UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

TESI DI LAUREA

L'italiano amministrativo dagli atti al Web: una risorsa per la semplificazione automatica

Relatore:
Prof. Alessandro Lenci

Correlatore:
Prof. Felice Dell'Orletta

Candidato:
Michele Papucci

ANNO ACCADEMICO 2020/2021

Indice

Introduzione

> Il brigadiere è davanti alla macchina da scrivere. L'interrogato, seduto davanti a lui, risponde alle domande un po' balbettando, ma attento a dire tutto quel che ha da dire nel modo più preciso e senza una parola di troppo: "Stamattina presto andavo in cantina ad accendere la stufa e ho trovato tutti quei fiaschi di vino dietro la cassa del carbone. Ne ho preso uno per bermelo a cena. Non ne sapevo niente che la bottiglieria di sopra era stata scassinata". Impassibile, il brigadiere batte veloce sui tasti la sua fedele trascrizione: "Il sottoscritto essendosi recato nelle prime ore antimeridiane nei locali dello scantinato per eseguire l'avviamento dell'impianto termico, dichiara d'essere casualmente incorso nel rinvenimento di un quantitativo di prodotti vinicoli, situati in posizione retrostante al recipiente adibito al contenimento del combustibile, e di aver effettuato l'asportazione di uno dei detti articoli nell'intento di consumarlo durante il pasto pomeridiano, non essendo a conoscenza dell'avvenuta effrazione dell'esercizio soprastante".

Così scrive Italo Calvino nel suo articolo *L'Antilingua* per *Il Giorno* nel 1965[1]. Calvino denuncia quello che definisce il "terrore semantico" caratteristico della lingua amministrativa, «cioè la fuga di fronte a ogni vocabolo che abbia di per se stesso un significato, come se "fiasco", "stufa", "carbone" fossero parole oscene, come se "andare", "trovare", "sapere" indicassero azioni turpi».

Da allora sono stati fatti diversi tentativi di semplificare il linguaggio della pubblica amministrazione, ma nonostante i libri e le pubblicazioni, dopo più di sessant'anni il linguaggio amministrativo è ancora percepito come oscuro e complesso.

Ad oggi grazie alle nuove tecnologie di *machine learning* è possibile l'addestramento di modelli per la semplificazione automatica del testo, ed è quindi teoricamente possibile risolvere il problema della complessità linguistica con una soluzione tecnologica. Per farlo

[1] Italo Calvino (1980). «L'antilingua». In: *Una pietra sopra*. Milano: Mondadori, pp. 150–155

è però necessario avere a disposizione una grande quantità di dati per la costruzione di *corpora di addestramento* per le reti neurali.

L'obbiettivo del presente studio è la costruzione di un corpus parallelo per la lingua della Pubblica Amministrazione italiana che può servire alla creazione di modelli di semplificazione linguistica automatica.

Nel primo capitolo saranno illustrate le conoscenze presenti nella letteratura scientifica per i principali settori di interesse del presente studio, nel secondo verrà illustrato il procedimento di costruzione del corpus parallelo, dalla scelta dei documenti alla semplificazione dei testi, e nel terzo saranno commentate le analisi effettuate sul corpus e sulla sua annotazione .

1. Stato dell'arte

In questo capitolo sarà fatta una breve introduzione alle nozioni generali presenti nella letteratura scientifica sulla complessità linguistica, sulle formule di leggibilità e sugli approcci di semplificazione basati su corpora.

1.1 Complessità Linguistica

La complessità linguistica può essere distinta, a seconda del punto di vista assunto, in *assoluta* o *relativa*. La prospettiva assoluta è oggettiva e basata sulla teoria linguistica, mentre la prospettiva relativa è soggettiva e si basa sulle evidenze empiricamente raccolte dall'esperienza dei parlanti (Blache, 2011). Nei prossimi paragrafi sarà esplorato questo secondo punto di vista, evidenziando quei fattori che specificatamente hanno un peso nell'aumentare la difficoltà di lettura di un testo.

1.1.1 Complessità Lessicale

La comprensione di un testo inizia dall'operazione che permette di associare ad ogni rappresentazione simbolica il suo significato nella memoria. Questa operazione è detta *word parsing* (Seidenberg, 1989). Sono diversi i fenomeni che possono facilitare o rallentare questa operazione, e di conseguenza, rendere una parola più o meno leggibile. Tra questi fenomeni, i più conosciuti sono:

- Lunghezza della parola: più una parola è lunga, più lentamente la si riconosce. Questa regola vale in maniera particolare per i nomi e meno per i verbi (Colombo e Burani, 2002);

- *Word frequency effect*: la velocità e la facilità con cui si riconosce una parola aumenta all'aumentare della frequenza della parola stessa, di conseguenza le parole meno frequenti sono più difficili da riconoscere (Segui et al., 1982). La parola *sto-*

ria che nel più grande corpus italiano[1] appare con una frequenza pari a 176592 sarà quindi più facilmente riconosciuta e interpretata della parola *giusnaturalismo* che nello stesso corpus appare solo 67 volte;

- *Root frequency effect*: la velocità con cui si riconosce una parola aumenta se la radice della parola è molto produttiva, ovvero se dalla stessa radice sono derivate parole ad alta frequenza. È quindi possibile per una parola con una bassa frequenza essere riconosciuta velocemente se la stessa radice produce parole che invece hanno un'alta frequenza (Colombo e Burani, 2002);

Oltre a questi fenomeni ne esistono altri che incidono in maniera più o meno importante sulla velocità di riconoscimento lessicale, come: l'*età di acquisizione*: che rappresenta l'età alla quale una data parola è stata imparata; *la concretezza*: le parole di significato astratto sono più complesse di quelle di significato concreto; la *disponibilità di contesto*: riconoscere una parola al di fuori di qualsaisi contesto è più difficile che riconoscerla inserita in una frase (Colombo e Burani, 2002).

1.1.2 Complessità Sintattica

Al livello della frase viene introdotto un nuovo meccanismo di comprensione detto *sentence parsing*. Lo scopo di questo processo è assegnare la corretta analisi sintattica a una frase, processo che secondo molti psicolinguisti è essenziale per capire il significato della frase letta o ascoltata (Kempen, 1998). Davanti ad alcune strutture grammaticali complesse o a frasi molto lunghe quest'operazione può rallentare o non funzionare bene. Il caso più famoso nel quale il parser umano ha difficoltà è quello di una frase ambigua, ovvero una frase che ha due (o più) letture a seconda dei ruoli che vengono assegnati alle componenti della frase.

(1) La bambina copre la bambola con la borsa

La frase (1) può essere interpretata sia come "La bambina copre con una borsa la bambola" sia come "La bambina copre con un oggetto indefinito la bambola che ha una borsa".

[1] Verena Lyding et al. (2013). *PAISÀ Corpus of Italian Web Text*. Eurac Research CLARIN Centre

Su come il parser sintattico umano riesca a disambiguare queste frasi sono stati proposti diversi modelli (Frazier, 1979; Kimball, 1973; Gibson et al., 1996; per citarne alcuni) ma ciò che è importante notare è che l'ambiguità è fattore di complessità sintattica (Ferstl e d'Arcais, 1999).
Altri fattori noti sono:

- la *profondità dell'albero sintattico*[2]: tra i vari fattori noti di difficoltà collegati ad esso, il più semplice è la *profondità dell'intero albero*. Quindi più un albero sintattico è profondo, più la frase viene percepita complessa (Dell'Orletta, Montemagni et al., 2011);

- la *lunghezza dei dependency link*, calcolata come il numero di parole tra la *head* sintattica e il *dependent*, ovvero tra la testa di un sintagma e gli altri elementi che dipendono da essa. Più questo valore è alto, più una frase è sintatticamente complessa (Gibson, 1998);

- Il rapporto tra le frasi subordinate e le principali, all'aumentare del numero di subordinate rispetto alle principali la complessità sintattica aumenta (Dell'Orletta, Montemagni et al., 2011).

1.2 Complessità del linguaggio burocratico italiano

Che il linguaggio burocratico sia complesso è fatto noto a tutti, Cortelazzo nel suo *Il linguaggio amministrativo. Principi e pratiche di modernizzazione* riporta come già nel 1540 Benedetto Varchi, studioso fiorentino, lamentasse di un ≪gergo a uso di lingua furfantina molto strano≫[3].
Nella storia recente, in Italia, sono stati fatti diversi tentativi di promozione di un linguaggio amministrativo più semplice: nel 1963 l'allora Ministro per la Funzione Pubblica

[2]Per albero sintattico si intende una forma di annotazione sintattica del testo, dove i nodi dell'albero sono composti dai diversi costituenti sintattici (sintagmi nominali, verbali, ecc.) e dai loro sottocostituenti opportunatamente gerarchizzati. (Lenci et al., 2005, p.214)

[3]Michele A. Cortelazzo (2021). *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*. Carrocci editore

Sabino Cassese pubblica il *Codice di Stile delle comunicazioni scritte ad uso delle pubbliche amministrazioni*; nel 1997 un gruppo di ricerca composto da esperti di legge, linguisti e pubblici ufficiali, coordinati dal giurista Alfredo Fioritto pubblica il *Manuale di Stile. Strumenti per semplificare le amministrazioni pubbliche*; nel 2011 l'Istituto ITTIG-CNR insieme all'Accademia della Crusca pubblicano la *Guida alla redazioni degli atti amministrativi. Regole e suggerimenti* e infine nel 2021 Michele Cortelazzo pubblica *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*.

Cortelazzo identifica come tratti salienti della scrittura amministrativa «la complessità, l'ampollosità, la ridondanza e la stereotipia», caratteristiche che si presentano a livello sintattico e lessicale.

Per quanto riguarda la complessità lessicale, alcuni dei fattori sono:

- tecnicismi collaterali, ovvero non propri del linguaggio amministrativo, ma caratteristici di altri ambiti settoriali. Non sono usati per effettiva necessità ma per elevare il registro del testo o della comunicazione (es: *diniego*, *inottemperanza*, *apporre*);
- sigle e abbreviazioni, che risultano totalmente oscure a chi non ha già un certo grado di familiarietà con l'ambito specialistico in esame (es: *ISEE*, *PRG*);
- forestierismi, ovvero l'utilizzo di prestiti linguistici da altre lingue (es: *lockdown*, *green pass*, *privacy*, *leader*);
- perifrasi costituite da verbo e sintagma nominale o sintagma preposizionale, ovvero una tendenza all'amplificazione, con perifrasi ridondanti e stereotipiche come: *dare corso all'apertura* al posto di *aprire*, *effettuare una verifica* al posto di *verificare*.

Per la complessità sintattica alcuni dei fattori sono:

- costrutti impersonali, nati dalla necessità di spersonalizzare e rendere collettive le comunicazioni istituzionali (es: *si allega*, *si porta a conoscenza del*, *viene trasmessa e comunicata*);
- nominalizzazione, processo di derivazione di un verbo in sostantivo o aggettivo (Treccani, n.d.) che contribuisce alla spersonalizzazione, passando la funzione predicativa dal verbo ai nomi (es: *ricevimento*, *accesso*, *trascrizione*);

- forme nominali del verbo, cioè verbi coniugati nei modi non finiti (participio, infinito e gerundio). Tra questi i più usati sono il participio presente «i gruppi di studenti *frequentanti* le lezioni», il gerundio «la fissazione di un termine, dalla cui scadenza, *difettando* l'emanazione del provvedimento, è possibile esplicare lecitamente l'attività» e il participio passato «*considerate* le funzioni assegnate ai Dirigenti centrali»;

- incisi, che nascono dalla tendenza di comporre frasi uniche, alle quali vengono poi aggiunte informazioni con frasi inserite tra virgole, trattini o parentesi. Gli incisi ostacolano la leggibilità del testo, interrompendo la linearità della frase: «Dichiara altresì di essere informato, *ai sensi e per gli effetti di cui all'art. 13 della legge 30 giugno 2003*, che i dati personali saranno trattati, *anche con strumenti informatici*, esclusivamente nell'ambito del procedimento per il quale la presente dichiarazione viene resa»

V. Tabella 2.4 per le motivazioni di complessità lessicale e sintattica individuate durante il lavoro di semplificazione del corpus.

1.3 Formule di leggibilità

Tradizionalmente la valutazione della leggibilità di un testo è stata ottenuta attraverso delle *formule di leggibilità*, ovvero delle equazioni matematiche che prendendo in considerazione vari parametri (solitamente non più di due) restituiscono un valore di difficoltà del testo (Brunato, 2015).

La prima formula di questo tipo è la *formula Winnetka* di Vogel e Washburne (1928). La formula nasce da un campione di 700 libri scelti tra le preferenze dei bambini partecipanti al questionario. I libri sono stati misurati secondo caratteristiche sia lessicali che morfosintattiche, e ognuna di esse è stata correlata con il livello medio di lettura dei bambini (attestato con lo Standford Achievement test). Le quattro caratteristiche con il miglior valore di correlazione (Numero di parole tipo in campione di 1000 parole; numero totale di preposizioni in un campione di 1000 parole; numero totali di parole non comuni, ovvero

non contenute nella Thorndike's list; numero di frasi semplici in un campione di 75 frasi) sono entrate in un'equazione con lo scopo di predire il livello di lettura necessario per leggere e capire un libro.

Successive alla Winnetka sono state teorizzate diverse formule di leggibilità, tra le quali la *formula Dall-Chall* (Chall e Dale, 1995), la *formula Flesch* (Flesch, 1948) e l'*indice GULPEASE* per l'italiano (Lucisano e Piemontese, 1988).

Il problema principale di queste formule è che sono una soluzione semplicistica ad un fenomeno sfaccettato e complesso come quello della lettura. Secondo Just e Carpenter (1980) ≪Non esiste solo un modo di leggere. Il modo in cui si legge cambia a seconda di chi legge, di cosa legge e del perché lo legge. [...] Gli obbiettivi del lettore sono probabilmente il fattore più determinante del processo di lettura.≫ (traduzione mia)

Alcuni studi hanno dimostrato come la riscrittura di testi secondo i requisiti delle formule tradizionali non migliorava la comprensione nei giovani studenti (Green e Olsen, 1988).

1.3.1 READ-IT

READ-IT è il primo strumento di valutazione di leggibilità basato su tecniche di Natural Language Processing (NLP) per l'italiano. Fa parte di una nuova generazione di formule di leggibilità, che, dai primi anni duemila, è emersa sfruttando le capacità computazionali dei sistemi di NLP. Infatti, grazie a questi sistemi di ultima generazione, è possibile effettuare molte misurazioni lessicali e sintattiche, con le quali calcolare il voto di leggibilità (Dell'Orletta, Montemagni et al., 2011).

READ-IT è stato pensato come strumento di supporto ad un processo di semplificazione automatica e prende in esame caratteristiche testuali divise in quattro classi:

- Caratteristiche *raw text*, ovvero quelle usate tradizionalmente dalle metriche di leggibilità. Comprendono quei valori risultanti dal processo di *tokenizzazione* (dividere in *token*[4]) come la *lunghezza media della frase* espressa in token e la *lunghezza media di parola* espressa in caratteri;

[4]I token sono le unità di base del testo digitale, che raggruppano oltre alle parole ortografiche tradizionali anche numeri, sigle, segni di punteggiatura, nomi propri, ecc. (Lenci et al., 2005)

- Caratteristiche lessicali, ovvero che hanno a che fare col vocabolario interno del testo, come la *Type/Token Ratio*[5] e la percentuale di parole appartenenti al Vocabolario di Base di De Mauro[6];

- Caratteristiche morfo-sintattiche come la *densità lessicale*, ovvero il rateo di parole significative (verbi, nomi, aggettivi e avverbi) rispetto al totale delle parole, e la distribuzione dei *tempi verbali*, un indice specifico per l'italiano che è stato introdotto per la prima volta in READ-IT;

- Caratteristiche sintattiche come la *profondità media dell'albero sintattico*, la *distribuzione di proposizioni subordinate rispetto alle proposizioni principali*.

Lo strumento restituisce un valore da 1 a 100, il cui significato è il grado di similarità, a livello di caratteristiche prese in esame, con uno dei due corpora sul quale è stato addestrato: il corpus *semplice* 2PAR (composto dai testi del periodico *Due Parole*) e il corpus *complesso* REP (composto dai testi del periodico *La Repubblica*).
Nel presente studio questo strumento è stato utilizzato durante le analisi sul corpus parallelo per valutare il lavoro di semplificazione effettuato (cfr. 3.4).

1.4 Corpus come strumento per la semplificazione

L'utilizzo di corpora come strumenti per la creazione di *modelli predittivi* è ben attestato in letteratura (Lenci et al., 2005, p. 38) e lo stesso vale per modelli specificamente creati con lo scopo di semplificare automaticamente. Saggion, in *Automatic Text Semplification*[7], parlando della nuova disponibilità di corpora di semplificazione per l'inglese, scrive:

[5]La type/token ration è il rapporto tra il numero di parole tipo (il numero di parole uniche nel testo) e il totale di token nel testo. Può essere interpretato come un indice della ricchezza lessicale di un testo (Lenci et al., 2005, p.133)

[6]Tullio De Mauro (2016). *Il Nuovo vocabolario di base della lingua italiana*. `https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana`. visitato il 22/10/2021

[7]Horacio Saggion (2017). ≪Automatic Text Simplification≫. In: *Synthethis Lectures on Human Language Technologies* 32

≪Questi corpora hanno reso possibile una nuova generazione di approcci alla semplificazione testuale, che si basa principalmente su tecniche di machine learning≫ (traduzione mia).

Un esempio di corpus parallelo per la semplificazione del linguaggio amministrativo è SimPA (Scarton et al., 2018), costruito per il linguaggio amministrativo inglese, composto da frasi recuperate automaticamente dal sito web del Sheffield City Council, che sono state poi semplificate a mano da degli annotatori volontari dal profilo piuttosto eterogeneo tra loro. Sono state recuperate circa 10000 frasi, tra di esse sono state scelte per la semplificazione 1100 tra le 5000 frasi più lunghe. La semplificazione è stata divisa in due step, prima una semplificazione lessicale e poi una sintattica.

Un esempio per l'italiano è invece SIMPITIKI (Tonelli et al., 2016), un corpus costruito a partire da 60 milioni di modifiche effettuate su pagine di Wikipedia italiane. Di queste modifiche sono state scelte solamente quelle segnalate come semplificazioni e sono state ricavate 4356 coppie di frasi. Di queste sono state esaminate da annotatori 2671 coppie di frasi, dalle quali ne sono state annotate 345. Come schema di annotazione è stato usato quello di Brunato et al. (2015). Alle coppie di frasi recuperate da Wikipedia sono poi state aggiunte altrettante frasi prese da documenti del municipio di Trento e semplificate a mano.

2. Costruzione del Corpus

In questo capitolo verrà illustrato il procedimento seguito per la costruzione di un corpus parallelo di italiano amministrativo e italiano semplificato. Nel primo paragrafo verrà illustrato il processo di selezione dei documenti da cui si è partiti per la costruzione del corpus, nel secondo paragrafo invece saranno spiegati i processi di filtraggio operati su questi documenti e nel terzo paragrafo sarà illustrato il procedimento di semplificazione e annotazione del corpus.

2.1 Scelta dei documenti

I linguaggi istituzionali sono varietà dell'italiano standard che si collocano lungo la direttrice diafasica[1], e hanno forme linguistiche distinte in base alla finalità della comunicazione (Vellutino, 2008). Per la costruzione del corpus parallelo sono stati individuati quattro diverse tipologie di testi scritti in diverse varietà di italiano istituzionale da cui iniziare:

- **PaWaC** (Passaro e Lenci, 2015): è un corpus contenente atti amministrativi scaricati dagli albi pretori di diversi comuni e presentato in formato CoNLL[2], quindi già tokenizzato e analizzato fino al *Part-of-Speech* (PoS) *tagging*[3]. Il linguaggio che caratterizza questi testi è quello normativo e processuale, la forma di linguaggio speciale più formale tra le varietà dell'italiano istituzionale;

[1] Le varietà di lingua vengono classificate sulla base della dimensione di variazione a cui fanno capo o su cui si collocano. Esistono quattro classi fondamentali di varietà: varietà diatopiche, che si differenziano in base al territorio (le varietà regionali), varietà diastratiche che sono differenziate in base all'appartenenza dei parlanti a diversi strati, fasce e gruppi sociali, varietà diafasiche differenziate in base alle situazioni di impiego della lingua, e varietà diacroniche, le cui differenziazioni si situano lungo l'asse del tempo

[2] CoNLL è un formato di rappresentazione digitale dell'annotazione testuale. I dati sono rappresentati da file di testo tradizionali in cui viene inserito sia il testo che le annotazioni (Dependencies, n.d.)

[3] Il PoS tagging è l'annotazione di ogni token con la propria *categoria grammaticale* o *parte del discorso* (Lenci et al., 2005, p. 213)

- **Social**: testi che prendono in esame il linguaggio della P.A. impiegato sui social network. È stata scelta una collezione di tweet pubblicati dagli account Twitter ufficiali di diversi comuni toscani del consorzio LineaComune, impiegato per l'addestramento di SEM il Chattadino, un chatbot che risponde alle domande sulla P.A. realizzato nell'ambito di un progetto finanziato dalla regione Toscana. Il linguaggio usato in questi testi è quello meno formale ed è classificabile come linguaggio pubblicitario.

- **Web**: pagine web dei siti ufficiali dei comuni. Questi testi sono stati estratti automaticamente attraverso un *crawler*[4] dalle pagine web di diversi comuni toscani. Anche questo corpus è stato impiegato nel progetto di SEM il Chattadino. I testi che si trovano in questa collezione utilizzano un linguaggio burocratico più semplice di quello di PaWaC, e sebbene si noti lo sforzo di rendere più leggibili le informazioni contenute sui siti comunali, il testo rimane piuttosto complesso per i non addetti ai lavori. Il linguaggio usato è quindi classificabile come linguaggio per la comunicazione pubblica, che lungo la direttrice diafasica si trova circa a metà.

- **FAQ**: per questa tipologia di testo è stato selezionato un dataset contenente *F.A.Q.* (Frequently Asked Questions) impiegato anche questo per il progetto di SEM. I testi contenuti in questa collezione sono più semplici di quelli contenuti nella collezione web, poiché l'idea stessa delle F.A.Q. è quella di dare risposte, comprensibili, alle domande più frequenti. Nonostante questo i testi mostravano comunque delle possibilità di semplificazione ulteriore. Il linguaggio utilizzato in questi testi, in termini di formalità, è ancora classificabile come linguaggio per la comunicazione pubblica, ma è significativamente più semplice rispetto al linguaggio usato nei testi Web.

2.2 Pulizia dei file

Per il filtraggio e l'analisi dei dati sono stati scritti degli script in Python. Per i corpora FAQ e Web è stata utilizzata la libreria in Python per l'analisi del linguaggio naturale

[4]Un crawler è un tipo di bot (programma o script che automatizza delle operazioni), che solitamente acquisisce una copia testuale di tutti i documenti presenti in una o più pagine web creando un indice che ne permetta, successivamente, la ricerca e la visualizzazione (Wikipedia, n.d.)

Stanza (Qi et al., 2020). Con essa i testi sono stati divisi prima in frasi (operazione di *sentence splitting*), poi ogni frase è stata tokenizzata e successivamente analizzata fino al PoS tagging. Per PaWaC non è stato necessario fare ulteriori analisi dato che il corpus è stato fornito già analizzato fino al PoS-tagging in formato CoNLL. Allo stesso modo i testi del corpus Social sono stati forniti già analizzati fino al PoS-tagging in formato JSON[5].
Come primo filtro per la pulizia dei testi selezionati è stato scelto di rimuovere tutte le frasi incomplete. È stata definita *incompleta* una frase che non presenta un verbo e un segno di punteggiatura finale e sono state rimosse da tutti i documenti le frasi che non rispettavano la condizione.

Dopo le prime analisi sono stati effettuati dei nuovi filtraggi ad hoc per ogni corpus con lo scopo di massimizzare il numero di frasi rappresentative della P.A. semplificabili. Quindi, seppur cercando di eliminare al meglio frasi inutili o non rappresentative del lingauggio amministrativo, la filosofia adottata è stata quella di massimizzare la quantità di frasi. In letteratura scientifica è comune riferirsi a questo approccio con il motto ≪There is no data like more data≫. Questo perché con i moderni sistemi di machine learning è più utile avere molti dati ”più sporchi” rispetto ad avere pochi dati molto ”puliti” (Lenci et al., 2005).
Di seguito sono riportate nel dettaglio le misure applicate ad ognuno dei quattro corpora.

2.2.1 PaWaC

Come si può vedere nella Tabella 2.1, è presente almeno una frase la cui lunghezza, misurata in token, è pari a 20860. Questo è dovuto a un errore di sentence splitting. Controllando queste frasi in ordine di lunghezza è stato notato che nel corpus sono presenti diverse frasi simili, quindi per risolvere questo problema è stato deciso di rimuovere tutte le frasi più lunghe di 100 token, le quali rappresentavano solo il 5% (24949 frasi) delle frasi filtrate. È stato poi deciso di eliminare tutte le frasi più corte di 5 token, in quanto difficilmente su una frase così piccola è possibile fare una buona semplificazione.

[5]JSON (JavaScript Object Notation) è un semplice formato per lo scambio di dati. Per le persone è facile da leggere e scrivere, mentre per le macchine risulta facile da generare e da analizzare. JSON è basato su due strutture: un insieme di coppie nome/valore e un elenco ordinato di valori; (json.org, n.d.)

Dato l'alto tasso di nominalizzazione tipico dei testi della Pubblica Amminstrazione, molte frasi senza verbi sono in realtà rappresentative e buone candidate per la semplificazione. Tuttavia, con la definizione di incompletezza inizialmente decisa, queste vengono automaticamente scartate dallo script. Alcuni esempi di queste frasi filtrate:

(1) a) l' individuazione dei soggetti destinatari degli alloggi e delle relative assegnazioni ordinarie , speciali , provvisorie , nonché i provvedimenti di emergenza ;

(2) - ambientale , attraverso la verifica di compatibilità ambientale dei singoli interventi , già nella fase di pianificazione urbanistica (Valutazione Ambientale Strategica = v.a.s.) , attraverso la riduzione del consumo del suolo , la raccolta dei rifiuti , l' abbattimento dei rumori , il disinquinamento delle acque , il risparmio energetico ed in generale attraverso il soddisfacimento dei criteri di sostenibilità ue .

Per questo motivo è stata rivista la definizione di incompletezza per questo specifico corpus, ed è stato deciso di eliminare solo le frasi senza verbi, che presentano però almeno un numero al loro interno. Questo perché la maggioranza delle frasi senza verbi, che sono effettivamente da scartare, sono solitamente riferimenti giudiziari, numeri di telefono o fax, firme con data, ecc.
Eliminando le frasi senza verbi ma con almeno un numero si riesce a mantenere le frasi interessanti evidenziate sopra, riuscendo però a eliminare correttamente frasi come:

(3) Il dpcm 4 luglio 2000 n.226 ;

(4) Il Testo Unico delle Leggi sull' ordinamento degli Enti Locali del 4.8.2000

(5) Comune di Abbadia Lariana Provincia di Lecco Uffici : 0341.731241 / Fax 0341.1881038 c.a.p. 23821 Ufficio tecnico : 0341.700423 codice fiscale 83007090133 Polizia municipale : 335/7202713 partita iva 00684170137 e-mail info @comune.abbadia-lariana.lc.it www.comune.abbadia-lariana.lc.it deliberazione di c.c. n. 08 del 19/04/2010 oggetto : Discussione e approvazione del " Documento di indirizzi al p.g.t .

Con questi nuovi filtri appliccati si può vedere nella Tabella 2.2 come si è riusciti a recuperare poco più di 62000 frasi. Inoltre il corpus ha raggiunto una lunghezza media e una

deviazione standard più ragionevole e sostanzialmente in linea con quanto riportato in altri studi sul *burocratese* italiano (Brunato, 2015, p. 94).

Corpus	Numero Frasi	Lunghezza media frasi	Mediana	Lunghezza minima frasi	Lunghezza massima frasi	Deviazione Standard	Frasi perse
PaWaC	431240	44.4	33	2	20860	76.2	368198
Social	8507	24.6	19	2	277	21.7	15592
Web	25630	41.8	32	2	974	37.7	16610
FAQ	458	17.6	12	4	236	16.7	184

Tabella 2.1: Analisi statistiche sui corpora dopo il primo filtraggio.

2.2.2 Social

Questo è il corpus nel quale sono state rimosse in proporzione più frasi tramite i filtri iniziali. Da un totale iniziale di 24099 frasi se ne sono perse 9641 (il 40% del corpus iniziale) rimuovendo le frasi senza verbi, e altre 5951 (il 24.7% del corpus iniziale) rimuovendo le frasi senza segni di punteggiatura finale, rimanendo così con sole 8507 frasi (il 35.3% del corpus iniziale).
Questo filtraggio così massivo è dovuto alle caratteristiche specifiche del linguaggio utilizzato sui social network come Twitter. Infatti sebbene sui social il ricorso alla nominalizzazione è piuttosto basso, le frasi sono molto più brevi (come si vede dalla lunghezza media nella Tabella 1) creando l'opportunità di creare frasi perfettamente corrette e complete senza inserire alcun verbo. Sempre il linguaggio tipico dei social spiega anche l'alto tasso di eliminazione sulle frasi senza punto finale: spesso questi *tweet* terminano con una coda di *hashtags* e link che non prevedono segni di punteggiatura finali. Alcuni esempi di frasi sono:

(6) Domani al via la pulizia delle rastrelliere del #Quartiere5

(7) @muoversintoscan #viabililiFI

(8) Da oggi l' asfaltatura di via delle Tre Pietre .

(9) Tutto pronto per la conferenza stampa di chiusura di @estatefi !
#StayTuned https://t.co/GfdKRv7uu3

Data la difficoltà che filtrare questo tipo di frasi correttamente avrebbe portato, e considerando che queste frasi sono già estremamente semplici, e difficilmente semplificabili ulteriormente, è stato deciso di escludere le frasi contenute in questa collezione per la costruzione del corpus parallelo.

2.2.3 Web

Le frasi per questo corpus sono state raccolte automaticamente da un web crawler, che ha scaricato le pagine web di diversi comuni toscani. Il software che si è occupato di fare questa operazione ha rilasciato in output ogni singola pagina web scritta su un'unica riga. Questo ha creato qualche problema al sentence splitter di Stanza che si occupa di divedere il testo in frasi per poi procedere alla tokenizzazione e infine al PoS-Tagging.
Per far sì che Stanza riuscisse a fare il sentence splitting delle frasi è stato quindi necessario implementare una fase di manipolazione anteriore, dove le frasi sono state automaticamente suddivise da un altro script che effettua una iniziale e cruda divisione delle frasi del testo basandosi sulla presenza di spazi, tab e punteggiatura. Questo risultato parziale è bastato poi alla libreria Stanza per continuare le analisi. Nonostante questo ulteriore passaggio, come per PaWaC, ci sono stati degli errori di sentence splitting, evidenziati dal fatto che la frase più lunga conta 974 token (vedi Tabella 2.1). È stato quindi deciso anche qui di applicare un filtro alla lunghezza minima e massima delle frasi che è stato impostato rispettivamente a 5 token minimi e 100 token massimi.
Nonostante i testi fossero caricati sul web, in un medium comunemente ritenuto semplice, i testi che compongono questo corpus hanno delle similarità importanti con quelli contenuti negli albi pretori appartentia PaWaC. Questo è evidente anche dalle analisi riportate nella Tabella 2.1 dove PaWaC e Web hanno valori di lunghezza media delle frasi molto vicini e mediana praticamente uguale. Rispetto a PaWaC il corpus Web appare più coerente, infatti la deviazione standard è più bassa (N.B.: la deviazione standard di PaWaC era fortemente falsata dalle frasi esageratamente lunghe riportate per gli errori di sentence

splitting).

Date le similitudini con PaWaC, sono state analizzate anche qui le frasi senza verbi, per capire se anche in questo caso un alto tasso di nominalizzazione potesse far scartare delle frasi interessanti. Le analisi hanno mostrato che la maggior parte delle frasi senza verbi che venivano scartate erano principalmente liste di contatti, riferimenti mail e riferimenti legislativi che non avevano un uso per la creazione del corpus. Inoltre all'interno del corpus parallelo finale, le frasi senza verbi con un alto tasso di nominalizzazione erano già rappresentate dai campioni presenti in PaWaC e quindi è stato possibile scartare le poche frasi del genere presenti in questo corpus per avere una pulizia migliore.

Durante queste analisi è stato però notato che molti degli elementi all'interno degli elenchi puntati non presentavano il segno di punteggiatura finale, e tra queste vi sono molte frasi meritevoli di essere semplificate, come:

(10) On-line: attraverso il sistema supplente messo a disposizione dalla Regione Toscana cui si può accedere attraverso il link indicato nella colonna a sinistra della presente scheda informativa sotto la voce "Link Esterni" (solo per alcuni procedimenti)

(11) Autocertificazione circa la fonte di sostentamento da parte del medesimo soggetto interessato o di chi presta l'aiuto economico

(12) Copia del certificato Asl attestante le condizioni di invalidità o handicap (solo in caso di invalidità o handicap superiore o uguale al 67%)

Come per PaWaC è stata quindi ridefinita la definizione di *incompletezza* per questo testo, andando a eliminare prima le frasi senza verbi, e poi quelle senza segno di punteggiatura finale che presentano almeno un numero al loro interno.

Il ragionamento è lo stesso di PaWaC: si usa la presenza di numeri per discriminare le frasi che contengono informazioni utili e semplificabili, da quelle che contengono riferimenti legislativi, numeri di telefono, fax, ecc.

Si può vedere nella Tabella 2.2 che con i nuovi filtri si sono recuperate circa 5400 frasi. Inoltre come per PaWaC i valori di lunghezza media e deviazione standard sono rientrati in valori in linea con quelli presenti in letteratura (Brunato, 2015, p.94).

2.2.4 FAQ

Per questo corpus, il più piccolo dei quattro, non sono stati necessari particolari accorgimenti. Per evitare problemi di sentence splitting è stato scelta 50 come lunghezza massima delle frasi in token (una perdita del 2.65%, 12 frasi, come si vede nella Tabella 2.2). I filtri scelti inizialmente hanno funzionato bene rimuovendo solo frasi poco interessanti da semplificare.

Corpus	Numero Frasi	Lunghezza media frasi	Mediana	Lunghezza minima frasi	Lunghezza massima frasi	Deviazione Standard	Differenza con primo filtraggio
PaWaC	493881	32.8	28	5	99	20.9	+ 62641
Web	31055	32.1	28	5	99	19.6	+ 5425
FAQ	446	15.8	12	4	49	10	- 12

Tabella 2.2: Analisi statistiche sui corpora dopo i filtraggi finali

2.3 Semplificazione e annotazione

In questo paragrafo sarà illustrato il procedimento di semplificazione e annotazione eseguito su un campione di testi prelevato da ognuno dei tre corpora con i quali è stato scelto di comporre il corpus parallelo.

2.3.1 Semplificazione Lessicale e Sintattica

Sono state selezionate 98 frasi (32 da FAQ, 32 da PaWaC e 34 da Web). Come per la costruzione di SIMPITIKI (Tonelli et al., 2016), abbiamo deciso di iniziare la semplificazione scegliendo tra le frasi più lunghe delle corrispettive tipologie di documenti. Come per la costruzione del corpus SimPA (Scarton et al., 2018), ognuna di queste frasi è stata semplificata in due passaggi consecutivi: prima una semplificazione lessicale della frase originale, poi una semplificazione sintattica a partire dalla nuova frase. Durante il primo

step si è cercato di evitare cambiamenti alla struttura sintattica e di sostituire soltanto parole o sintagmi con corrispettivi più semplici. Come indice di semplicità per le parole è stato utilizzato il Nuovo Vocabolario di Base dell'Italiano di Tullio de Mauro. Quindi, quando possibile, si è cercato di sostituire parole che non fanno parte del VdB con sinonimi che ne facessero parte. Per quanto riguarda la semplificazione sintattica i principali accorgimenti sono stati: ridurre il numero di frasi subordinate; convertire le frasi impersonali in frasi nel modo indicativo e le frasi negative in frasi affermative; eliminare le ripetizioni superflue e i riferimenti intertestuali ad altre leggi che non ricoprissero un ruolo fondamentale nel discorso; ridurre la lunghezza delle frasi.

2.3.2 Schema di annotazione XML

Inizialmente per l'annotazione è stato impiegato uno schema XML. Ogni frase era rappresentata dalla voce originale, seguita da una semplificazione lessicale e poi una sintattica. Queste due macro-categorie erano composte da singole operazioni di semplificazione, che venivano costruite l'una a partire dalla precedente per arrivare al testo semplificato finale. All'interno di ogni semplificazione venivano indicate le parti di testo modificate per la semplificazione (v. Listing 2.1)

Listing 2.1: Esempio di semplificazione lessicale annotata in XML

```
<simplification type="31" desc="4" v="1" from="0">
    <before>
        E' possibile fare specifica richiesta di uscita pomeridiana
            anticipata per permettere la frequenza ad attività <del>
            extrascolastiche</del> (sport, musica, danza...) tramite un
             modulo da richiedere alle insegnanti.
    </before>
    <after>
        E' possibile fare specifica richiesta di uscita pomeridiana
            anticipata per permettere la frequenza ad attività <ins>
            esterne alla scuola</ins> (sport, musica, danza...) tramite
             un modulo da richiedere alle insegnanti.
    </after>
</simplification>
```

In ognuna di queste semplificazioni è presente l'informazione sul tipo di operazione effettuata[6], seguendo lo schema utilizzato da Brunato et al. (2015). Le operazioni sono divise in sei macrocategorie, che sono:

- **Split** si usa per indicare una semplificazione avvenuta dividendo una frase in più frasi. Solitamente è usata per separare dalla principale le proposizioni subordinate o coordinate.

- **Merge** è l'operazione opposta a Split, si usa per unire due proposizioni principali in un'unica frase semplificata.

- **Reordering** è l'operazione con la quale si indica una semplificazione avvenuta modificando l'ordine delle parole. Solitamente infatti le frasi che seguono l'ordinamento SVO sono associate ad una complessità inferiore (Slobin e Bever, 1982).

[6] v. Listing 2.1: attributo 'type' del tag simplification

- **Insert** indica un inserimento rispetto alla frase originale. Questo può risultare in una frase semplificata, che per essere più chiara, è più lunga di una frase originale.
- **Delete** è l'operazione opposta alla Insert. Rimuovere parti del testo può aiutare a rendere una frase più semplice da capire, togliendo informazioni ripetute o inutili.
- **Transformation** è la macro-categoria che comprende più operazioni, che servono a specificare meglio la natura della trasformazione avvenuta nella semplificazione. In ogni caso l'operazione indica una trasformazione di uno o più elementi della frase in corrispettivi più semplici (v. Tabella 2.3 per le operazioni).

Oltre all'operazione di semplificazione è stata annotata anche la motivazione dietro ad ogni operazione, per farlo sono state costruite due liste di motivazioni possibili, una lista per le motivazioni di semplificazioni lessicali e una lista per le motivazioni di semplificazione sintattiche. Queste sono state scelte a parite dalla tabella dei tratti tipici della lingua Pubblica Amministrazione in Brunato (2015) e dall'analisi dei tratti tipici della P.A. in Cortelazzo (2021). Le motivazioni sono riportate nella Tabella 2.4.

Categorie	Operazioni	Breve descrizione	ID
Split		Divide una proposizione principale	1
Merge		Unisce due proposizioni principali	2
Reordering		Cambia l'ordine delle parole	3
Insert	Verb	Inserimento di un verbo	11
	Subject	Inserimento del soggetto	12
	Other	Inserimento di elementi generici	13
Delete	Verb	Eliminazione di un verbo	21
	Subject	Eliminazione del soggetto	22
	Other	Eliminazione di elementi generici	23
Transformation	Lexical Substitution (Word Level)	Una singola parola è sostituita da altro testo	31
	Lexical Substitution (Phrase Level)	Un sintagma è sostituita da altro testo	32
	Anaphoric Replacement	Sostituzione di un pronome con il suo antecedente lessicale	33
	Noun to Verb	Una nominalizzazione è sostituita da un verbo	34
	Verb to Noun	Un verbo è sostituito da una nominalizzazione o da un verbo supporto	35
	Verbal Voice	Un verbo, e la frase, passano dall'attivo al passivo o viceversa	36
	Verbal Features	Un verbo cambia modo o tempo	37

Tabella 2.3: Riassunto delle operazioni tracciate con l'annotazione XML di Brunato et al. (2015)

Semplificazione Lessicale		Semplificazione Sintattica	
ID	Motivazione	ID	Motivazione
1	Parole lunghe	41	Frasi Lunghe
2	Tecnicismi collaterali e termini ambigui	42	Alta frequenza di frasi impersonali e implicite
3	Forestierismi	43	Alta frequenza frasi passive
4	Latinismi	44	Stile nominale
5	Arcaismi	45	Frasi preposizionali complesse
6	Nomi astratti	46	Frasi congiuntive complesse
7	Nomi deverbali	47	Predominanza dell'ipotassi sulla paratassi
8	Verbi denominali		
9	Abbreviazioni e acronimi	49	Largo uso di frasi negative
10	Nomi formali e poco comuni	50	Ordine non canonico delle parole
11	Espressioni pleonastiche e stereotipiche	51	Alta frequenza dell'uso verbale del participio presente
21	Errore ortografico		
		52	Pronomi enclitici con verbo finito
23	Ripetizioni	53	Numero elevato di modificatori del nome e di sintagmi nominali pesanti
24	Aggettivi formali e poco comuni	54	Frasi parentetiche e incisi con marcata intertestualità
25	Verbi formali e poco comuni	55	Uso eccessivo di collegamenti anaforici e cataforici

Tabella 2.4: Motivazioni delle semplificazioni linguistiche

2.3.3 Brat strumento di annotazione rapido

Nonostante XML sia un formato più che valido per l'annotazione di testi, annotare a mano dovendo scrivere la struttura XML si è rivelato estremamente lento, e per velocizzare le operazioni è stato deciso di cambiare metodo di annotazione, passando a Brat.
Brat è uno strumento basato su tecnologie web per l'annotazione di testi (*mini-introduction to brat* n.d.). Invece di semplificare e annotare contemporaneamente si è deciso di semplificare prima le frasi e di annotarle successivamente. Sono state semplificate le 98 frasi scelte, dividendole per tipologia di documento di origine in 3 file di testo diversi. Ognuna di queste frasi è stata rappresentata da 4 linee all'interno del testo. La prima (OL) indica la frase originale, la seconda (SL) indica la frase semplificata lessicalmente, la terza (OS) indica la frase semplificata lessicalmente come nuova frase di partenza (la frase è uguale a quella in SL, serve ripeterla per migliorare la leggibilità delle annotazioni) e la quarta (SS) indica la frase semplificata sintatticamente. Un esempio dove sono evidenziate le modifiche ad ogni step:

OL: La pesca dilettantistica può essere esercitata da chiunque abbia provveduto al versamento della tassa di concessione regionale per una delle licenze di tipo 'B' di durata annuale per l'esercizio della pesca con canna con mulinello, con tirlindana, la mazzacchera e la bilancia di € 35,00 o licenza di tipo 'C' della durata di quindici giorni per la pesca dilettantistica con canna, anche con mulinello, con la tirlindana, la mazzecchera e la bilancia di € 10,00.

SL: La pesca dilettantistica è possibile per chiunque abbia pagato la tassa di concessione regionale per una delle licenze di tipo 'B' di durata annuale per l'esercizio della pesca con canna con mulinello, con tirlindana, la mazzacchera e la bilancia di € 35,00 o licenza di tipo 'C' della durata di quindici giorni per la pesca dilettantistica con canna, anche con mulinello, con la tirlindana, la mazzecchera e la bilancia di € 10,00.

OS: La pesca dilettantistica è possibile per chiunque abbia pagato la tassa di concessione regionale per una delle licenze di tipo 'B' di durata annuale per l'esercizio della pesca con canna con mulinello, con tirlindana, la mazzacchera e la bilancia di € 35,00 o licenza di tipo 'C' della durata di quindici giorni per la pesca dilettantistica con canna, anche con mulinello, con la tirlindana, la mazzecchera e la bilancia di € 10,00.

SS: La pesca dilettantistica con canna, con mulinello, tirlindana, mazzacchera e bilancia è possibile per chiunque abbia pagato la tassa di concessione regionale per una delle licenze: Licenza tipo 'B' (€ 35,00) di durata annuale o licenza tipo 'C' (€ 10,00) della durata di 15 giorni.

Una volta organizzati i tre file seguendo queste regole, questi sono stati caricati all'interno di Brat.

Durante la configurazione dello strumento sono stati inseriti dei campi personalizzati per le annotazioni *type* (Tabella 2.3) e *desc* (Tabella 2.4), che corrispondono al tipo di operazione effettuata e alla sua motivazione.

Selezionando una parte di frase su brat è possibile inserire quindi i due campi appena descritti ed eventuali note. All'interno di quest'ultime sono stati inseriti due id, uno per la frase e uno per la semplificazione, per facilitare i successivi lavori di analisi (v. Figura 2.1).

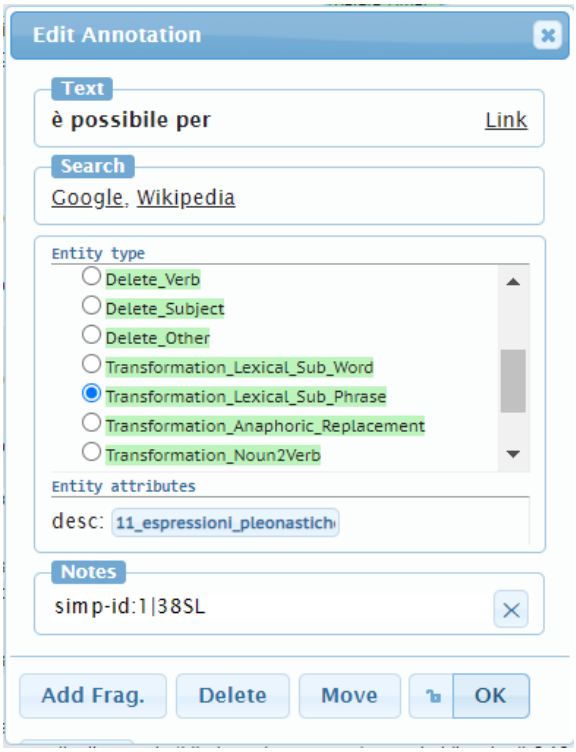


Figura 2.1: Schermata di annotazione di una parte di testo selezionata

In ogni semplificazione, come era stato fatto in XML, viene riportata sia sulla frase originale, selezionando ciò che verrà modificato, sia sulla frase di arrivo, selezionando ciò

che è stato modificato. Delle 98 frasi semplificate sono state annotate 30 frasi, 10 per ogni tipologia di documento.

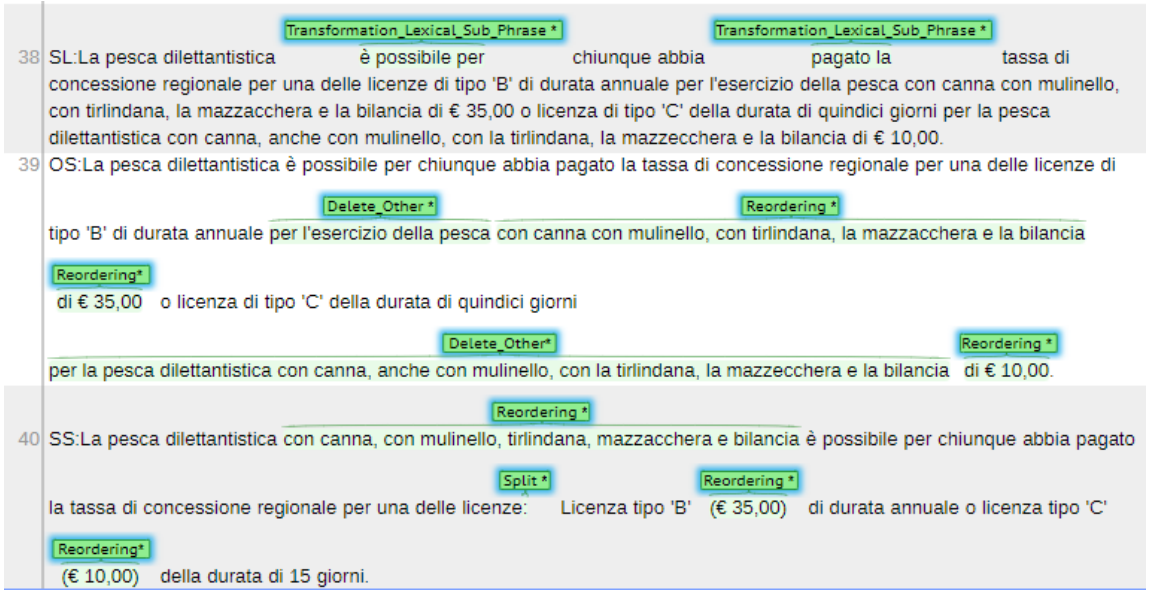


Figura 2.2: Esempio di frase annotata

3. Analisi statistiche sul Corpus e sulle annotazioni

In questo capitolo saranno illustrate e commentate le analisi effettuate sul corpus parallelo e sulla sua annotazione. Le analisi saranno divise in: analisi sulle caratteristiche *raw*[1] del testo, analisi sulle caratteristiche lessicali e analisi sulle caratteristiche sintattiche. Verranno poi commentate anche le analisi sui valori di leggebilità ottenuti con Read-IT (cfr. 1.3.1) e le analisi effettuate sulle annotazioni.

3.1 Analisi sulle caratteristiche *raw*

Le analisi sono state effettuate sulle 98 frasi semplificate, dividendole per documento di origine e per step di semplificazione, quindi sia per PaWaC, che per Web e FAQ sono state analizzate e confrontate le frasi originali con quelle semplificate lessicamente e quelle semplificate sintatticamente.

Per queste analisi è stato impiegato uno script in Python con la libreria Stanza (Qi et al., 2020), che ha analizzato le frasi fino al PoS-Tagging. A partire dalla tokenizzazione del corpus sono state calcolate la lunghezza media di parola espressa in caratteri e la lunghezza media della frase, la mediana della lunghezza delle frasi e la deviazione standard della lunghezza delle frasi espresse in numero di token (v. Tabella 3.1).

Dai dati su questa analisi si può notare come la lunghezza media delle parole sia pressoché uguale tra tutti e tre i tipi di documenti e che con l'avanzare degli step di semplificazione la situazione sostanzialmente non migliora. Questo non implica però una cattiva semplificazione lessicale: i valori sono in linea con quelli presenti in letteratura[2] che variano da 4.9 a 5.6 a seconda del genere del testo, e, come si vedrà nel prossimo paragrafo, altri indici di semplificazione lessicale migliorano con gli step di semplificazione.

[1] Caratteristiche che derivano dal processo di tokenizzazione (Dell'Orletta, Montemagni et al., 2011)

[2] Felice Dell'Orletta et al. (2013). ≪Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose≫. In: *RANLP*, p.5

La media, la mediana, e la deviazione standard della lunghezza della frase si abbassano molto su tutti e tre i tipi di documenti, specialmente durante la fase di semplificazione sintattica: questo è in linea con il *modus operandi* adottato per la semplificazione, infatti è solo durante l'ultimo passaggio che vengono rimosse parti superflue della frase e trasformate le frasi più lunghe in forme più corte e semplici. Essendo il corpus composto da semplificazioni delle frasi più lunghe presenti alla fine dei filtraggi (v. Capitolo 2) i valori sono superiori ai valori presenti in letteratura.

Origine frase	Versione frase	Lunghezza media parola	Lunghezza media frase	Mediana lunghezza frase	Deviazione standard frase
F.A.Q.	Originale	5.1	29.5	31	9.4
	S. lessicale	4.9	29.3	30	9.7
	S. sintattica	4.9	22.4	24	9.8
Web	Originale	5.1	81.8	83	12.5
	S. lessicale	4.9	81.1	82	16.9
	S. sintattica	5	41.7	37	21.3
PAWAC	Originale	5.2	66.5	84	33.7
	S. lessicale	5.2	66.1	82	33.1
	S. sintattica	5.2	39.3	37	23

Tabella 3.1: Valori calcolati a partire dalla tokenizzazione dei corpora semplificati

3.2 Analisi lessicali e morfo-sintattiche

Anche queste analisi sono state effettuate sulle 98 frasi semplificate, divise per documento di origine e per step di semplificazione.

Per queste analisi è stato impiegato uno script in Python con la libreria Stanza (Qi et al., 2020), che ha analizzato le frasi fino al PoS-Tagging. I valori calcolati sono la Type/Token Ratio, la percentuale di lemmi appartenenti al Vocabolario di Base di Tullio De Mauro (De Mauro, 2016) e l'età media di acquisizione, calcolata con il vocabolario di age of

acquisition di Montefinese et al. (2019) (v. Tabella 3.2).
Le analisi sulla TTR mostrano un indice piuttosto basso rispetto sia ai valori riportati in letteratura per altri generi sia a quelli calcolati su testi amministrativi[3].
L'analisi sul vocabolario di base è stata ulteriormente suddivisa in base all'appartenenza dei vari lemmi alle tre sotto-categorie del VdB:

- FO - Vocabolario Fondamentale: contiene 1.991 parole. Le più usate in assoluto in italiano (esempi: amore, lavoro, pane).
- AU - Vocabolario di Alto Uso: contiene 2.750 parole. Molto usate, ma meno di quelle del Vocabolario fondamentale (esempi: palo, seta, toro).
- AD - Vocabolario di Alta Disponibilità: contiene 2.337 parole. Poco usate nella lingua scritta, ma molto in quella parlata (esempi: mensa, lacca, tuta).

Dalle analisi risulta come nei vari step di semplificazione la percentuale di lemmi appartenenti al Vocabolario Fondamentale aumenti. Questo perché durante la semplificazione, specialmente quella lessicale, è stata fatta molta attenzione a trovare dei sinonimi all'interno del VdB per sostituire termini complessi ritrovati nel testo. Si è anche cercato di sostituire termini del Vocabolario di Alto Uso e di Alta Disponibilità con termini del Vocabolario Fondamentale, e questo è evidente dall'abbassamento di percentuale di questi vocabolari negli step Lessicali e Sintattici, sia in Web che in F.A.Q.
Sulla età di acquisizione (AoA) si può vedere come questa diminuisca leggermente con il passare degli step. È importante notare come il vocabolario di AoA utilizzato comprenda solo poco meno di 2000 parole e quindi il valore di età di acquisizione media sia calcolato solo sulle parole presenti nel vocabolario, la cui percentuale sul testo è indicata nel campo (% app. AoA).

[3]Dominique Brunato (2015). ≪A study on linguistic complexity from a computational linguistics perspective. A corpus-based investigation of Italian bureaucratic texts≫. Tesi di dott.

Origine frase	Versione frase	TTR	Vocabolario di base				AoA	% app. AoA
			TOT	FO	AU	AD		
F.A.Q.	Originale	0.47	83.4%	65.5%	15.2%	2.7%	5.4	12.6%
	S. lessicale	0.46	85.8%	67.9%	15%	2.8%	5.4	14.6%
	S. sintattica	0.47	85.6%	68.1%	14.9%	2.6%	5.3	14.9%
Web	Originale	0.39	82%	62.9%	15.1%	4.1%	5.8	9.6%
	S. lessicale	0.39	83.4%	64.6%	14.3%	4.4%	5.6	10.6%
	S. sintattica	0.39	84%	66%	13.8%	4.2%	5.6	10.8%
PAWAC	Originale	0.40	82%	62.1%	16%	3.9%	6.5	10.2%
	S. lessicale	0.40	84.6%	64.6%	16%	4%	6.4	11.2%
	S. sintattica	0.40	85.3%	65.1%	16.1%	4.1%	6.4	11.5%

Tabella 3.2: Analisi lessicali effettuate dall'annotazione morfosintattica del corpus

3.3 Analisi sintattiche

Anche queste analisi sono state effettuate sulle 98 frasi semplificate, divise per documento di origine e per step di semplificazione.

Per queste analisi è stato impiegato uno script in Python con la libreria Stanza (Qi et al., 2020), che ha analizzato le frasi fino al *dependency parsing*[4]. Da queste analisi (v. Tabella 3.3) si può notare un netto miglioramento su tutti i valori nella fase di semplificazione sintattica. In particolar modo è interessante notare come nonostante le frasi dalle quali è partita la semplificazione fossero molto lunghe, tra le più lunghe dei rispettivi corpora (cfr 2.3.1), il valore di profondità media dell'albero sintattico al termine della semplificazione sia in linea con i valori presenti in letteratura per l'italiano amministrativo semplificato (Brunato, 2015). Questo implica che è possibile giungere a una buona semplificazione sintattica anche partendo dalle frasi "peggiori".

I valori delle analisi sui dependency link anche dopo la semplificazione rimangono leggermente più alti dei valori presenti in letteratura, con solo le frasi derivate dalle FAQ che

[4]Il dependency parsing costruisce un albero sintattico a partire da una frase, mettendo in evidenza le relazioni tra i nodi (*Dependency Parsing* n.d.)

raggiungono i valori più vicini a quelli presenti negli altri studi, come quello di Brunato (2015).

Origine	Versione frase	Lunghezza media dependency link	Lunghezza media max dependency link	Profondità media albero sintattico
F.A.Q.	Originale	3.5	25.6	6.2
	S. lessicale	3.6	25	6.1
	S. sintattica	3.3	18	5.1
Web	Originale	4.1	69.2	12.3
	S. lessicale	4.1	65.5	11.8
	S. sintattica	3.8	35.5	7.5
PAWAC	Originale	4	58.9	9.9
	S. lessicale	4	56.7	9.5
	S. sintattica	3.7	34.7	6.9

Tabella 3.3: Analisi effettuate dall'annotazione morfosintattica e sintattica effettuate sul corpus

3.4 Valori di leggibilità con READ-IT

Una delle analisi effettuate è stata calcolare la media del valore di leggibilità per frase di Read-IT (Dell'Orletta, Montemagni et al., 2011) per i vari step di semplificazione dei tre tipi di documenti. Il punteggio di Read-IT (cfr. 1.3.1) è diviso in quattro categorie:

- raw: punteggio calcolato sulle sole caratteristiche derivate dalla tokenizzazione del testo;

- lex: punteggio calcolato sulle sole caratteristiche lessicali prese in esame;
- syn: punteggio calcolato sulle sole caratteristiche sintattiche prese in esame;
- all: punteggio calcolato su tutte le caratteristiche prese in esame da Read-IT;

In tutte e quattro le tipologie di analisi e per tutti e tre i tipi di documenti si può vedere come i valori calcolati sulle frasi originali siano più alti rispetto ai valori calcolati sulle frasi semplificate (v. Tabella 3.3).

I valori calcolati sulla base di tutte le caratteristiche, anche al termine della semplificazione, rimangono piuttosto alti. Questo perché il punteggio di Read-IT è una misura di similarità ad un corpus modello scelto come ”complesso”, che è il corpus REP (composto dai testi del periodico *La Repubblica*). Considerando che le misure di leggibilità sono fortemente influenzate dal genere letterario (Dell'Orletta, Venturi et al., 2012) e che in questo caso parliamo di genere amministrativo, che è generalmente più complesso del genere giornalistico[5], questi valori alti di leggibilità non sorprendono.

Un altro dato interessante è l'aumento della difficoltà di leggibilità calcolata sui parametri lessicali (lex) nello step tra semplificazione lessicale e sintattica in F.A.Q. Questo implica che nella fase di semplificazione sintattica sono entrati in gioco dei fattori che mediamente hanno alzato la complessità lessicale delle frasi.

Origine frase	Versione frase	Read-IT			
		raw	lex	syn	all
F.A.Q.	Originale	69	65	96.4	97.4
	S. lessicale	69	61.3	95.5	96
	S. sintattica	66.6	61.6	89.7	91.6
Web	Originale	87.1	68.7	95.6	97.8
	S. lessicale	87.2	65.2	95.4	97.4
	S. sintattica	74.3	63.5	90.8	93.9
PAWAC	Originale	89.6	71.6	97	98.8
	S. lessicale	88.4	66.4	96.7	98.2
	S. sintattica	70.6	64.8	89.5	93

Tabella 3.4: Valori di leggibilità generati con Read-IT

[5]Dominique Brunato (2015). ≪A study on linguistic complexity from a computational linguistics perspective. A corpus-based investigation of Italian bureaucratic texts≫. Tesi di dott., cap.3

3.5 Analisi sull'annotazione

Dalle distribuzioni di frequenza delle operazioni nella Tabella 3.4 si può notare come per ogni corpus le sostituzioni lessicali a livello di parola o frase siano le due operazioni più frequenti. Anche il resto delle operazioni è piuttosto omogeneo come distribuzione in tutte e tre le tipologie di documenti, fatta eccezione per le operazioni di Split e di Merge: in Pawac e Web, dove le frasi sono mediamente più lunghe (v. Tabella 3.1), si fa un uso maggiore dell'operazione di split, mentre in FAQ, dove le frasi sono mediamente più corte, si fa un uso maggiore dell'operazione di Merge.

Categorie	Operazioni	Frequenza			
		FAQ	Web	Pawac	TOT
Split		7	21	31	59
Merge		11	3	3	17
Reordering		12	12	8	32
Insert	Verb	0	0	0	0
	Subject	0	0	0	0
	Other	2	6	0	8
Delete	Verb	0	0	1	1
	Subject	0	0	0	0
	Other	7	18	20	45
Transformation	Lexical Substitution (Word Level)	25	21	41	87
	Lexical Substitution (Phrase Level)	20	19	42	81
	Anaphoric Replacement	0	0	0	0
	Noun to Verb	0	0	0	0
	Verb to Noun	0	0	0	0
	Verbal Voice	2	4	0	6
	Verbal Features	1	2	0	3

Tabella 3.5: Frequenza delle operazioni effettuate divise per tipologia di documento

Confrontando la frequenza relativa delle operazioni con la frequenza relativa delle operazioni di SIMPITIKI (Tonelli et al., 2016, dati sul corpus del municipio di Trento, cfr. 1.4), si può notare che sia nel nostro corpus che in SIMPITIKI la maggior parte delle operazioni sono di trasformazione lessicale (v. Figura 3.1). Nella nostra annotazione c'è un uso più frequente delle operazioni di Reordering, Merge, Split, e Delete Other, che può essere spiegato dal diverso metodo di annotazione, avendo noi annotato in due step successivi, uno lessicale seguito da uno sintattico. È comunque bene notare che la distribuzione di frequenza delle annotazioni è fortemente dipendente dal modo di annotare, che può variare di molto da annotatore ad annotatore.

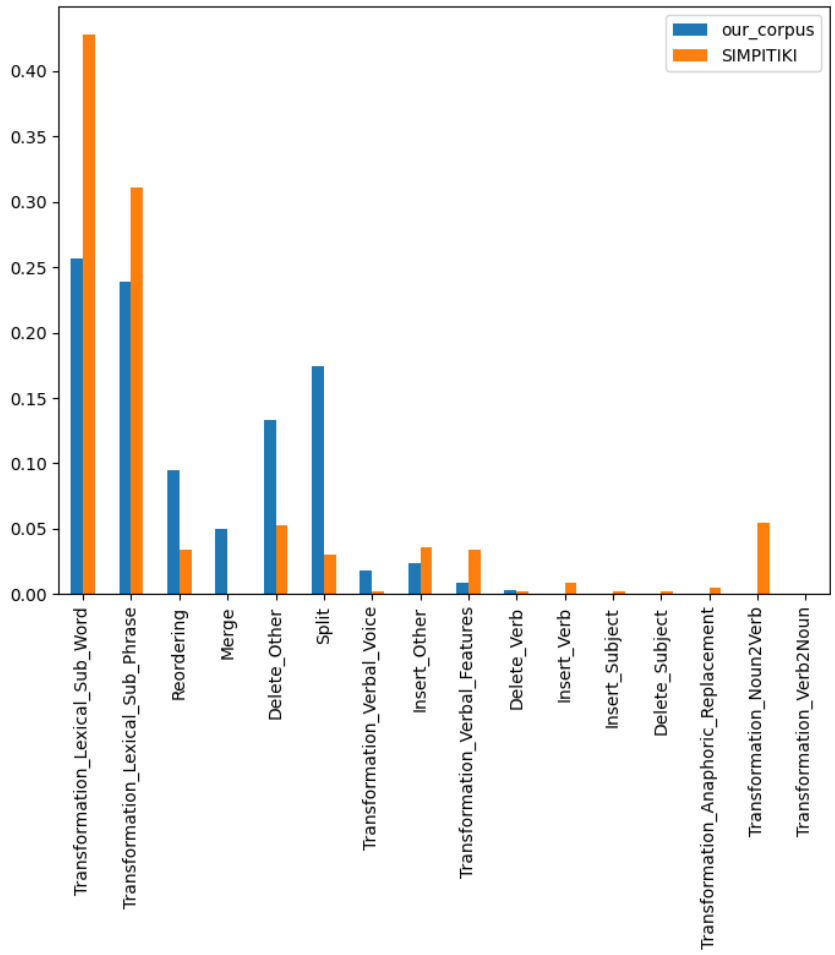
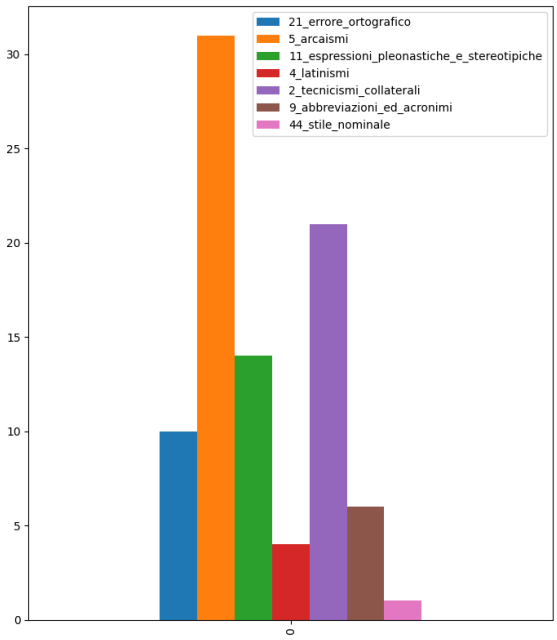
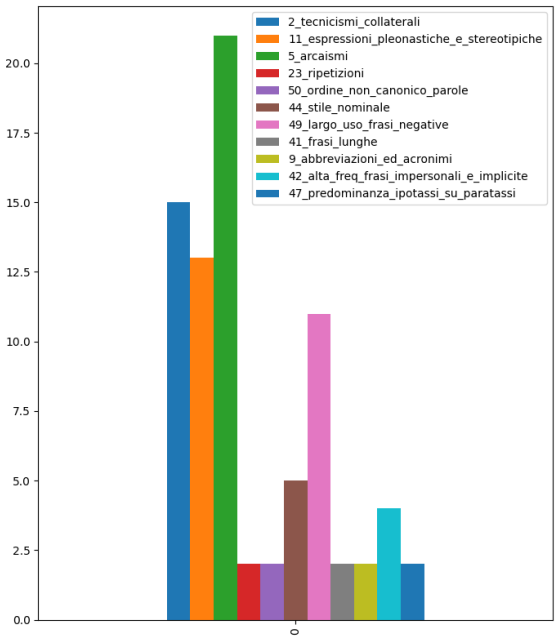


Figura 3.1: Differenza tra le operazioni effettuate nel nostro corpus e quelle effettuate in SIMPITIKI nei documenti del municipio di Trento

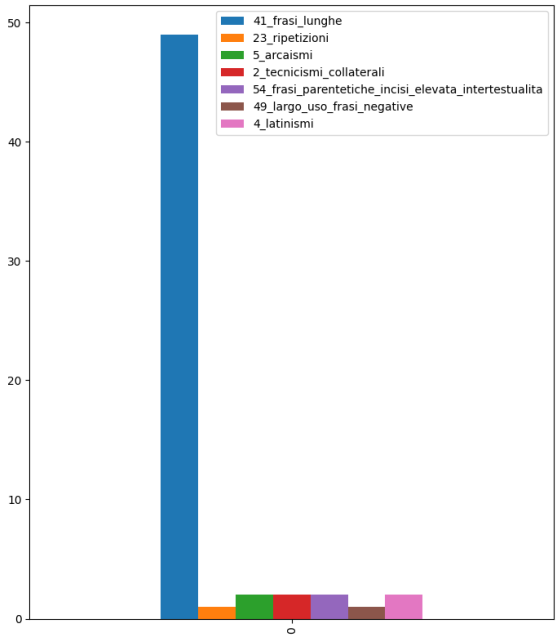
Per quanto riguarda le motivazioni delle semplificazioni è interessante notare come per le operazioni di trasformazione lessicale le motivazioni principali siano tutti fattori di complessità fortemente tipici del genere amministrativo (cfr. 1.2). In particolare si può vedere come le tre motivazioni principali per una sostituzione lessicale a livello di parola siano la presenza di arcaismi, tecnicismi collaterali e espressioni pleonastiche e stereotipiche. Lo stesso vale anche per la sostituzione a livello di frase, ma con frequenze diverse (v. Figura 3.2 a e b). Non sorpende che essendo partiti dalle frasi più lunghe disponibili, la terza operazione più usata nell'annotazione sia quella di Split per accorciare le frasi lunghe (v. Figura 3.2 c). Nella figura 3.3 si può vedere la correlazione tra tutte le operazioni e le motivazioni del loro uso su tutto il corpus annotato.



(a) Lexical Substitution (Word Level)



(b) Lexical Substitution (Phrase Level)



(c) Split

Figura 3.2: Distribuzione di frequenza delle motivazioni per le tre operazioni più frequenti nella semplificazione

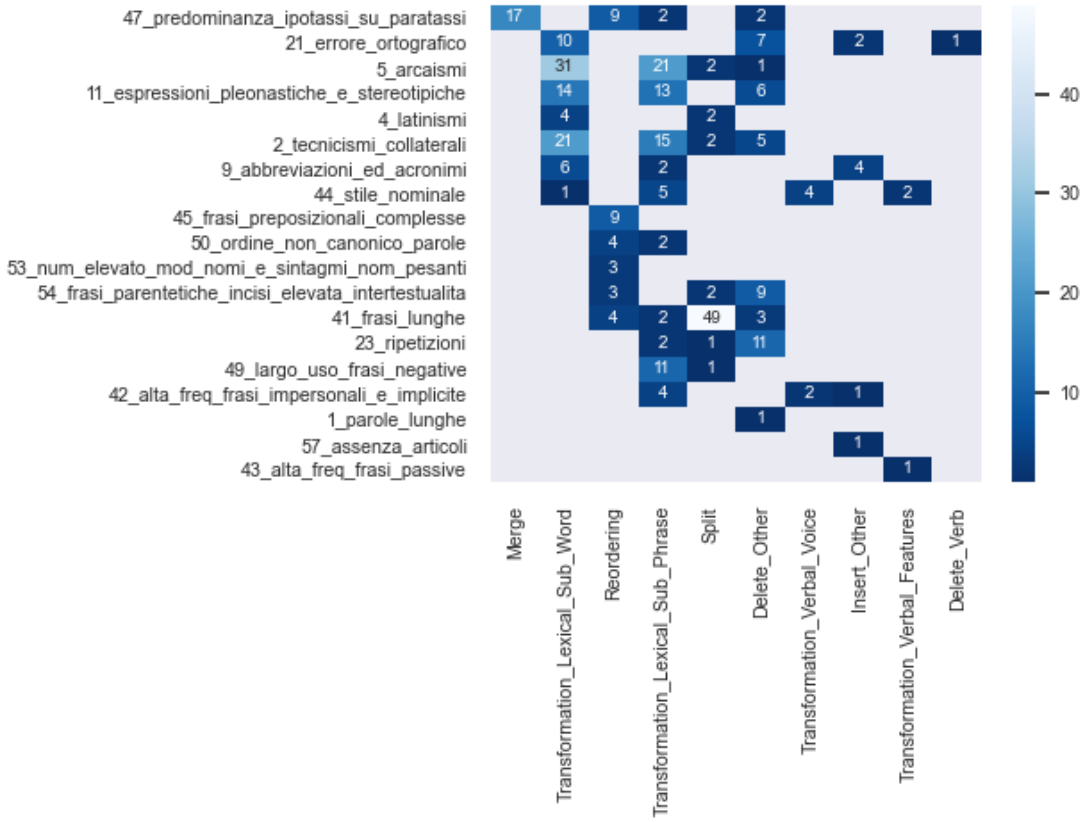


Figura 3.3: Correlazione tra le operazioni e le motivazioni su tutto il corpus annotato

Conclusioni

Con il presente studio è stata presentata una modalità operativa di creazione di un corpus parallelo ed è stata iniziata la creazione dello stesso.
Allo stato attuale sono stati scelti e analizzati i documenti da cui recuperare le frasi, delle quali ne sono state manualmente semplificate 98, e 30 di queste sono già state annotate.
Dalle analisi effettuate su questo campione si sono ottenuti dati incoraggianti sulla qualità delle frasi scelte, della semplificazione su esse effettuate e sulle modalità di annotazione.
Sarà possibile quindi continuare il processo di costruzione del corpus per accrescere il numero di frasi e realizzare una risorsa fondamentale per l'addestramento di reti neurali alla semplificazione di frasi del dominio della Pubblica Amministrazione. Questo può essere il primo passo per una soluzione tecnologica al problema della complessità del linguaggio burocratico, rendendo il linguaggio della Pubblica Amministrazione alla portata anche di chi non è addetto ai lavori, ma sopratutto di coloro che non sono madrelingua italiana o che hanno un livello di scolarizzazione basso. Persone che sono, comunque, interessate dai documenti scritti in questa varietà dell'italiano standard.

Bibliografia

Blache, Philippe (2011). ≪A computational model for linguistic complexity≫. In: *Biology, Computation and Linguistics*.

Brunato, Dominique (2015). ≪A study on linguistic complexity from a computational linguistics perspective. A corpus-based investigation of Italian bureaucratic texts≫. Tesi di dott.

Brunato, Dominique, Felice Dell'Orletta, Giulia Venturi e Simonetta Montemagni (giu. 2015). ≪Design and Annotation of the First Italian Corpus for Text Simplification≫. In: *Proceedings of The 9th Linguistic Annotation Workshop*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 31–41.

Calvino, Italo (1980). ≪L'antilingua≫. In: *Una pietra sopra*. Milano: Mondadori, pp. 150–155.

Chall, Jeanne S e Edgar Dale (1995). *Readability revisited: the new Dale-Chall readability formula*. Cambridge, Mass: Brookline Books.

Colombo, Lucia e Cristina Burani (2002). ≪The Influence of Age of Acquisition, Root Frequency, and Context Availability in Processing Nouns and Verbs≫. In: *Brain and Language* 81.1, pp. 398–411.

Cortelazzo, Michele A. (2021). *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*. Carrocci editore.

De Mauro, Tullio (2016). *Il Nuovo vocabolario di base della lingua italiana*. `https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana`. visitato il 22/10/2021.

Dell'Orletta, Felice, Simonetta Montemagni e Giulia Venturi (2011). ≪READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification≫. In: SLPAT '11. Edinburgh, Scotland: Association for Computational Linguistics, pp. 73–83.

Dell'Orletta, Felice, Giulia Venturi e Simonetta Montemagni (dic. 2012). ≪Genre-oriented Readability Assessment: a Case Study≫. In: *Proceedings of the Workshop on Speech and Language Processing Tools in Education*. Mumbai, India:

The COLING 2012 Organizing Committee, pp. 91–98. URL: https://aclanthology.org/W12-5812.

Dell'Orletta, Felice, Simonetta Montemagni e Giulia Venturi (2013). «Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose». In: *RANLP*.

Dependencies, Universal (n.d.). *CoNLL-U Format*. https://universaldependencies.org/format.html. visitato il 18/10/2021.

Dependency Parsing (n.d.). https://stanfordnlp.github.io/stanza/depparse.html. visitato il 08/11/2021.

Ferstl, Evelyn e Giovanni Flores d'Arcais (1999). «The Reading of Words and Sentences». In: *Language Comprehension: A Biological Perspective*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 175–210.

Flesch, Rudolf Franz (1948). «A new readability yardstick.» In: *The Journal of applied psychology* 32 (3), pp. 221–233.

Frazier, Lyn (1979). «On Comprehending Sentences: Syntactic Parsing Strategies». In: *ETD Collection for University of Connecticut*.

Gibson, Edward (1998). «Linguistic complexity: locality of syntactic dependencies». In: *Cognition* 68.1, pp. 1–76.

Gibson, Edward, Neal Pearlmutter, Enriqueta Canseco-Gonzalez e Gregory Hickok (1996). «Recency preference in the human sentence processing mechanism». In: *Cognition* 59.1, pp. 23–59.

Green, Georgia M. e Margaret S. Olsen (1988). «Preferences for and Comprehension of Original and Readability Adapted Materials». In: *Linguistic complexity and text comprehension: Readability issue reconsidered*. Hillsdale, NJ.

json.org (n.d.). *Introduzione a JSON*. https://www.json.org/json-it.html. visitato il 17/10/2021.

Just, Marcel Adam e Patricia A. Carpenter (1980). «A theory of reading: from eye fixations to comprehension.» In: *Psychological review* 87 (4), pp. 329–54.

Kempen, Gerard (1998). ≪Sentence Parsing≫. In: *Language Comprehension: A Biological Perspective*, pp. 213–228.

Kimball, John (1973). ≪Seven principles of surface structure parsing in natural language≫. In: *Cognition* 2.1, pp. 15–47.

Lenci, Alessandro, Simonetta Montemagni e Vito Pirrelli (2005). *Testo e Computer. Elementi di Linguistica Computazionale*. corso Vittorio Emanuele II, 229, 00186, Roma: Carrocci Editore.

Lucisano, Pietro e Maria Emanuela Piemontese (1988). ≪Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana≫. In: *Scuola e città* 34, pp. 110–124.

Lyding, Verena, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci e Vito Pirrelli (2013). *PAISÀ Corpus of Italian Web Text*. Eurac Research CLARIN Centre.

mini-introduction to brat (n.d.). `https://brat.nlplab.org/introduction.html`. visitato il 23/10/2021.

Montefinese, Maria, David Vinson, Gabriella Vigliocco e Ettore Ambrosini (2019). ≪Italian Age of Acquisition Norms for a Large Set of Words (ItAoA)≫. In: *Frontiers in Psychology* 10, p. 278.

Passaro, Lucia C. e Alessandro Lenci (2015). ≪Extracting terms with EXTra≫. In: *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pp. 188–196.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton e Christopher D. Manning (2020). ≪Stanza: A Python Natural Language Processing Toolkit for Many Human Languages≫. In: *"Association for Computational Linguistics (ACL) System Demonstrations*.

Saggion, Horacio (2017). ≪Automatic Text Simplification≫. In: *Synthethis Lectures on Human Language Technologies* 32.

Scarton, Carolina, Gustavo Paetzold e Lucia Specia (2018). ≪SimPA: A Sentence-Level Simplification Corpus for the Public Administration Domain≫. In: *LREC*.

Segui, Juan, Jacques Mehler, Uli Frauenfelder e John Morton (1982). ≪The word frequency effect and lexical access≫. In: *Neuropsychologia* 20.6, pp. 615–627.

Seidenberg, Mark S. (1989). ≪Reading Complex Words≫. In: *Linguistic Structure in Language Processing*. A cura di Michael K. Carlson Greg N.and Tanenhaus. Dordrecht: Springer Netherlands, pp. 53–105.

Slobin, Dan I. e Thomas Bever (dic. 1982). ≪Children use canonical sentence schemas: A crosslinguistic study of word order and inflections≫. In: *Cognition* 12, pp. 229–265.

Tonelli, Sara, Alessio Palmero Aprosio e Francesca Saltori (2016). ≪SIMPITIKI: a Simplification corpus for Italian≫. In: *Proceedings of CLiC-it*.

Treccani (n.d.). *Nominalizzazione*. `https://www.treccani.it/enciclopedia/nominalizzazione_\%28La-grammatica-italiana\%29/`. visitato il 20/10/2021.

Vellutino, Daniela (2008). *L'italiano istituzionale per la comunicazione pubblica*. Il Mulino.

Vogel, Mabel e Carleton Washburne (1928). ≪An Objective Method of Determining Grade Placement of Children's Reading Material≫. In: *The Elementary School Journal* 28.5, pp. 373–381.

Wikipedia (n.d.). *Crawler*. `https://it.wikipedia.org/wiki/Crawler`. visitato il 20/10/2021.

Ringraziamenti

Vorrei ringraziare la dott.ssa Miliani per il supporto tecnico e morale datomi durante la fase di tirocinio e di tesi.
Vorrei ringraziare Sofia e i miei amici di sempre, Gabriele e Matteo, per essermi sempre stati vicini, e i miei genitori per avermi supportato durante tutto il percorso di studi.
Vorrei ringraziare i brillanti colleghi, e ormai amici, che in questi anni mi hanno accompagnato lungo il percorso universitario, tra i quali: Lorenzo, Marco, Agnese, Elena, Chiara e Alice.
Vorrei ringraziare Francesco, per essere sempre stato un modello di tenacia e per il supporto che mi ha fornito nonostante la lontananza.
Vorrei infine ringraziare Silvia per l'aiuto tecnico fornitomi per la comprensione dei più difficoltosi testi amministrativi.