UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Fenomeni di complessità sintattica: uno studio linguistico-computazionale sull'ordinamento delle strutture grammaticali all'interno di diverse varietà linguistiche

Candidato: *Giulia Pieri*

Relatore: *Dott. Felice Dell'Orletta*

Correlatore: *Prof. Alessandro Lenci*

Anno Accademico 2014-2015

A nonno

Indice

Capitolo 1

Introduzione

Nel corso degli ultimi due secoli si è dibattuto a lungo circa la nozione di complessità linguistica e la possibilità di individuare una metrica universalmente valida con la quale poter classificare le lingue secondo una scala di complessità.
Nel XIX secolo gli studiosi tentarono di elaborare una classificazione linguistica ricalcando le classificazioni razziali dell'epoca: partendo dal presupposto che la lingua fosse un "organismo" in continua evoluzione, essi cercarono di riprodurre, sulle orme di Schlegel, la presunta inferiorità delle popolazioni che venivano man mano sottomesse dalla colonizzazione europea distinguendo due classi principali: le lingue indoeuropee, coincidenti con le lingue flessive, che costituivano il livello più evoluto del linguaggio umano e quindi erano adatte a elaborare ed esprimere il pensiero complesso, e le lingue non indoeuropee, ovvero le lingue non flessive, considerate primitive e inferiori (Gallissot *e altri*, 2001). Nel corso del XX secolo la prospettiva è cambiata, e si fa avanti l'idea, dal punto di vista linguistico, di ritenere tutte le lingue ugualmente complesse e di conseguenza, dal punto di vista biologico, di considerare gli uomini dotati delle stesse capacità cognitive e di linguaggio, indipendentemente dal luogo d'origine e d'appartenenza razziale.
La teoria della pari complessità delle lingue è stata a lungo obiettata e oggi superata da molti studiosi, i quali hanno riconosciuto che le lingue presentano livelli diversi di complessità, anche se questa diversità è direttamente dipendente dalla lingua madre, e che la semplicità di alcuni livelli comporta la complessità di altri livelli. Nonostante ciò, il dibattito sulla possibilità di adottare una metrica universale rimane acceso: infatti, mentre alcuni linguisti prevedono che sia impossibile attuare una valutazione oggettiva e unica delle lingue (Tavosanis, 2009), altri perseguono l'idea che esistano dei tratti che possono essere valutati come indici di complessità "universalmente validi" (Fiorentino, 2009).
Per quanto riguarda il livello della sintassi, uno degli indicatori riconosciuti che riveste importanza nella complessità è l'ordinamento lineare dei costituenti. In proposito,

esistono due principali linee di pensiero: alcuni studiosi opinano che l'ordine degli elementi linguistici sia determinato dalla struttura dell'informazione (Diessel, 2005), invece altri ritengono che sia causato dalla capacità di elaborare più velocemente l'informazione (Hawkins, 1994).

Il *focus* di questo elaborato sarà lo studio linguistico-computazionale dell'ordinamento dei costituenti in due differenti generi linguistici, narrativo e giornalistico, secondo due gradi di complessità, "semplice" e "complesso". In particolare, per il genere narrativo saranno analizzate due raccolte contenenti testi per bambini in forma originale e semplificata (*"Terence"* e *"Teacher"*), che verranno suddivise in modo da ottenere testi solo originali e testi solo prodotti dai processi di semplificazione, mentre per il genere giornalistico verranno proposte una raccolta di testi giornalistici adeguati a un livello culturale medio-alto (*"La Repubblica"*) e una raccolta di testi di facile lettura appositamente creati per un pubblico adulto con un livello di alfabetizzazione primitivo o con lievi disabilità intellettuali (*"Due Parole"*), che non derivano dunque, come nel primo caso, da un processo di semplificazione.

L'analisi verrà effettuata su due livelli, il genere testuale e il grado di complessità: l'obiettivo è verificare quali ordinamenti dei costituenti dipendono dal genere testuale, quali dal processo di semplificazione, e quali da entrambi. Si prevede di riscontrare in alcuni casi una sostanziale tendenza dell'ordinamento dei costituenti in relazione al genere analizzato, mentre in altri di ritrovare dati statistici simili nel confronto tra testi semplici, così come tra testi complessi, dei due differenti generi: infatti, indipendentemente dal genere, si suppone che i testi semplici si attengano principalmente all'ordinamento canonico degli elementi nella lingua italiana, mentre i testi complessi utilizzino con maggiore frequenza forme di focalizzazione dei costituenti.

Questa indagine linguistico-computazionale sarà resa possibile grazie all'utilizzo di strumenti di annotazione linguistica del testo, ad oggi stato dell'arte, e allo sviluppo di un sistema di estrazione delle informazioni linguistiche, che permetteranno di ricostruire il profilo linguistico di ciascun testo e verificare i casi di ordinamento introdotti nel capitolo 2. Nel dettaglio, in questo capitolo verrà approfondita la nozione di complessità linguistica nei diversi livelli linguistici e verrà effettuata un'analisi dettagliata della complessità a livello sintattico sui casi di ordine non marcato e marcato di soggetto, oggetto, aggettivo, avverbio e subordinata; dopodiché, nel capitolo 3 verranno descritte le quattro raccolte di testi che saranno l'oggetto dell'analisi dell'elaborato, mentre nel capitolo 4 verrà affrontato il tema dell'importanza delle tecnologie linguistico-computazionali, grazie alle quali è stato possibile annotare i testi in maniera automatica e, di conseguenza, monitorare le caratteristiche linguistiche desiderate. In seguito, nel capitolo 5 si analizzeranno i dati statistici prodotti dal monitoraggio dei testi giornalistici e narrativi; infine, nel capitolo 6 verrà discusso il successo dell'obiettivo previsto e verranno dedotte le conclusioni del lavoro presentato.

Capitolo 2

La complessità linguistica

In questo capitolo verrà illustrata come viene affrontata dalla ricerca linguistica moderna la nozione di complessità. Nello specifico, verranno analizzate la concezione della complessità secondo criteri interni alla lingua e in relazione all'utente, e le metriche sulla base delle quali poter misurare la complessità di una lingua nei diversi livelli linguistici. Dopodiché, un'analisi dettagliata della complessità a livello sintattico mostrerà come i casi di ordine non marcato e marcato degli elementi implichino un differente grado di difficoltà di elaborazione dell'informazione da parte del parlante e dell'ascoltatore e, di conseguenza, cosa questo comporti nella pianificazione di un discorso. Infine, nella parte conclusiva verrà affrontato il tema della variazione della complessità in relazione alla differente modalità di trasmissione dell'informazione, scritta e parlata.

2.1 Introduzione alla nozione di complessità linguistica

La nozione di *complessità linguistica* è un tema ancora largamente dibattuto dai linguisti di diversi orientamenti. Essa, infatti, può essere intesa da diversi punti di vista: il punto di vista psicolinguistico basa il concetto di complessità sul costo di processamento dell'informazione, il punto di vista matematico rimanda a una misura della complessità riferita alla teoria dell'informazione, mentre il punto di vista empirico si concentra sulla difficoltà di acquisizione di un linguaggio da parte di un *outsider* (Kusters, 2003).

Nel XIX secolo gli studiosi tentarono di definire la complessità linguistica proponendo una classificazione delle lingue che rispecchiasse le classificazioni razziali dell'epoca: partendo dal presupposto che la lingua fosse un "organismo" in continua evoluzione, essi cercarono di riprodurre, sulle orme di Schlegel, la presunta inferiorità delle popolazioni che venivano man mano sottomesse dalla colonizzazione europea distinguendo due classi principali: le lingue indoeuropee, coincidenti con le lingue flessive, che co-

stituivano il livello più evoluto del linguaggio umano, e le lingue non indoeuropee, ovvero le lingue non flessive, considerate primitive e inferiori (Gallissot *e altri*, 2001). In proposito, Fiorentino (2009) argomenta che "Fin dalla sua comparsa nella storia del pensiero linguistico agli inizi del XIX secolo la nozione di *complessità linguistica* viene messa in relazione con la *complessità del pensiero*": l'autrice spiega che le lingue più complesse, corrispondenti alle lingue indoeuropee flessive, sono state ritenute fin dai primi studi come le più adatte a elaborare ed esprimere il pensiero complesso rispetto alle lingue isolanti o agglutinanti; dunque, la sua supposizione è che la nozione di complessità sembrerebbe potersi associare anche a tratti linguistici determinati.
Nel corso del XX secolo la prospettiva è cambiata, e si fa avanti l'idea, dal punto di vista linguistico, di ritenere tutte le lingue ugualmente complesse e di conseguenza, dal punto di vista biologico, di considerare gli uomini dotati delle stesse capacità cognitive e di linguaggio, indipendentemente dal luogo d'origine e d'appartenenza razziale. La teoria della pari complessità delle lingue è stata a lungo obiettata e oggi superata da molti studiosi, i quali hanno riconosciuto che le lingue presentano livelli diversi di complessità, anche se questa diversità è direttamente dipendente dalla lingua madre, e che la semplicità di alcuni livelli comporta la complessità di altri livelli.
Secondo McWhorter (2001), un linguaggio può definirsi complesso se, comparato ad uno più semplice, contiene maggiori distinzioni fonetiche, morfologiche, sintattiche e semantiche al di là della necessità comunicativa, mentre per Ferguson (1982) la struttura linguistica che risulta più semplice corrisponde a ciò che è più diffuso nelle lingue naturali e che viene immagazzinato durante l'acquisizione del linguaggio. Hawkins (2009), invece, nega che la teoria della complessità linguistica possa essere basata sull'intuizione che più unità strutturali e regole implichino maggiore complessità, poiché spesso la semplificazione di un livello linguistico comporta la complessificazione di un altro. Questo fa sì che vi siano diversi problemi riguardo alla definizione di una metrica universale della complessità nel suo insieme: dunque, Hawkins preferisce risolvere il concetto di complessità nella più ampia teoria dell'efficienza comunicativa.

2.1.1 Complessità nel sistema e complessità per l'utente

Il linguaggio è un'abilità complessa in quanto dipende da sistemi complessi come la mente e il cervello, e tale complessità viene analizzata dai linguisti secondo due modalità, come complessità nel sistema e come complessità per l'utente.

La complessità nel sistema Il linguista compara sistemi linguistici e strutture linguistiche definendo ciò che è complesso sulla base di criteri interni alle lingue.
Secondo Cangelosi e Turner (2002), il sistema linguistico è caratterizzato da una serie

di elementi che interagiscono tra loro in maniera distribuita, autonoma e gerarchica. Infatti, le abilità linguistico-comunicative sono organizzate dal basso verso l'alto, per cui quelle del livello inferiore, come ad esempio le abilità fonetiche, hanno influenza su quelle dei livelli superiori, come il livello semantico-lessicale: il processo di interazione e autoorganizzazione di queste componenti comporta la nascita di strutture linguistiche e comportamenti complessi, come la sintassi e la comunicazione linguistica tra gruppi di individui. La valutazione di tale complessità da parte del linguista verte su alcuni criteri come, per esempio, il numero di regole che servono per produrre un certo output, il numero di eccezioni alle regole, il numero di unità previste in un certo livello linguistico e la mancata trasparenza nella relazione forma-significato.
Inoltre, la nozione di complessità nel sistema viene studiata a livello sociale ed evoluzionista perché si ipotizza che le trasformazioni della complessità di un sistema linguistico, ovvero l'insieme dei processi adattivi, sociali, e neurali che ha portato alla graduale emergenza di facoltà socio-comunicative (basata su abilità vocali o gestuali) fino ad abilità cognitivo-linguistiche sempre più complesse, facciano parte di schemi universali evolutivi che descrivono il modo in cui le lingue possono cambiare diacronicamente la loro complessità strutturale.

La complessità per l'utente A differenza della complessità nel sistema, la complessità per l'utente si misura in base all'efficienza comunicativa tra parlanti: in particolare, la complessità viene a dipendere dalle strategie cognitive messe in atto dall'utente per la produzione o ricezione di un messaggio. Dunque, ciò che viene definito complesso è ciò che è difficile da produrre o comprendere, che richiede più passaggi di elaborazione, dando un carico maggiore per la memoria di lavoro e un impegno cognitivo più costoso. Secondo Hawkins (2009), la comunicazione è efficiente quando un messaggio pianificato dal parlante è trasmesso all'ascoltatore in tempo rapido e con uno sforzo minimo di rielaborazione. Inoltre, egli propone tre princìpi generali responsabili dell'efficienza comunicativa:

- minimizzare i domini: concentrare le sequenze connesse sintatticamente e semanticamente, rendendo i domini più brevi possibili, alleggerisce la memoria di lavoro del processore;
- massimizzare il processamento on-line: il processore umano elabora più velocemente quando le proprietà di un elemento X si trovano tutte insieme e possono essere assegnate all'elemento mano a mano che X viene processato; un'assegnazione successiva, oltre a richiedere un maggiore sforzo, può aumentare il margine di errore durante la fase di processamento;
- minimizzare le forme: l'elaborazione delle forme linguistiche e le proprietà di cui esse sono portatrici richiede uno sforzo che può essere ridotto minimizzando

l'utilizzo di queste forme e sfruttando le informazioni extralinguistiche che sono già attive nella comunicazione, fra cui la frequenza delle parole e le inferenze.

In conclusione, una comunicazione efficiente implica semplicità strutturale e grammaticale, che richiede un impegno cognitivo poco costoso e un processamento rapido da parte degli utenti. Solo in alcuni casi l'efficienza si risolve in una complessità maggiore, e ciò avviene quando fattori addizionali determinano le selezioni strutturali del parlante.

2.2 Metrica della complessità nei diversi livelli della lingua

La nozione di "complessità" richiama il problema di individuare una metrica sulla base della quale misurare essa stessa, al fine di classificare le lingue dotate di minore o maggiore complessità strutturale. Secondo McWhorter (2001), non vi sono metriche concordate convenzionalmente per misurare la complessità nelle grammatiche per diverse motivazioni: partendo dalla consuetudine errata di ritenere che tutte le lingue sono allo stesso modo complesse, egli sostiene che è un compito difficile effettuare una diagnosi onnicomprensiva e completa di ogni lingua naturale per una classificazione precisa su una scala di complessità.
Tuttavia, secondo Fiorentino (2009) egli propone una metrica che può essere ritenuta come "universalmente valida" che si concentra maggiormente sui fenomeni fonologici e morfologici (2.2.1), la cui presenza potrebbe essere indice di un sistema linguistico complesso. L'intuizione guida della metrica di McWorther è che un'area della grammatica è più complessa della stessa area in un'altra grammatica nella misura in cui essa comprende maggiori distinzioni e/o regole di un'altra grammatica.

2.2.1 Metrica fonologica e morfologica

Fonologia Il primo parametro di complessità fonologica secondo McWhorter (2001) riguarda l'*inventario fonemico*: esso è più complesso se ha membri marcati. I fonemi marcati sono quelli incontrati meno frequentemente nelle lingue rispetto ad altri, convenzionalmente ritenuti non marcati: per esempio, le consonanti eiettive, clic e labializzate sono considerate suoni marcati rispetto ai suoni più frequenti e non marcati delle consonanti occlusive, vocali arrotondate posteriori o semivocali. La complessità di un inventario fonemico deriva dal fatto che esso contiene membri marcati oltre a quelli non marcati: infatti, non esistono inventari fonemici con solo suoni marcati, dunque, questi implicano l'esistenza simultanea di quelli non marcati. Un grande in-

ventario fonemico richiede, quindi, il mantenimento di distinzioni intersegmentali[1] più fini.
Il secondo parametro di complessità fonologica è incentrato sui *tonemi*: un sistema tonale è più complesso quando ha più toni perché questa fonologia richiede la padronanza e l'elaborazione di un insieme più ampio di contrasti e il mantenimento di distinzioni intertonali più sottili.

Morfologia La metrica morfologica di McWhorter (2001) si basa sullo studio e la comparazione della complessità della morfologia flessiva rispetto alle altre morfologie. Infatti, la flessione comporta lo sviluppo di processi morfofonologici che costituiscono una componente aggiuntiva della grammatica da imparare; questi stessi processi, a loro volta, causano processi fonetici imprevedibili (ad esempio, il mutamento delle consonanti celtiche o la dieresi delle lingue germaniche). La flessione rende più complessa una grammatica quando si vanno a codificare le distinzioni tra le classi di nomi e di verbi: infatti, a differenza della morfologia di una lingua isolante, alcune flessioni, come la marca di genere e le declinazioni delle classe nominali, non corrispondono a concetti espressi da tutte le grammatiche. Dunque, la flessione può avere ampie ripercussioni nella grammatica poiché, essendo un fattore di complessità, esercita un carico maggiore sulla processabilità.
Kusters (2003) ribadisce che la flessione è un dominio adatto per esaminare la complessità grazie a diversi fattori, fra cui la sua stabilità, la sua varietà interlinguaggio e la sua indipendenza dalla semantica lessicale[2]. In particolare, egli analizza la complessità morfologica studiando e comparando la morfologia flessionale e derivazionale: la morfologia derivazionale è più semplice perché la regola derivazionale in un linguaggio corrisponde solitamente a diverse regole lessicali di un'altra lingua, regolarizza il lessico e semplifica la formazione delle parole, mentre la morfologia flessionale è indice di complessità in quanto una regola flessionale non corrisponde solitamente a un insieme di regole in un'altra lingua e rende il lessico più complesso.

2.2.2 Metrica semantica e pragmatica

Semantica La semantica si può ritenere essere strettamente connessa alla sintassi. Berruto (1990) sostiene che la semplificazione di un livello linguistico può portare complessità ad altri livelli: ad esempio, egli constata che la semplificazione dei tratti sintattici, come l'utilizzo di un ridotto paradigma flessionale verbale, può determinare una complessificazione dei tratti semantici, causando una maggiore polisemia nelle

[1] La fonetica articolatoria "intersegmentale" studia i fenomeni prodotti nel passaggio da una configurazione articolatoria alla successiva, che avviene senza soluzione di continuità.

[2] La *semantica lessicale* è un sottocampo della semantica che studia il significato di espressioni linguistiche a livello di parola o di lessema.

forme del parlato.
Secondo Voghera (2001), esistono tre tratti semantici indici di maggiore complessità del lessico:

- il significato *astratto*, non percepibile fisicamente ma conoscibile soltanto attraverso la mente, è più complesso del significato concreto, accessibile attraverso i nostri cinque sensi;
- la proprietà di poter attribuire più significati ad una stessa parola, definita *polisemia*, è indice di maggior complessità rispetto alla monosemia[3];
- il lessico *funzionale*, composto da parole "vuote" (congiunzioni, articoli, preposizioni..) si definisce più complesso del lessico referenziale, costituito da parole "piene" (nomi, aggettivi, verbi, avverbi..).

Pragmatica La pragmatica è una disciplina che studia come e per quali scopi la lingua viene utilizzata e in che misura soddisfi esigenze e scopi comunicativi: in particolare, si occupa di come il contesto extralinguistico influisca sull'interpretazione dei significati. Affinché la comunicazione tra utenti di una lingua funzioni appropriatamente, gli interlocutori devono essere in possesso non solo di conoscenze relative a fonetica, morfologia, sintassi e lessico di una determinata lingua, ma anche del contesto sociale, ambientale e psicologico entro cui il discorso viene collocato. Secondo Hellö (2005), tali conoscenze, infatti, permettono di gestire adeguatamente precisi fenomeni linguistici di seguito descritti, che altrimenti risulterebbero incomprensibili:

- l'ambiguità di singole parole o di interi enunciati mette l'ascoltatore in dovere di disambiguare tra i diversi significati possibili;
- esprimere un significato non letterale con tono umoristico o sarcastico mette in dubbio l'intenzione comunicativa reale da quella apparente;
- l'eventuale mancanza di riferimenti diretti e specifici al contesto non mette in condizione di capire l'intenzione del parlante;
- la possibilità di emettere messaggi indiretti va intesa come una una richiesta indiretta nascosta dentro le parole da parte del parlante;
- l'utilizzo dei verbi *performativi* (come *giurare, comandare, vietare..*) all'interno degli enunciati vuole determinare direttamente alcuni effetti nella realtà;

[3]In realtà, Voghera (2001) sostiene che "la monosemia non è più semplice della polisemia in assoluto" poiché "la riduzione di materiale sintagmatico e la monosemia possono essere fattori di semplificazione per il produttore, ma un fattore di straordinaria complessità per il ricevente." Dunque, la monosemia è preferibile in testi scritti formalizzati, scientifici e nelle istruzioni, mentre nella trasmissione orale, come meglio descritto in 2.3, prevale la polisemia.

In conclusione, per poter capire il vero significato del messaggio il parlante e l'ascoltatore devono conoscere le *convenzioni comunicative*, ovvero capire il significato letterale, il messaggio indiretto e collegare l'enunciato in una situazione reale. "La somma di competenza pragmatica e di competenza linguistica produce la cosiddetta 'competenza comunicativa"' (Wikipedia, 2015b).

2.2.3 Metrica sintattica: un'analisi specifica della complessità basata sull'ordine degli elementi

> "I messaggi linguistici, a differenza dei messaggi di altri codici naturali, possono presentare un alto grado di elaborazione strutturale; i rapporti fra gli elementi o parti del segno, danno luogo a una fitta trama plurima, percepibile nella sintassi del messaggio. Questa proprietà si può chiamare complessità sintattica." (Berruto, 2004)

Già Ferguson (1982) indicava una serie di tratti sintattici dotati di diverso grado di complessità: la paratassi, l'ordine fisso degli elementi e l'assenza di parole funzionali (copula, preposizioni, pronomi) sono più semplici rispetto alla subordinazione, all'ordine variabile dei costituenti e alla presenza delle parole funzionali.
Secondo McWhorter (2001), la sintassi è più complessa quando richiede il processamento di più regole, come le asimmetrie tra la frase principale e le frasi subordinate (ad esempio il *verb-second (V2) word order*[4] in tedesco), o quando ammette più sistemi esistenti (ad esempio, ergativo/assolutivo e nominativo/accusativo).
Givón (1979) definisce il processo di *sintatticizzazione* come una serie di passaggi che vanno dall'emergere della sintassi alla presenza di strutture sintattiche complesse, seguendo una scala progressiva di complessità: dal tema al soggetto, dalle frasi con tema alle clausole relative, dalla complementazione alla subordinazione dentro un sintagma verbale, la nascita di verbi causativi[5], l'emergere delle costruzioni complesse col genitivo, la nascita delle frasi scisse e, infine, la presenza della flessione.
In conclusione, gli aspetti che hanno rilevanza nella complessità sintattica possono essere riassunti come segue (Wikipedia, 2015a):

- l'ordine lineare degli elementi di una frase, che permette di evitare le possibili ambiguità di significato;
- le relazioni e le dipendenze che vigono fra elementi non contigui;

[4] *Verb-second word order* è una restrizione specifica sul posizionamento del verbo di modo finito all'interno di una data frase o clausola. Il principio V2 richiede che il verbo appaia in seconda posizione di una clausola principale, per cui la prima posizione è occupata da un unico importante costituente che funziona come clausola argomento.

[5] I verbi *causativi*, detti anche fattitivi, esprimono un'azione non compiuta dal soggetto ma fatta compiere ad altri (ad esempio, *addormentare* rispetto a *dormire*).

- il grado di incassatura degli elementi;
- la ricorsività, che conferisce una particolare complessità interna;
- le parti del discorso, che danno informazioni sulla sua strutturazione interna (ad esempio, le congiunzioni coordinanti e subordinanti);
- la discontinuità, ossia la possibilità che elementi o parti strettamente unite semanticamente o sintatticamente non siano linearmente adiacenti.

Origine dell'ordine dei costituenti Soggetto-Verbo-Oggetto Secondo Gell-Mann e Ruhlen (2011), recenti studi svolti nel campo della linguistica comparativa mostrano come tutti (o quasi) i linguaggi umani derivino da un unico linguaggio antenato: questo linguaggio non è da identificarsi, come si pensava, nell'ordine degli elementi Soggetto-Verbo-Oggetto (d'ora in poi *SVO*), ma bensì nell'ordine Soggetto-Oggetto-Verbo (d'ora in poi *SOV*). Un processo lungo e graduale avrebbe portato da un'alta frequenza di ordine SOV e una bassa di ordine SVO ad invertirsi, ricorrendo sempre più a quest'ultimo ordine di parole; questo cambiamento avrebbe portato poi allo sviluppo degli altri ordini più comuni, VSO e VOS.
Alcuni studiosi hanno indagato i motivi che hanno portato l'ordine SVO a essere così comune. Gibson *e altri* (2013) ritiene che la variazione SOV/SVO possa essere spiegata in relazione alla sensibilità degli utenti di una lingua al "rumore" che potrebbe contaminare il segnale linguistico: egli, dunque, sostiene che il linguaggio umano sia un esempio di quello che Shannon chiamava *noisy channel* e che le lingue, di conseguenza, avrebbero elaborato determinate regole, relative all'ordine delle parole, in modo da ridurre al minimo il rischio di errori nella comunicazione. Secondo Shannon, infatti, l'efficacia di una comunicazione può essere compromessa da eventuali "rumori", ovvero fattori del contesto che possono impedire al messaggio di arrivare correttamente al destinatario; essi potrebbero risultare da errori da parte del producente, da interferenze esterne o da errori di comprensione da parte dell'ascoltatore.
Per comprendere al meglio l'ipotesi del canale rumoroso è necessario comparare un evento semanticamente irreversibile con uno reversibile: nel primo caso, esemplificato dalla frase *la ragazza calcia il pallone*, l'agente (la ragazza) agisce sul paziente (il pallone) e può esserci una sola interpretazione corretta semanticamente (un pallone non può calciare un ragazza); dunque, anche se l'ordine utilizzato fosse SOV, *la ragazza il pallone calcia*, il significato della frase non sarebbe ambiguo. Invece, nel caso di un evento semanticamente reversibile, ad esempio *la ragazza colpisce il ragazzo*, utilizzare l'ordine SOV causerebbe un'ambiguità nell'individuare l'agente dell'azione (*la ragazza il ragazzo colpisce*); dunque, è necessario in questo caso prediligere l'ordine SVO per disambiguare l'agente e il paziente e per far sì che il significato della frase

non cambi qualora l'esposizione al rumore eliminasse il soggetto o l'oggetto e si perdesse dunque informazione. In ogni caso, sia che si parli di eventi reversibili o non reversibili, l'ordine SVO ha una migliore possibilità di preservare le informazioni se il canale di comunicazione è rumoroso. Dunque, SOV tende a essere un ordine marcato rispetto all'ordine SVO. Tuttavia, la posizione finale del verbo rimane preferibile quando, sia il parlante di lingua SOV che di lingua SVO, devono esprimere a gesti un evento: questo significa che il linguaggio gestuale segue un ordine delle parole che prescinde da quello utilizzato nella lingua nativa.

Ordine marcato e non marcato degli elementi nella lingua italiana Esistono due principali linee di pensiero riguardo l'ordine dei costituenti in una frase: mentre alcuni studiosi opinano che l'ordine degli elementi linguistici sia determinato dalla struttura dell'informazione (Diessel, 2005), altri ritengono che l'ordine sia causato dalla capacità di elaborare più velocemente l'informazione (Hawkins, 1994).
L'ordine degli elementi viene definito "non marcato" nella lingua italiana quando il posizionamento dei costituenti corrisponde all'ordine basico SVO, come riassunto nella Tabella 2.1. Tuttavia, per esigenze comunicative, spesso si è soliti ricorrere a un cambiamento dell'ordine dei costituenti, allo scopo di focalizzare l'attenzione sull'informazione che ci preme prima comunicare: in questo caso, l'ordine viene definito "marcato".

Funzione grammaticale	Soggetto	Verbo/Predicato verbale	Oggetto
Ruolo semantico	Agente	Azione	Paziente
Struttura informativa	Tema	Rema	
Ruolo pragmatico	Dato	Nuovo	

Tabella 2.1: Ordine non marcato degli elementi nella lingua italiana

Per definire meglio il concetto di marcatezza, Corpina (2009) fa riferimento a tre diversi livelli di analisi: fonologico, sintattico e pragmatico.

- marcatezza fonologica: una frase è marcata dal punto di vista fonologico quando la melodia intonativa ad essa associata non può essere rappresentata come una curva continua, ma presenta interruzioni, pause o picchi intonativi;
- marcatezza sintattica: una frase si dice marcata sintatticamente quando i costituenti che la compongono non occupano le loro posizioni canoniche, ma sono dislocati al fine di focalizzare una particolare informazione. Una frase marcata sintatticamente è generalmente caratterizzata da un'intonazione particolare, dunque marcatezza sintattica e fonologica sono strettamente correlate;

- marcatezza pragmatica: una frase è marcata pragmaticamente quando non si adatta ad un numero molto alto di contesti e di situazioni linguistiche; tipicamente, è una frase in cui l'informazione data precede quella nuova.

Per quanto riguarda la marcatezza sintattica, la lingua italiana presenta molteplici deviazioni rispetto all'ordine canonico (Treccani, 2011): infatti, i sintagmi maggiori hanno una notevole libertà di movimento (anche se minore rispetto ad altre lingue) che permette diverse forme di focalizzazione, mentre è più ridotta la libertà di spostamento dei sintagmi minori e dei loro componenti.

Soggetto Solitamente il soggetto occupa una posizione preverbale e, a differenza degli altri argomenti del verbo, non può essere sottoposto a tematizzazione mediante dislocazione a sinistra sia per l'assenza di forme di clitico soggetto nella lingua italiana, sia perché non è particolarmente necessario focalizzare un costituente che di norma svolge il ruolo di tema[6] (Treccani, 2011). Tuttavia, un forte indizio significativo della libertà di spostamento è la possibilità di omissione del soggetto, così come una certa flessibilità della sua posizione per essere evidenziato come elemento nuovo o inatteso. Ad esempio, la collocazione del soggetto in posizione postverbale si può riscontrare negli *enunciati tetici*, costituiti solamente dal rema che veicola informazione nuova (ad esempio, *"È arrivata Lucia"*), mentre le *frasi scisse* offrono la possibilità di rematizzare il soggetto a inizio frase mediante l'utilizzo del verbo essere in funzione di copula e di tratti prosodici ("*È la polizia che* si occupa di questi problemi").

Oggetto L'accentuata possibilità di manipolazioni sintattiche nella lingua italiana si può considerare un'eredità dal latino classico, caratterizzato da un ordine basico SOV, anche se l'ordine dei costituenti è meno libero di quanto lo sia nelle lingue con casi (Treccani, 2011).
Il complemento oggetto in italiano si può trovare in posizione preverbale o postverbale, a seconda della priorità dell'intenzione comunicativa. Ad esempio, nel caso della *dislocazione a sinistra*, che si verifica quando avviene un movimento di un costituente in posizione di tema, normalmente l'oggetto viene preposto al verbo e ripreso con un pronome clitico (ad esempio, "*la cena, la* prepara Giovanna"), mentre nel caso della *dislocazione a destra* l'oggetto viene collocato a destra della frase, con un ordine marcato rema-tema (ad esempio, "Giovanna non *la* prepara, *la cena*"). Inoltre, un altro caso nella lingua italiana di ordinamento OV, come in latino, è la *rematizzazione a sinistra* (o frase focalizzata), che consiste nello spostamento di un costituente, tipicamente l'oggetto, in posizione preverbale senza clitico di ripresa, come *focus* della frase (ad esempio, "*Te* cercavo").

[6]Esistono casi in cui si può avere una costruzione come "*Marco, lui* ha sempre comprato auto usate", con ripresa del soggetto dislocato da parte di un pronome tonico, ma sono assai sporadici.

Modificatori del nome Si può riscontrare poca flessibilità nella lingua italiana nel caso di alcuni modificatori del nome (Treccani, 2011): infatti, gli articoli, i determinanti, i quantificatori e i numerali si trovano antecedenti alla testa, contraddicendo l'ordine canonico italiano con testa a sinistra. In questo caso, la deviazione rispetto alla testa non permette eccezioni e non può, evidentemente, essere soggetta a condizionamenti di natura pragmatica.
La spiegazione di questo fenomeno si può trovare nella *"Branching direction theory"* (Dryer, 2009): essa prevede che, nelle costruzioni non marcate, i costituenti formati da più parole (dunque con struttura sintattica) tendano a conformarsi in modo quasi categorico alla matrice prevalente nella lingua, mentre i costituenti di tipo puramente lessicale, composti in genere da un'unica parola (quindi privi di struttura sintattica) paiono meno propensi ad adattarsi alla matrice prevalente e, talvolta, a occupare rigidamente una specifica posizione nella struttura di frase. I rispettivi esempi sono, rispetto al sintagma nominale, la frase relativa e il genitivo, modificatori con struttura sintattica che sono dislocati rigorosamente alla sua destra, gli aggettivi, costituenti lessicali privi di struttura sintattica con posizione variabile (descritti nel prossimo paragrafo), e alcuni modificatori lessicali (articoli, determinanti, quantificatori e numerali) che, pur appartenendo a costituenti di tipo lessicale privi di struttura sintattica, occupano invece una posizione rigida iniziale all'interno del sintagma. Secondo la teoria, la ramificazione dovrebbe avvenire sempre nella stessa direzione, o a destra o a sinistra della testa; invece, gli elementi che non ramificano, come gli ultimi modificatori analizzati, sono esenti da questa restrizione.

Aggettivi L'ordine tra l'aggettivo e il sostantivo non è fisso: solitamente, la tendenza è quella di porre l'aggettivo dopo il nome se l'intento è di attribuirgli una funzione *restrittiva*, ovvero se indica una qualità distintiva del soggetto rispetto ad altri della categoria di appartenenza (una casa *bella*), mentre assume una funzione *descrittiva*, cioè fornisce un dato oggettivo caratterizzante il nome a cui si riferisce, se viene preposto al nome (una *bella* casa). Generalmente, la posizione *non marcata* dell'aggettivo è dopo il nome cui si riferisce perché, quando un aggettivo qualificativo precede il nome, esso indica di solito una maggiore soggettività di giudizio in chi parla o scrive, una particolare enfasi emotiva o ricercatezza stilistica (Treccani, 2010a).
Tuttavia, alcune categorie di aggettivi hanno un ordine fisso: gli aggettivi alterati (una casa *piccolina*), che reggono un complemento (una casa *piena* di mobili), che derivano da un participio presente o passato (un edificio *ristrutturato*) e che indicano colore (la macchina *rossa*), forma (la palla *rotonda*) o nazionalità (il film *italiano*) seguono sempre il nome, mentre gli aggettivi possessivi (il *suo* cane) e gli aggettivi usati in senso figurato (un *alto* magistrato) sono posti prima del nome (ad eccezione, per i possessivi, di una dislocazione post-nominale per motivi di focalizzazione).

Avverbi La posizione degli avverbi rispetto al verbo è variabile. Infatti, mentre gli avverbi di modo possono essere collocati in qualunque posizione senza alterare il significato della frase (Andavo *velocemente* / *Velocemente* andavo), in altri casi la posizione dell'avverbio segue alcune regole che dipendono dal tipo di elemento a cui si riferisce: ad esempio, quando il verbo è coniugato in un tempo composto, l'avverbio si colloca dopo il verbo (Nadia ha lavorato *duramente*), mentre alcuni avverbi di tempo (*ancora, appena, finalmente, già, mai, sempre, spesso, subito, talvolta*) e di giudizio (*certamente, forse, neanche, nemmeno, neppure, probabilmente, proprio, sicuramente*) possono essere collocati tra l'ausiliare e il participio passato (Non sono *mai* andato a Roma). Inoltre, se il verbo è accompagnato da complementi, l'avverbio può variare la sua posizione mantenendo il suo significato, collocandosi subito dopo il verbo (Maria parla *fluentemente* l'italiano), oppure in fondo alla frase (Maria parla l'italiano *fluentemente*).
In alcuni casi, è possibile osservare come gli avverbi ricevano una particolare interpretazione a seconda della posizione che occupano rispetto agli altri elementi della frase: ad esempio, nelle frasi "Ho risposto *semplicemente*" e "Ho *semplicemente* risposto", lo stesso avverbio in posizione postverbale modifica il predicato con valore modale (primo caso), mentre in posizione preverbale viene usato come avverbio di tipo limitativo (secondo caso).
Gli avverbi *focalizzatori* sono specializzati nel modificare l'elemento della frase maggiormente saliente e informativo, ovvero il focus: essi sono *anche*, *solo*, *perfino*, *soprattutto*, *specialmente*, *proprio*, *mica*, *affatto*. Ad esempio, nella seguente sequenza di frasi l'avverbio *solo*, a seconda della collocazione, conferisce alla frase un diverso significato (Treccani, 2010b):
- *Solo* Marco ha giocato a calcio con Luca (e non Paolo);
- Marco ha *solo* giocato a calcio con Luca (e non ha fatto altro);
- Marco ha giocato *solo* a calcio con Luca (e non a tennis);
- Marco ha giocato a calcio *solo* con Luca (e non con Paolo).

Il caso delle subordinate In questa sezione verranno esaminate le posizioni che le subordinate possono assumere rispetto alla clausola principale e quali sono i criteri che causano i diversi ordinamenti.
Diessel (2005) basa i suoi i studi su corpora in lingua inglese parlata e scritta e mostra come la posizione iniziale o finale delle subordinate rispetto alla principale dipenda da fattori funzionali e cognitivi in competizione: infatti, se da un lato le subordinate posposte alla principale si rivelano più velocemente processabili, dall'altro risultano più efficienti se preposte alla principale, grazie a un' efficacia comunicativa maggiore. In proposito, vi sono tre linee di pensiero:

- la prima teoria, strettamente correlata alla funzione pragmatica, predilige la di-

slocazione a sinistra delle subordinate rispetto alla principale. Alcuni studiosi ritengono, infatti, che l'ordine delle subordinate dipenda dalla struttura dell'informazione: anteporre le subordinate significa rendere note delle informazioni già conosciute all'ascoltatore necessarie per introdurre l'argomento nuovo che si sta per affrontare;

- la seconda linea di pensiero invece favorisce l'ordinamento posposto delle subordinate rispetto alla principale in quanto più facilmente processabile. Hawkins (1994) sostiene che la priorità della funzione pragmatica subentra solo se due alternative di ordine sono ugualmente difficili da processare;

- in linea con Wasow (2002), Diessel propone una terza teoria che favorisce la priorità di entrambi i fattori, mostrando come l'ordinamento sia determinato dall'interazione fra *processing*, pragmatica e semantica.

La seconda linea di pensiero può essere esemplificata dalla *Performance theory of order and constituency* di Hawkins (1994), con la quale sostiene che l'ordine lineare è funzionale al più veloce ed economico riconoscimento della struttura in costituenti di una frase da parte del *parser*. In una lingua SVO, solitamente, sono le subordinate a seguire la frase reggente poiché questo ordine risulta più semplice per i princìpi di processamento e pianificazione:

- minor processamento da parte del *parser*: il processore non sa di avere di fronte una frase complessa, ma una volta raggiunta la congiunzione subordinante sarà in grado di individuare il nodo della frase complessa senza avere carico sospeso nella memoria, in quanto la reggente è già stata analizzata. Nel caso opposto, invece, il *parser* riconoscerebbe subito che si trova davanti ad una frase complessa, ma non saprebbe quando la subordinata ha fine e dunque dovrebbe processare l'intera frase con a carico nella memoria la subordinata;

- minimizzare i domini: il processore preferisce ordini di sintagmi che realizzino il più breve dominio di riconoscimento del costituente; questo spiega come, nel caso delle subordinate preposte alla clausola principale, esse siano più brevi rispetto a quelle che la seguono;

- minor complessità per la pianificazione del discorso: anteporre la subordinata rispetto alla principale comporta uno sforzo maggiore da parte del parlante poiché implica che egli abbia già in mente l'intera frase complessa, e richiede maggiore impegno cognitivo anche da parte dell'ascoltatore.

Tuttavia, in accordo con la prima linea di pensiero, è necessario tenere conto di alcuni princìpi semantici e pragmatici (Diessel, 2005), per cui, ad esempio:

- la subordinata costituisce lo sfondo tematico dell'evento principale,conferendo così la funzione di orientamento, collegamento tematico e introduzione per la reggente;

- l'ordinamento subordinata+principale è preferibile se la subordinata corrisponde ad una clausola condizionale, in quanto viene dichiarata la condizione affinché l'evento della principale si possa realizzare;

- le subordinate temporali precedono solitamente la principale per un principio di tipo "iconico": infatti, si trovano solitamente in posizione antecedente alla reggente sia che designino un'azione precedente all'avvenimento principale, sia che esprimano un evento successivo alla principale.

Per quanto riguarda la lingua italiana, Fiorentino (2009) ha analizzato 2200 clausole avverbiali di modo finito appartenenti a cinque aree semantiche diverse (temporali, causali, condizionali, finali e concessive) che componevano un corpus di italiano scritto elettronico e ha rivelato che l'italiano, come lingua VO, ammette sia subordinate preposte alla reggente che posposte ad essa, anche se l'ordine posposto ha prevalso largamente (circa il 67 %). In particolare, le finali rispettano nel 97.7% dei casi questo ordine, seguono le condizionali (68.6%), poi le temporali (67.1%), ancora le concessive (63.5%) ed infine le causali (58.2%). Mentre le clausole finali sono le uniche con posizione fissa (posposte alla reggente), nelle altre clausole si osserva una variazione della posizione a seconda delle congiunzioni utilizzate. Nel dettaglio, possono essere fatte le seguenti affermazioni:

- le clausole con dislocazione quasi fissa corrispondono alle causali introdotte da *siccome* (prima della reggente) e *perché* e *poiché* (dopo la reggente), alla finale introdotta da *affinché* e alla concessiva introdotta da *purché* (sempre posposta alla principale);

- le clausole che ammettono una posizione variabile sono introdotte dalle congiunzioni *sebbene*, *quando*, *anche se*, *benché*, *qualora*, *mentre*, *appena*, *se*;

- le proposizioni condizionali seguono la principale se sono introdotte da *qualora* e *purché*, mentre con *se* è più probabile che ricorrano prima della reggente. Quindi, in antitesi con i princìpi pragmatici, l'ordinamento essenziale della subordinata condizionale preposto alla principale non viene seguito letteralmente, poiché, in ogni caso, la condizione è logicamente e semanticamente precedente rispetto al contenuto della reggente;

- le clausole temporali seguono la principale, soprattutto se introdotte da *finché* e *mentre*, mentre la precedono se introdotte dalla congiunzione *appena*;

- le clausole concessive seguono perlopiù la reggente, soprattutto se introdotte da *anche se* e *benché*, mentre con *sebbene* ci sono altrettante possibilità di una posizione preposta;

- le proposizioni causali solitamente seguono la reggente, ma in un numero altrettanto ampio di casi la antecedono. In particolare, se le subordinate sono introdotte da *perché* e *poiché* seguono la reggente, mentre con *siccome* la posizione è fissa prima della reggente.

In accordo con Diessel, Fiorentino arriva alla conclusione che la lunghezza delle clausole a volte vìola il principio di minimizzazione dei domini, e questo implica una maggiore complessità cognitiva che viene però superata da motivazioni di natura pragmatica e semantica.
La terza linea di pensiero condotta da Diessel (2005) riguardo l'interazione fra diverse forze all'origine dell'ordinamento di un costituente per la lingua inglese, dunque, può essere ritenuta valida anche per le altre lingue. Questo, però, non significa che tutte le lingue seguono il modello distribuzionale inglese o italiano, che si avvale di subordinate preposte e posposte alla principale: esistono lingue, come ad esempio il giapponese, che tendono a porre la subordinata in posizione precedente alla principale. Il motivo principale può essere rivisto nella teoria di processamento di Hawkins[7]: Diessel (2005) ha appurato, infatti, che tutte le lingue in cui le subordinate tendono a precedere la principale sono lingue con ramificazione a sinistra, suggerendo dunque che la ramificazione a sinistra implica una differente strategia di processamento rispetto alle lingue con ramificazione a destra. Nelle lingue in cui le subordinate sono preposte alla reggente, il nodo madre della subordinata (ovvero la congiunzione o l'affisso subordinante) compare sempre in posizione finale: dunque, è deducibile che se per la lingua inglese, data la congiunzione iniziale della frase, la posizione della subordinata può essere preposta o posposta, per le lingue con ramificazione a sinistra (come nel caso del Giapponese) la congiunzione finale implica che la posizione della subordinata sia preposta alla reggente. Infatti, la posizione preposta della clausola e la congiunzione finale sono fattori ottimali che permettono un più veloce riconoscimento del dominio poiché, in questo caso, le subordinate sono più facili da processare.

2.3 La complessità nella lingua scritta e parlata

Secondo Voghera (2001), "la valutazione della semplicità o complessità della struttura sintattica non può avvenire *ceteris paribus*, ma deve necessariamente tener conto dell'insieme dei vincoli imposti dalla modalità di trasmissione usata.". La studiosa riflette sul fatto che non è possibile valutare *in loco* la complessità di una struttura impiegata

[7]La teoria è descritta in 2.1.1 paragrafo "La complessità per l'utente"

in un testo parlato e in un testo scritto poiché, essendo testi prodotti con modalità di trasmissione differenti, è impossibile ritrovare le stesse condizioni enunciative. Infatti, a ogni condizione concorrono diversi fattori che devono essere tenuti in considerazione, quali i diversi meccanismi di elaborazione e ricezione connessi all'uso del sistema fonico-uditivo e grafico-visivo, la diversa pianificabilità consentita nelle due modalità, il diverso impegno cognitivo e il differente ruolo del ricevente.
Si ritiene che le differenze nello scritto e nel parlato risiedano principalmente nella sintassi e nel lessico:

- **pianificazione dell'organizzazione del discorso e impacchettamento dell'informazione:** scrivere è un processo lento, ponderato e modificabile, che porta a una struttura delle frasi altamente complessa e che richiede tempo per essere scritto, letto e interpretato. Parlare, invece, è un processo spontaneo, non pianificato e realizzato tramite l'utilizzo di frasi corte e indipendenti: essendo prodotto "al volo", è anche maggiormente disorganizzato, con false partenze, riformulazioni, ripetizioni e correzioni. Chafe (1985, citato in Radić-bojanić (2006)) introduce a proposito il concetto di *unità ideazionale*, ovvero un' unità che contiene tutte le informazioni che il parlante può focalizzare e gestire in un unico tempo. L'informazione frammentata, tipica del parlato, include un insieme di unità ideazionali brevi e senza connettivi, mentre l'informazione integrata, tipica dello scritto, comporta una catena più complessa e ricca di unità ideazionali lunghe;

- **scelta sintattica e grado di incassatura di clausole dipendenti:** il parlato è caratterizzato da una sintassi ridotta, che si avvale di un numero esiguo di forme disponibili nei paradigmi e di marche più semplici e generiche. Tuttavia, anche la subordinazione, presente in modo consistente nello scritto, viene utilizzata in misura minore anche nel parlato. Infatti, ciò che costituisce un forte elemento di complessità non è la presenza della subordinata in sé, ma la combinazione tra subordinazione e altri fattori, fra cui l'ordine rispetto alla principale, la sequenza di clausole rispetto all'ordine temporale degli eventi e il grado di incassatura (Voghera, 2001). Nel parlato, ad esempio, si è soliti utilizzare una sintassi a concatenazione lineare di brevi clausole verbali e connetterle tramite congiunzioni subordinanti ad alta frequenza (ad esempio il *che*), invece, nello scritto, il grado di incassatura è solitamente elevato e si tende a utilizzare congiunzioni subordinanti differenti e più ricercate;

- **specificità del lessico:** nella trasmissione orale la scelta dei lessemi è soggetta alla rapidità della pianificazione, dunque, il lessico risulterà generico e di alta frequenza; per lo scritto, invece, si può attingere anche a parole meno abituali e più sofisticate. Per entrambi i casi si preferisce utilizzare parole che abbiano

più accezioni, ovvero che coprono uno spazio semantico ampio, in modo tale da avere una maggiore possibilità d'uso in un numero ampio di contesti (Voghera, 2001). In particolare, la scelta di parole polisemiche si può riscontrare maggiormente nel parlato perché le parole frequenti ad alta flessibilità funzionale permettono costruzioni sintattiche aperte che sopportano bene il peso di interruzioni, ripetizioni ed incisi, assai frequenti durante una conversazione.

Queste differenze dimostrano come il parlato, in molti aspetti, sia più semplice dello scritto. Il fatto però che il parlato sia più semplice e che includa solo un sottoinsieme rispetto al sistema complessivo in una data lingua, non implica che esso vada ritenuto una varietà semplificata dello scritto: infatti, la semplicità del parlato è messa in relazione con fattori come la pianificazione a breve raggio, l'egocentrismo, l'emotività e la dipendenza del discorso dal contesto extralinguistico (Voghera, 2001).
Tuttavia, la trasmissione orale ha bisogno di maggior ridondanza poiché è più esposta al rumore rispetto allo scritto e la non-permanenza delle fonìe, rispetto ai segni grafici, comporta una certa ripetizione al fine di poter reperire più facilmente le informazioni e "fissare" i concetti. Inoltre, come già spiegato in 2.2.2 (paragrafo "Semantica"), la semplificazione di un livello può portare alla complessità di altri livelli: per esempio, la riduzione del materiale sintagmatico, considerato un elemento di semplificazione da un lato, può avere l'effetto opposto nella produzione di testi molto densi lessicalmente (indice di complessità semantica) e a bassa ridondanza, di forte complicazione per il ricevente; per giunta, la pesantezza lessicale non implica necessariamente una maggiore leggerezza sintattica.
In conclusione, è difficile trovare un "compromesso linguistico" che risulti semplice per entrambe le modalità di trasmissione in quanto è chiaro che una stessa struttura può assumere un grado di difficoltà differente a seconda del tipo di testo in cui occorre e a seconda che si consideri il punto di vista del produttore o del ricevente (Voghera, 2001).

Capitolo 3

I corpora analizzati

In questo capitolo verranno descritte due macro-raccolte di testi (d'ora in poi denominate anche *corpora*) appartenenti a due differenti generi testuali, narrativo e giornalistico. Per ciascuna macro-raccolta, sono state selezionate due collezioni di testi che esemplificano due diversi gradi di complessità del linguaggio di ciascun genere rappresentato, in accordo al tipo di destinatario previsto; ogni macro-raccolta, dunque, è composta da una collezione di testi "complessi" e una di testi "semplificati".
In particolare, per ogni collezione verranno illustrati l'origine, la dimensione, il pubblico al quale è indirizzata e, nel caso dei testi semplici, le strategie di semplificazione. Nel quinto capitolo, i testi saranno oggetto di analisi e confronto per studiare la similarità o la diversità degli ordinamenti dei costituenti sia nel processo di semplificazione che nella variazione di genere.

3.1 Introduzione al corpus

Un *corpus* è una collezione sistematica di testi utilizzata per ricavare informazioni sul linguaggio. Creare e progettare un corpus, dunque, significa selezionare e organizzare i testi secondo precisi criteri, allo scopo di fornire un campione rappresentativo delle varietà e tendenze linguistiche di una specifica popolazione linguistica che si intende analizzare.
Anche se "corpus" può riferirsi a qualsiasi raccolta strutturata di testi, oggi il termine è spesso utilizzato solo in riferimento a raccolte che sono state digitalizzate (Nesselhauf, 2005). Infatti, l'avvento dell'era informatica ha rivoluzionato anche l'uso stesso dei corpora, in quanto il computer oggi permette di immagazzinare quantità di dati testuali prima inimmaginabili, creando numerosi vantaggi fra cui un notevole miglioramento dell'accessibilità ai dati. Inoltre, l'emergere dei linguaggi standard di marcatura del

testo, ad esempio XML [1], ha reso possibile l'annotazione linguistica[2] su ampia scala: un testo riccamente annotato costituisce un ottimo filtro di ricerca che permette a strumenti informatici specializzati di interrogare in maniera avanzata e rapida il contenuto del corpus.
Esistono vari tipi di corpora e possono essere classificati secondo diversi criteri, fra cui i più importanti sono (Lenci *e altri*, 2005):

- Grado di generalità: i corpora si definiscono generali o specializzati secondo il grado di specificità con cui viene descritta una lingua, la quale può essere studiata nel suo quadro complessivo o come rappresentante di una particolare varietà linguistica;
- Cronologia: i corpora diacronici raccolgono testi appartenenti a periodi diversi, mentre i corpora sincronici includono testi che appartengono a una stessa finestra temporale;
- Lingua: si distinguono i corpora monolingua, contenenti testi di una sola lingua, dai corpora multilingua, che comprendono testi di almeno due lingue differenti. Questi ultimi si differenziano a loro volta in:
  - paralleli: comprendono testi sia nella loro lingua originaria (L1) sia in traduzione in un'altra lingua (L2). In particolare, se ciascuna frase della lingua L1 è esplicitamente collegata col suo traducente nella lingua L2 si parla di corpora paralleli "allineati";
  - comparabili: contengono testi originali in lingue diverse;
- Modalità: i corpora possono essere di lingua parlata, scritta o misti a seconda della modalità di produzione;
- Integrità dei testi: i corpora possono contenere testi interi oppure porzioni di testi di lunghezza prefissata;
- Codifica dei testi:
  - Corpora codificati: i testi sono arricchiti con etichette (codici) che rendono espliciti vari tipi di informazione come, ad esempio, la struttura testuale e la composizione;
  - Corpora annotati: le informazioni codificate sul testo riguardano la struttura linguistica del testo a livelli diversi di rappresentazione (morfologica, sintattica, semantica..).

[1] Extensible Markup Language (XML) è un linguaggio di markup che definisce un insieme di regole per codificare i testi in un formato leggibile per l'uomo e la macchina.

[2] "Annotare" un testo significa rendere esplicite informazioni (meta-)linguistiche mediante l'attribuzione di una etichetta a una porzione specifica del testo.

I corpora narrativi e i corpora giornalistici analizzati nelle seguenti sezioni sono esempi di corpora specializzati e monolingua, in quanto rappresentano entrambi una particolare tipologia testuale (giornalistica e narrativa infantile) e comprendono testi in una sola lingua. In realtà, nel caso dei corpora narrativi si parlerà di corpora monolingua allineati, poiché le frasi dei testi originali ("L1") sono state allineate con le frasi dei testi semplificati ("L2")[3].

3.2 I corpora narrativi

I due corpora narrativi descritti in questa sezione costituiscono la prima risorsa italiana per la semplificazione automatica e semi-automatica dei testi (Brunato *e altri*, 2015), creata dall'ItaliaNLP-LAB dell'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa.
I testi, indirizzati a differenti *target* di lettori, sono stati allineati manualmente e rappresentano due differenti strategie di semplificazione manuale: la strategia "strutturale", che implica una semplificazione cumulativa su diversi livelli linguistici (descritta in 3.2.1 nel caso del primo corpus *Terence*), e la strategia "intuitiva", che si avvale invece dell'intuizione e dell'esperienza dell'insegnante (descritta in 3.2.2 nel caso del secondo corpus *Teacher*).
L'ItaliaNLP-LAB ha definito un nuovo schema di annotazione capace di intercettare le semplificazioni manuali a differenti livelli della struttura linguistica e di affrontare le diverse strategie di semplificazione (Brunato *e altri*, 2015). Una volta annotati i corpora con il precedente schema, lo scopo del progetto sarà utilizzare questa risorsa annotata per addestrare un classificatore supervisionato che sia in grado di semplificare i testi in modo semi-automatico, riuscendo a identificare le aree di complessità linguistica della frase e suggerendo all'autore la regola di semplificazione più appropriata per il contesto.

3.2.1 Terence

Terence è una collezione comprendente 32 racconti brevi per bambini e le rispettive versioni semplificate manualmente.
Il corpus trae il nome dal "Progetto Terence", un progetto dell'Unione Europea (Terence_Corsortium, 2012) nato alla fine del 2010 rivolto a bambini e bambine dai sette agli undici anni con deficit uditivi o con difficoltà nella comprensione dei testi. Il progetto si basa sulla creazione di differenti livelli e tipi di complessità del testo che possono risultare necessari per favorire la comprensione del racconto a seconda delle differenti

[3]Sebbene i corpora contengano testi in sola lingua italiana, essi risultano paralleli in quanto viene attuata una divisione dell'italiano in due varietà, una corrispondente a un italiano "complesso" (L1), e l'altra a un italiano "semplificato" (L2).

capacità cognitive dei lettori. Perciò, come già accennato, in *Terence* si può notare un tipo di strategia per la semplificazione dei testi definita "strutturale", in quanto il gruppo di esperti autore della semplificazione segue una linea guida predefinita che affronta la semplificazione su tre livelli testuali distinti (successivamente descritti) e pensata secondo i bisogni del *target* specifico ai quali sono indirizzati i testi.
Il sistema di semplificazione sviluppato in *Terence* è basato sull'integrazione di una prospettiva linguistico-cognitiva per la comprensione del testo, ritenuta di primaria importanza rispetto alla semplificazione sintattico-lessicale. Difatti, i tre obiettivi posti dalla semplificazione sono i seguenti (Terence_Corsortium, 2012):

- migliorare la comprensibilità: il progetto si propone di migliorare la coerenza del testo, chiarendo al meglio possibile la coesione fra gli eventi in relazione alle necessità dei lettori;
- offrire testi "multilivello": l'obiettivo della strategia strutturale è produrre testi a diversi livelli di semplificazione per soddisfare i bisogni dei differenti *target* ai quali essi sono indirizzati;
- ridurre al minimo le modifiche: il processo di semplificazione tenta di preservare la struttura testuale e linguistica della storia autentica e di apporre modifiche solo quando strettamente necessario alla comprensione finale degli eventi per far sì che i bambini imparino comunque ad utilizzare un lessico ricco e strutture sintattiche che possano migliorare il loro linguaggio.

Per fare ciò, ogni storia originale è stata riscritta in maniera cumulativa secondo tre livelli discendenti di difficoltà. Di seguito si possono osservare i passaggi:

- Livello 4. Versione originale del testo scritto dall'autore e non semplificato;
- Livello 3. Coerenza globale: il primo psicolinguista, partendo dalla storia originale, produce una prima semplificazione della storia rendendo esplicite delle informazioni necessarie alla comprensione del significato generale di essa, che altrimenti il lettore avrebbe dovuto dedurre;
- Livello 2. Coerenza locale: il secondo psicolinguista, partendo da un'analisi del livello 3, produce una nuova versione della storia semplificata a livello della coerenza interna, migliorando la comprensibilità grazie all'introduzione di connettivi espliciti o altre informazioni necessarie a rendere le relazioni tra le frasi più esplicite e chiare;
- Livello 1. Lessico e grammatica: il linguista, partendo da un'analisi del livello 2, produce una nuova versione semplificata della storia a livello di vocabolario,

eliminando espressioni metaforiche, frasi troppo lunghe e complesse e costruzioni sintattiche insolite, sostituendo il lessico poco familiare con parole comuni e mettendo in ordine le frasi secondo il reale ordine temporale degli eventi.

Per allineare il corpus, l'ItaliaNLP-LAB si è avvalso solamente dei testi semplificati a livello della coerenza locale e a livello lessicale/grammaticale: il livello 2, infatti, è diventato l'equivalente del testo originario, mentre il livello 1 l'equivalente del testo semplificato. Così, sono state allineate manualmente 1036 frasi originali con 1060 frasi semplificate. La decisione di dedicarsi solo ai livelli 2 e 1 della semplificazione deriva dalla necessità di concentrarsi solamente sulle categorie linguistiche lessicali, grammaticali e sintattiche che interesseranno la semplificazione automatica.

3.2.2 Teacher

Teacher è un corpus allineato formato da 24 coppie di testi originali e semplificati raccolti da siti web educativi specializzati che forniscono risorse gratuite per gli insegnanti (Brunato *e altri*, 2015). I generi testuali rappresentati in questa raccolta sono vari e riguardano, oltre a generi letterari, anche testi scolastici di storia o geografia.
Teacher costituisce un tipo di strategia per la semplificazione dei testi denominata "intuitiva": infatti, a differenza di *Terence*, ogni testo è stato semplificato indipendentemente da differenti insegnanti e pensato per essere indirizzato ad un unico *target* di persone, ovvero gli studenti L2 (corrispondente almeno al livello B2 per la lingua Italiana). Perciò, la semplificazione è stata trattata su diversi livelli linguistici e senza tener conto di regole o gerarchie predefinite proprio perché ogni insegnante ha intuitivamente adattato il testo come riteneva più opportuno in base ai bisogni e alle esigenze del destinatario.

3.3 I corpora giornalistici

Per il genere giornalistico verranno illustrati due corpora monolingua che rappresentano rispettivamente una collezione di testi complessi, *"La Repubblica"*, e una collezione di testi semplici, *"Due Parole"*; a differenza dei corpora narrativi, questi testi non sono stati allineati.

3.3.1 La Repubblica

La prima versione de *La Repubblica* consiste in un ampio corpus di testi giornalistici composto da circa 326 milioni di *token*[4] e include tutti gli articoli pubblicati dal 1985 al 2000 sul giornale italiano "La Repubblica".

[4]Un "token" è l'unità minima di analisi di una frase

Il corpus, sviluppato presso la "SSLMIT" dell'Università degli Studi di Bologna, già nel 2004 contava più di 175 milioni di *token* e si riteneva essere il più grande corpus italiano liberamente accessibile a quel tempo disponibile (Biagini et al.(2000) citato in Baroni *e altri* (2004)).
La collezione di testi che invece verrà analizzata nel quinto capitolo comprende solamente gli articoli scritti tra gli anni 2000 e 2005, per un totale di circa 232.000 *token*.

3.3.2 Due Parole

Due Parole è un corpus che trae il nome dall'omonimo quotidiano "Due Parole" (DueParole, 2002), un giornale italiano d'informazione di facile lettura studiato e scritto da linguisti esperti in semplificazione dei testi utilizzando un linguaggio controllato per un pubblico adulto con un livello di alfabetizzazione primitivo o con lievi disabilità intellettuali (Piemontese 2006, citato in Brunato *e altri* (2015)). Il progetto del giornale, sviluppato a partire dal 1983 presso l'Università di Roma "La Sapienza", si pone come obiettivo quello di scrivere articoli in una lingua il più possibile leggibile e comprensibile fornendo solo informazioni essenziali, per una comunicazione semplice ed efficace (DueParole, 2002). Infatti, attraverso lo studio del processo di comprensione, della leggibilità e della comprensibilità dei testi è stato possibile per i linguisti definire alcuni criteri di scrittura controllata essenziali per coloro che hanno poca familiarità con la lingua o che hanno difficoltà di comprensione dei testi, quali, ad esempio, la brevità dei testi, la semplicità delle frasi, la scelta di parole più comuni della lingua e un'accurata organizzazione logico-concettuale dei testi.
Il corpus colleziona tutti gli articoli scritti durante gli anni 2001-2006 e contiene circa 73.000 *token*.

Capitolo 4

Le tecnologie linguistico-computazionali per l'analisi linguistica automatica

In questo capitolo verranno mostrati i processi ai quali la lingua viene sottoposta al fine di effettuare diverse analisi linguistiche: l'annotazione linguistica è il passaggio fondamentale che rende possibile il monitoraggio delle caratteristiche linguistiche che si desidera analizzare in una lingua, presupposto a sua volta per compiti applicativi quali la valutazione di leggibilità di un testo sulla combinazione di diversi parametri linguistici. In particolare, verrà sottolineato quanto sia stato importante l'avvento delle tecnologie linguistico-computazionali e quanti benefici abbia apportato nello sviluppo degli strumenti atti all'analisi linguistica e, dunque, negli studi stessi. Tali strumenti di analisi, *LinguA* (4.2.1), *Monitor-IT* (4.3.1) e *READ-IT* (4.4.1), verranno descritti nelle sottosezioni indicate.

4.1 Introduzione al trattamento automatico della lingua

> "*La lingua è veicolo e chiave di accesso alla conoscenza*, e oggi più che mai è urgente la realizzazione di una infrastruttura consolidata di tecnologie linguistiche.(...) Gli sviluppi più recenti della *linguistica computazionale* e del *natural language engineering* hanno creato soluzioni tecnologiche dalle enormi potenzialità per migliorare la ricerca e gestione intelligente dell'informazione contenuta nei documenti testuali. Le nuove tecnologie della lingua, infatti, permettono ai sistemi informatici di accedere ai contenuto digitale attraverso il *Trattamento Automatico della Lingua* (TAL) o *Natural Language Processing* (NLP)." (Calzolari e Lenci, 2004)

Nonostante il fatto che elaborare il linguaggio naturale mediante un calcolatore elettronico possa essere reso particolarmente difficile e complesso a causa delle caratteristiche intrinseche di ambiguità del linguaggio umano, è ormai appurato che dotare il computer di conoscenze sulla struttura interna del linguaggio potenzi la capacità di estrarre informazione dai testi, migliori il processo di gestione e condivisione della conoscenza e, soprattutto, promuova una svolta significativa nell'accessibilità e nell'usabilità dei contenuti digitali grazie all'elaborazione di grandi quantità di documenti, per far fronte a un'utenza sempre più in crescita ed eterogenea. I fattori determinanti che hanno permesso questa svolta metodologica e tecnologica del TAL sono costituiti dai corpora di grandi dimensioni e dai lessici computazionali[1] (Calzolari e Lenci, 2004).
In questo capitolo verrà mostrato come, grazie a strumenti di trattamento automatico del linguaggio, oggi sia possibile effettuare analisi linguistiche approfondite molto utili al fine di specifici compiti di monitoraggio linguistico che permettono, ad esempio, la valutazione della leggibilità di un testo: in particolare, questi strumenti possono essere determinanti per monitorare la capacità di lettura e lo sviluppo della sintassi nel linguaggio infantile, identificare deficit cognitivi attraverso misure di complessità sintattica e misurare la leggibilità di testi per studenti L1 e L2 (Montemagni, 2013).

4.2 L'annotazione linguistica

Un corpus *annotato* è un corpus al quale sono state esplicitate le informazioni linguistiche e metalinguistiche mediante l'attribuzione di etichette (denominate *mark-up* o *tag*) a porzioni specifiche del testo. L'identificazione della struttura linguistica del testo avviene in modo incrementale attraverso l'annotazione di diversi livelli linguistici (Wikipedia, 2015c):

- segmentazione delle frasi e identificazione dei singoli *token*;
- annotazione morfologica: dopo il processo di *lemmatizzazione*[2], a ciascun *token* viene assegnata la rispettiva categoria grammaticale;
- annotazione sintattica: fornisce una descrizione della frase in termini di relazioni di dipendenza tra parole indicanti relazioni grammaticali (soggetto, oggetto, modificatore..);

[1] I "lessici computazionali" sono dizionari *machine-readable* per la rappresentazione del significato lessicale e utilizzabili per compiti di NLP, come la disambiguazione sintattica, la traduzione e il recupero ed estrazione dell'informazione: ne sono un esempio "WordNet" e "FrameNet".

[2] La "lemmatizzazione" è il processo di riduzione di una forma flessa di una parola alla sua forma canonica (non marcata), detta "lemma".

- annotazione semantica: viene codificato il significato delle espressioni linguistiche del testo secondo diversi criteri: la *"Named Entity Recognition"* assegna ai nomi propri di entità la loro categoria semantica in contesto (persona, luogo, ora..); i ruoli semantici descrivono la funzione semantica svolta da un sintagma nell'evento espresso dal verbo (agente, strumento, paziente..); le relazioni semantiche assegnano a coppie di parole nel testo la relazione semantica che le lega (composti, nominali complessi, modificatori aggettivali..) ;

- annotazione pragmatica: vengono evidenziati vari fenomeni riguardanti la funzione comunicativa di un enunciato o relazioni fra elementi linguistici che vanno al di là della singola frase come, ad esempio, la funzione illocutoria degli enunciati [3] o le relazioni di anafora e catafora.

L'annotazione manuale è oggi forte oggetto di discussione in confronto all'utilizzo dell'annotazione (semi-) automatica: infatti, l'annotazione manuale presenta una serie di svantaggi poiché, oltre ad essere altamente costosa e lenta, presenta notevoli incoerenze tra annotatori, che usano criteri diversi per annotare anche seguendo lo stesso schema di annotazione, e in ciascun annotatore, che può usare di volta in volta criteri diversi; tuttavia, risulta ancora necessaria per alcuni tipi di annotazione (ad esempio, pragmatica e prosodica).
L'introduzione delle tecnologie linguistico- computazionali rappresenta una svolta fondamentale anche nello studio della variazione linguistica della lingua italiana: queste tecnologie, infatti, permettono di accedere al contenuto informativo dei testi attraverso l'individuazione della struttura linguistica sottostante e la sua rappresentazione esplicita, riuscendo a identificare una vasta gamma di parametri utili al fine del monitoraggio linguistico. L'identificazione della struttura linguistica del testo avviene in modo incrementale, attraverso analisi linguistiche a livelli di complessità crescente: "tokenizzazione"; analisi morfo-sintattica e lemmatizzazione del testo "tokenizzato"; analisi della struttura sintattica della frase in termini di relazioni di dipendenza (Montemagni, 2013).
Entrambi i metodi di annotazione arricchiscono il testo codificando le informazioni dei vari livelli linguistici che permettono di condurre analisi approfondite, ma è proprio grazie all'utilizzo di sistemi informatici se oggi è possibile effettuare ricerche ed elaborazioni ancora più avanzate e rapide, in quanto l'annotazione costituisce un ottimo filtro di ricerca per interrogare in maniera avanzata e rapida il contenuto del corpus da parte del computer.
Si può constatare che lo "stato dell'arte" nei compiti di annotazione linguistica è ad

[3] Per "atto illocutorio" si intende un atto la cui esecuzione è da rendere nota ad altre persone e la cui prestazione coinvolge la produzione di 'conseguenze convenzionali' come, ad esempio, diritti, impegni o obblighi.

oggi rappresentato da sistemi basati su algoritmi di apprendimento automatico supervisionato (Montemagni, 2013): infatti, a partire da un corpus di addestramento annotato con informazione morfo-sintattica e sintattica, viene costruito un modello probabilistico per l'annotazione linguistica del testo e, a ogni passo di computazione, il sistema sceglie l'annotazione più probabile data la parola in input, i suoi tratti descrittivi, il contesto e le annotazioni linguistiche già identificate.
In conclusione, è possibile affermare che, nonostante l'annotazione automatica includa inevitabilmente dei margini di errore, essa risulta comunque uno strumento certamente affidabile nella ricostruzione del profilo linguistico del testo.

4.2.1 LinguA: Linguistic Annotation pipeline

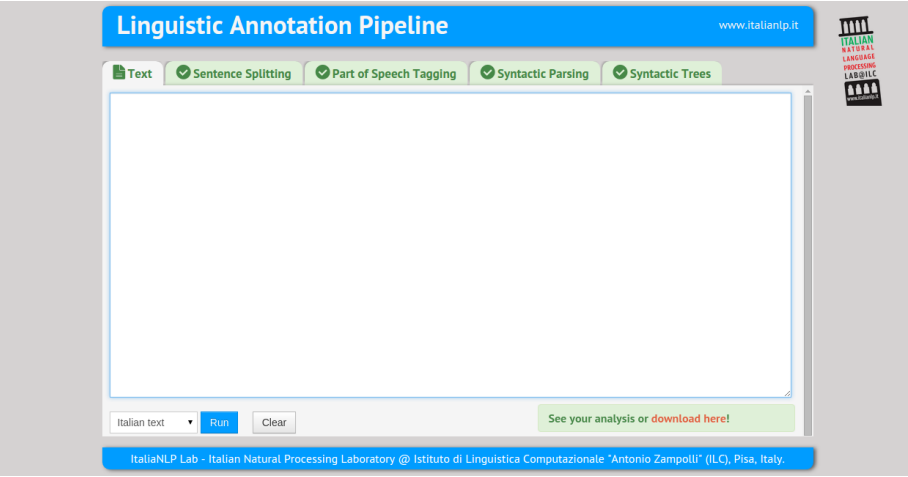

Figura 4.1: Schermata della pagina web del dello strumento LinguA

Lo strumento che ci permette di annotare automaticamente i testi combinando algoritmi di apprendimento automatico e basati su regole si chiama *LinguA* [4]. Una volta inserito il testo che si desidera analizzare, LinguA, come possibile notare nella Figura 4.1, esegue le seguenti fasi di annotazione:

- *Sentence splitting*: individua le frasi che compongono il testo;
- *Part-of-speech tagging*: identifica la struttura linguistica in modo incrementale secondo i processi di tokenizzazione, lemmatizzazione e analisi morfo-sintattica, attribuendo a ciascun *token* il lemma corrispondente e la categoria morfo-sintattica di appartenenza. Il sistema ILC-POS-Tagger (Dell'Orletta, 2009) ha un'accuratezza del 96,34% nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati;
- *Syntactic parsing*: tramite l'analisi della struttura sintattica, identifica le dipendenze all'interno della frase, specificando il tipo di relazione tra *token* e le te-

[4] http://linguistic-annotation-tool.italianlp.it/

ste sintattiche. Lo strumento DeSR (Attardi *e altri*, 2009) risulta affidabile per l'87,71% secondo la metrica *UAS*[5] e per l'83,38% secondo la metrica *LAS*[6]: questo significa che DeSR è molto affidabile nel ricostruire le relazioni di dipendenza che collegano le parole della frase (UAS), mentre appare più problematica l'identificazione simultanea del tipo di dipendenza e della testa sintattica (LAS) (Montemagni, 2013);

- *Syntactic Trees*: rappresenta graficamente l'albero delle dipendenze sintattiche.

4.3 Il monitoraggio linguistico

Per *monitoraggio linguistico* si intende ricostruire il profilo linguistico di un testo a vari livelli linguistici e in relazione alle possibili variabili sociolinguistiche:

- diacronia: in una prospettiva dinamica ed evolutiva, studia i fatti linguistici secondo il loro divenire nel tempo;
- diamesia: relativa al mezzo materiale adottato per comunicare, distingue principalmente testi orali e testi scritti;
- diafasia: determinata dal mutare della situazione nella quale il parlante si trova a comunicare, tiene di conto del contesto, degli interlocutori, delle circostanze o delle finalità della comunicazione;
- diastratia: relativa alla situazione dei parlanti, studia la provenienza socio-culturale, l'età, il sesso e il livello di istruzione.

Uno studio di monitoraggio dovrebbe basarsi, quindi, su una collezione "aperta" di testi, definita *monitor corpus*, che muta nel tempo secondo criteri prestabiliti per poter tenere traccia dei processi di variazione nei livelli linguistici tra diverse varietà d'uso nei generi testuali e nel tempo (Montemagni, 2013).
Fino ad oggi, gli studi per la lingua italiana sulla varietà diamesica e dei generi testuali si sono tipicamente basati su lessici di frequenza realizzati a partire da corpora il cui processo di lemmatizzazione e annotazione morfo-sintattica è stato condotto in modo manuale o semi-automatico. Nel panorama italiano, l'unico corpus a essere stato creato invece a supporto dello studio delle differenze a livello sintattico è costituito dal *Corpus Penelope*, una risorsa di dimensioni contenute (poco più di 30.000 parole), che

[5] *Unlabelled Attachment Score* (UAS) misura la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione della testa sintattica.

[6] *Labelled Attachment Score* (LAS) misura la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda sia la testa sintattica sia il tipo di relazione di dipendenza svolto in rapporto a essa.

ha rappresentato per anni il riferimento più importante per studi e ricerche di tipo sintattico per tener traccia delle tendenze e delle trasformazioni sincroniche e diacroniche della lingua. Dalla metà degli anni Ottanta, la lingua inglese fu la prima a concentrare gli studi su *"register variation"* basati su corpora di maggiori dimensioni e su un livello di annotazione morfo-sintattica "potenziato" con regole ad hoc operanti sulla sequenza delle categorie grammaticali e sui lemmi per l'identificazione di particolari costruzioni sintattiche e strutture semantiche (Montemagni, 2013).

Tuttavia, mediante il ricorso a tecnologie linguistico-computazionali oggi è possibile monitorare in modo affidabile un ampio spettro di parametri di diversi livelli della lingua e in relazione a corpora testuali di dimensioni sempre più vaste, e intercettare anche informazioni della struttura sintattica che ad oggi sembravano impensabili in quanto difficilmente attingibili mediante un'analisi manuale del testo, fornendo così un livello di analisi più avanzato sia rispetto alla tradizione di studi sulla lingua italiana, sia rispetto alla letteratura corrente sulla *register variation*.

Alcune caratteristiche linguistiche estrapolate da testi annotati in modo automatico sono già state oggetto di monitoraggio in letteratura: attraverso lo studio delle categorie morfo-sintattiche, infatti, è possibile rintracciare differenze e somiglianze tra diversi generi testuali e tra scritto e parlato, calcolare la distribuzione delle congiunzioni coordinanti e subordinanti, monitorare l'ordine relativo tra la clausola principale e le subordinate e i livelli di incassamento gerarchico; inoltre, è possibile mettere in relazione la frequenza di occorrenze di certe categorie grammaticali rispetto ad altre, ad esempio misurando la "densità lessicale", calcolata come la proporzione delle parole piene rispetto al totale delle occorrenze, oppure il rapporto tra diverse categorie morfo-sintattiche (per esempio, il rapporto tra nomi e verbi).

Ciò che Montemagni (2013) dimostra tramite i suoi studi è un importante elemento di novità riguardante la tipologia di parametri di monitoraggio indagati, basati su "microprelievi" che si possono effettuare sul testo arricchito con informazione sintattica e morfo-sintattica; nonostante lo strumento di annotazione includa un inevitabile margine di errore (presenta un'affidabilità del 87.71%), esso rende possibile l'osservazione di aspetti della struttura linguistica altrimenti difficilmente investigabili.

In particolare, lo studio si incentra sulla questione della distribuzione di nomi e verbi nella variazione diamesica dello scritto e del parlato e in relazione a generi testuali differenti[7], e dimostra come l'arricchimento di ulteriori evidenze linguistiche abbia permesso di studiare nuove dimensioni di analisi e variazione. Utilizzando uno strumento di annotazione morfo-sintattica addestrato su corpora di lingua scritta, l'analisi della distribuzione di nomi e verbi è stata circoscritta alle sole parole incluse nel dizionario

[7]Lo studio è stato effettuato su sei corpora di lingua scritta corrispondenti a differenti generi testuali e livelli di complessità linguistica: quattro corpora, uno di genere giornalistico, uno narrativo, uno legislativo e uno fantastico, sono stati affiancati a due corpora di linguaggio semplificato, uno di genere giornalistico e l'altro corrispondente a un sussidiario della scuola primaria.

morfologico di riferimento, in modo da eliminare le forme linguistiche ed extralinguistiche che ricorrono nel parlato non standard e non ottenere, così, risultati falsati. In questo modo, l'analisi iniziale ha ottenuto un significativo ridimensionamento, con una notevole riduzione della preponderanza accentuata dei verbi rispetto ai nomi che ora invece rispecchia la predominanza "reale", significativamente più bassa. Montemagni (2013) dimostra come, nella forma scritta, una maggiore frequenza di nomi è associata a testi caratterizzati da un'alta densità informativa (quali i giornali e le leggi), mentre generi testuali più vicini alla lingua parlata (quali la narrativa e composizioni di scrittura creativa) sono caratterizzati da una maggiore frequenza di verbi. Quanto registrato appare in linea con la letteratura sulla variazione linguistica a livello diamesico, diafasico e testuale e, infatti, è comprovato da una forte correlazione con i dati risultanti dal monitoraggio di corpora la cui annotazione è stata rivista manualmente.
Dopodiché, Montemagni (2013) attua delle analisi in riferimento alle sottocategorie che suddividono la classe dei verbi in ausiliari, modali e principali, e arriva a distinguere quelli che svolgono il ruolo di testa verbale. Inoltre, dopo aver definito il numero medio di clausole per periodo, l'autrice raffina la misura del rapporto nomi/verbi calcolando la ricorrenza media di nomi per clausola e verifica in che misura il parametro dei dipendenti per testa verbale si rapporta alla misura dei nomi per clausola.
In conclusione, grazie alla nuova tipologia di parametri utilizzati, oggi è possibile considerare sincronicamente maggiori varietà delle dimensioni di analisi, rendendo così possibile il monitoraggio di aspetti della struttura linguistica fino a oggi inesplorati, grazie ai quali è possibile monitorare la competenza linguistica di studenti L1 o L2 della lingua italiana o definire un indice di leggibilità "avanzato" basato su parametri riguardanti l'uso della lingua in tutte le sue componenti (Dell'Orletta *e altri*, 2011).

4.3.1 Monitor-IT

I risultati derivanti dalle metodologie utilizzate nell'annotazione linguistica (descritte nella sezione 4.2.1) vengono successivamente implementati da *Monitor-IT*, [8] uno strumento di monitoraggio linguistico online grazie al quale è possibile, mediante l'inserimento di testi, estrarre statistiche linguistiche utili per un'analisi dettagliata della forma linguistica degli stessi (Marinelli, 2015). In particolare, Monitor-IT conduce le sue statistiche effettuando un confronto delle caratteristiche linguistiche (Figura 4.2) tra i testi inseriti e cinque corpora di riferimento che rappresentano diversi livelli di difficoltà equivalenti a cinque differenti gradi di apprendimento scolastico, dalle elementari alle superiori. Nel dettaglio, i corpora sono:

- Biennio elementari: 7270 periodi per 93171 *token*, testi di normale utilizzo per studenti dei primi due anni delle elementari;

[8] http://monitor-it.italianlp.it/

- Triennio elementari: 10736 periodi per 183125 *token*, testi degli ultimi tre anni delle elementari;
- Medie: 3167 periodi per 63509 *token*, testi delle scuole medie;
- Biennio superiori: 2521 periodi per 55835 *token*, testi del primo biennio delle scuole superiori;
- Triennio superiori: 5389 periodi per 176561 *token*, testi degli ultimi tre anni delle scuole superiori.

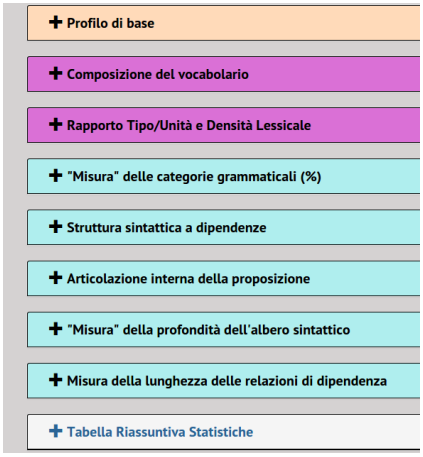

Figura 4.2: Le caratteristiche linguistiche misurate in Monitor-IT

Monitor-IT mostra i risultati del confronto delle caratteristiche linguistiche principalmente mediante grafici a barre, nei quali il testo inserito dall'utente è rappresentato dalla barra orizzontale, mentre ogni corpora di riferimento è costituito da una barra verticale e rappresentato cromaticamente da un colore: il livello più semplice delle elementari è costituito dal rosa, il livello più difficile delle superiori dal rosso.

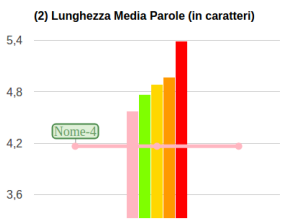

Figura 4.3: Esempio di grafico tratto dalla statistica "Profilo di base"

Ad esempio, il grafico proposto nella Figura 4.3 mostra la lunghezza media delle parole nella frase "Questo è un esempio di grafico" in rapporto a quella dei corpora di riferimento: il risultato è che la lunghezza media delle parole corrisponde a quella del livello del biennio delle elementari; dunque, la barra orizzontale assume lo stesso colore della barra del livello corrispondente.

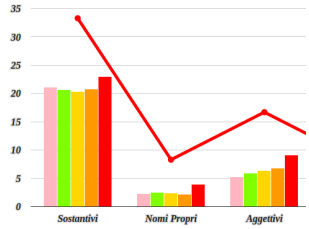

Figura 4.4: Esempio di grafico tratto dalla statistica "Misura delle categorie grammaticali"

Se invece si vuole analizzare la distribuzione delle categorie morfo-sintattiche nella frase "Antonio il pescivendolo buttò duecento pesci vivi nel piccolo lago del paese", è possibile notare nella Figura 4.4 come, nel caso per esempio dell'aggettivo, la distribuzione sia più alta rispetto al numero di aggettivi presenti in un testo del triennio delle superiori.

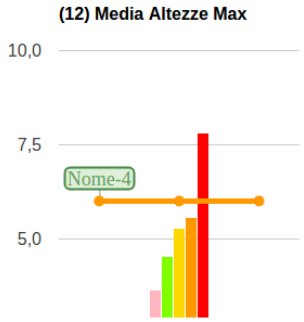

Figura 4.5: Esempio di grafico tratto dalla statistica "Misura della profondità dell'albero sintattico"

Infine, si può analizzare a livello sintattico la frase "Simo decise di iscriversi nella Categoria Strumenti, quindi lui ed Effy si precipitarono entrambi fino alla piazza del paese".
Nella Figura 4.5 è possibile notare come la media delle altezze massime dell'albero sintattico della frase analizzata, essendo la barra di colore arancione, sia più vicina a quella dei testi del biennio delle superiori, e addirittura tendente verso la difficoltà massima dei testi del triennio.

4.4 La leggibilità

E' ormai riconosciuto che la semplificazione del linguaggio, in entrambe le forme dello scritto e del parlato, faciliti la comprensione di un testo per uno studente principiante della lingua (Tweissi, 1998). Ridurre la complessità di un testo e semplificarlo in modo da renderlo maggiormente comprensibile per il lettore significa operare su due livelli:

- livello linguistico: il destinatario può avere delle mancanze di conoscenze linguistiche, quindi le parole inusuali e le costruzioni sintattiche complesse devono essere sostituite con parole più consuete e periodi più semplici;
- livello cognitivo: il destinatario può avere difficoltà a rielaborare le nuove informazioni e dedurre conclusioni solo con la sua conoscenza di base, quindi è necessario rendere esplicite alcune informazioni implicite, affinché l'ascoltatore possa conoscere il *background* comunicativo e possa riuscire a costruirsi una mappa mentale nella quale collocare gli eventi narrati.

Ciò ha dato l'impulso agli studiosi di sviluppare metodi per la semplificazione dei testi e definire misurazioni che potessero quantificare la difficoltà di un testo e la sua comprensione, come per esempio la leggibilità, intesa come scorrevolezza della lettura in funzione della struttura linguistica di un testo (Marinelli, 2015).
I primi misuratori di leggibilità, ad esempio l'indice di *Flesch-Kincaid* o l'indice di

Gunning's Fog per la lingua inglese e l'indice di *Gulpease* per la lingua italiana, si affidavano a caratteristiche testuali, come la lunghezza delle parole e delle frasi (ItaliaNLPLab, 2014). I misuratori di leggibilità di seconda generazione, invece, fanno ricorso al progressivo avanzamento delle tecnologie linguistico-computazionali, come, ad esempio, *Coh-Metrix*[9] (Crossley *e altri*, 2011) o altri strumenti di annotazione linguistica automatica, grazie ai quali è possibile definire la leggibilità di un testo sulla base di parametri linguistici più complessi che spaziano tra vari livelli di analisi linguistica, ottenendo risultati molto migliori rispetto a quelli ottenuti con i classici metodi di valutazione della leggibilità.
Per quanto riguarda la lingua italiana, READ-IT (Dell'Orletta *e altri*, 2011), descritto in 4.4.1, è ad oggi il primo e unico strumento avanzato che si avvale di caratteristiche estratte dai differenti livelli linguistici per attuare una duplice valutazione della leggibilità, sul documento e sulla singola frase: la valutazione di quest'ultima rappresenta un'importante novità perché permette a READ-IT di contribuire alla semplificazione dei testi, grazie all'identificazione dei luoghi di complessità della frase e ad una classificazione semantica.
L'incontro tra l'interesse per la semplificazione dei testi e il trattamento automatico della lingua ha portato alla nascita e a un'attenzione sempre maggiore negli ultimi anni per la *Semplificazione Automatica dei Testi* (ATS): infatti, il suo impiego favorisce sia la fase di processamento migliorando l'efficienza di *parsing* ed estrazione delle informazioni (Brunato *e altri*, 2015), sia l'accessibilità al testo. Per questo è utilizzata sempre più in scenari educativi e nelle tecnologie assistite per adattare un testo a particolari lettori, come, ad esempio, studenti L2 o con disabilità cognitive: lo scopo è di ridurre la complessità lessicale e sintattica cercando di preservare il significato originale del testo.

4.4.1 READ-IT: Assessing Readability of Italian Texts

READ-IT [10] è uno strumento online di valutazione di leggibilità per testi italiani che si basa sulla combinazione di tratti linguistici appartenenti a tre diversi livelli di descrizione linguistica: lessicale, morfo-sintattico e sintattico. Questo strumento implementa un indice di leggibilità "avanzato" su analisi linguistica del testo condotta da un insieme di tecnologie che rappresentano lo "stato dell'arte" per il trattamento della lingua italiana, sviluppate presso l'Istituto di Linguistica Computazione "Antonio Zampolli" (ItaliaNLPLab, 2014).
L'approccio seguito nei sistemi avanzati per la valutazione della leggibilità di un testo rispecchia quello dei sistemi di annotazione linguistica, basati su algoritmi di appren-

[9] http://cohmetrix.com/
[10] http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest

dimento supervisionato: anch'esso, infatti, si incentra su un insieme di categorie linguistiche da assegnare (i livelli di leggibilità), un corpus di apprendimento, costituito da un insieme di esempi pre-classificati a mano rispetto alle categorie di leggibilità, e un insieme di tratti descrittivi selezionati sulla base del compito di classificazione da svolgere.

Infatti, READ-IT, dopo aver identificato la struttura linguistica, opera sul testo annotato e conduce una classificazione probabilista del testo rispetto alle classi "leggibile" e "complesso", sulla base dei diversi livelli di descrizione linguistica. In particolare, lo strumento propone un'analisi dell'intero documento che valuta la comprensibilità del testo secondo diversi livelli di complessità (lessicale, morfosintattico e sintattico) e, inoltre, permette di analizzare la complessità delle singole frasi fornendo un supporto-guida nel processo di revisione e semplificazione del testo da parte del redattore grazie all'individuazione del grado di difficoltà nei livelli base, sintattico, lessicale e globale delle frasi, che viene rappresentato cromaticamente da colori che vanno dal verde (testo leggibile) al rosso (testo particolarmente difficile).

Capitolo 5

Analisi dei dati estratti

In questo capitolo verranno analizzati i dati statistici relativi all'ordinamento dei costituenti che sono stati estratti dai corpora narrativi e giornalistici descritti nel capitolo 3. Dopo aver fornito una panoramica dei dati generali di ciascun corpus (5.1), verranno esaminati i dati statistici dell'ordinamento preposto e posposto al verbo di soggetto (5.1.2), oggetto (5.1.3) e avverbio (5.1.5), della posizione dell'aggettivo rispetto al sostantivo (5.1.4), e dell'ordinamento preposto e posposto alla clausola principale delle subordinate (5.1.6) per indagare, attraverso uno studio linguistico-computazionale, sugli stessi casi di ordinamento descritti teoricamente nel capitolo 2.
Il confronto tra dati avverrà su due livelli: la variazione di genere e il grado di complessità. L'analisi consisterà nel verificare quali sono gli ordini degli elementi che vengono condizionati dal genere testuale e quali dipendono dal grado di complessità: ci si aspetta di ritrovare una certa somiglianza dell'ordine degli elementi in relazione al genere, ma soprattutto di appurare che, indipendentemente dal genere, i testi semplici siano più fedeli a seguire l'ordinamento canonico degli elementi nella lingua italiana, mentre i testi complessi abbiano una più alta percentuale di casi di ordine marcato; si prevede, dunque, di ottenere dati statistici molto simili tra testi di diverso genere testuale ma con stesso grado di complessità.
Infine, nella sezione 5.2 si discuteranno i risultati delle statistiche nel complesso.

5.1 Una panoramica dei dati

In questa sezione verranno presentati i dati del profilo di base delle quattro raccolte dei testi, ovvero, per ciascun corpus verrà indicato il numero di *token*, la media del numero di *token* per documento e la media del numero delle frasi per documento, e verranno confrontanti la media di *token* per periodo e la media del numero di caratteri per *token*; dopodiché, verranno mostrati e paragonati i dati statistici della distribuzione delle categorie grammaticali quali sostantivo, verbo, aggettivo e avverbio e delle frasi

principali e subordinate. Alcune categorie e le frasi subordinate verranno riprese nelle successive sottosezioni per un'analisi specifica sull'ordinamento di esse rispetto alla testa sintattica o alla frase principale.

Corpus	Genere	Num. token	Media token/doc.	Media frasi/doc.
Terence e Teacher originali	narrativo	26.311	469,24	24,21
Terence e Teacher semplificati	narrativo	24.083	430,05	24,73
Repubblica	giornalistico	232.908	721,98	28,95
Due Parole	giornalistico	73.314	226,62	12,14

Tabella 5.1: Alcuni parametri del profilo di base dei quattro corpora

La Tabella 5.1 riporta, per ogni corpus analizzato, il genere testuale, il numero totale di *token*, la media del numero di *token* per documento e la media del numero delle frasi per documento.

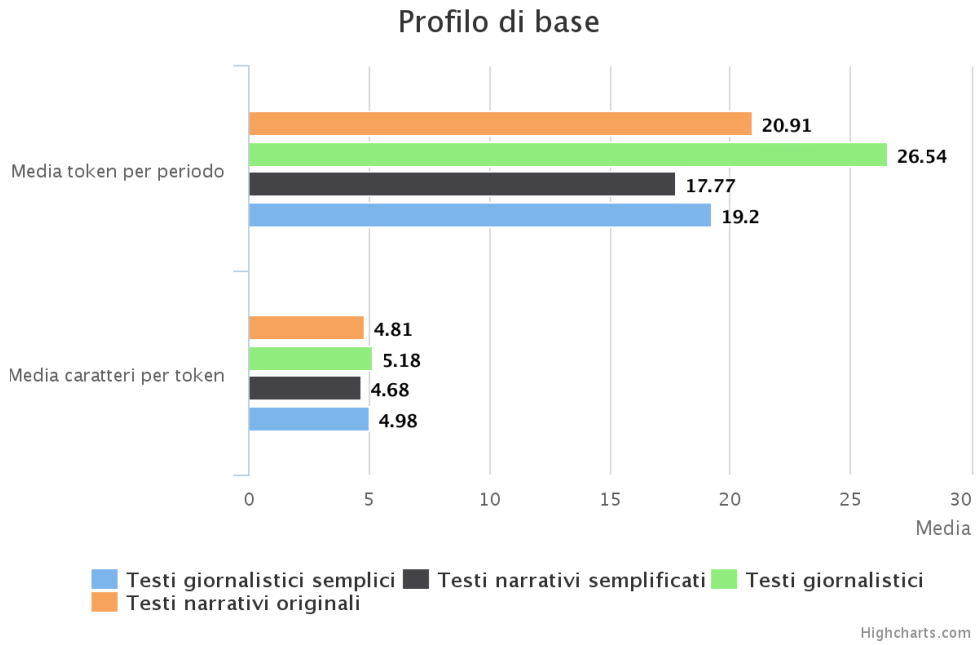

Figura 5.1: Alcuni parametri del profilo di base dei quattro corpora a confronto

Nella Figura 5.1 sono mostrati la media di *token* per periodo e la media del numero di caratteri per *token*. Per quanto riguarda il primo caso, è possibile notare una riduzione del numero di *token* nei testi semplici di entrambi i generi testuali (19,2 la media per i testi giornalistici e 17,77 la media per i testi narrativi) rispetto ai testi complessi (26,54

la media per i testi giornalistici e 20,91 la media per i testi narrativi). Le statistiche rispecchiano dunque la preferenza nei testi di facile comprensione a utilizzare frasi corte e con poche parole rispetto ai testi complessi, che invece prediligono frasi più lunghe. Per quanto riguarda la media del numero di caratteri per *token*, invece, si può affermare che nei testi giornalistici si utilizzano parole più lunghe rispetto ai testi narrativi: infatti, al primo posto si trovano i testi giornalistici (5,18), seguiti dai testi giornalisti semplici (4,98), poi dai testi narrativi complessi (4,81), e infine dai testi narrativi semplificati (4,68).

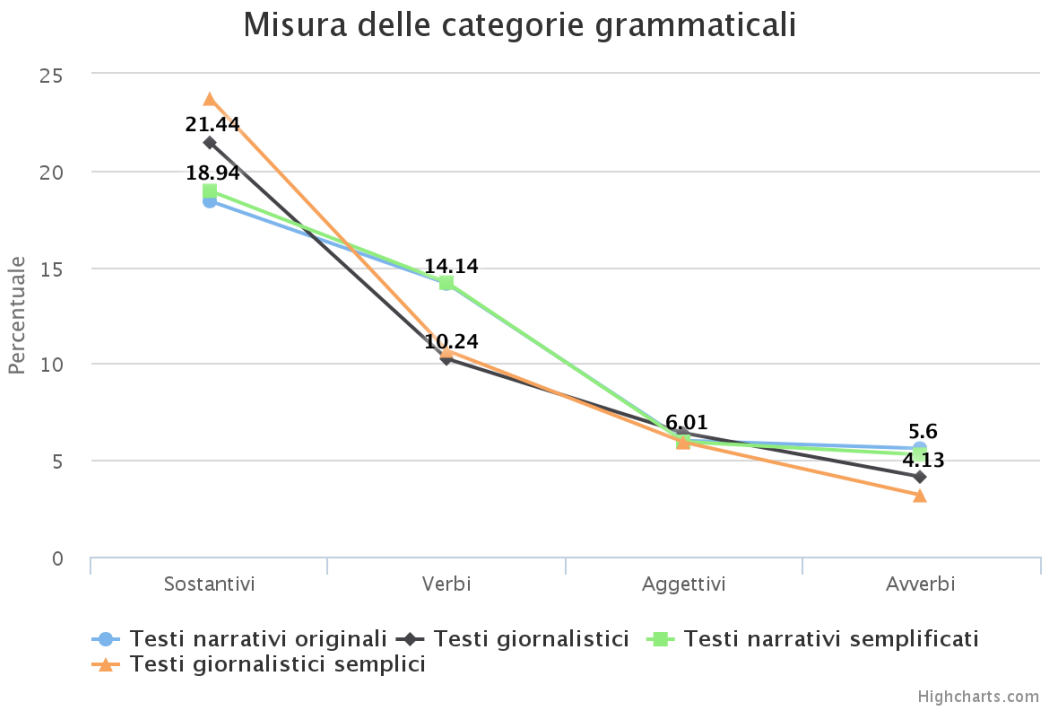

Figura 5.2: Distribuzione di sostantivi, verbi, aggettivi e avverbi nei quattro corpora

Nella Figura 5.2 viene mostrata la media delle distribuzioni delle categorie grammaticali nei quattro corpora.
Nel caso dei sostantivi, è possibile notare un aumento delle percentuali nei testi giornalistici (in media, 23,72% nei testi semplici e 21,44% nei testi "complessi") rispetto ai testi narrativi (in media, 18.94% nei testi semplificati e 18,41% nei testi originali); inoltre, confrontando il passaggio dai testi complessi ai testi semplici, si può osservare, in entrambi i generi, un aumento dell'utilizzo dei sostantivi. Dunque, si può dedurre che, nel complesso, i nomi vengono adoperati più frequentemente nel genere giornalistico, che essendo un genere conciso si avvale frequentemente delle nominalizzazioni, e nei testi semplici, in quanto i sostantivi rendono più comprensibile un testo come,

per esempio, un soggetto esplicitato piuttosto che omesso.
Per quanto riguarda i verbi, invece, si può osservare un'alta frequenza di utilizzo nel caso dei testi narrativi: infatti, i testi semplificati (14,19%) e i testi originali (14,14%) sono quelli che adoperano maggiormente questa categoria, mentre nei testi giornalistici è possibile notare una riduzione notevole (10.24% nel caso dei testi "complessi" e 10.67% nel caso dei testi semplici). In questo caso, dunque, possiamo notare una netta tendenza nella varietà del genere piuttosto che nel grado di complessità dei testi: infatti, il genere giornalistico è più sintetico, mentre quello narrativo necessita dell'utilizzo dei verbi per raccontare e spiegare le azioni.
Nel caso degli aggettivi, si può notare un utilizzo di essi poco maggiore nei testi complessi piuttosto che nei testi semplici: infatti, mentre nei testi giornalistici e narrativi complessi si osserva una media rispettiva di 6,40% e 6,01%, nei testi narrativi semplificati e in quelli giornalistici semplici si nota una media di 5,95% e 5,92%. Sebbene la differenza tra i dati sia minima, la spiegazione può essere data dal fatto che, introducendo attributi al nome, si forniscono informazioni aggiuntive e opzionali che possono rendere più complesso il testo e, di conseguenza, la sua comprensione.
Infine, per quanto riguarda gli avverbi, è possibile notare una differenza della frequenza di questa categoria tra i testi narrativi e i testi giornalistici: infatti, i testi che maggiormente si avvalgono dell'utilizzo degli avverbi sono quelli narrativi complessi (5,60%) e semplificati (5,28%), mentre i testi giornalistici presentano una rispettiva media di 4,13% e 3,18%. Inoltre, è possibile osservare come la distribuzione di avverbi diminuisca nei testi semplici: come già spiegato nel caso degli aggettivi, anche gli avverbi introducono informazioni nuove perlopiù superflue e costituiscono, perciò, un ostacolo in più alla comprensione del testo, che nei testi semplici si cerca di ridurre al minimo.
In conclusione, si può constatare che le differenze tra generi si possono notare nel caso dei sostantivi, molto più frequenti in ambito giornalistico, dei verbi, adottati maggiormente nel genere narrativo e, in misura minore, degli avverbi, utilizzati anch'essi con più alta frequenza nei testi narrativi; anche le differenze tra grado di complessità si osservano maggiormente nel caso dei sostantivi, in percentuale più alta nei testi semplici, e degli avverbi, più frequenti nei testi complessi.

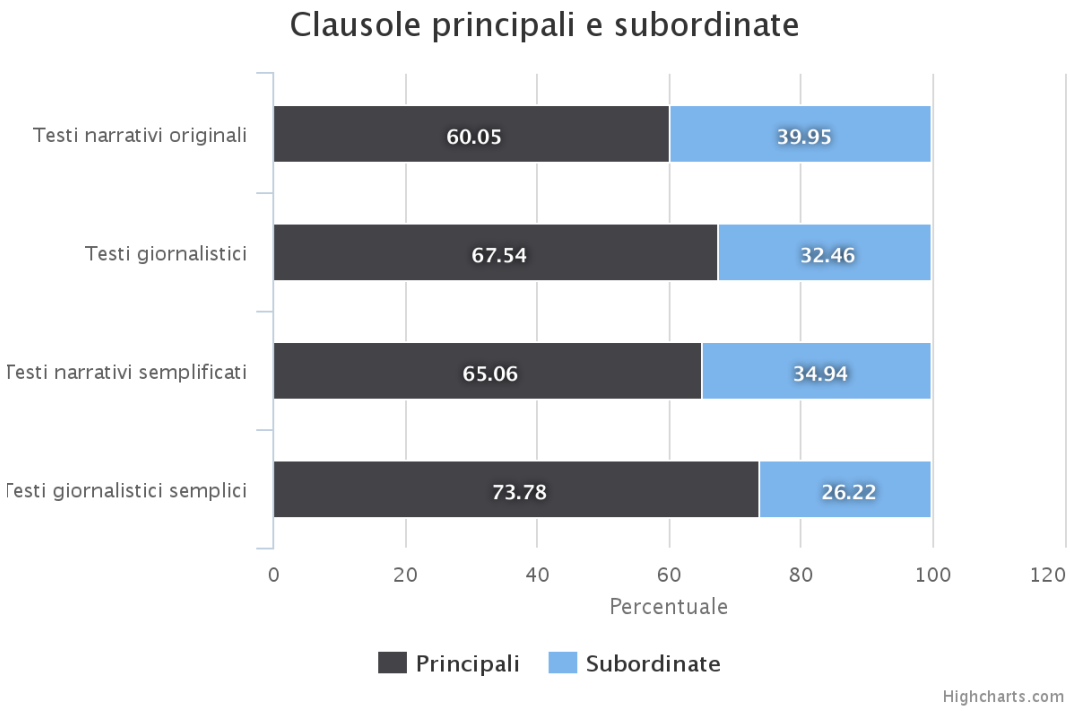

Figura 5.3: Distribuzione delle frasi principali e subordinate nei quattro corpora

Nella Figura 5.3 sono mostrate le percentuali di distribuzione delle frasi principali e subordinate nei quattro corpora: in media, nel 66,61% dei casi vengono utilizzate frasi principali, mentre nel 33,39% sono adoperate proposizioni subordinate. È possibile notare che nel passaggio dai testi complessi ai testi semplici la percentuale di principali aumenta a discapito delle subordinate: questo dato conferma il fatto che i testi complessi, essendo composti da un numero maggiore di subordinate, risultano dotati di una maggiore complessità strutturale, mentre i testi semplici, che si avvalgono in più casi di frasi principali o proposizioni coordinate, sono costituiti da una struttura grammaticale interna semplice.

In particolare, i testi narrativi complessi sono quelli che nel 39,95% dei casi utilizzano proposizioni subordinate, seguiti dai testi giornalistici (32,46% dei casi), dai testi narrativi semplificati (34,94% dei casi) e, infine, dai testi giornalistici semplici, che si avvalgono delle clausole subordinate nel 26,22% dei casi. I dati rispecchiano il fatto che il genere giornalistico è più essenziale e diretto e quindi necessita di frasi brevi che possono trasmettere l'informazione in modo rapido, a differenza del genere narrativo che esige un numero maggiore di proposizioni subordinate per arricchire la trama del testo.

In conclusione, si può dedurre che la distribuzione delle clausole principali e subordinate varia sia a seconda del genere testuale che del grado di complessità: infatti, le frasi principali aumentano nel caso del genere giornalistico e nel caso dei testi semplici.

5.1.1 Il soggetto

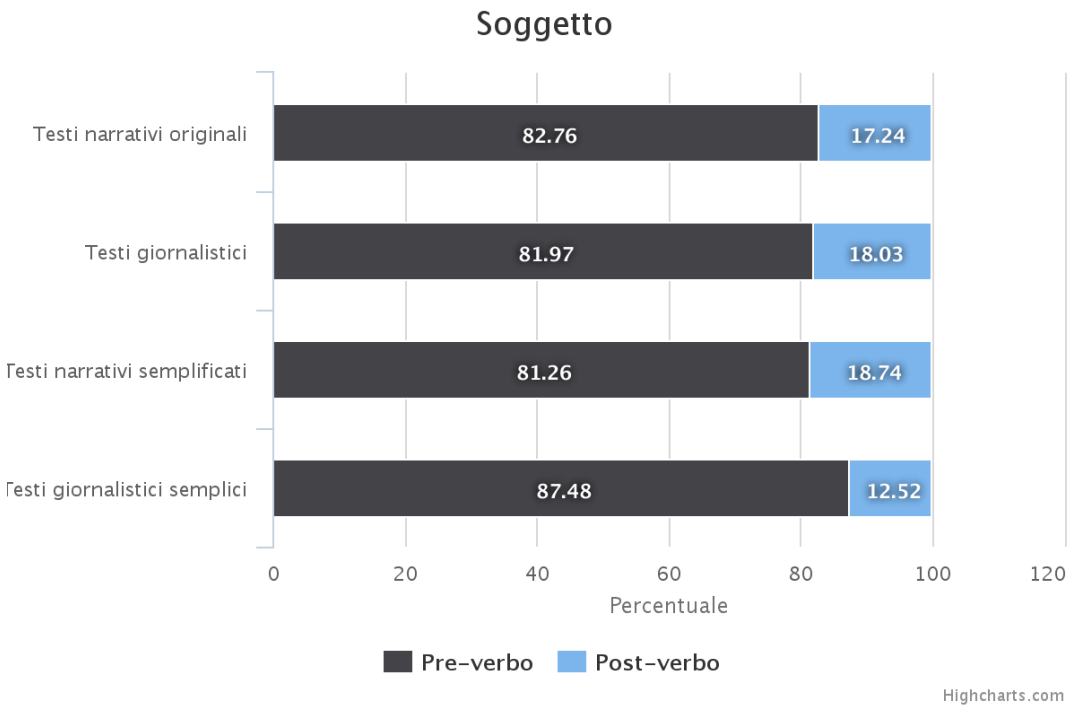

Figura 5.4: Dati statistici della posizione preverbale e postverbale del soggetto al verbo

I risultati mostrati nella Figura 5.4 indicano una netta prevalenza del soggetto in posizione preverbale, in media nell' 83,37% dei casi, rispetto alla posizione postverbale, in media nel 16,63% dei casi. La scelta predominante di collocare il soggetto in posizione preverbale, dunque, rispecchia la posizione non marcata del soggetto nella lingua italiana. Confrontando i risultati, è possibile notare che i dati statistici che si attengono maggiormente all'ordine canonico Soggetto-Verbo sono quelli dei corpora giornalistici semplici (87,48%) e dei corpora narrativi complessi (82,76%), seguiti poi dai corpora giornalistici (81,97%) e, infine, dai corpora narrativi semplificati (81,26%).
È interessante notare come, mentre nel passaggio dai testi giornalistici "complessi" a quelli semplici ci sia stato un aumento della collocazione del soggetto in posizione preverbale, dai testi narrativi complessi a quelli semplificati ci sia stata una leggera diminuzione del numero di casi di ordine non marcato del soggetto: questo può essere dovuto al fatto che i testi narrativi presentano solitamente una struttura più flessibile rispetto ai testi giornalistici e che, nel processo di semplificazione, non ci sia stato bisogno di anteporre il soggetto in posizione preverbale. Infatti, i testi narrativi presentano molti dialoghi, e questo può aver influenzato la scelta di utilizzare, ad esempio, enunciati tetici[1] come *"È arrivata Lucia"*, piuttosto che scegliere la forma non marcata *"Lucia è arrivata"*, "autorizzando" gli autori a lasciare il costrutto non canonico poiché risultava, in ogni caso, di facile comprensione o più adatto al contesto narrativo.

[1]Per approfondimenti, vedere il paragrafo "Soggetto" della sottosezione 2.2.3.

5.1.2 L'oggetto

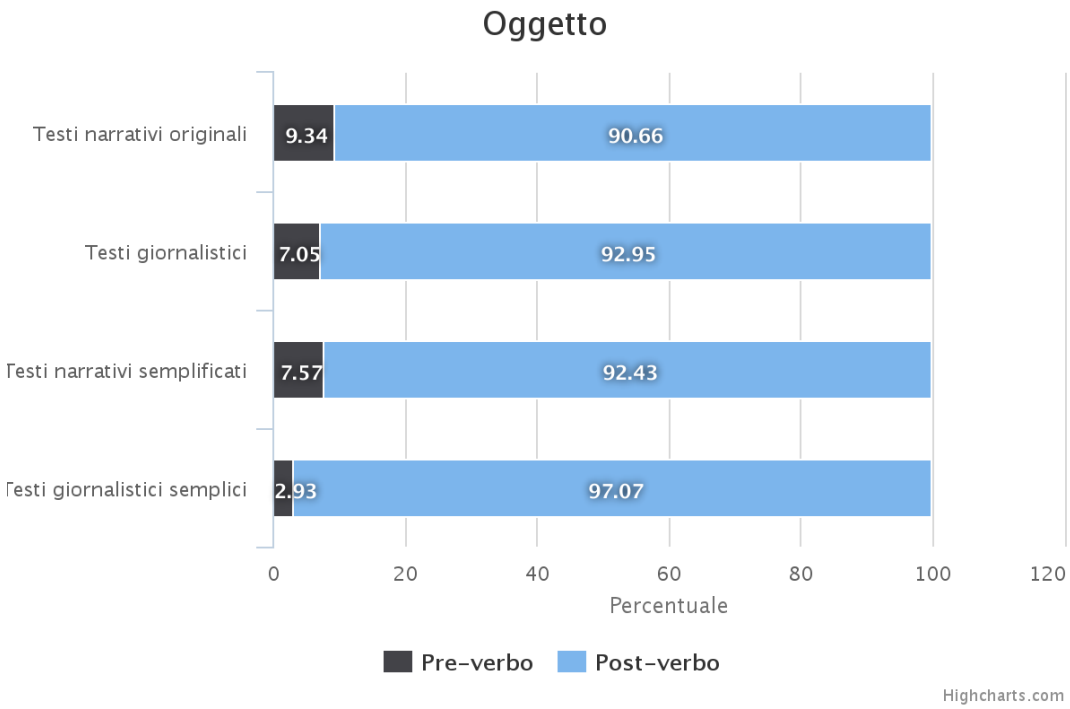

Figura 5.5: Dati statistici della posizione preposta e posposta dell'oggetto al verbo

I risultati mostrati nella Figura 5.5 evidenziano una chiara tendenza a collocare l'oggetto in posizione postverbale, in media nel 93,28% dei casi, piuttosto che in posizione preverbale, in media nel 6,72% dei casi. L'ordinamento canonico Verbo-Oggetto della lingua italiana[2], dunque, viene scelto nella maggior parte dei casi. In particolare, i testi giornalistici semplici sono quelli che per il 97,07% dei casi rispecchiano questo ordinamento, seguiti dai testi giornalistici (92,95% dei casi), poi dai testi narrativi semplificati (92,43% dei casi) e, infine, dai testi narrativi complessi (90,66% dei casi). Dunque, è possibile dedurre che i testi giornalistici si attengono maggiormente a seguire l'ordine canonico dell'oggetto in posizione postverbale, mentre nei testi narrativi aumentano il numero dei casi nei quali è possibile incontrare l'ordine marcato dell'oggetto in posizione preverbale. Inoltre, si può notare come la differenza tra i dati vari a seconda del grado del complessità, in quanto i testi semplici si avvalgono dell'oggetto in posizione preverbale in percentuale minore: infatti, il complemento oggetto preposto al verbo rappresenta un ordinamento più marcato rispetto al soggetto in posizione postverbale, ed è per questo che i testi semplici sono maggiormente ligi a seguire l'ordine canonico VO rispetto al caso del soggetto analizzato nella sottosezione precedente.

[2]Per approfondimenti, vedere il paragrafo "Oggetto" della sottosezione 2.2.3.

5.1.3 L'aggettivo

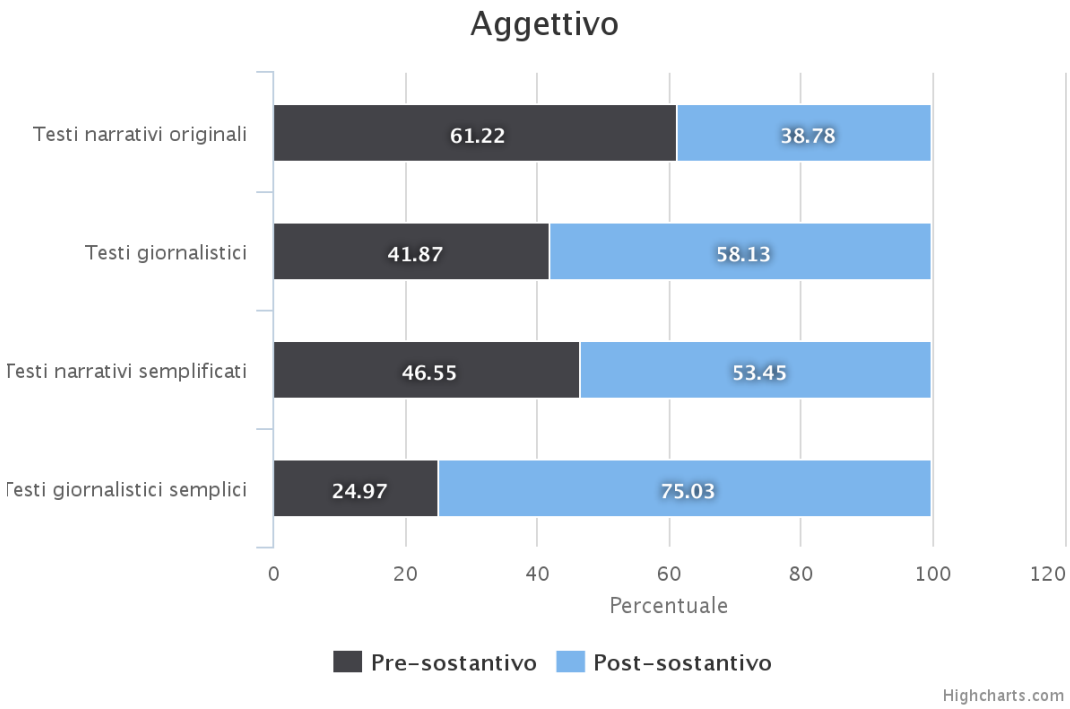

Figura 5.6: Dati statistici della posizione preposta e posposta dell'aggettivo al sostantivo

Nella Figura 5.6 possiamo notare i dati percentuali della disposizione dell'aggettivo nel sintagma nominale. In media, nel 56,35% dei casi l'aggettivo è posposto al sostantivo, mentre nel 43,65% dei casi viene anteposto: la preferenza di collocare l'aggettivo in posizione postsostantivale, dunque, rispecchia l'ordine non marcato nella lingua italiana sostantivo+aggettivo. Ciò può essere confermato nel passaggio dai testi complessi ai testi semplici, dove si può notare una diminuzione significativa dei casi di ordine marcato dell'aggettivo rispetto al sostantivo.
Tuttavia, è importante sottolineare che l'aggettivo presenta una posizione meno rigida rispetto a quella di soggetto e oggetto in quanto detiene una funzione lessicale: infatti, la tendenza è quella di porre l'aggettivo dopo il nome se l'intento è di attribuirgli una funzione restrittiva, ovvero se indica una qualità distintiva del soggetto rispetto ad altri della categoria di appartenenza (*una casa bella*), mentre assume una funzione descrittiva, cioè fornisce un dato oggettivo caratterizzante il nome a cui si riferisce, se viene preposto al nome (*una bella casa*). Dunque, anche se la posizione non marcata dell'aggettivo è dopo il nome cui si riferisce (Treccani, 2010a), bisogna tenere conto anche della funzione che deve svolgere nel contesto e delle restrizioni grammaticali che alle volte presentano[3].

[3]Esistono alcune categorie di aggettivi che presentano ordine fisso, come gli aggettivi alterati, che indicano nazionalità, che reggono un complemento o che derivano da un participio presente o passato. Per approfondimenti, vedere il paragrafo "Aggettivo" della sottosezione 2.2.3.

È possibile osservare che sono i testi di genere narrativo ad adottare maggiormente l'ordine marcato dell'aggettivo (61,22% dei casi nei testi narrativi complessi e 46,55% nei testi narrativi semplificati) rispetto ai testi di genere giornalistico, nei quali viene preferito l'ordine non marcato, anche se in misura diversa: quest'ultimi, infatti, presentano il 41,87% dei casi di ordine marcato per quanto riguarda i testi "complessi", mentre solo il 24,97% per quanto riguarda i testi semplici.
In conclusione, si può argomentare che l'ordine dell'aggettivo nel sintagma nominale dipende sia dal genere testuale che dal grado di complessità: infatti, è possibile notare un'alta percentuale di aggettivi in posizione presostantivale nel genere narrativo piuttosto che in quello giornalistico, ma anche una forte diminuzione di questa posizione a favore dell'ordine non marcato sostantivo+aggettivo nel passaggio dai testi complessi ai testi semplici.

5.1.4 L'avverbio

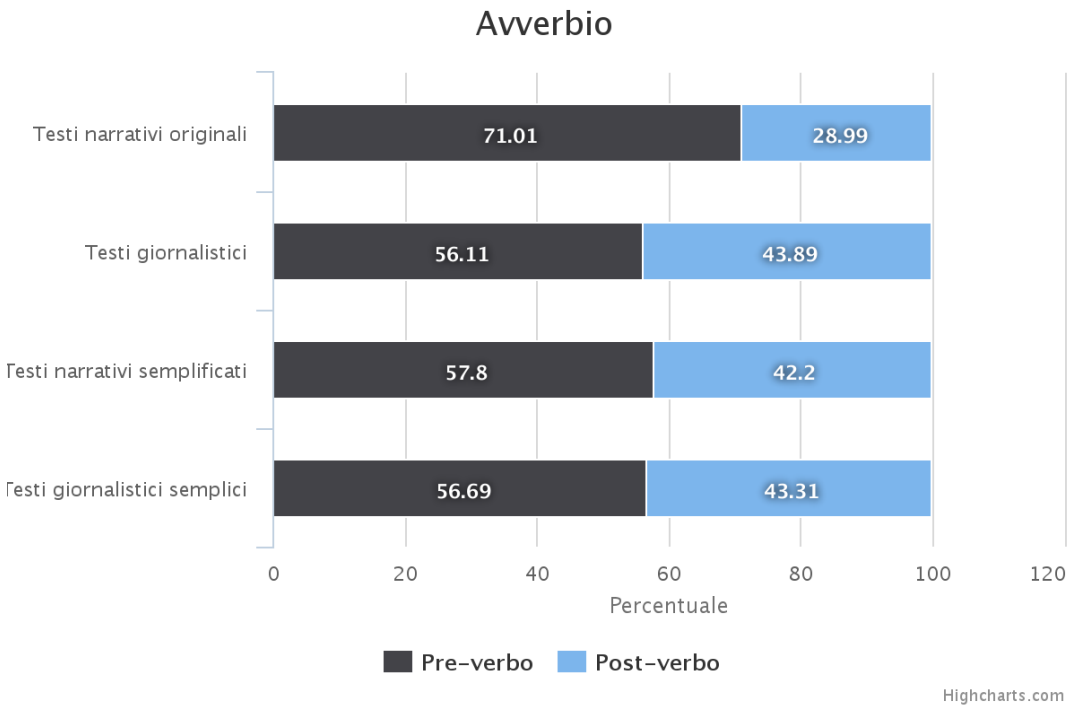

Figura 5.7: Dati statistici della posizione preposta e posposta dell'avverbio al verbo

Nella Figura 5.7 sono mostrati i dati statistici in percentuale della posizione preposta e posposta dell'avverbio rispetto alla testa verbale. In media, nel 60,40% dei casi gli avverbi vengono collocati in posizione preverbale, mentre nel 39,60% dei casi in posizione postverbale.
Gli avverbi che modificano verbi codificano per lo più informazioni relative al tempo, al luogo e alla modalità nella quale è svolta l'azione espressa dal verbo e, in un ordinamento non marcato, si trovano in posizione postverbale. Tuttavia, come è stato

argomentato per l'aggettivo, è necessario aggiungere che anche l'avverbio è legato a restrizioni grammaticali, e soprattutto a una funzionalità semantica, perciò la sua posizione può variare a seconda della grammaticalità della frase e del significato che deve ricoprire nel contesto: ad esempio, è possibile notare come nelle frasi *"Ho risposto semplicemente"* e *"Ho semplicemente risposto"* lo stesso avverbio in posizione postverbale modifichi il predicato con valore modale (primo caso), mentre in posizione preverbale venga usato come avverbio di tipo limitativo (secondo caso).
Si può notare come in tutti e quattro i corpora prevalga la posizione dell'avverbio preposta al verbo: in particolare, i testi narrativi complessi sono quelli che maggiormente utilizzano l'avverbio in posizione preverbale (71,01%), seguiti dai testi narrativi semplificati (57,80%), poi dai testi giornalistici semplici (56,69%) e, infine, dai testi giornalistici (56,11%).
Questo significa che in tutti e quattro i casi è preferito l'ordine marcato dell'avverbio, soprattutto nei testi narrativi, e in particolare è possibile notare che mentre nel passaggio dai testi complessi ai testi semplificati di genere narrativo i casi di ordine marcato diminuiscono, nel passaggio dai testi "complessi" a quelli semplici di genere giornalistico i casi aumentano (anche se in maniera esigua); come spiegato sopra, i risultati possono essere influenzati dalla funzione semantica che l'avverbio deve svolgere nel contesto del racconto e che, dunque, può comportare la scelta della sua collocazione in una posizione non canonica rispetto alla testa verbale.

5.1.5 La subordinata

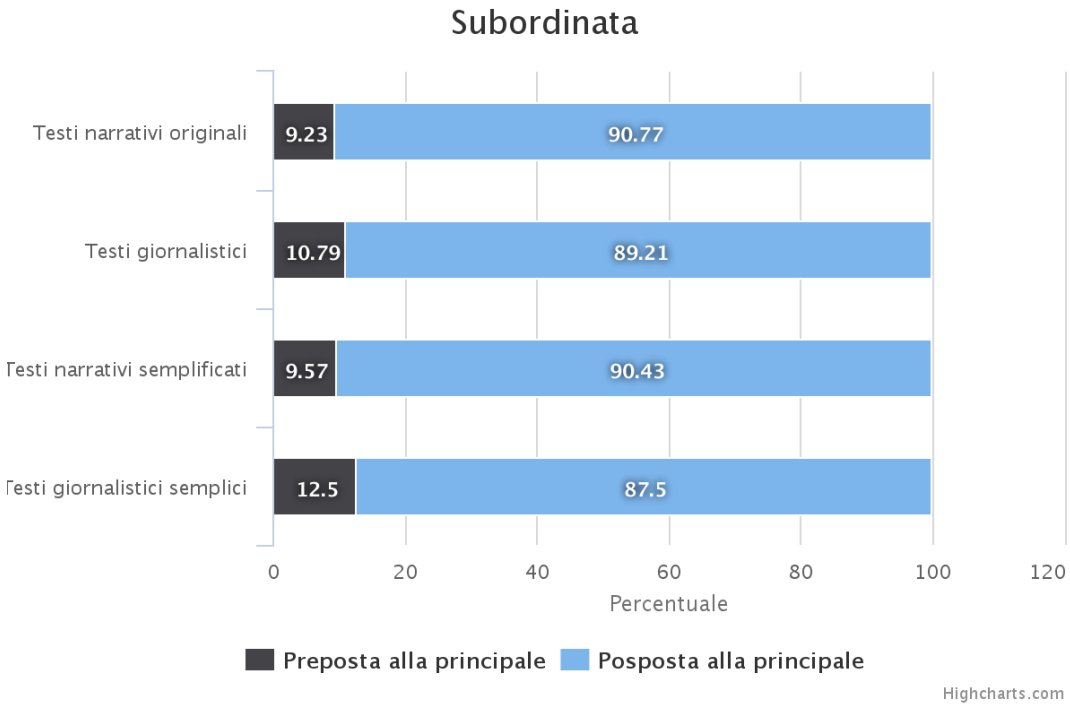

Figura 5.8: Dati statistici della posizione preposta e posposta delle subordinate alle principali

I risultati mostrati nella Figura 5.8 indicano una netta prevalenza della posizione posposta della subordinata alla principale, in media nell' 89,48% dei casi, rispetto alla posizione preposta, in media nel 10,52% dei casi.
Come descritto nella sezione 2.2.3 (paragrafo "Il caso delle subordinate"), l'ordinamento principale+subordinata è preferibile perché risulta più semplice per i princìpi di processamento e pianificazione, in quanto risulta un ordinamento più veloce da processare per la mente e comporta un impegno cognitivo minore da parte del parlante e dell'ascoltatore nella pianificazione del discorso (Hawkins, 1994). È possibile notare che il genere testuale giornalistico si avvale maggiormente della posizione preposta della subordinata (12,5% nel caso dei testi semplici e 10,79% nel caso dei testi "complessi") rispetto al genere narrativo (9,57% nel caso dei testi semplificati e 9,2% nel caso dei testi complessi).
In conclusione, si potrebbe dunque dedurre che in ambito giornalistico possa essere necessario, in un numero maggiore di casi, anteporre la subordinata poiché costituisce lo sfondo tematico dell'evento principale e conferisce la funzione di orientamento, collegamento tematico e introduzione per l'informazione nuova (Diessel, 2005); altrimenti, un'altra spiegazione potrebbe essere data dal fatto che il genere giornalistico, essendo più conciso, si avvale di subordinate più corte e che quindi, anche se preposte alla principale, non richiedono un maggiore impegno cognitivo da parte degli interlocutori. Tuttavia, i risultati dimostrano la preferenza a seguire l'ordine più velocemente processabile dalla mente principale+subordinata, che dunque potrebbe rivelarsi l'ordinamento che rispecchia maggiormente l'ordine non marcato.

5.2 Discussione

In questo capitolo sono stati analizzati i dati estratti dai quattro corpora descritti nel terzo capitolo: riepilogando, per il genere narrativo sono state scelte due raccolte contenenti testi per bambini in forma originale e semplificata (*"Terence"* e *"Teacher"*), che sono state suddivise in modo da ottenere testi solo originali e testi solo prodotti dai processi di semplificazione, mentre per il genere giornalistico sono state proposte una raccolta di testi giornalistici adeguati a un livello culturale medio-alto (*"Repubblica"*) e una raccolta di testi di facile lettura appositamente creati per un pubblico adulto con un livello di alfabetizzazione primitivo o con lievi disabilità intellettuali (*"Due Parole"*), che non derivano dunque da un processo di semplificazione, come nel primo caso. In questo modo, è stato possibile confrontare i dati sia in base alle differenze di genere testuale, sia in base al grado di complessità dei testi, cosicché potessero essere individuati quali fossero gli ordinamenti dei costituenti condizionati dal genere testuale, quali dal processo di semplificazione, e quali da entrambi.
Nel complesso si può constatare che, in media, i quattro corpora presentano una per-

centuale più alta di ordini non marcati dei costituenti. In particolare, i dati dei testi giornalistici semplici sono quelli che, per quattro analisi su cinque, rispecchiano maggiormente l' ordine canonico degli elementi, seguiti dai testi giornalistici "complessi", che solo nel caso dell'avverbio si attengono più volte a un ordine non marcato dei costituenti rispetto agli altri, poi dai testi narrativi semplificati e, infine, dai testi narrativi complessi, che invece risultano essere i testi con il maggior numero di casi di ordine marcato. Questi dati confermano il fatto che, solitamente, i testi giornalistici vengono scelti come esemplificazione grammaticale e strutturale dell'italiano standard, mentre i testi narrativi presentano maggiore flessibilità perché la narrazione è soggettiva e comporta, a seconda del contesto, scelte strutturali complesse.
Inoltre, è interessante notare che non tutti i dati statistici ottenuti rispecchiano i risultati attesi: infatti, nei casi di soggetto, oggetto e aggettivo si riscontrano sorprendentemente valori più simili tra i testi giornalistici "complessi" e i testi narrativi semplificati; nel caso dell'avverbio, i dati più simili si trovano tra i testi giornalistici "complessi" e quelli semplici; infine, nel caso della subordinata, i valori che maggiormente si avvicinano sono quelli tra i testi narrativi complessi e quelli semplificati.
In particolare, si può osservare che l'avverbio è la categoria che viene maggiormente condizionata dal genere testuale, l'aggettivo, l'oggetto e la subordinata sono quelle influenzate sia dalla differenza di genere che dal processo di semplificazione, mentre il soggetto è l'unica categoria che non sembrerebbe seguire un particolare andamento sia nel caso del genere che del grado di difficoltà; quest'ultimo caso, ancora in fase di studio, rientra in un cerchio più ampio di casi nei quali confluiscono un insieme di fattori che devono essere tenuti di conto nell'analisi complessiva.
Infine, facendo un confronto complessivo tra i dati statistici dei testi complessi e quelli dei testi semplici è possibile notare che, anche se i valori non sono molto simili, i dati dei testi semplici dei due generi testuali si avvicinano di più rispetto a quelli dei testi complessi: infatti, la media delle differenze dei valori dei testi semplici risulta 7,30%, mentre quella dei testi complessi 7,78%; inoltre, facendo un confronto tra i dati dei testi narrativi e quelli dei testi giornalistici, si può constatare che la media delle differenze dei valori dei testi giornalistici, 5,77%, è minore rispetto a quella dei testi narrativi, 6,3%: questo significa che, a prescindere dal grado di difficoltà, il genere giornalistico cerca di attenersi a un ordine dei costituenti fisso in un maggior numero dei casi rispetto al genere narrativo.
In conclusione, è possibile affermare che l'ordinamento dei costituenti varia sia in relazione al genere testuale che al grado di difficoltà, ma si è appurato che talvolta entrano in gioco una serie di fattori contestuali, come è stato visto nel caso delle molteplici funzionalità semantiche di aggettivo e avverbio, che possono essere la causa di una differente scelta dell'ordine degli elementi, a prescindere da ciò che sarebbe stato adeguato per il livello di complessità e il genere testuale.

Capitolo 6

Conclusioni

Questo elaborato è uno studio su un particolare fenomeno relativo alla complessità sintattica, ovvero l'ordinamento dei costituenti, che é stato analizzato all'interno di testi semplici e complessi di due varietà linguistiche, narrativa e giornalistica.
Inizialmente è stata fatta una panoramica della nozione di complessità linguistica e di come viene affrontata dalla ricerca linguistica moderna e, successivamente, sono stati presentati, per ogni livello linguistico, i criteri che possono essere utili al fine di classificare una lingua come dotata di minore o maggiore complessità strutturale. In particolare, per lo studio della complessità sintattica sono stati esaminati accuratamente i casi di ordinamento di soggetto, oggetto, aggettivo e avverbio rispetto alla testa sintattica e della subordinata rispetto alla frase reggente.
Sono stati quindi presentati i corpora sui quali sono stati analizzati, attraverso uno studio linguistico-computazionale, gli ordinamenti sopra descritti: infatti, grazie agli strumenti di annotazione automatica del testo e ai programmi sviluppati appositamente per il monitoraggio linguistico desiderato, é stato possibile estrarre i dati relativi alla posizione preposta o posposta di soggetto, oggetto e avverbio rispetto al verbo, di aggettivo rispetto al nome e della subordinata rispetto alla clausola principale.
Il confronto tra dati è avvenuto su due livelli, la variazione di genere e il grado di complessità: l'obiettivo era verificare quali ordinamenti dei costituenti dipendessero dal genere testuale, quali dal processo di semplificazione, e quali da entrambi. La previsione era che si potessero riscontrare ordinamenti simili dovuti al genere testuale, ma soprattutto, che si potessero ritrovare nel confronto tra testi complessi dati statistici molto simili così come nel confronto tra testi semplici: ci si aspettava che i testi complessi riportassero un maggior numero di casi di ordine marcato degli elementi, mentre i testi semplici si attenessero principalmente a un ordine non marcato.
È stato interessante notare come non tutti i dati statistici ottenuti rispecchiassero i risultati attesi: infatti, nei casi di soggetto, oggetto e aggettivo sono stati riscontrati sorprendentemente valori più simili tra i testi giornalistici complessi e i testi narrativi

semplificati; nel caso dell'avverbio i dati più simili sono stati trovati tra i testi giornalistici complessi e quelli semplici; infine, nel caso della subordinata, i valori che maggiormente si sono avvicinati sono stati quelli tra i testi narrativi complessi e quelli semplificati. In particolare, si è appurato che l'avverbio è la categoria che maggiormente è stata influenzata dalla differenza di genere, l'aggettivo, l'oggetto e la subordinata sono state quelle condizionate sia dalla differenza di genere che dal processo di semplificazione, mentre il soggetto è l'unica categoria che non è sembrata seguire un particolare andamento sia nel caso del genere che del grado di difficoltà.
Nel complesso, è stato possibile constatare che i testi giornalistici semplici sono risultati i testi che maggiormente si sono attenuti all'ordinamento canonico degli elementi, mentre i testi narrativi complessi sono stati quelli che hanno utilizzato con maggiore frequenza un ordinamento marcato. Inoltre, facendo un confronto complessivo tra i dati statistici dei testi complessi e quelli dei testi semplici, è stato osservato che i dati dei testi semplici dei due generi testuali erano più vicini rispetto a quelli dei testi complessi: infatti, la media delle differenze dei valori dei testi semplici è risultata 7,30%, mentre quella dei testi complessi 7,78%; dopodiché, facendo un confronto tra i dati dei testi narrativi e quelli dei testi giornalistici, si è potuto constatare che la media delle differenze dei valori dei testi giornalistici, 5,77%, fosse minore rispetto a quella dei testi narrativi, 6,3%; questo significa che, a prescindere dal grado di difficoltà, il genere giornalistico si è attenuto a un ordine dei costituenti fisso con più alta frequenza rispetto al genere narrativo. Questi dati confermano il fatto che solitamente i testi giornalistici vengono scelti come esemplificazione grammaticale e strutturale dell'italiano standard, in quanto il genere stesso richiede linearità, schematicità e chiarezza per la funzione che deve adempiere, mentre i testi narrativi vengono indicati come quelli che presentano maggiore flessibilità, perché la narrazione, essendo soggettiva, viene sottoposta a molteplici variazioni in relazione al contesto narrativo.
Casi di ordinamento interessanti sono stati riscontrati durante le analisi di aggettivo e avverbio, del soggetto nel genere narrativo, e delle subordinate in ambito giornalistico. Nel primo caso, è stato sottolineato come le posizioni di aggettivo e avverbio fossero più strettamente correlate alla funzionalità semantica piuttosto che grammaticale: una posizione marcata di queste categorie, infatti, può risultare indispensabile se il significato della frase richiede che venga assunta una posizione differente di esse rispetto all'ordine canonico; nella seconda circostanza, è stato osservato che nel passaggio dai testi narrativi complessi a quelli semplificati ci fosse stata una leggera diminuzione del numero di casi di ordine non marcato del soggetto: ancora in fase di studio, la spiegazione potrebbe essere ricollegata al fatto che il genere narrativo presenta una struttura più flessibile e che quindi, nel processo di semplificazione, non ci fosse stato bisogno di anteporre il soggetto in posizione preverbale. Infine, nell'ultimo caso è stata rilevata una più alta percentuale di subordinate preposte alla principale: la motivazione

si potrebbe ritrovare nei princìpi pragmatici della struttura dell'informazione, oppure nella lunghezza delle subordinate poiché, essendo il genere giornalistico conciso, si potrebbe avvalere di subordinate più corte che, quindi, potrebbero non richiedere un maggiore impegno cognitivo da parte degli interlocutori.
In conclusione, si può dedurre che l'ordinamento dei costituenti varia in relazione al genere testuale e al grado difficoltà, ma talvolta entrano in gioco una serie di fattori contestuali che possono essere la causa di una differente scelta dell'ordine degli elementi, a prescindere da ciò che sarebbe stato adeguato per il livello di complessità e il genere testuale. I limiti che sono stati riscontrati durante alcuni casi di analisi possono destare interesse verso una nuova direzione della ricerca, nella quale si potrebbero studiare i casi di ordinamento di aggettivo e avverbio tenendo conto delle spiegazioni legate alla semantica, in modo da poter confrontare la distribuzione di essi all'interno della frase con la funzionalità lessicale svolta nel contesto; altrimenti, si potrebbero approfondire i casi di ordinamento della subordinata tenendo conto della lunghezza della stessa nel caso di anteposizione e posposizione alla principale, in modo da verificare l'effettiva influenza dei princìpi pragmatici e di processamento in relazione alla collocazione e alla lunghezza della proposizione.

Bibliografia

Attardi G.; Dell'Orletta F.; Simi M.; Turian J. (2009). Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009*, (Reggio Emilia, Italia, Dicembre 2009).

Baroni M.; Bernardini S.; Comastri F.; Piccioni L.; Volpi A.; Aston G.; Mazzoleni M. (2004). Introducing the la repubblica corpus: A large, annotated, tei(xml)-compliant corpus of newspaper italian. In *Proceedings of LREC 2004*, Lisbona, Portogallo, Maggio 2004.

Berruto G. (2004). *Nozioni di linguistica generale*. Linguistica e linguaggi. Liguori, Napoli.

Brunato D.; Dell'Orletta F.; Venturi G.; Montemagni S. (2015). Design and annotation of the first italian corpus for text simplification. In *Proceedings of LAW IX - The 9th Linguistic Annotation Workshop*, (Denver, Colorado, Giugno 2015).

Calzolari N.; Lenci A. (2004). Linguistica computazionale - strumenti e risorse per il trattamento automatico della lingua. *Mondo Digitale*, **2**, 56–69.

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

Corpina B. (2009). *Topic e Focus in Hdi. Strategie a confronto e analisi dei testi*. Università degli Studi di Roma Tre, Roma (Italia).

Crossley S. A.; Allen D. B.; McNamara D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, **23**(1), 84–102.

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

Dell'Orletta F.; Montemagni S.; Venturi G. (2011). Read–it: Assessing readability of italian texts with a view to text simplification. In *SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, Edimburgo, UK, Luglio 2011.

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 (RANLP 2013)*, (Hissar, Bulgaria, Settembre 2013).

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

Dryer M. S. (2009). The branching direction theory of word order correlations revisited. In *Universals of Language Today*. A cura di Scalise S., E.Magni, A.Bisetto, volume 76 di *Studies in Natural Language and Linguistic Theory*, pp. 185–207. Springer, Berlin.

DueParole (2002). Due parole, mensile di facile lettura. http://www.dueparole.it/. Ultima visita: 28/10/2015.

Ferguson C. (1982). Simplified registers and linguistic theory In *Exceptional Language and Linguistics*. A cura di Obler L. K., Menn L., pp. 49–66. Academic Press, New York.

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

Gallissot R.; Kilani M.; Rivera A. (2001). *L'imbroglio etnico in quattordici parole-chiave*. Ed. Dedalo srl, Bari.

Gell-Mann M.; Ruhlen M. (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(42), 17290–17295.

Gibson E.; Piantadosi S.; Brink K.; Bergen L.; Lim E.; Saxe R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, **24**(7), 1079–1088.

Givón T. (1979). From discourse to syntax: Grammar as a processing strategy. *Syntax and Semantics*, (12), 81–112.

Hawkins J. A. (1994). *A performance theory of order and constituency* in Cambridge studies in Linguistics. Numero 73. Cambridge University Press., Cambridge.

Hawkins J. A. (2009). An efficiency theory of complexity and related phenomena. In *Language Complexity as an Envolving Variable*. A cura di Sampson G., Gil D., Trudgill P., volume 13 di *Studies in the Evolution of Language*, pp. 252–268. Oxford University Press, Oxford.

Hellö S. (2005). *La gestualità forma alternativa di comunicazione*. Università di Lund, Lund (Svezia).

ItaliaNLPLab (2014). Read-it. documentazione demo online. http://www.italianlp.it/wp-content/uploads/2014/06/Demo-Documentation.pdf. Ultima visita: 4/11/2015.

Kusters W. (2003). *Linguistic Complexity The Influence of Social Change on Verbal Inflection* in LOT Dissertation Series. Numero 77. LOT, Utrecht.

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

Marinelli L. (2015). *Studio della complessità e della semplificazione linguistica a partire da un'analisi computazionale di un corpus parallelo di testi italiani*. Università di Pisa, Pisa (Italia).

McWhorter J. H. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology*, **5**, 125–166.

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

Nesselhauf N. (2005). *Corpus Linguistics: A Practical Introduction*. http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf. Ultima visita: 28/10/2015.

Radić-bojanić B. (2006). Fragmentation/integration and involvement/detachment in chatroom discourse. *Skase Journal of Theoretical Linguistics*, **3**(1), 38–46.

Tavosanis (2009). *Linguaggio e scrittura*. http://linguaggiodelweb.blogspot.it/2009/04/ma-si-puo-misurare-la-complessita.html. Ultima visita: 28/10/2015.

Terence_Corsortium (2012). Story simplification: User guide. Restricted Distribution.

Treccani (2010a). voce *"Aggettivi"*. http://www.treccani.it/enciclopedia/aggettivi_(altro)/. Ultima visita: 29/10/2015.

Treccani (2010b). voce *"Focalizzazioni"*. http://www.treccani.it/enciclopedia/focalizzazioni_(Enciclopedia_dell'Italiano)/. Ultima visita: 29/10/2015.

Treccani (2011). voce *"Ordine degli elementi"*. http://www.treccani.it/enciclopedia/ordine-degli-elementi_(Enciclopedia_dell'Italiano)/. Ultima visita: 28/10/2015.

Tweissi A. I. (1998). The effects of the amount and type of simplification on foreign language reading comprehension. *Reading in a Foreign Language*, **11**(2), 191–204.

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.

Wasow T. (2002). *Postverbal Behavior*. CSLI Publications, Stanford.

Wikipedia (2015a). voce *"Proprietà della lingua"*. https://it.wikipedia.org/wiki/Propriet%C3%A0_della_lingua. Ultima visita: 28/10/2015.

Wikipedia (2015b). voce *"Pragmatica"*. https://it.wikipedia.org/wiki/Pragmatica. Ultima visita: 28/10/2015.

Wikipedia (2015c). voce *"Linguistica computazionale"*. https://it.wikipedia.org/wiki/Linguistica_computazionale. Ultima visita: 28/10/2015.

Ringraziamenti

Ringrazio i miei genitori, *Maria Rosaria* e *Giampaolo*, per tutto quello che fanno per me ogni giorno, per avermi permesso di studiare, per i loro sforzi e l'amore che mi trasmettono incondizionatamente, da sempre. Il mio traguardo di oggi, è merito loro.

Ringrazio i miei fratelli, *Luca* e *Chiara*, i miei punti fermi, che mi hanno sempre tenuta per mano e mi hanno sempre seguita e accompagnata in ogni strada che intraprendessi, con amore e giudizio. Se oggi sono qui, lo devo anche a loro.

Ringrazio le mie nonne, *Lidia* e *Amalia*, che mi sono state vicine, mi hanno sempre dimostrato un affetto infinito e non hanno mai perso occasione di ribadirmi l'importanza dello studio. Sono felice che oggi possano assistere al mio traguardo, da loro tanto sognato.

Ringrazio i miei zii, *Elisabetta*, *Paolo* e *Francesco*, che mi hanno sempre sostenuta in tutti questi anni e si sono preoccupati per me e per i miei studi.

Ringrazio il mio fidanzato, *Federico*, l'angelo che ogni giorno mi sostiene e mi guida con amore e pazienza e che, nonostante la lontananza, riesce sempre a trasmettermi quanto creda in me, tanta serenità e una forza incredibile. Sono felice di condividere questo mio traguardo insieme.

Ringrazio tutti colori che in questi anni hanno incrociato il mio percorso e condiviso con me anche solo un piccolo passo di questo sentiero, ma soprattutto, un doveroso ringraziamento va ai miei amici di sempre, *Marco*, *Anna*, *Ilaria*, *Vanessa*, *Giulia*, *Andrea*, *Giulia*, *Federica*, *Marta*, *Elena*, *Parge* e *Isabella*, che mi sono stati vicini, mi hanno sostenuta e hanno creduto in me, sempre.

Ringrazio i miei compagni di studi, *Davide*, *Gabriele* e *Simona*, con i quali ho passato quattro anni universitari indimenticabili, tra amicizia e risate.

Ringrazio i miei amici dell'associazione *Four For Africa*, in particolar modo *Ilaria*, *Diana*, *Giulia* e *Francesca*, ma soprattutto *Andrea* che, nonostante alti e bassi, è sempre stato una guida e un grande punto di riferimento per me.

Ringrazio i miei amici capoeiristi sparsi nel mondo, la famiglia de ouro, ma soprattutto *Giacomo*, *Ginevra*, *Philip*, *Mauro* ed *Emily*, che mi hanno accolta, cresciuta e accompagnata in tutti questi anni.

Ringrazio i miei amici della *Comunità di Sant'Egidio*, ormai una seconda famiglia, e i miei amici di strada, *Lino*, *Marcello*, *Giorgio* e *Christian*, che ogni lunedì, a loro modo e nel loro piccolo, non hanno mai mancato di chiedermi come andavano gli esami e quando sarebbero dovuti andare a comprarsi il vestito per la mia laurea.

Ringrazio i miei splendidi colleghi, *Sara*, *Yessica*, *Gaia*, *Simone*, *Mone*, *Diddo*, *Michele*, *Betti*, *Monni*, *Lucy*, *Veronica* e *Giovanni*, che in tutti questi anni si sono preoccupati per me e non hanno mai mancato di strapparmi un sorriso e starmi vicina.

Infine, ultimo ma non certo per importanza, il mio ringraziamento più sentito va all'Istituto di Linguistica Computazionale *"A.Zampolli"* del CNR di Pisa, in particolar modo al mio relatore dott. *Felice Dell'Orletta* e alla dott.ssa *Dominique Brunato*, due persone eccezionali che mi hanno dato l'opportunità di intraprendere questo percorso con loro e che mi hanno seguita e guidata con professionalità e amichevolezza fino a oggi, rendendo possibile la realizzazione di questa tesi.