dalla sua testa	Numero di figli	Numero di fratelli	Distanza dalla radice
Corpora semplici vs corpora semplici	✓	☑	☑	☑
Corpora complessi vs Corpora complessi	☑	☑	☑	✓
Tutti i corpora vs Tutti i corpora	☑	☑	☑	☑

Tabella 5.13: calcolo del p-value per determinare la significatività statistica tra le informazioni estratte dai sostantivi più facili e più difficili in tutti i corpora.

Per concludere questa sezione dedicata ai fenomeni sintattici che co-occorrono con i sostantivi più facili e difficili degli otto corpora utilizzati, è utile operare dei confronti di marcatura più generale. Nella tabella 5.13 sono mostrati i confronti tra tutti i sostantivi facili e tutti i sostantivi difficili, considerati nei casi già esaminati, nei corpora semplici, in quelli difficili e in tutti i corpora come se fossero un unico grande testo. In tutti e tre i casi, mediamente, le variazioni possono essere definite estremamente significative. Questo ci dimostra come, in una visione più ampia, le strutture in cui sono distribuiti i sostantivi variano significativamente sia nei corpora rappresentanti di un livello di complessità basso che in quelli rappresentanti di un livello di complessità più alto e, di conseguenza, vale lo stesso anche se consideriamo tutti i corpora insieme.

5.3 – Lessico e sintassi: brevi considerazioni

I dati statistici analizzati in questo capitolo sono relativi al secondo esperimento descritto, ovvero uno studio su alcuni fenomeni linguistici ricavati dall'analisi sintattica a dipendenze (tipo di dipendenza, distanza dalla testa del singolo lemma, numero di suoi dipendenti, numero di suoi fratelli, distanza dalla radice) che co-occorrono con i sostantivi estratti dai corpora, divisi per genere: giornalistico (*2Parole* e *Repubblica*), scientifico (*Wikipedia* e *Articoli Scientifici*), scolastico (testi di *Scuola Elementare* e testi di *Scuola Superiore*) e narrativo (testi di *Terence* e *Teacher* semplificati e originali).

Considerando tutti i sostantivi per ogni corpus è venuto fuori che le differenze di genere non influenzano di molto le caratteristiche sintattiche esaminate nei corpora; lo fanno invece nelle varietà "semplici". Un esempio è il dato relativo alla distanza dalla radice dei sostantivi in *2Parole* rispetto a quella in *Wikipedia* o nei testi di *Scuola Elementare* e *Terence e Teacher* semplificati.

Invece, considerando i sostantivi più facili e difficili all'interno dei vari corpora, sembra esserci similarità nella distribuzione delle occorrenze nei corpora dello stesso genere e di complessità differente; il confronto di genere, però, ha dato luogo a un'eccezione: considerando i sostantivi facili, la distribuzione non cambia, mentre per quelli difficili a volte sì.

Per quanto riguarda le strutture in cui compaiono i sostantivi più difficili nei corpora confrontati per genere, abbiamo variazioni significative tranne per qualche eccezione

nel caso in cui si confrontino corpora complessi tra di loro; invece, per i sostantivi più facili, nel confronto per complessità riscontriamo variazioni poco significative, mentre nel confronto per generi abbiamo situazioni più varie.

In generale, dunque, e mediamente come ci si aspetterebbe, il lessico è legato alle strutture sintattiche esaminate e varia significativamente al variare di generi e livelli di complessità.

La significatività dei confronti è stata verificata grazie al calcolo del p-value con il Wilcoxon rank-sum test.

CAPITOLO 6

Conclusioni

In questo elaborato è stato condotto uno studio sul lessico rispetto a due dimensioni, il genere testuale e la complessità linguistica, e uno studio sul lessico legato alla sintassi. L'obiettivo che è stato perseguito è quello di capire da un lato se e come possa variare la distribuzione del lessico rispetto alle due dimensioni d'analisi e, dall'altro, come questa distribuzione influenzi le strutture sintattiche e in particolare i fenomeni sintattici quali la distanza di un sostantivo dalla testa, il numero di fratelli e di dipendenti e la distanza dalla radice.

Per prima cosa è stato presentato uno stato dell'arte sulla complessità linguistica, descrivendo le diverse teorie che hanno contribuito ad ampliare il dibattito sul tema; è stata discussa la differenza tra lingua scritta e lingua parlata, con particolare riferimento alle riflessioni prodotte da Voghera e Fiorentino; oggetto di discussione è stata anche la differenza tra complessità nel sistema e complessità per l'utente oltre che la metrica della complessità proposta da McWhorter (1998, 2001).

Successivamente sono stati descritti gli otto corpora utilizzati per portare avanti l'indagine e gli strumenti utili, nonché il Nuovo Vocabolario di Base della Lingua Italiana (NVdB) di Tullio De Mauro, il corpus di riferimento itWaC e il test non parametrico per campioni indipendenti di Wilcoxon (Wilcoxon rank-sum test), che è servito a stabilire il livello di significatività delle variazioni calcolate.

Sono stati fatti due tipi di analisi dei dati estratti. Il primo è uno studio sulla distribuzione del lessico all'interno degli otto corpora considerati, volta a indagare quanto varia la complessità lessicale al variare sia del genere sia del livello di complessità del testo all'interno dello stesso genere. L'ipotesi di partenza prevedeva di trovare variazioni significative al variare del livello di complessità, perché generalmente è lecito attendersi un più alto numero di parole comuni in testi scritti in una varietà di italiano semplificata o, comunque, più semplice di quello presente in testi più settoriali o complessi. È verosimile, ad esempio, che il corpus di testi *2Parole* possa contenere con più frequenza termini semplici e facenti parte del vocabolario di base.

È stato analizzato come variano le tre categorie del Nuovo Vocabolario di Base (NVdB) di Tullio De Mauro, ovvero il vocabolario Fondamentale, quello di Alto Uso

e quello di Alta Disponibilità nel confronto tra generi e complessità. Ciò che ne risulta è che all'interno dei corpora analizzati, indipendentemente dal genere o dal livello di complessità, il vocabolario Fondamentale ricopre la quasi totalità delle forme e dei lemmi; invece il vocabolario di Alta Disponibilità è protagonista di leggere variazioni significative rispetto al genere, mentre quello di Alto Uso riflette variazioni anche al variare della complessità. Sono state inoltre indagate le variazioni di frequenza delle forme e dei lemmi di itWaC all'interno dei corpora esaminati, da un lato considerando il tipo di part of speech, dall'altro escludendo questa informazione. Questo tipo di analisi è stata effettuata anche per singola part of speech. Come prevedibile, l'evidenza più ovvia riscontrata è che nomi e verbi sono le forme più diffuse e che variano significativamente al variare dei generi ma non al variare del livello di complessità.

L'obiettivo dell'indagine era di stabilire se e quanto potesse variare la complessità lessicale al variare sia del genere sia del livello di complessità del testo all'interno dello stesso genere, e alla luce delle analisi fatte possiamo affermare che in generale la complessità lessicale varia, anche se in alcuni casi solo leggermente.

Ad esempio, se prendiamo il corpus *2Parole* rispetto a *Repubblica* (i due rappresentanti del genere giornalistico), abbiamo che le variazioni dei tre repertori lessicali del NVdB nel confronto fra complessità sono tutte significative, anche se minime. Una conclusione simile viene fuori dal confronto tra *2Parole* e i corpora semplici di altro genere.

Il secondo esperimento si è proposto di studiare il lessico nominale, in particolare allo scopo di capire se e come la sua distribuzione sia influenzata da particolari fenomeni sintattici tipicamente correlati alla complessità linguistica quali la distanza di un sostantivo dalla testa, il numero di fratelli e di dipendenti e la distanza dalla radice. Queste caratteristiche sono state estratte sia rispetto a tutti i sostantivi che rispetto ai sostantivi più facili e difficili di ciascuno dei corpora. L'obiettivo era quello di verificare se esistesse una correlazione tra complessità lessicale e sintattica. Ci si aspettava una risposta affermativa, in quanto non è irrealistica l'ipotesi che in testi più semplici si possano riscontrare strutture più corte e meno complesse, quindi con sostantivi mediamente non troppo distanti dalla radice o dalla loro testa. I risultati hanno dimostrato quanto si supponeva, quindi la risposta è sì: *la correlazione tra complessità lessicale e sintattica esiste*. Più nel dettaglio, considerando tutti i sostantivi per ciascun corpus,

è emerso come il numero di occorrenze di sostantivi che hanno una relazione di dipendenza di tipo modificazione con la loro testa (di tipo *Mod* nel tagset utilizzato) è molto più alto rispetto ai soggetti e agli oggetti all'interno dei testi rappresentanti di un livello di complessità più elevato, ad eccezione del corpus di *Terence* e *Teacher* formato dai testi originali e includendo il corpus di *Wikipedia*. Questo perché nonostante il primo sia il rappresentante del livello di complessità alto, all'interno del genere di narrativa, si tratta comunque di testi per l'infanzia, quindi molto semplici; invece, *Wikipedia*, pur rappresentando un livello di complessità basso all'interno del genere di prosa scientifica, rispetto ad altri corpora considerati semplici presenta testi molto più complessi.
Sempre considerando tutti i sostantivi, è emerso anche che le differenze di genere influenzano, sì, le caratteristiche sintattiche esaminate nei corpora complessi, ma non di molto; l'evidenza è molto più marcata, invece, per le varietà semplici.
In generale, nel confronto tra due corpora dello stesso genere e di complessità differente c'è similarità nella distribuzione delle dipendenze sintattiche relative ai sostantivi. Se consideriamo il confronto di genere rispetto solo ai sostantivi più facili e difficili di ciascun corpus, per quelli facili vale la stessa conclusione; invece, per quanto riguarda i sostantivi difficili ci sono variazioni.
Al contrario, considerando i fenomeni sintattici, è evidente, alla luce dell'analisi fatta, come parole semplici e difficili occorrano in strutture linguistiche diverse e che questa correlazione vari al variare dei generi e del livello di complessità del testo.

Bibliografia

Berruto G.; Cerruti M. (2011). La linguistica. Un corso introduttivo. UTET Università, 19 maggio 2011.

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

Chiari I.; De Mauro T. (2014). The New Basic Vocabulary of Italian as a Linguistic Resource.

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

Fiorentino G. (2009). Complessità linguistica e variazione sintattica. Studi Italiani di Linguistica Teorica e Applicata (SILTA), (2), 281–312.

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

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

Pallotti G. (2008). Una nuova misura della complessità linguistica: l'indice di complessità morfologica (ICM). Università di Modena e di Reggio Emilia, 2008.

Terence_Corsortium(2012). Story simplification: Userguide. RestrictedDistribution

Voghera M. (2001). Riflessioni su semplificazione, complessità e modalità di trasmissione: sintassi e semantica. In Scritto e parlato. Metodi, testi e contesti. A cura di Dardano M., Pelo A., Stefinlongo A., pp. 65–78. Aracne, Roma.

Wild Chris (1997). The Wilcoxon Rank-Sum Test, University of Auckland, Department of Statistics.

Sitografia

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

Wikipedia, su Mann-Whitney U test, https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

Wikipedia, su p-value, https://en.wikipedia.org/wiki/P-value

Wikipedia, su Statistical Significance, https://en.wikipedia.org/wiki/Statistical_significance

Ringraziamenti

Questo elaborato rappresenta la chiusura di un percorso, forse il più breve, e l'inizio di una strada ben più lunga e tortuosa. Qualcuno potrebbe affermare che sono arrivato sin qui con le mie sole forze, ma non è così. Questa pagina è la mia argomentazione più forte in tal senso, perché è il posto in cui posso ringraziare tutti coloro i quali mi hanno aiutato, intenzionalmente o senza pensarci, a diventare ciò che sono e a raggiungere un traguardo così importante.

Tu che stai leggendo questa pagina, un sincero ringraziamento va soprattutto a te. Potresti essere un professore che mi ha ispirato con le sue lezioni, un amico che mi è stato vicino o che anche nella distanza si è ricordato di me; potresti essere un parente, un conoscente; potresti essere qualcuno che, in un modo o nell'altro, si sia interessato a questo lavoro.

Ringrazio la mia famiglia, in particolare mia madre e mio padre, Wanda e Antonio, mia sorella Mariarosa e mia zia Sabrina.

Ringrazio i miei colleghi di Informatica Umanistica, Chiara, Serena, Simone e Nicolò, che ormai considero più che amici.

Ringrazio il gruppo di persone con cui passo il tempo a studiare e non solo; li ringrazio perché, consapevoli o meno, vecchi o nuovi arrivati, sono per me diventati una seconda famiglia, mi hanno aiutato e continuano a farlo: Pietro, Fabrizio, Giulia, Ilaria, Luigi, Giacomo e Umberto.

Ringrazio gli amici del Laboratorio Batubanda di Pisa, perché quell'ora e mezza a settimana in compagnia della musica e del ritmo sono stati e continuano a essere essenziali, Marco, Andrea, Delia, Samanta, Hari, Melissa e gli altri: siete troppi per potervi citare tutti.

Ringrazio i miei amici più cari, Gabriele, Damiano e Andrea e tutti i miei amici di Trani, perché soprattutto nella distanza sono riusciti a infondermi forza e motivazione: Mary, Francesca, Ilaria, Sergio, Luca, Michele, Corinne, Gabriel, Ylenia, Arianna, Nicholas, Vladimir e tutti gli altri. In particolar modo, un sentito ringraziamento va a Carmela, perché, così come tante altre cose, senza di lei probabilmente non avrei raggiunto questo traguardo con l'entusiasmo e la determinazione necessari.

Infine, ultimo ma non meno importante, un grazie sentito e sincero va all'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa, ai miei relatori, dott. Felice Dell'Orletta e dott.ssa Dominique Brunato, e alla mia correlatrice dott.ssa Simonetta Montemagni: è merito loro se sono arrivato sino a qui. Li ringrazio per l'opportunità che mi hanno offerto, l'aiuto che mi hanno dato e per l'entusiasmo, la professionalità, l'attenzione e la simpatia con cui hanno seguito me e il mio lavoro.
Ancora grazie.

Appendice

Qui di seguito sono rese disponibili le tabelle con i dati discussi all'interno dell'elaborato.

Forme	2Parole	Repubblica	Wikipedia	Articoli scientifici	Scuola elementare	Scuola superiore	Terence Teacher sempl.	Terence Teacher orig.
Aggettivi	0.638	0.736	0.986	1.025	0.8055	1.031	0.724	0.75
Avverbi	0.232	0.342	0.32	0.287	0.407	0.399	0.44	0.492
Congiunzioni	0.066	0.069	0.086	0.074	0.094	0.103	0.122	0.127
Determinanti	0.135	0.069	0.093	0.104	0.085	0.094	0.098	0.082
Preposizioni	0.36	0.411	0.425	0.436	0.383	0.393	0.32	0.347
Interiezioni	0.002	0.003	0.003	0.009	0.009	0.003	0.007	0.009
Numeri	0.22	0.219	0.159	0.221	0.087	0.068	0.082	0.073
Pronomi	0.088	0.179	0.14	0.131	0.28	0.28	0.26	0.326
Articoli	0.237	0.198	0.207	0.191	0.24	0.202	0.201	0.196
Nomi	3.163	3.036	3.2004	3.195	2.667	2.48	2.611	2.546
Predeterminanti	0.0185	0.007	0.0107	0.007	0.01	0.013	0.022	0.023
Verbi	1.724	1.659	1.388	1.321	1.831	1.557	1.783	1.751
Altro	0.0005	0.0003	0.002	0.019	0	0.0009	0	0

Distribuzioni di frequenza medie delle forme presenti in itWaC suddivise per POS.

Lemmi	2Parole	Repubblica	Wikipedia	Articoli scientifici	Scuola elementare	Scuola superiore	Terence Teacher sempl.	Terence Teacher orig.
Aggettivi	0.557	0.624	0.929	0.973	0.684	0.843	0.515	0.531
Avverbi	0.242	0.348	0.316	0.294	0.402	0.397	0.433	0.466
Congiunzioni	0.094	0.093	0.098	0.09	0.115	0.132	0.159	0.159

Determi-nanti	0.102	0.054	0.071	0.081	0.067	0.076	0.078	0.063
Preposizioni	0.204	0.245	0.246	0.254	0.247	0.223	0.223	0.229
Interiezioni	0.0012	0.0026	0.003	0.0012	0.008	0.002	0.004	0.005
Numeri	0.084	0.13	0.087	0.117	0.055	0.052	0.057	0.061
Pronomi	0.096	0.181	0.138	0.13	0.27	0.278	0.261	0.321
Articoli	0.001	0.00078	0.001	0.001	0.001	0.0007	0.044	0.044
Nomi	2.785	2.6707	2.847	2.811	2.346	2.208	2.268	2.181
Predetermi-nanti	0.0133	0.005	0.008	0.005	0.007	0.009	0.015	0.015
Verbi	0.917	0.903	0.75	0.797	1.072	0.911	1.12	1.189
Altro	0.0005	0.0003	0.002	0.018	0	0.0009	0	0

Distribuzioni di frequenza medie dei lemmi presenti in itWaC suddivise per POS.