INDICE

dedica

0. INTRODUZIONE

Questo elaborato si propone di analizzare come le tecnologie linguistico-computazionali possano essere impiegate per favorire lo sviluppo di sistemi di semplificazione semiautomatica di testi.

Il punto di partenza di questa ricerca è l'annotazione[1] di un corpus, costituito in modo tale da rappresentare una tipologia di semplificazione definita come *"strutturale"*.

La caratteristica di questo corpus è di essere costituito da due versioni allineate: una contenente i testi nella loro forma *"originale"* e l' altra nella loro forma *"semplificata",* riadattata da linguisti esperti per una specifica categoria di persone (in questo caso bambini con difficoltà di comprensione del testo o *" poor comprehender "*) .

Il processo di annotazione del testo è stato realizzato applicando uno schema di annotazione costituito da regole di semplificazione per classificare le diverse tipologie di intervento sul testo.

Dopo aver introdotto e spiegato le diverse regole, verranno discussi i dati relativi alla loro distribuzione nel corpus, ricavati dall' applicazione di uno script per comparare la loro frequenza.

1. I testi annotati sono testi in cui viene codificata dell'informazione linguistica in associazione al testo. L'unità di annotazione è il tag, una parola chiave o un termine associato a un'informazione, che descrive l'oggetto rendendo possibile la classificazione e la ricerca di informazioni basata su parole chiave; i tags sono generalmente scelti in base a criteri informali e personalmente dagli autori/creatori dell'oggetto dell'indicizzazione.

Si discuteranno, in seguito, gli effetti dell' applicazione delle diverse tipologie di regole di semplificazione sul livello di leggibilità del testo, misurato attraverso il software **READ-IT** (Dell'Orletta et al, 2011) che assegna un punteggio di leggibilità ai testi sulla base di caratteristiche linguistiche di varia natura (*di base, lessicali, morfo-sintattiche e sintattiche)*. Insieme ai risultati di READ-IT, verranno anche comparati i risultati estratti dalle misure tradizionali di leggibilità (*Indice Gulpease*).

Una volta discussa l' analisi qualitativa delle regole di semplificazione, si approfondiranno i profili linguistici delle frasi alla quale sono state applicate due distinti gruppi di regole di semplificazione, così da capire il loro impatto sulle variazioni di leggibilità del testo.

Infine, verrà condotta una valutazione qualitativa di una delle regole di semplificazione maggiormente utilizzata nel corpus (la regola *Split,* ovvero la divisione di una frase in due o più frasi autonome).

Nel dettaglio, la tesi si articola in 3 capitoli:

Il **primo capitolo** è una breve introduzione sulla nozione di complessità linguistica e di semplificazione del testo e discute due possibili metodologie attuabili quando si parla di semplificazione dei testi: il metodo *"strutturale"* e il metodo *"intuitivo"*. Inoltre verranno descritti gli strumenti di monitoraggio linguistico usati nella parte sperimentale.

Nel **secondo capitolo** verrà illustrato il corpus utilizzato in questo elaborato, gli strumenti usati per le varie analisi. Verranno inoltre spiegate le varie regole di

annotazione sul corpus allineato. In seguito, verrà condotta un' analisi distribuzionale delle regole di semplificazione sul corpus, registrando le frequenze di applicazione delle regole di annotazione, il numero delle frasi alle quali sono state applicate le regole e la frequenza di combinazione delle stesse.

Nel **terzo capitolo**, verrà condotta un' analisi qualitativa delle regole di semplificazione: il monitoraggio linguistico del testo verrà fatto attraverso il software READ-IT, descrivendo in dettaglio i vari livelli di analisi e mostrando i risultati statistici degli stessi. Inoltre, verrano presi in considerazione i risultati del monitoraggio linguistico di due gruppi di regole di annotazione per verificare il loro effetto sul livello di leggibilità del corpus.

Infine, in appendice, vengono riportati i codici e gli script dei programmi creati per l'estrazione dell' informazione d'interesse.

1. TECNOLOGIE LINGUISTICO-COMPUTAZIONALI PER L' ANALISI DELLA COMPLESSITA' DEI TESTI

Questo capitolo introduce gli strumenti di trattamento automatico del linguaggio utilizzati in questo elaborato per analizzare i testi, estrarre conoscenza linguistica e valutare la complessità del testo. Inoltre verranno introdotte le possibili strategie adottate per la semplificazione del testo.

L'intuizione di partenza riguardo al *" potere diagnostico "*delle tecnologie linguistico-computazionali in ambiti di monitoraggio linguistico trova conferma in un ciclo di studi avviato a livello internazionale e, all' interno dei quali, vi sono analisi linguistiche generate da strumenti di trattamento automatico del linguaggio usate per :

Il monitoraggio e lo sviluppo della sintassi nel linguaggio infantile (Sagae *et al.,* 2005; Lu, 2008);

L' identificazione di deficit cognitivi attraverso misure di complessità sintattica (Roark *et al.,* 2007);

Le misure sulla leggibilità di testi per studenti di L1 e L2 (Heiman *et al.,* 2007; Collins - Tompson,2005);

Il monitoraggio sulla capacità di lettura come componente centrale della competenza linguistica (Schwarm *et al.,* 2005; Petersen et al.,2009).

Tra le possibili applicazioni basate sul trattamento automatico della lingua che richiedono un' analisi delle caratteristiche di complessità dei testi, questa tesi affronta la semplificazione semi- automatica[2] dei testiche è un ambito molto studiato negli ultimi anni a livello internazionale.

Nell' indagare sulla potenzialità di queste tecnologie per il monitoraggio della lingua a partire dall' analisi automatica di testi rappresentativi di diverse varietà linguistiche, gli studi condotti si sono soffermati sulla definizione di una metodologia che possa servire sia su un versante teorico sia su contesti applicativi (come ad esempio il monitoraggio delle competenze linguistiche in ambito scolastico, cfr. Dell' Orletta e Montemagni, 2013).

Le tecnologie linguistiche utilizzate nel presente lavoro sono state sviluppate dal laboratorio ItalianNLp (www.italianlp.it) (ILC – CNR), che realizza piattaforme consolidate e ampiamente sperimentate di metodi e strumenti per il trattamento automatico della lingua che ha ricevuto non solo un' ampia validazione nell' ambito di ricerca dedite a l' estrazione di informazioni linguistiche di corpus testuali, ma anche progetti di carattere applicativo finalizzati a l' estrazione di conoscenza di dominio.

2. Montemagni S., 2013 *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana.*

1.1 Monitoraggio delle caratteristiche linguistiche dei testi

In questo paragrafo sarà presentata brevemente la metodologia di monitoraggio dei testi e gli strumenti utilizzati nell' analisi del corpus oggetto di questo elaborato.

Gli studi recenti sull' analisi del testo hanno riguardato diversi livelli della struttura linguistica (lessicale, morfosintattica e sintattica).

Lo stato dell' arte nei compiti di annotazione linguistica è rappresentato da sistemi basati su algoritmi di apprendimento automatico supervisionato, che funzionno in questo modo: ad ogni passo di computazione, infatti, il sistema, in base alla parola in input, ai suoi tratti descrittivi, al contesto e alle annotazioni linguistiche già identificate, sceglie l' annotazione più probabile.

Questo è anche l'approccio seguito nei sistemi " avanzati " per la valutazione della leggibilità di un testo che viene così riformulato come un compito di classificazione probabilistica.

Tre sono gli "ingredienti" fondamentali di questo tipo di approccio rispetto alla valutazione della leggibilità: l'insieme delle categorie linguistiche da assegnare (i livelli di leggibilità, nel nostro caso); il corpus di apprendimento, ovvero un insieme di esempi pre- classificati a mano rispetto alle categorie (di leggibilità, in questo caso) da riconoscere automaticamente e un insieme di tratti descrittivi accuratamente selezionati sulla base del compito di classificazione da svolgere.

Gli strumenti software e i tool di annotazione alla base del monitoraggio linguistico utilizzatti in questa tesi sono stati:

LinguA (http://www.italianlp.it/demo/linguistic-annotation-tool/)

È una pipeline dell'annotazione linguistica allo stato dell'arte che combina algoritmi basati su regole e algoritmi di apprendimento automatico.

Esso comprende le seguenti fasi di annotazione:

- *divisione delle frasi;*
- *tokenizzazione;*
- *analisi grammaticale e lemmatizzazione;*
- *analisi a dipendenza.*

LinguA consente di analizzare i testi italiani e inglesi in input , nonché di visualizzare e scaricare l' analisi generale in formato CoNLL[3]dove:

- le frasi sono separate da una riga vuota;
- ogni token inizia una nuova linea ed è annotata con le seguenti informazioni linguistiche (lemma, macro e micro elementi grana *Part of-Speech,* morfologiche, sintattiche[4]).

Il *parser morfo-sintattico* ha un' accuratezza del 96,34 % nell' identificare simultaneamente le categorie grammaticali.

Invece, il *parser sintattico* ha un' accuratezza del 87 % circa.

3. Dell'Orletta F. " Ensemble system for Part-of-Speech tagging ". In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009).L' accuratezza è stata calcolata come il rapporto tra il numero di token classificati correttamente e il numero totlae di token analizzati .
4. Attardi G., Dell'Orletta F. " *Reverse Revision and Linear Tree Combination for* Dependency Parsing ". In: NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies (Boulder, Colorado, June 2009). Proceedings, pp. 261 – 264. Association for Computational Linguistics, 2009.

MONITOR-IT *(* http://monitor-it.italianlp.it/ *)* strumento che crea il profilo linguistico di un testo, estraendo caratteristiche linguistiche del testo a differenti livelli di descrizione (lessico, morfo-sintassi, sintassi) partendo dall'analisi linguistica. *MONITOR-IT* ha la capacità di analizzare il testo preso in esame e collocarlo rispetto ai vari livelli di istruzione scolastica.

READ-IT[5] è uno strumento per la valutazione automatica della leggibilità che è stato utilizzato per l' analisi distribuzionale e qualitativa dei testi presi in esame in questo elaborato (successivamente trattato nel dettaglio).

Per quanto i risultati dell' annotazione linguistica automatica includano inevitabilmente un margine d' errore (che può avere delle variazioni in base al livello e al tipo di informazione linguistica considerata), se esplorati in modo appropriato possono fornire indicazioni affidabili nella ricostruzione del profilo linguistico di un testo.

5. http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest

1.2 Strategie di semplificazione testuale

La semplificazione del testo può seguire due metodi principali già precedentemente citati: il metodo *intuitivo*[6] e il metodo *strutturale*[7].

Il metodo *intuitivo,* è un tipo di semplificazione testuale raggiunta normalmente dagli insegnanti, seguendo la loro conoscenza del contesto scolastico e delle abilità linguistiche dei propri studenti.

Per indirizzare i propri allievi alla comprensione e alla lettura di un' opera,

l' insegnante cerca di realizzare una semplificazione basata su un livello morfologico e sintattico molto semplice: ad esempio, accorciare frasi contenenti un alto livello di subordinate ; utilizzare verbi a tempo presente o imperfetto, semplificare lessicalmente un testo, utilizzando un vocabolario di base al fine di utilizzare termini elementari e intuitivi (così da rendere il testo accessibile a studenti stranieri o con difficoltà cognitive).

In più, per indirizzare meglio gli allievi alla comprensione e alla lettura di alcune opere, gli insegnanti possono introdurre anche dati bibliografici sugli autori e protagonisti delle storie per un approccio più empatico del testo.

6. Crossley Scott A. , Allen David B., McNamara D., *Text readability and intuitive simplification: A comparision of readability formulas,* April 2011, pp. 84- 101
7. Allen D., *A study of the role of relative clauses in the simplification of news texts for learners of English,* April 2011.

Durante la comprensione di un testo vengono messe in atto diverse abilità mentali non solo linguistiche, ma soprattutto cognitive, quali:

- selezionare;
- analizzare;
- generalizzare;
- classificare;
- dedurre;
- fare previsioni e ipotesi, che vengono poi ridefinite nel corso della lettura;
- collegare le informazioni che vengono presentate nel testo.

Tenendo conto di questi processi, l' insegnate attua la sua semplificazione testuale volta ad una maggior comprensione dei suoi alunni.

Il metodo *strutturale* è oggetto dell' analisi di questa tesi. Segue regole e strutture definite a priori da esperti in materia, ovvero da linguisti capaci di identificare i problemi di comprensione testuale in relazione ad una specifica categoria di lettori.
In questo elaborato, è stato selezionato un corpus che contiene esempi di semplificazione di testi narrativi indirizzati a bambini di età compresa tra i 7 e i 10 anni di età con difficoltà cognitive di comprensione del testo..

Questo metodo è stato seguito dal ***" Progetto europeo Terence "***[8] finalizzato alla pianificazione, allo sviluppo e alla valutazione di un sistema adattivo di apprendimento per *"poor comprehenders"* sia per la lingua italiana che per quella inglese.

8. http://www.terenceproject.eu

Ad esempio, in questo contesto, le costruzioni che sono state tipicamente semplificate nei testi per bambini sono quelle relative alle voci passive, alle proposizioni relative ed ipotetiche e al lessico più complesso dal momento che, ricerche psicolinguistiche sulla comprensione, hanno evidenziato una maggiore coerenza e relazione fra gli elementi in un testo piuttosto che alla semplice somma delle caratteristiche linguistiche delle parole o delle frasi individuali in esso.

Inoltre è stato evidenziato che, durante la lettura, i bambini sono portati a riconoscere e usare i cosiddetti *"coesive links"*, ovvero degli elementi che fanno sì che un bambino (o anche una persona adulta), dopo aver letto un testo, riconosca in questo un dato che gli è particolarmente familiare o di sua appartenenza e, di conseguenza, capace di apprendere le relazioni semantiche nei testi.

Il processo della semplificazione del testo tende a conservare quanto più possibile la struttura linguistica e testuale della storia autentica (e quindi il significato che c'è alla base di qualsiasi storia narrata) .

I bambini che si sforzano di leggere hanno bisogno di leggere testi con un vocabolario sufficientemente stimolante e una sintassi che migliori le loro abilità di lingua e di lettura.

In linea con questo principio (differentemente dagli altri sistemi esistenti) il sistema di semplificazione offre ai lettori livelli graduali di difficoltà, accostandosi progressivamente alla difficoltà che i lettori incontrano nel testo autentico. A tutti i livelli però l'attenzione è posta sulla struttura globale e sulla coerenza del testo, così che anche la versione più semplice del corpus conservi quanto più possibile la struttura narrativa e lo stile della storia originale.

2. ANALISI E DESCRIZIONE DELLE REGOLE DI SEMPLIFICAZIONE E ANNOTAZIONE APPLICATE AI CORPORA

In questo capitolo verranno descritte le regole di semplificazione e analizzate la loro distribuzione nel corpus in esame . Con il termine *annotazione* intendiamo la marcatura di parole e/o porzioni di frasi tramite regole grammaticali prestabilite da linguisti esperti.

Gran parte del lavoro svolto per la stesura di questo elaborato è stato annotare, infatti, le regole di semplificazione, mettendo a confronto i corpora allineati (*originale – semplificato*). Il lavoro di annotazione è avvenuto tramite una piattaforma chiamata *Brat Rapid Annotation Tool* che ha permesso di marcare (tramite specifiche regole di annotazione adottate da esperti) le variazioni grammaticali riscontrate nei testi ogirinali in base alla sua corrispettiva versione semplificata.

Nello specifico, in questo capitolo verranno di seguito spiegati :

1. il corpus analizzato ;
2. il tipo di regole usate ;
3. i risultati della distribuzione quantitativa di queste regole;

2.1 Descrizione dei corpora utilizzati per la semplificazione

In questo elaborato utilizzeremo i testi prodotti all'interno del ***progetto europeo Terence***[9]*,* che, infatti, realizza 4 livelli di semplificazione testuale con la volontà di rendere, nei vari passaggi, una semplificazione testuale coerente con le storie trattate nei testi:

Livello 4 : è la storia autentica di un autore inglese. Questo testo viene ottimizzato da psico-linguisti che attuano un *Livello 3* ;

Livello 3 : ***Global coherence:*** ovvero la coerenza della storia originale, contenente le informazioni necessarie per capire il significato generale della storia, la sequenzialità degli eventi, le locations o la morale della storia.
Questo livello di analisi è curato dagli psico-linguisti che producono la prima semplificazione della storia, producendo così un' informazione proposizionale esplicita che il lettore può desumere. Questo lavoro svolto dagli psico-linguisti produce una semplificazione del testo di *Livello 2* ;

Livello 2 : ***Local coherence***: il testo, semplificato al livello globale, viene semplificato ad un livello locale, aumentando le connessioni logiche tra le frasi, migliorando la coesione e ridimensionando le ambiguità riferite a oggetti, luoghi e

9. Arfè, B., Oakhill, J., Pianta, E., Alrifai, M.: *Story simplification user guide*, Technical report D .2.2, TERENCE Project (2012)

caratteri. I linguisti attuano una revisione del livello 2, producendo così una nuova versione della storia semplificata di *Livello 1*;

Livello 1 : ***Lexicon-grammar*** : il testo semplificato a livello globale e locale, viene semplificato dai linguisti anche a livello lessicale e grammaticale, usando più parole concrete, riducendo espressioni idiomatiche e metaforiche, semplificando sintatticamente le frasi.

Queste semplificazioni possono essere necessarie per quei bambini che trovano maggiori difficoltà nel riconoscere le parole o frasi più complesse.

Le quattro versioni della storia sono state successivamente tradotte dall' inglese all' italiano da parte dei parlanti nativi italiani, facendo un accurato intervento linguistico : come ad esempio l' uso dei pronomi, un livello di vocabolario familiare e semplificazione della sintassi più complessa.

Stando a questi livelli adottati dal progetto Terence, in questo elaborato si sono presi in considerazione i livelli due e uno (rispettivamente l' *"originale"* e il *"semplificato"* del medesimo corpus *)* per determinare le regole d' annotazione in base a ciò che , dall' originale al semplificato, viene cambiato.

I corpora (plurale di corpus che indica un insieme di testi) sono stati selezionati dal progetto europeo Terence ed estrapolati da gli ultimi due livelli di semplificazione testuale redatta dai linguisti (*Livello 2* : *Local coherence*, ovvero i nostri corpora originali e il *Livello 1* : *Lexicon- grammar*, ovvero i nostri corpora modificati).

I gruppi di storie trattate per la nostra semplificazione sono state 5 : per ogni storia avevamo a disposizione dai 5 ai 9 files (in formato *.txt*) in base alla storia esaminata.

Inizialmente i vari files sono stati allineati parallelamente: ogni frase[10], (ovvero l' unità tipica di allineamento), è stata numerata secondo l' ordine di successione nella storia.

Successivamente, ogni frase originale è stata allineata alla sua rispettiva frase modificata, creando così un corpus mono-lingua parallelo.

Tipicamente un corpus parallelo allineato comprende testi nella loro lingua originale definita come *L1*[11], e nella loro traduzione in un'altra lingua (*L2*); nel caso qui esaminato invece la versione allineata rappresenta una versione semplificata, ma sempre nella lingua di partenza.

L' allineamento delle frasi può seguire 2 tipologie principali : può essere, infatti, di tipo *uno a uno :* per ogni frase originale, ne equivale una frase semplificata; oppure di tipo *uno a molti:* per ogni frase originale, ne equivalgono due o più frasi semplificate. Questa ultima tipologia di allineamento può avvenire in seguito ad una divisione della frase originale (risultabile più complicata per via di un numero maggiore di subordinate) in più frasi semplificate (risultabili più semplici e immediate in quanto costituite da un numero maggiore di principali).

10. Il delimitatore di ogni frase è il punto.

11http://it.wikipedia.org/wiki/Quadro_comune_europeo_di_riferimento_per_la_conoscenza_delle_lingue

Di seguito, riportiamo alcuni esempi di allineamento.

Esempio 0. Allineamento delle frasi dal testo originale al semplificato

- *Esempio frase 1 a 1 :*

  frase originale
  <*frase id="9" frase_al="9"*>La barca sembrava abbandonata, così Simo fece un grosso balzo e atterrò sul ponte con un tonfo.<*/frase*>

  frasi semplificata
  <*frase id="9"*>La barca sembrava abbandonata, così Simo fece un grosso balzo e atterrò sul ponte con un tonfo.<*/frase*>

- *Esempio frase 1 a 2 :*

  frase originale

  <*frase id="32" frase_al="32;33"*> Ernesta sistemò la strada in modo da rendere più semplice raggiungere il paese.<*/frase*>

  frasi semplificata

  <*frase id="32"*>Ernesta aggiustò la strada.<*/frase*>
  <*frase id="33"*> Così, grazie a lei, era più facile arrivare in paese. <*/frase*>

- *Esempio frase 1 a molti :*

frase originale

<*frase id="8" frase_al="7;8;9"*>Ernesta si ricordò allora del suo amico Mauro, un famoso scienziato che aveva inventato molte strane macchine, e così corse da Mauro per chiedere il suo aiuto “Mauro! Mauro! Il Presidente Clip ha perso il cappello! Dobbiamo cercarlo.”<*/frase*>

frasi semplificata

<*frase id="7"*>Allora Ernesta pensò di chiedere aiuto al suo amico Mauro.<*/frase*>
<*frase id="8"*>Mauro era un famoso scienziato e aveva inventato molte strane macchine.<*/frase*>
<*frase id="9"*>“Mauro! Mauro! Il Presidente Clip ha perso il cappello! Dobbiamo cercarlo.”<*/frase*>

Tramite i tags <*frase*> <*/frase*> definiscono le marcature delle frasi annotate. L' attributo *“ id= " " ”* identifica il numero della frase del testo originale che andiamo ad annotare in
base al suo ordine nella storia; mentre l' attributo *“ al=" " ”* identifica la frase semplificata di riferimento a quella originale .

I corpora utilizzati in questo elaborato per l' analisi sulla semplificazione sono racconti per bambini di età compresa tra i cinque e i dieci anni di età :

Ernesta Sparalesta Esploratrice di Monica Massaro;

Le avventure di Sofia e Benedetto di Adel Varzegi;

Muoversi di Suzanna Drew- Edwards ;

Ugo Scellino Giramondo di Monica Massaro;

Un'estate da ricordare di Nykki Irvin***;***

2.2 Regole utilizzate per l'annotazione del corpus allineato

In questo paragrafo sono elencate tutte le regole utilizzate per l'annotazione del corpus allineato, le quali specificano quale regola di semplificazione testuale è stata utilizzata. Queste regole sono state redatte dal laboratorio di ricerca *Italian Natural Language Processing Lab*[12] dell'Istituto di Linguistica Computazionale *"Antonio Zampolli"* all'interno del Centro Nazionale delle ricerche di Pisa, vengono annotate utilizzando dei tags (etichette)[13]nella annotazione XML[14].

Tabella 1: Schema delle regole di annotazione per la semplificazione

Classi	**Sotto-classi**
Split	
Marge	
Insert	*Verb* *Subject* *Other*
Delete	*Verb* *Subject* *Other*

12. www.italianlp.it
13. Da ora in avanti considereremo equivalenti regola e tag.
14. L'XML (eXtensible Markup Language) è un linguaggio di markup, ovvero un linguaggio marcatore basato su un meccanismo sintattico che consente di definire e controllare il significato degli elementi contenuti in un documento o in un testo.

Le regole citate in *Tabella 1* sono state descritte nell' articolo di D. Brunato , F. Dell' Orletta G. Venturi e S. Montemagni *" Defining an annotation scheme with a view to automatic text simplification"* (2014). Questi tags (regole di annotazione) sono stati inseriti secondo le tipologie di applicazione e non secondo la loro frequenza di utilizzo. Sono state evidenziate regole di annotazione che attuano una **trasformazione** di parola o una porzione di testo; regole di **inserzione** di un elemento mancante nella frase come una parola oppure il cambiamento radicale della frase stessa (o un pezzo della medesima); infine l'ultima applicazione delle regole di annotazione è stata la **rimozione** di una parola o di una parte intera della frase .

In dettaglio, elenchiamo di seguito tutte le regole d' annotazione precedentemente schematizzate con relativi esempi:

<split> : questa regola viene utilizzata per segnalare che una frase *originale* (es. *proposizione coordinata*) è stata divisa in due o più frasi nel testo semplificato.

Esempio 1.

Frase originale: " Ernesta si ricordò allora del suo amico Mauro, **un famoso** scienziato che aveva inventato molte strane macchine, e così corse da Mauro per chiedere il suo aiuto "Mauro! Mauro! Il Presidente Clip ha perso il cappello! **Dobbiamo cercarlo."**"

Frase semplificata: " Allora Ernesta pensò di chiedere aiuto al suo amico Mauro. **Mauro era un famoso scienziato e aveva inventato molte strane macchine.** "Mauro! Mauro! Il Presidente Clip ha perso il cappello! Dobbiamo **cercarlo**." "

<merge> : questa regola viene utilizzata per marcare due o più frasi autonome nella versione *originale* che sono state rese in una singola frase nella versione *semplificata*.

Esempio 2.
Frase originale: "Ugolino pensò che il suo amico Elio era troppo impegnato per chiacchierare e passò oltre. Andò da Giacomo il fruttivendolo ma anche il fruttivendolo non gli **rivolse neanche una parola.** "

Frase semplificata: " Ugolino pensò che il suo amico Elio era troppo impegnato per chiacchierare **e andò a trovare Giacomo il fruttivendolo, ma anche il fruttivendolo non gli rivolse neanche una parola**."

<spostamento> : questa regola viene utilizzata per marcare uno spostamento di una parte o più parti della frase (es. una *frase subordinata* che nell'*originale* precede la principale mentre nel *semplificato* segue la *principale*)

Esempio 3.

Frase originale: " Dopo essere entrati nel Complesso Sportivo, **la Mamma e la sua amica** si sedettero per prendere una tazza di caffè, mentre Sofia e Michele allargavano i loro asciugamani sulla terrazza accanto alla piscina."

Frase semplificata:“ Dopo che **la Mamma e la sua amica** Tina entrarono nel Complesso Sportivo, si misero sedute per prendere una tazza di caffè, mentre Sofia e Michele allargavano i loro asciugamani sulla terrazza accanto alla piscina.”

1) Trasformazione

< *Lexical Subtitution* > : questa regola viene utilizzata per marcare una sostituzione lessicale (es. uso di un sinonimo più semplice e immediato al fine di far si che i bambini associno più velocemente quella parola ad un determinato oggetto, sensazione emotiva, ecc...) dall'*originale* al *semplificato*. Indica i tratti come attributi delle regole di annotazione.

Esempio 4.

Frase originale: “ Il Nano dell’Isola non era troppo ottimista riguardo a questa idea, perché era ancora **giù di morale** per il risultato del suo esperimento.”

Frase semplificata: “ Il Nano dell’Isola non era troppo entusiasta di questa idea, perché era **stufo e stanco** del suo esperimento fallito.”

< *Anaphoric relacement* > : questa regola viene utilizzata per marcare i casi in cui un pronome è stato sostituito da un sintagma nominale lessicale.

Esempio 5.

Frase originale: " Il giorno in cui i genitori partirono, i bambini **li salutarono** durante la colazione."

Frase semplificata: " Il giorno della partenza, i bambini **salutarono i loro genitori** durante la colazione."

< *Verbal Voice* > : questa regola viene utilizzata per marcare i casi in cui il la radice verbale è stata mantenuta, ma sono cambiati alcuni dei suoi tratti (t*empo, modo, persona* es: *dal passato remoto al passato prossimo o imperfetto*). Anche in questo caso, indica i tratti come attributi dellle regole di annotazione.

Esempio 6.

Frase originale: " "Forse mi **piacciono** le vacanze", pensò tra sé."

Frase semplificata: " Ida pensò tra sé che le **piacevano** proprio le vacanze."

<pass_attivo> : cambiamento della *diatesi verbale* (da passivo ad attivo) (tag da marcare sul verbo).

Esempio 7.

Frase originale: " Solo il papà di Luisa, "Crispino mangia cracker", era dispiaciuto, perché **era stato battuto** da Tonio Battaglia, che aveva mandato giù quattro cracker; un record!"

Frase semplificata:" Solo il papà di Luisa era triste, perché Tonio Battaglia **lo**

aveva battuto. Tonio aveva mangiato quattro cracker in un minuto; un record! ”

< *Verb_to_Noun (nominalization)* > : questa regola viene utilizzata per marcare il caso in cui un verbo, nella versione *semplificata,* diventi un sostantivo.

Esempio 8.

Frase originale: “ Lei **sorrideva** mentre diceva a Sofia e Benedetto che il suo nome era Annabella. ”

Frase semplificata:“ Lei fece **un sorriso** e disse a Sofia e Benedetto che si chiamava Annabella.”

< *Noun_to_ Verb* > : questa regola viene utilizzata per marcare l' annullamento di una nominalizzazione o di una perifrasi nominale, trasformata nella corrispondente struttura verbale.

Esempio 9.

Frase originale: “ “Tu, siediti qui e stai lontano dai guai” minacciò la signora Perticoni, “o niente **pattinata!**” A Luisa e Clara non piaceva Tonio Battaglia, perché faceva sempre brutti scherzi ”

Frase semplificata:“ “Tu, siediti qui e stai lontano dai guai” disse a Tonio, “ o non andrai **a pattinare**!” A Luisa e Clara non piaceva Tonio Battaglia, perché faceva sempre scherzi molto cattivi.”

2) Inserimento

<sogg_espl> : questa regola viene utilizzata per marcare i casi in cui nella frase *originale* ci fosse un soggetto sottinteso che è stato esplicitato nella frase *semplificata*.

Esempio 10.

Frase originale: "**(X)**Sembrava tanto grosso – ma in realtà era un gigante buono."

Frase semplificata:" **Alberto** sembrava tanto grosso – ma era veramente buono."

<verbo_piu> : questa regola viene utilizzata per marcare i casi in cui nella frase *originale* non si presente un verbo che è stato inserito nella frase *semplificata* (può avere gli attributi *"tempo", "modo", "persona")*.

Esempio 11.

Frase originale: " Era una ragazza intelligente e di spirito con **(X)** gli stessi occhi brillanti del fratello."

Frase semplificata: " Era una ragazza intelligente e di spirito e **aveva** gli stessi occhi vivaci del fratello. "

<insert> : questa regola viene utilizzata per marcare altri tipi di inserimento (parole che non sono *soggetto* o *verbo*) oppure sequenze di più parole.

Esempio 12.

Frase originale: "Sofia e Benedetto rivolsero lo sguardo verso la Mamma **con ansia**, sperando di poter portare gli animali a casa. "

Frase semplificata:" Sofia e Benedetto rivolsero uno sguardo **pieno di speranza** verso la Mamma, sperando di poter portare gli animali a casa."

- **Rimozione**

<verbo_meno> : questa regola viene utilizzata per marcare i casi in cui un verbo nella frase *originale* sia stato eliminato nella frase *semplificata*.

Esempio 13.

Frase originale: " Luisa **cercava di non dimostrarlo**, ma a volte tutto quello che desiderava era essere di nuovo nella sua vecchia casa con la sua migliore amica Emma."

Frase semplificata:" **(X)** Qualche volta Luisa avrebbe voluto essere di nuovo nella sua vecchia casa, con la sua migliore amica Emma."

<sogg_sott> : questa regola viene utilizzata per marcare i casi in cui nella frase originale c'è un soggetto esplicito che è stato sottinteso nella frase semplificata.

Esempio 14.

Frase originale: " Grazie a questi calcoli e a queste misure, **Mauro** scoprì che il cappello doveva essere finito in cima al campanile."

Frase semplificata: "**(X)** Così scoprì che il cappello era finito sul campanile."

<delete> : questa regola viene utilizzata per marcare che una frase originale (o una parte di frase) è stata completamente rimossa nella versione semplificata.

Esempio 15.

Frase originale: " La corrente aveva una grande forza e proprio quando i bambini pensarono che Tonio Battaglia sarebbe stato trasportato via **lontano da loro, sentirono lo strattone** di Tonio che afferrava la corda e tutti insieme lo tirarono fuori!"

Frase semplificata:" Il fiume era molto veloce, perciò i bambini pensarono che Tonio Battaglia sarebbe stato portato via **(X)** dall'acqua, ma lui prese la corda e tutti insieme lo tirarono fuori! "

È stato previsto l'uso della regola di annotazione **<manca_regola>** quando nessuna delle regole precedenti poteva essere applicata ai testi per intercettare il tipo di riscrittura o semplificazione.

Per esempio:

Esempio 16:

Frase originale: " **Si stava dibattendo con forza** e sembrava che stesse annegando!"

Frase semplificata:" Sembrava che **non sapesse nuotare** e che stesse annegando! "

Come si evince dall' *Esempio 16* sarebbe più opportuno rimuovere la porzione di frase in grassetto dell' originale (*DELETE*) e inserire la parte di frase semplificata (*INSERT*).

Considerando che " *Si stava dibattendo con forza"* è una conseguenza del "*non sapesse nuotare* ", risulta difficile stabilire un' abolizione della prima porzione di frase a favore dell' inserimento della seconda.. Per questa ragione, non siamo riusciti in questo specifico caso (come in altri rari e particolari casi) ad assegnare la regola di annotazione giusta.

Un' altra considerazione che risulta utile fare riguarda le *nominalizzazioni*: esse sono state successivamente concepite durante il lavoro di annotazione, in quanto si è riscontrata una frequenza abbastanza alta di verbi che, nel testo semplificato, si sono trasformati nel loro rispettivo sostantivo (*nominalizzazione_più*) o, nel caso contrario, nel loro rispettivo verbo (*nominalizzazione_meno*).

Durante il compito d' annotazione di questi corpora, si è usato uno strumento *open source* molto facile e immediato per marcare le varie parole o le porzioni di frase

con la loro regola di annotazione appropriata. Questo tool si chiama *"Brat Annotation Rapid"*[15] che dà l' opportunità di caricare i testi che vogliamo prendere in esame e annotare con facilità le parole o porzioni di frasi contenute nel testo.

Con una sola sottolineatura, infatti, è possibile annotare le entità grazie ad una finestra di visualizzazione delle regole di annotazione inserite precedentemente in un file *.conf*, selezionando successivamente quella di nostro interesse. Brat, inoltre, è considerato un tool *stand-off,* ovvero capace di creare automaticamente un file a parte (in formato *.ann)* per registrare tutte le annotazioni che si vanno a marcare nel testo d' interesse, dandone anche la precisa collocazione. Infatti nel file *.ann* non solo è possibile sapere quale annotazione è stata utilizzata, ma anche dove è stata posizionata.

15. http://brat.nlplab.org/

2.3 Analisi distribuzionale delle regole annotate al testo

Per verificare la frequenza distribuzionale delle regole, è stato sviluppato un programma in *Python*[16] che ha estratto dai corpora le regole di annotazione descritte nel paragrafo 2.2.

Qui sotto viene riportato un estratto di output di un programma che ci ha consentito di recuperare le seguenti statistiche:

1. La frequenza applicativa delle regole di annotazione;

2. Numero di frasi nei corpora analizzati alle quali sono state applicate le varie regole;

3. La frequenza di combinazione di regole nelle frasi in cui è stata applicata solo una combinazione specifica .

Sono stati selezionati solo le prime regole di annotazione ordinate per frequenza. L'output completo è riportato nel capitolo *7. Appendice* .

16. Python è un linguaggio di programmazione sviluppato all'inizio degli anni '90. È particolarmente adatto allo sviluppo di sistemi per il trattamento di dati testuali, e molte librerie e molte funzioni di base sono già presenti nelle librerie standard del linguaggio

1) Frequenza di applicazione delle REGOLE:

```
SOST_LEX 891.0

DELETE   534.0

INSERT   329.0

SPOSTAMENTO 182.0

VERBO_PIU    128.0

TRATTI_VERBO  110.0

SOGG_ESPL    52.0

SPLIT41.0

VERBO_MENO

25.0

Nominalizzazione-22.0

(…)
```

2) Numero di frasi alle quali sono state applicate le varie REGOLE:

```
SOST_LEX 525.0

DELETE   353.0

INSERT   251.0

SPOSTAMENTO 157.0

VERBO_PIU    117.0

TRATTI_VERBO  98.0

SOGG_ESPL    49.0
```

```
SPLIT37.0

VERBO_MENO

24.0

Nominalizzazione-22.0

(…)
```

3) Frequenza di combinazione di regole secondo il numero delle frasi :

```
SOST_LEX 144.0

DELETE SOST_LEX   67.0

DELETE  INSERT  SOST_LEX

43.0 DELETE    36.0

INSERT SOST_LEX   19.0

INSERT   18.0

DELETE SOST_LEX SPOSTAMENTO     15.0

SOST_LEX SPOSTAMENTO   15.0

 DELETE  INS  SOST_LEX  VERBO_P  13.0

 DELETE INSERT 12.0

DELETE INSERT SOST_LEX SPOSTAMENTO   9.0

(…)
```

Osservando le *frequenze assolute*[17] delle regol si notano subito quali sono le regole maggiormente applicate dai linguisti per la semplificazione di un testo.

17. Data una parola, contare quante volte ricorre all'interno del testo.

Analizzando per esempio l'utilizzo della regola <*sost_lex*>, che ricorre ben 891, si può osservare come i linguisti abbiano sfatto ampio ricorso all' uso di sinonimi semplici rispetto a parole complesse.

A seguire si riportano qualche esempi :

frase originale: " Il cappotto e le scarpe di Clara davano l'idea di non essere ***mai asciutti*** da settimane e la bambina aveva dovuto persino ritirare fuori i suoi vecchi stivali di gomma."

frase semplificata: " Il cappotto e le scarpe di Clara erano **sempre bagnati** a causa della pioggia e la bambina dovette indossare i suoi vecchi stivali di gomma. "

frase originale: "Aidan sbirciò dalla porta della roulotte e, nell'oscurità, vide all'interno una scatola ***sudicia*** con dentro due minuscoli cuccioli."

frase semplificata: " Aidan sbirciò dalla porta della roulotte e, nell'oscurità, vide all'interno una scatola ***sporchissima*** con dentro due minuscoli cuccioli."

Un' altra regola di annotazione molto ricorrente è il <*verbo_piu*> che ha una frequenza di 534 . Questa regola è stata usata sopratutto per aumentare il livello di comprensione di alcune frasi che determinavano un' azione da parte del soggetto o del personaggio trattato : come ad esempio, nella frase originale, *" Infine telefonò a Mario l'idraulico per farsi portare trecento tubi di plastica."* oltre ad essere stato introdotto un soggetto *esplicito,* è stato inserito il verbo *"domandare"* nella frase semplificata per precisare l' azione compiuta dal soggetto : " **Ernesta** telefonò anche a Mario l'idraulico e gli ***domandò*** di portare trecento tubi di plastica."

Inoltre, un' altra regola di annotazione molto ricorrente è lo <*spostamento*>. Con una frequenza di 182, la regola ha aiutato la comprensione testuale spostando parole o porzioni di frasi già presenti, in modo da chiarire i legami all' interno della frase: ad esempio nella frase originale " Dopo essere entrati nel Complesso Sportivo, ***la Mamma e la sua amica*** si sedettero per prendere una tazza di caffè, mentre Sofia e Michele allargavano i loro asciugamani sulla terrazza accanto alla piscina." diventa nella frase semplificata: "Dopo che ***la Mamma e la sua amica*** Tina entrarono nel Complesso Sportivo, si misero sedute per prendere una tazza di caffè, mentre Sofia e Michele allargavano i loro asciugamani sulla terrazza accanto alla piscina. ". Due altre regole di annotazione molto ricorrenti sono lo <*delete*> e il <*insert*> .

Per quanto riguarda la prima regola (con frequenza di 524) , sono state eliminate parole o porzioni di frasi talvolta rindondanti o compesse per il target di lettori di riferimento :

frase originale: " Poi, ***chiese*** *in prestito* il telefono di Luigi e telefonò a Nicola l'imbianchino chiedendogli di procurare dieci secchi di vernice bianca."

frase semplificata: " Poi, ***con*** il telefono di Luigi, chiamò Nicola l'imbianchino e gli chiese di comprare dieci secchi di vernice bianca. " o semplicemente pleonastiche:

frase originale: "Ernesta Sparalesta l'esploratrice, durante uno dei suoi numerosi viaggi, era partita ***in missione*** nello spazio sulla sua bicicletta di legno trasformata in un razzo superveloce."

frase semplificata: "Ernesta Sparalesta l'esploratrice, durante uno dei suoi numerosi viaggi, era partita **(X)** per lo spazio sulla sua bicicletta di legno trasformata in un razzo superveloce."

frase originale: "***Ebbe inoltre la brillante idea di*** passare in pescheria per chiedere ad Antonio, il pescivendolo, di procurare del pesce vivo."

frase semplificata: " **(X)**Poi andò in pescheria per chiedere ad Antonio, il pescivendolo, di portare un sacco pieno di pesci vivi."

Per quanto riguarda la seconda regola (che ricorre 329), è bene sottolineare il suo molteplice utilizzo sia nei riguardi di inserimenti minimali (come congiunzioni subordinative e coordinative) che nelle porzioni vere e proprie di frasi allo scopo di esplicitare il contenuto come da esempio:

> *frase originale*: " Ida si svegliò di soprassalto per un terribile rumore ***come*** di alberi che cadevano."
>
> *frase semplificata*: " Ida si svegliò di soprassalto per un terribile rumore ***che sembrava di un tuono oppure*** di alberi che cadevano."

Un' altra regola di annotazione molto ricorrente è il <*tratti_verbo*> con una frequenza di 110. Andando a indagare tra i valori degli attributi di questa regola possiamo notare che molti verbi hanno cambiato il loro tempo da passato remoto a presente o imperfetto, ovvero tempi e modi più frequenti e semplici.

> *frase originale*: " Benedetto **nuotò** a cane, senza stile, più veloce che poteva, ma ancora una volta non ce la fece ad arrivare al bordo vasca opposto della piscina prima di Luca. "
>
> *frase semplificata*: Benedetto **nuotava** più veloce che poteva, ma non ce la fece ad arrivare al bordo vasca dalla parte opposta della piscina prima di Luca. "

In altri casi, inoltre, da un costrutto verbale composto si è dato spazio ad un semplice tempo imperfetto come ad esempio in queste due frasi seguenti:

frase originale: " Emanuele pensò che ***sarebbe stato*** divertente prendersi cura di un cane così grosso! ";

frase semplificata: " Emanuele pensò che ***era*** divertente portare a spasso quel cane grosso! "

Al posto di un participio passato si è dato maggior rilievo ad un semplice infinito come ad esempio in queste due frasi:

frase originale: "Il mattino seguente, il sole era quasi ***sorto*** nel cielo quando gli abitanti del paese riuscirono a radunare i cani e i gatti al negozio del paese."

frase semplificata: "Il mattino seguente, gli abitanti del paese riuscirono a radunare i cani e i gatti al negozio del paese al ***sorgere*** del sole."

In altri casi ancora, infine, nella frase originale è stata solo cambiata il tratto di persona, come ad esempio:

frase originale: "Il tacchino ripieno era il piatto preferito della Mamma e poiché tutta la famiglia la stava festeggiando, volevano che tutto ***fosse*** come piaceva a lei."

frase semplificata: “ Il tacchino ripieno era il piatto preferito della Mamma e tutta la famiglia voleva che le cose ***fossero*** come piaceva a lei, perché la festa di compleanno era la sua.”.

Nel capitolo successivo si effettuerà un' analisi qualitativa delle regole semplificazione, considerando i risultati ottenuti.

3. ANALISI LINGUISTICO-COMPUTAZIONALE DELLE REGOLE DI SEMPLIFICAZIONE

Dopo aver descritto le varie regole di semplificazione adottate nei corpora analizzati e aver introdotto alcuni esempi rilevanti (mettendo a confronto i corpora allineati sulla base della loro versione originale / semplificata), in questo capitolo verranno di seguito spiegati :

1. L' analisi e i risultati della semplificazione dal punto di vista qualitativo;

2. L' analisi e i risultati di alcune regole di semplificazione adottate su una base qualitativa, indagando sopratutto sull' eventuali risultati statistici che quest' ultimi hanno avuto nella semplificazione testuale;

3. Infine si indagherà sulla regola *Split* dandone un' accurata valutazione qualitativa sull' effetto, riportando esempi pratici ;

In questo contesto verrà introdotto nell' analisi uno strumento per la valutazione automatica della leggibilità, **READ-IT** .

3.1 Analisi qualitativa delle regole di semplificazione

Fino adesso è stato osservato come sono state distribuite le regole e quali combinazioni si possono trovare all'interno del corpus annotato; ora si valuterà il monitoraggio linguistico del testo, comparando corpora originali e semplificati rispetto alle caratteristiche linguistiche monitorate. Inoltre, si analizzerà l'effetto delle regole di semplificazione rispetto alla leggibilità del testo.

Per esaminare l' effetto della leggibilità del testo è stato utilizzato un tool chiamato **READ-IT**[18](Dell'Orletta et al, 2011), un'applicazione web capace di valutare la leggibilità di un testo e di estrarne il profilo linguistico.

READ-IT è basato su una combinazione di tratti linguistici che spaziano tra diversi livelli di

descrizione linguistica: l*essicale, morfo-sintattico e sintattico.*

Le caratteristiche di READ-IT consistono in una valutazione della leggibilità articolata su due livelli: il documento e la singola frase.

Attraverso l'identificazione dei luoghi di complessità del testo (in termini di frasi) che necessitano di revisione e semplificazione, accompagnata da una classificazione semantica del tipo di difficoltà rilevata, READ-IT può anche essere utilizzato come ausilio per la semplificazione del testo.

18. www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest

Prima operazione svolta con READ-IT è stata la comparazione tra il totale complessivo dei due testi originali e semplificati, per verificare se effettivamente i suoi indici identificano gli interventi di semplificazione.

Per spiegare maggiormente i risultati ottenuti dall' inserimento dei nostri corpus in READ-IT, è doveroso descrivere i vari profili che costituiscono l' applicazione come i vari indici contenuti in essi, mostrando i dati estrapolati dai corpus in esame.

READ-IT restituisce un monitoraggio delle caratteristiche linguistiche rispetto a diversi profili :

Profilo di base;

Profilo lessicale;

Profilo sintattico;

I suddetti profili sono spiegati nella S*ezione 3.1.1.* di questo capitolo.

In tutti i casi, oltre al valore numerico del parametro, viene fornita in *READ-IT* una rappresentazione grafica di confronto dati.

La rappresentazione grafica mette confronto il dato relativo al testo analizzato che corrisponde alla (barra azzurra) con la corrispondente informazione rilevata nei corpora di riferimento facile (barra verde) e difficile lettura (barra rossa).

Un valido esempio di corpus di difficile lettura è *" La Repubblica*[19]*"*, (del Gruppo Editoriale L' Espresso); mentre il giornale *" 2 parole"*[20]*,* creato e diretto da Tullio De Mauro , risulta un valido esempio di corpus di facile lettura e comprensione .

3.1.1 Risultati del monitoraggio linguistico del testo

La sezione della scheda corrispondente alla valutazione globale della leggibilità del documento fornisce i risultati del monitoraggio di un sottoinsieme delle caratteristiche linguistiche utilizzate da READ-IT nella misurazione della leggibilità.

Di seguito vengono riportate le informazioni dettagliate dei profili[21] di *base, lessicale e*
sintattico del testo, organizzate in tre sezioni :

1. **Profilo di base :** Questo profilo sfrutta le caratteristiche di base del testo valutate secondo :

   - *Numero totale periodi*: ovvero, il numero di periodi (o frasi) in cui si articola il testo analizzato, considerando sempre la punteggiatura e il ritorno a capo come separatori di periodo ;

19. http://www.repubblica.it/
20. http://www.dueparole.it/default_.asp
21. Dell'Orletta F., Montemagni S., Venturi G. "*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, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

- *Numero totale parole (tokens)*: ovvero, il numero di occorrenze di parole (o tokens) in cui si articola il testo preso in analisi;
- *Lunghezza media dei periodi (in tokens)*: questo dato calcola la lunghezza media dei periodi, espressa in tokens, all'interno analizzato;
- *Lunghezza media delle parole (in caratteri)*: questo dato calcola la lunghezza media delle parole, espressa in caratteri, all'interno del corpus analizzato.

Figura 1. output di READ-IT riguardante i testi *originali* sulla base del profilo di base.

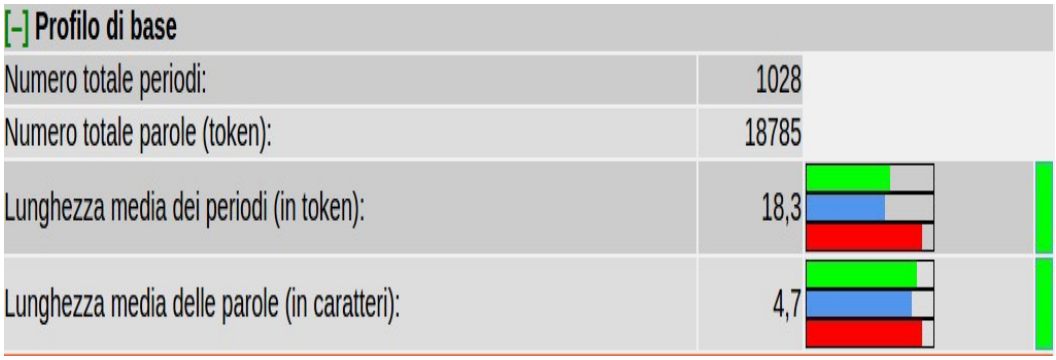
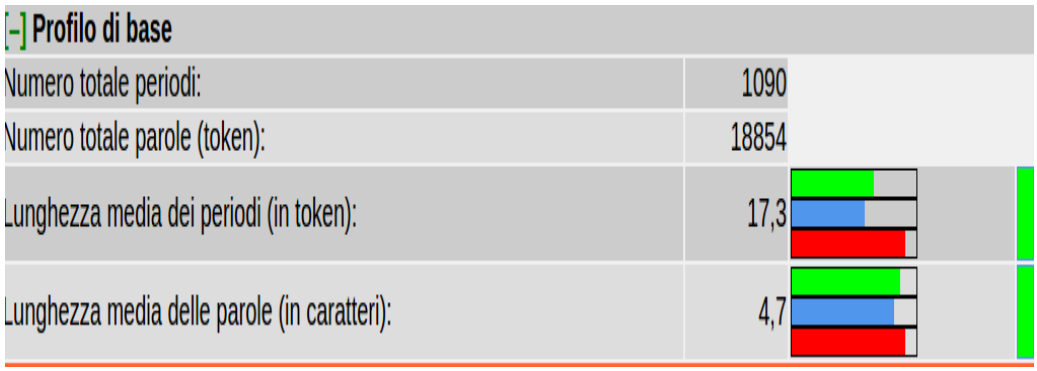


Figura 2. output di READ-IT riguardante i testi **s***emplificati* sulla base del profilo di base.



Come si può osservare dalla *figura 1* e *figura 2* e in *Tabella 2*, c'è stata una diminuzione dei valori nella versione semplificata.

Tabella 2. Percentuali READ-IT, nell'ordine testi originali e semplificati del profilo di base.

Profilo di base	**Testo originale**	**Testo semplificato**
Numero totale periodi	1028	1090
Numero totale parole (in token)	18785	18854
Lunghezza media dei periodi (in token)	18,3	17,3
Lunghezza media delle parole (in	4,7	4,7

In particolare, notiamo che il *numero totale di periodi* nel corpus semplificato è del 1090. rispetto al 1028 del corpus originale. Questo fenomeno è dovuto a l' uso frequente dello *SPLIT* che ha favorito un aumento dei periodi brevi per una comprensione maggiore del testo come si evince dalla *lunghezza media in termini di periodi* che passa da 18.3 dei testi originali al 17.3 dei testi semplificati.
Dovuto proprio a l' adozione della regola *SPLIT, la lunghezza media dei periodi in termini di tokens* è diminuita mentre *la lunghezza media delle parole in caratteri* è rimasta la stessa: questo ultimo dato è molto interessante perché dimostra che, nonostante ci siano state moltissime sostituzioni lessicali, inserzioni e rimozioni (come già visto nel paragrafo *2.3* sulla frequenza di applicazione delle regole) le

tre regole di annotazione di riferimento non hanno avuto un impatto sulla lunghezza delle parole in termine di caratteri.

Inoltre, notiamo che *il numero totale delle parole in termini di tokens* è aumentato : questo aumento è relazionato al fatto che una frase divisa in due o più frasi grazie all' uso dello *SPLIT* determina, molto spesso, un' inserzione di parola (*INSERT)* o l' inserimento di un *SOGGETTO ESPLICITO* o di un *VERBO PIÙ* (anche se in tempi verbali più comprensibili per bambini con difficoltà cognitive come presente, imperfetto e infinito).

Infine, come possiamo evincere dalla *fugura 1* e *figura 2,* la rappresentazione grafica del confronto dati rivela che il testo originale sia già di per se semplice da leggere (barra blu) confrontato con la barra verde (livello stimato a 19,2) in base ai paramentri di facile lettura (rappresentate dallestatistiche estratte dal giornale " *2 parole* " di T. De Mauro precedentemente citato nel paragrafo 3.1) e difficile lettura dei testi (livello stimato a 100) (il quotidiano *La Repubblica*) . Da questo si evince come il corpus preso in esame, scritto originariamente per un pubblico di lettori di età inferiore ai 10 anni, sia stato in seguito semplificato ulteriormente dai linguisti.

2. Profilo lessicale :

Il profilo utilizza dizionari per valorizzare il lessico presente nei testi. La ricerca per indagare sulla ricchezza lessicale è stata valutata secondo:

- *Composizione del vocabolario:* un parametro riguardante la tipologia del vocabolario usato, ovvero l'insieme delle parole tipo[22] che ricorrono all'interno del documento.

Come dizionario di riferimento, si è preso in considerazione il ***Grande Dizionario Italiano dell'uso*** (VdB) (GRADIT, De Mauro, 2000)[23]; nella prima riga, è riportata la percentuale di vocabolario del testo appartenente al VdB. Si tratta di una risorsa lessicale della lingua italiana, creata dal linguista *Tullio de Mauro*, che comprende circa 7.000 parole, quelle che hanno la maggiore frequenza statistica di utilizzo nella nostra lingua, ovvero quelle maggiormente utilizzate e di nostra familiarità.

Il vocabolario di base si divide in:

1. *Vocabolario fondamentale*, composto da 1.991 parole. Sono le più usate in assoluto 31 nella nostra lingua (esempi: *amore, lavoro, pane*) .
2. *Vocabolario di alto uso*, composto da 2.750 parole. Sono molto usate, ma meno di quelle del Vocabolario fondamentale (esempi: *palo, seta, toro*).

22. Si definisce vocabolario di un testo l'insieme delle parole tipo che ricorrono al suo interno.
23. http://it.wikipedia.org/wiki/Vocabolario_corrente .

*3. Vocabolario di alta disponibilit*à, composto da 2.337 parole. Sono poco usate nella lingua scritta, ma molto in quella parlata (esempi: *mensa, lacca, tuta*).

• *Rapporto tipo/unità* (calcolato rispetto alle prime 100 parole del testo): questa misura, conosciuta anche come *"Type/Token Ratio"* (o TTR) è uno dei metodi più diffusi per misurare la varietà lessicale di un testo. Questo indice mette in rapporto il numero delle occorrenze delle unità del vocabolario di un testo (al denominatore) con il numero di parole tipo (al numeratore)[24]i valori oscillano tra 0 e 1, dove valori vicini allo 0 indicano che il vocabolario del testo è meno vario lessicalmente mentre valori vicini a 1 caratterizzano testi particolarmente variegati .

Essendo la TTR un indice sensibile alla lunghezza del testo, quest'ultimo viene calcolato rispetto a campioni di testo della stessa lunghezza (nella versione corrente tale limite è stato fissato alle prime 100 parole unità di un testo).

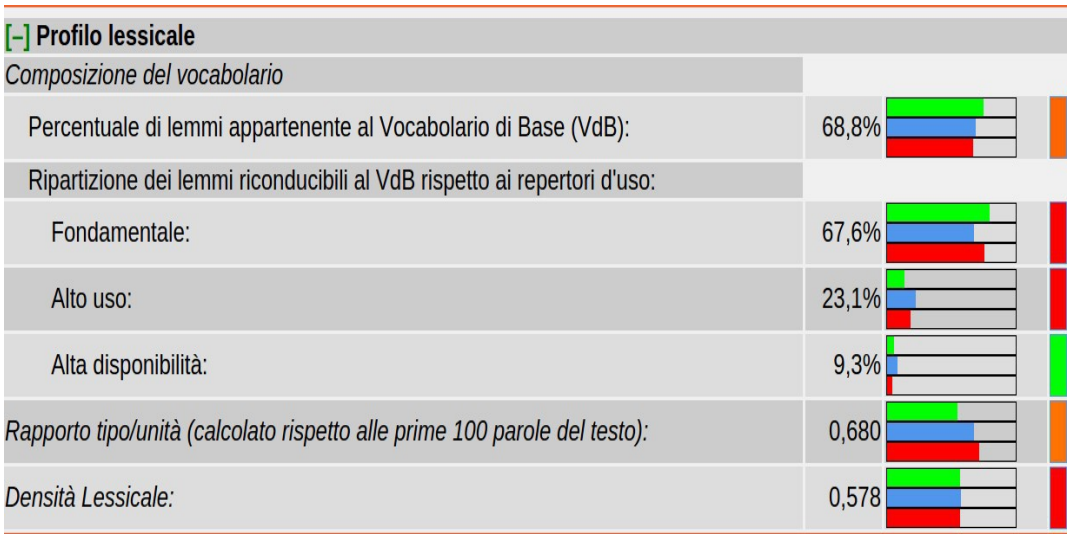
• *Densità Lessicale*: questo parametro riguarda il rapporto tra parole piene (ovvero portatrici di significato) e parole funzionali all'interno di un testo, in modo particolare la sua "densità lessicale" (abbreviata come DL) calcolata come la proporzione delle parole semanticamente "piene"(ovvero, nomi,

24. Dell'Orletta F., Montemagni S., Venturi G. "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, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

aggettivi, verbi e avverbi) rispetto al totale delle occorrenze di parola all'interno del testo.

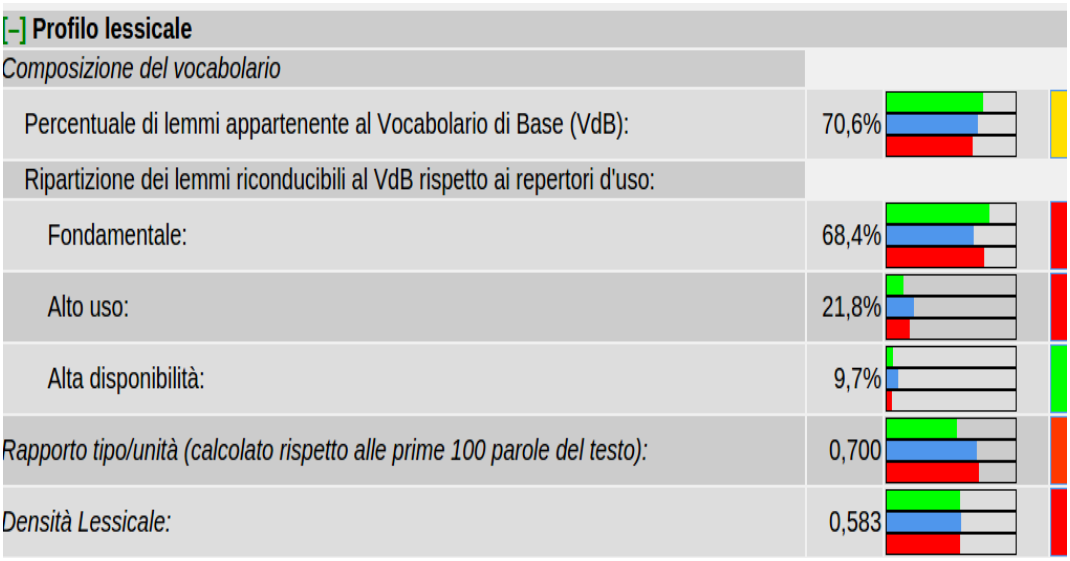
Stando alla letteratura, valori più alti di DL corrispondono in linea di massima a maggiore leggibilità. In tutti i casi, oltre al valore numerico del parametro viene fornita una rappresentazione grafica che mette a confronto il dato relativo al testo oggetto dell'analisi (corrispondente alla barra azzurra) con la corrispondente informazione rilevata nei corpora di riferimento di facile (barra verde) e difficile (barra rossa) lettura.

Figura 3. output di READ-IT riguardante i testi *originali* sulla base del profilo lessicale.



[–] Profilo lessicale	
Composizione del vocabolario	
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	68,8%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:	
Fondamentale:	67,6%
Alto uso:	23,1%
Alta disponibilità:	9,3%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,680
Densità Lessicale:	0,578

Figura 4. output di READ-IT riguardante i testi *semplificati* sulla base del profilo lessicale.



[–] Profilo lessicale	
Composizione del vocabolario	
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	70,6%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:	
Fondamentale:	68,4%
Alto uso:	21,8%
Alta disponibilità:	9,7%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,700
Densità Lessicale:	0,583

Come si può notare dalla *Tabella 3, la percentuale di lemmi*[25] *appartenente al VdB* è aumentata per via di una notevole adozione di *INSERT* , *VERBO_PIÙ e SOST_LEX* (se si vanno ad inserire lemmi più familiari) nei testi semplificati che hanno favorito un aumento delle parole appartenenti al vocabolario di De Mauro. Alla stessa maniera, il *Vocabolario fondamentale* e quello di *alta disponibilità* hanno aumentato (se pur di poco) il loro valore nei testi semplificati : questo ha determinato un incremento maggiore di parole più usate nella nostra lingua e, molte di esse, sono legate alla lingua parlata.

I *rapporto tipo / unità* per le prime 100 parole , inoltre, è aumentato dello 0,02 in quanto la divisione tra le parole del vocabolario e le parole tipo del testo ha dato un risultato che si avvicina ad 1 (quindi c'è stata una maggiore varietà lessicale) .
Questo dato è dovuto a l' adozione di sostituzioni lessicali (*SOST_LEX)* e inserzioni (*INSERT)* più alte che hanno favorito un aumento lessicale, nonostante ci sia stata una frequenza notevole di DELETE come visto nel paragrafo 2.3 sull' analisi distribuzionale delle regole di annotazione. Alla stessa maniera, l' aumento dei verbi (*VERBO_PIÙ)* nel testo ha determinato un incremento maggiore di parole (nonostante i tratti verbali siano stati più semplici e immediati).
Nonostante ci sia stata una frequenza considerevole di DELETE, l' adozione di trasformazione (*SOST_LEX)* e di inserzione (*INSERT ;VERBO_PIÙ)* hanno favorito un aumento della varietà lessicale.

25. Il lemma è anche, per estensione, l'articolo o la voce che, in un'enciclopedia o un dizionario, spiega il significato e l'uso di una parola o del suo lessema.

Tabella 3. Percentuali READ-IT , nell'ordine testi originali e semplificati nel profilo lessicale.

Profilo lessicale	**Testo originale**	**Testo semplificato**
Percentuale di lemmi appartenente al VdB	68,80 %	70,60%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d' uso :		
Fondamentale	67,60%	68,40%
Alto uso	23,10%	21,80%
Alta disponibilità / lingua parlata	9,30%	9,70%
Rapporto tipo / unità (prime 100 parole)	0,68	0,7
Densità lessicale	0,58	0,58

Questi dati sono significativi in quanto dimostra la volontà di incrementare parole diverse (se pur non eccessivamente lessicalmente complesse) per incitare il lettore ad acquisire termini nuovi e aumentare così il suo livello di istruzione. Questa considerazione appena espressa è maggiormente avvalorata dal dato sulla *densità lessicale* che è aumentato dello 0,005 proprio per via di un' aumento di parole con un grado comprensione accessibile .

Infine, il *Vocabolario di alto uso* è passato dal 23,1 % nel testo originale al 21,8% nel testo semplificato: questo dato è coerente con gli altri criteri di valutazione

precedentemente espressi in quanto rivela la diminuzione di parole discretamente diffuse nella lingua italiana e una maggiore inserzione di *" parole fondamentali "* (ovvero le più usate in assoluto) e di *"parole con alta disponibilità "* (ovvero le più usate nel linguaggio parlato).

3. Profilo sintattico :

Il profilo sfrutta le caratteristiche linguistiche più sofisticate relativa alla struttura grammaticale.

Il profilo sintattico di un testo si articola in due parti, a seconda che l'informazione monitorata riguardi :

1) l'analisi morfo-sintattica del testo, ovvero il livello in cui a ogni "token" del testo viene associata la categoria grammaticale che la parola ha nel contesto specifico ;

2) la struttura sintattica sottostante basata su una descrizione della frase in termini di relazioni di dipendenza tra parole, come "soggetto", "oggetto diretto", "modificatore", etc..

3) "Distribuzione" delle categorie grammaticali: questo dato permette di valutare se ci sono analogie e/o differenze tra diversi generi testuali[26]sulla base della distribuzione di categorie grammaticali.

In questa sede, si riporteranno i valori relativi a un sottoinsieme di categorie grammaticali, ovvero sostantivi (distinguendo tra nomi comuni e propri), aggettivi, verbi e congiunzioni.

26 La lingua scritta e parlata è un fatto ampiamente riconosciuto nella letteratura linguistica.

Per quanto riguarda le congiunzioni, viene riportata la ripartizione in coordinanti e subordinanti.

- *Articolazione interna del periodo*: questo costituisce un parametro complesso volto a caratterizzare l'organizzazione interna del periodo.

Esso include informazioni come:

a. *il numero medio di proposizioni per periodo*: si tratta di un dato elementare, costituito dal rapporto tra proposizioni e periodi.

Chiaramente,con l' aumento di proposizioni cresce la complessità sintattica del corpus. Con l' aumento di questo rapporto, cresce la complessità del testo.

b. *proposizioni principali vs subordinate*: questo dato registra la proporzione di principali e subordinate. Chiaramente, l' aumento di costruzioni subordinate contribuisce in modo significativo alla complessità grammaticale del testo;

- *Articolazione interna della proposizione*: è descritta nei termini di

a. numero medio di parole per proposizione;

b. numero medio di dipendenti per testa verbale;

- *"Misura" della profondità dell'albero sintattico*: un altro aspetto rilevante per misurare la complessità del testo riguarda i livelli gerarchici : in presenza

di più di una proposizione subordinata all'interno dello stesso periodo, diventa importante ricostruire quale tipo di rapporto sussista tra di esse, ovvero se siano *"ricorsivamente incassate l'una all'interno dell'altra"(Documentazione manuale di read- it)*.

Una prima e approssimativa indicazione dei livelli di incassamento gerarchico all'interno della struttura sintattica può essere ricostruita a partire dall'altezza massima dell'albero, che misura la massima distanza che c'è tra una foglia (rappresentata da parole del testo senza dipendenti) e la radice dell'albero, espressa come numero di archi (ovvero relazioni di dipendenza) attraversati nel cammino foglia-radice. Il parametro "media delle altezze massime" riporta il valore medio delle altezze massime degli alberi a dipendenza registrate all'interno del testo analizzato.

Questa misura viene approfondita maggiormente rilevando ulteriori tipi di costrutti sintattici:

a. *la ricorrenza di strutture nominali complesse;*

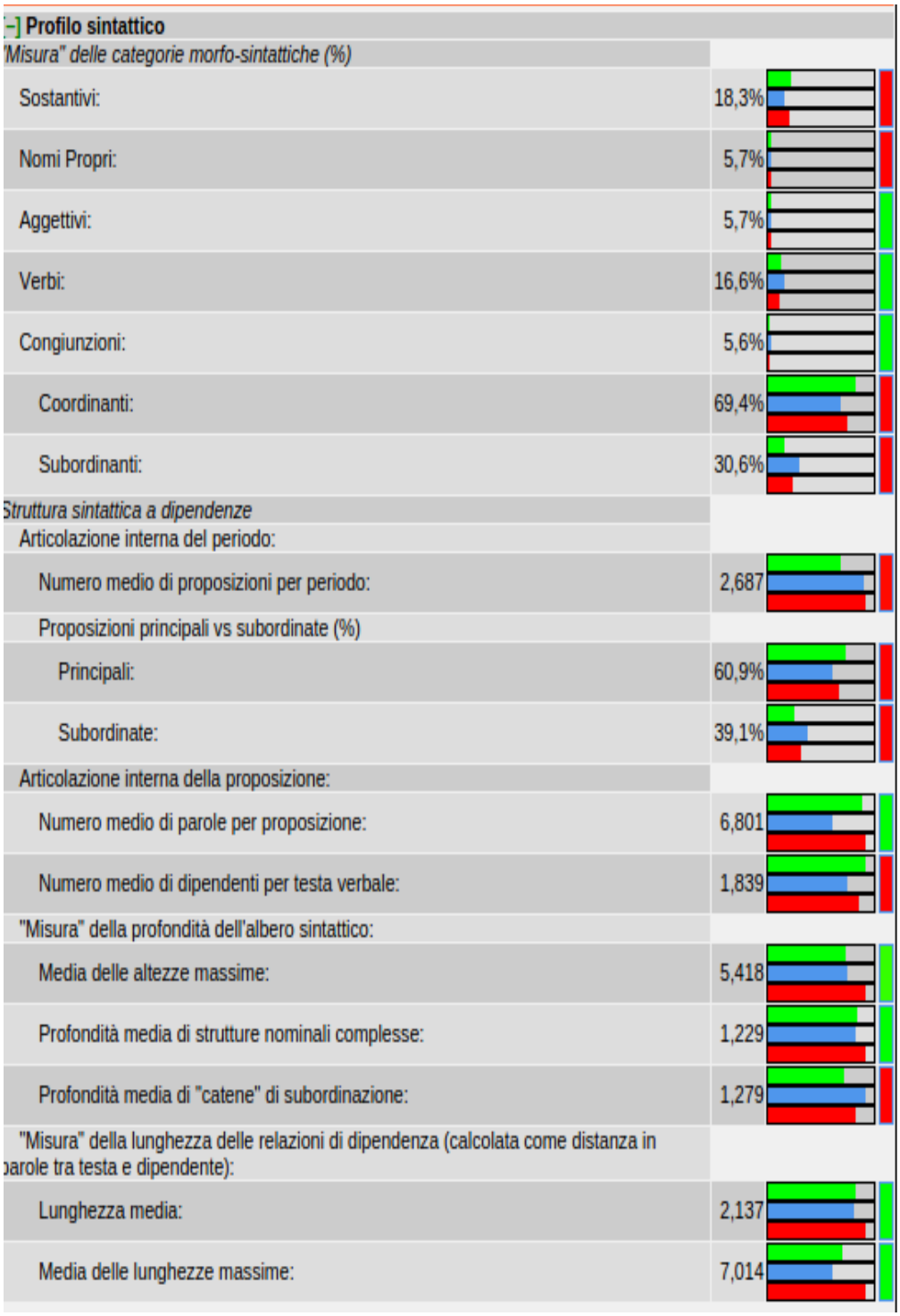
b. *la ricorrenza di proposizioni subordinate ricorsivamente incassate* (questo dato è riportato come " Profondità media di 'catene' di subordinazione ").

"Misura" della lunghezza delle relazioni di dipendenza: è chiaro che la contiguità di elementi semanticamente e/o sintatticamente 'vicini' consente un' immediata recuperabilità e accessibilità dei rapporti che intercorrono tra le parole.

> La “lunghezza” delle relazioni di dipendenza, calcolata come la distanza in parole tra la testa e il dipendente, rappresenta quindi un fattore di complessità ampiamente riconosciuto nella letteratura linguistica, psicolinguistica e linguistico - computazionale.

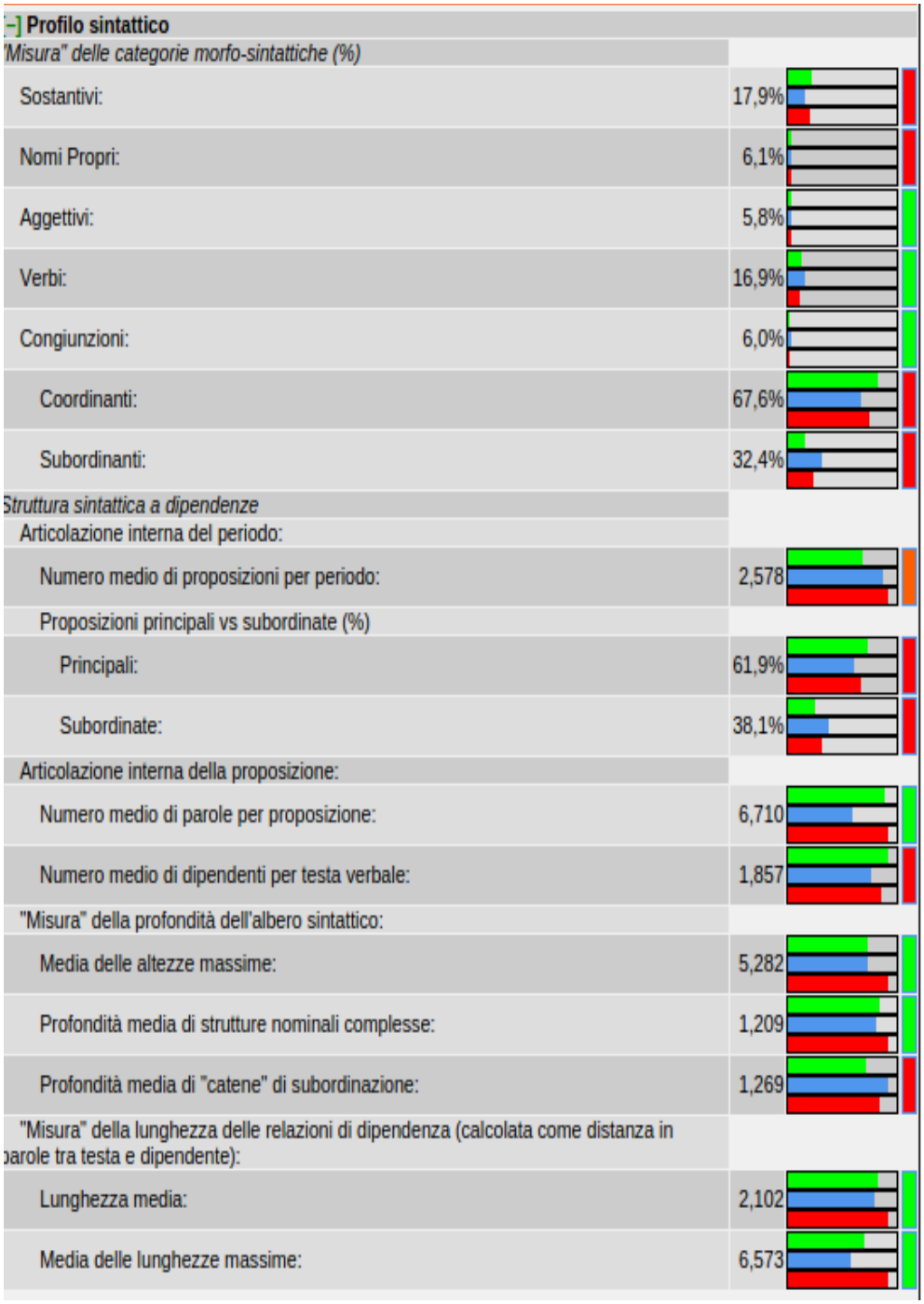
Questo aspetto della struttura sintattica viene monitorato attraverso due parametri, corrispondenti alla media della lunghezza di tutte le relazioni di dipendenza e alla media dei legami di dipendenza più lunghi per ciascuna frase.

Figura 5. output di READ-IT riguardante i testi *originali* sulla base del profilo sintattico



[-] Profilo sintattico	
"Misura" delle categorie morfo-sintattiche (%)	
Sostantivi:	18,3%
Nomi Propri:	5,7%
Aggettivi:	5,7%
Verbi:	16,6%
Congiunzioni:	5,6%
Coordinanti:	69,4%
Subordinanti:	30,6%
Struttura sintattica a dipendenze	
Articolazione interna del periodo:	
Numero medio di proposizioni per periodo:	2,687
Proposizioni principali vs subordinate (%)	
Principali:	60,9%
Subordinate:	39,1%
Articolazione interna della proposizione:	
Numero medio di parole per proposizione:	6,801
Numero medio di dipendenti per testa verbale:	1,839
"Misura" della profondità dell'albero sintattico:	
Media delle altezze massime:	5,418
Profondità media di strutture nominali complesse:	1,229
Profondità media di "catene" di subordinazione:	1,279
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente):	
Lunghezza media:	2,137
Media delle lunghezze massime:	7,014

Figura 6. output di READ-IT riguardante i testi *semplificati* sulla base del profilo sintattico



Come possiamo notare dai risultati mostrati nella *Tabella 4*, i *Nomi Propri,* gli *aggettivi* e i *verbi* hanno aumentato il loro valore percentuale per via di un notevole utilizzo di regole come *INSERT*, *VERBO_PIÙ* , *ANAFORA* e *SOGG_ESP* .

Questi risultati sono maggiormente avvalorati se notiamo l' incremento di *frasi principali* nei testi semplificati : logicamente, una volta adottato lo *SPLIT* per dividere le frasi è necessaria un' attività di inserzione verbale e sostantivale.

In relazione al dato sull' incremento delle frasi principali, le *frasi subordinate* sono diminuite nei testi semplificati, proprio per l' attuazione di regole come *SPLIT e DELETE.*

Il *numero medio di dipendenti per testa verbale*, inoltre, è più alto : una delle ragioni che ha favorito l' aumento di questo dato da 1,839 (testo originale) al 1,857 (testo semplificato) è l' inserzione del SOGG_ESP che ha favorito un numero maggiore di dipendenti in un discorso e di DELETE che ha diminuito il numero delle dipendenti rendendo così più lineare e comprensibile le relazioni di dipendenza.

Inoltre, notiamo che *il numero medio di proposizioni per periodo* e *il numero di parole per proposizione* sono diminuiti grazie alle regole di annotazione precedentemente espresse.

L' articolazione interna della proposizione, in linea di massima, è stata ridotta drasticamente nei testi semplificati: questo dato ci dimostra quanto la “misura” della profondità dell' albero sintattico si sia ridimensionata.

In dettaglio, notiamo infatti come la *profondità media di strutture nominali complesse* e *la media di “catene” di subordinazione* siano diminuite

rispettivamente dello 0,02 (testo originale) e dello 0,01 (testo semplificato) : questo dato è ricondotto ad un aumento delle frasi principali nei testi semplificati e all' utilizzo del *DELETE* che ha ulteriormente favorito la riduzione della catena di subordinazione delle frasi. Infine, la *" Misura " della lunghezza delle relazioni di dipendenza* (calcolata come la distanza in parole tra testa e dipendente) in relazione alla *lunghezza media* e alla *media delle lunghezze massime* si è ridotta notevolmente il che spiega la riduzione della distanza tra la testa e la dipendente all' interno di una stessa proposizione. Questo dato è sicuramente ricondicubile all' utilizzo non solo delle regole SPLIT e DELETE ma anche alla regola di SPOSTAMENTO in quanto ha creato un avvicinamento delle parole tra la testa e la dipendente.

Tabella 4. Percentuali READ-IT , rispetto ai testi originali e semplificati nel profilo sintattico

Profilo sintattico	**Testo originale**	**Testo semplificato**
Sostantivi	18,3 %	17,9 %
Nomi propri	5,7 %	6,1 %
Aggettivi	5,7 %	5,8 %
Verbi	16,6 %	16,9 %
Congiunzioni	5,6 %	6,0 %
Coordinati	69,4 %	67,6 %
Subordinati	30,6%	32,4 %
Numero medio di proposizioniper periodo:	2,687	2,578
Proposizioni principali VS subordinate (%)		
Principali	60,9%	61,9 %

Subordinate	39,1%	38,1 %
Articolazione interna della proposizione		
Numero medio di parole per proposizione	6,801	6,710
Numero medio di dipendenti per testa verbale	1,839	1,857

Profilo sintattico	**Testo originale**	**Testo originale**
Media delle altezze massime	5,42	5,28
Profondità media di strutture nominali complesse	1,229	1,209
Profondità media di "catene"di subordinazione	1,279	1,269
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come la distanza in parole tra testa e dipendente)		
Lunghezza media	2,14	2,1
Media delle lunghezze massime	7,014	6,573

3.1.2 Analisi globale della leggibilità

A partire dall' insieme delle caratteristiche linguistiche monitorate nei paragrafi precedenti, READ-IT calcola un indice di leggibilità articolato su quattro profili,ognuno dei quali definito su diverse configurazioni caratteristiche.

Qui si andrà a verificare l' effetto delle regole di semplificazione rispetto alla leggibilità del testo come misurata in READ-IT, intesa come il risultato dell'analisi condotta in relazione al singolo documento. Tale risultato si articola in due sezioni distinte dedicate alla valutazione della leggibilità del documento effettuata da diversi modelli di analisi basati su diversi tipi di informazione, che potremmo considerare come diversi indici di leggibilità;

READ-IT si basa su questo tipo di approccio: esso opera sul testo arricchito con informazione linguistica e conduce una classificazione probabilistica del testo rispetto a due classi (leggibile vs complesso) sulla base di informazione lessicale, morfo-sintattica e sintattica.

Per ciascun livello, READ-IT restituisce un valore compreso in un range di 0/100 dove 0 indica il valore dei leggibilità più alto e 100 quello più basso sulle diverse configurazioni di caratteristiche del testo:

READ-IT BASE: in questo modello, le caratteristiche considerate sono quelle tipicamente usate nelle misure tradizionali della leggibilità di un testo, ovvero la lunghezza della frase (calcolata come numero medio di parole per frase) e la lunghezza delle parole, (calcolata come numero medio di caratteri per parola).

Questo modello può essere visto come un'approssimazione delle misure tradizionali di leggibilità, in particolare dell'*indice Gulpease*[27], specificamente concepito per la lingua italiana; (Piemontese, Lucisano, 1988).

READ-IT LESSICALE: questo modello si focalizza sulle caratteristiche lessicali del testo, costruite dalla composizione del vocabolario così come dalla sua ricchezza lessicale ;

READ-IT SINTATTICO: questo modello si basa su informazione di tipo grammaticale, ovvero sulla combinazione di tratti morfo-sintattici e sintattici desunti dai corrispondenti livelli di analisi linguistica ;

READ-IT GLOBALE: si tratta di un modello basato sulla combinazione di tratti di varia natura, che spaziano dalle caratteristiche generali del testo del modello *READ-IT BASE* a quelle lessicali e sintattiche degli altri due modelli.

Per ciascun modello, la percentuale esprime il livello di difficoltà, ovvero si riferisce alla probabilità di appartenenza del testo in esame alla classe dei testi di difficile leggibilità:

27. L'Indice Gulpease è un indice di leggibilità di un testo tarato sulla lingua italiana. Rispetto ad altri ha il vantaggio di utilizzare la lunghezza delle parole in lettere anziché in sillabe, semplificandone il calcolo automatico.

Figura 7. output di READ-IT riguardante i testi *originali* sulla base dell' indice di leggibilità e del livello di difficoltà

indice di leggibilità	livello di difficoltà	
Dylan BASE	23,3%	
Dylan LESSICALE	72,6%	
Dylan SINTATTICO	19,7%	
Dylan GLOBALE	90,3%	
indice di leggibilità	**livello di semplicità**	
GULPEASE	59,5	

Figura 8. output di READ-IT riguardante i testi *semplificati* sulla base dell' indice di leggibilità e del livello di difficoltà

indice di leggibilità	livello di difficoltà	
Dylan BASE	17,9%	
Dylan LESSICALE	79,7%	
Dylan SINTATTICO	14,0%	
Dylan GLOBALE	79,2%	
indice di leggibilità	**livello di semplicità**	
GULPEASE	61,3	

Inoltre, READ-IT restituisce anche il valore di **Gulpease** che considera due variabili linguistiche: la lunghezza della parola e la lunghezza della frase rispetto al numero delle lettere.

Formula:

$$89 + \frac{300 * (numero\ delle\ frasi) - 10 * (numero\ delle\ lettere)}{numero\ delle\ parole}$$

I risultati sono compresi tra 0 e 100, dove il valore "100" indica la leggibilità più alta e "0" la leggibilità più bassa. In generale risulta che testi con un indice inferiore a 80 sono difficili da leggere per chi ha la licenza elementare, con un indice inferiore a 60 sono difficili da leggere per chi ha la licenza media, con un indice inferiore a 40 sono difficili da leggere per chi ha un diploma superiore.

Tabella 5. Percentuali di READ-IT, nell'ordine testi originali e semplificati su l' indice di leggibilità e il livello di difficoltà e Gulpease.

Indice di leggibilità	**Livello di difficoltà**	**Livello di difficoltà**
Read-it BASE	23,3 %	17,9 %
Read-it LESSICALE	72,6 %	79,7 %
Read-it S I N T A T T I C O	19,70%	14,00%
Read-it GLOBALE	90,30%	79,20%
Read-it Gulpease	59,50%	61,30%

Come possiamo notare delle percentuali estratte, il *Profilo di Base* rispecchia una riduzione della lunghezza delle frase e delle parole dal testo originale a quello semplificato.

Nonostante il testo di partenza fosse sintatticamente semplice, il *Profilo Sintattico* è passato dal 19,7 % nel testo originale al 14,0% nel testo semplificato, ovvero una riduzione degli aspetti morfo-sintattici e sintattici desunti dai corrispettivi livelli d' analisi linguistica.

Come già espresso nel paragrafo precedente, il *Profilo Lessicale* ha avuto un aumento in relazione all' uso delle regole di inserzione e trasformazione che hanno favorito una maggior ricchezza lessicale, adottando parole di alto uso nel *Vocabolario fondamentale* redatto da De Mauro, senza però aumentare drasticamente il livello di difficoltà e utilizzando tempi verbali e parole di linguaggio parlato e colloquiale accessibili al target di riferimento.

Nel *Profilo Globale* che racchiude i tre precedenti modelli, notiamo che il livello di difficoltà del testo originale (90,3 %) è stato ridotto nel testo semplificato (79,2 %) per uno scarto di 11,1% .

Infine, l' *Indice di Gulpease* identifica un livello di semplicità che, dai testi originali a quelli semplificati, è passato dal 59,5 al 61,3 : questa variazione non è molto significativa come si può ipotizzare dal fatto che le parole mantegnono mediamente sempre la stessa lunghezza in termini di caratteri (4,7 sia per i testi originali che per i testi semplificati) visto nella sezione 3.1.1.

3.2 Osservazioni per regole raggruppate

Nella precedente sezione di questo capitolo, sono state descritte le più importanti regole di annotazione sui testi che sono oggetto dell' analisi. In qesta sezione si condurràun altro tipo di confronto, selezionando dai corpora originali e semplificati le frasi contenenti determinati tipi di regole di annotazione . Tra le varie estrazioni statistiche eseguite, l'analisi si è concentrata su due combinazioni di regole. Una combinazione prende in esame quelle che riducono la lunghezza media della frase che sono le classiche regole di semplificazione del testo, o perché viene adottata una riduzione di frase (*split*) o perché viene eliminato un qualche elemento all' interno della frase (*delete, verbo_meno*); la seconda combinazione raggruppa tutte le regole di inserimento e trasformazione della frase (v*erbo_piu, sogg_espl, insert, merge, spostamento, sogg_sott, sost_lex, anafora, nominalizzazione_piu, nominalizzazione_meno, pass_attivo, att_passivo).*

Questa sezione avrà lo scopo di verificare se anche le altre regole hanno un effetto sulla leggibilità dei testi . Inotre, si verificherà se le tre regole raggruppate abbiano una maggior influenza sulla leggibilità del testo rispetto alle altre. Per questo motivo, riportiamo sotto le analisi condotte con READ-IT sulle frasi originali e semplificate alle quali sono state applicate regole di " riduzione "(*delete, verbo_meno, split*) e le analisi sulle frasi originali e semplificate alle quali sono state applicate tutte le altre regole (v*erbo_piu, sogg_espl, insert, merge, spostamento, sogg_sott, sost_lex, anafora, nominalizzazione_piu, nominalizzazione_meno, pass_attivo, att_passivo*), in modo da poter osservare le variazioni di leggibilità del testo. La *Figura 9* riporta le analisi di READ-IT

condotte sulle frasi originali alle quali sono state annotate le regole di riduzione, mentre la *Figura 10* riporta le analisi sulle rispettive frasi semplificate.

Figura 9. risultato del confronto in READ-IT delle frasi *originali*, alle quali **sono** state applicate le regole *split, delete, e verbo_meno.*

indice di leggibilità	livello di difficoltà	
Dylan BASE	43,7%	
Dylan LESSICALE	90,5%	
Dylan SINTATTICO	31,3%	
Dylan GLOBALE	97,5%	

indice di leggibilità	livello di semplicità	
GULPEASE	56,0	

[+] [–] Caratteristiche estratte dal testo

[–] Profilo di base	
Numero totale periodi:	399
Numero totale parole (token):	8725
Lunghezza media dei periodi (in token):	21,9
Lunghezza media delle parole (in caratteri):	4,8

[–] Profilo lessicale	
Composizione del vocabolario	
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	71,8%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:	
Fondamentale:	71,6%
Alto uso:	20,2%
Alta disponibilità:	8,2%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,750
Densità Lessicale:	0,573

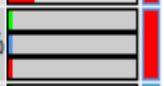
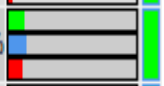
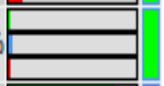
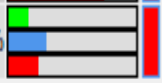
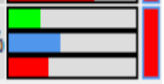
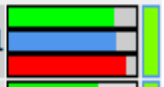
[–] **Profilo sintattico**	
"Misura" delle categorie morfo-sintattiche (%)	
Sostantivi:	18,4% 
Nomi Propri:	5,5% 
Aggettivi:	5,8% 
Verbi:	16,4% 
Congiunzioni:	5,6% 
Coordinanti:	69,2% 
Subordinanti:	30,8% 
Struttura sintattica a dipendenze	
Articolazione interna del periodo:	
Numero medio di proposizioni per periodo:	3,143 
Proposizioni principali vs subordinate (%)	
Principali:	57,9% 
Subordinate:	42,1% 
Articolazione interna della proposizione:	
Numero medio di parole per proposizione:	6,958 
Numero medio di dipendenti per testa verbale:	1,860 
"Misura" della profondità dell'albero sintattico:	
Media delle altezze massime:	6,042 
Profondità media di strutture nominali complesse:	1,247 
Profondità media di "catene" di subordinazione:	1,262 
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente):	
Lunghezza media:	2,221 
Media delle lunghezze massime:	8,509 

Figura 10. risultato del confronto in READ-IT delle frasi *semplificate*, alle quali *sono* state applicate le regole *split, delete, e verbo_meno*.

indice di leggibilità	livello di difficoltà
Dylan BASE	21,0%
Dylan LESSICALE	74,7%
Dylan SINTATTICO	12,3%
Dylan GLOBALE	55,8%

indice di leggibilità	livello di semplicità
GULPEASE	60,6

[+] [–] Caratteristiche estratte dal testo

[–] Profilo di base	
Numero totale periodi:	410
Numero totale parole (token):	7401
Lunghezza media dei periodi (in token):	18,1
Lunghezza media delle parole (in caratteri):	4,7

[–] Profilo lessicale	
Composizione del vocabolario	
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	75,1%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:	
Fondamentale:	74,7%
Alto uso:	16,7%
Alta disponibilità:	8,6%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,740
Densità Lessicale:	0,578

-] Profilo sintattico	
Misura" delle categorie morfo-sintattiche (%)	
Sostantivi:	17,5%
Nomi Propri:	6,3%
Aggettivi:	5,9%
Verbi:	17,1%
Congiunzioni:	6,2%
Coordinanti:	65,4%
Subordinanti:	34,6%
Struttura sintattica a dipendenze	
Articolazione interna del periodo:	
Numero medio di proposizioni per periodo:	2,671
Proposizioni principali vs subordinate (%)	
Principali:	63,1%
Subordinate:	36,9%
Articolazione interna della proposizione:	
Numero medio di parole per proposizione:	6,759
Numero medio di dipendenti per testa verbale:	1,910
"Misura" della profondità dell'albero sintattico:	
Media delle altezze massime:	5,311
Profondità media di strutture nominali complesse:	1,191
Profondità media di "catene" di subordinazione:	1,219
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente):	
Lunghezza media:	2,144
Media delle lunghezze massime:	6,902

L'effetto sulla variazione delle *caratteristiche di leggibilità* in questo caso delle regole di "riduzione" si focalizza principalmente nei *periodi* che aumentano di 11 nei testi semplificati a causa dell'effetto della regola *split* (410 periodi nel testo semplificato e 399 nel testo originale), il quale ha diviso diverse frasi, aumentando il numero delle stesse.

Un' ulteriore effetto dello SPLIT è la divisione della frase che si riscontra anche nella diminuzione della *lunghezza media dei periodi in termini di tokens* dei testi semplificati, per un valore di 18,1 contro 21,9 dei testi originali (3,8 di scarto). Questo dato dimostra la notevole influenza delle regole prese in esame, in quanto hanno certamente ridotto il numero delle parole per periodo (in modo particolare le regole di annotazione *DELETE* e *VERBO_MENO)* .

Ovviamente le variazioni statistiche appena elencate incidono fortemente sulla semplificazione delle frasi : secondo l'indice READ-IT *Globale,* le frasi originali hanno una complessità del 97,5% mentre le frasi semplificate hanno il 55,8%. Come prevedibile, anche l' *indice GULPEASE* intercetta questo effetto, como dimostra il valore da 56,0 a 60,6 (originale e semplificato).

Per quanto riguarda il *Profilo Lessicale*, inoltre, si può notare come le tre regole (in modo particolare DELETE e VERBO_MENO) abbiano diminuito il dato relativo al *Vocabolario di Alto Uso,* passando dal 16,7% nei testi originali all 20,2%. Da questo ultimo dato si può dedurre che le regole di riduzione abbiano diminuito parole di uso scritto, incrementando leggermente quelle di uso parlato (*Alta disponibilità* : 8, 2% nei testi originali e 8,6% nei testi semplificati).

Oltre alla diminuzione delle lunghezze medie delle frasi, si può notare la variazione della distribuzione delle *proposizioni principali e subordinate.*

Come si nota in *Tabella 6 ,* lo scarto di entrambe le coppie di valori mostrate è equivalente: 4,9%; questo dato spiega come le *proposizioni principali* dei testi semplificati aumentino mentre diminuiscano le *proposizioni subordinati* (nei testi semplificati). Il risultato appena citato dimostra che l' aumento delle proposizioni principali è stato significativo grazie all' adozione delle tre regole di annotazione prese in esame (in modo particolare con la regola SPLIT) e che la diminuzione di

periodi più complessi ha aumentato il numero di periodi più brevi con minor ipotassi nel testo.

Tabella 6. confronto delle percentuali delle *proposizioni* subordinate e principali, nell'ordine testi originali e semplificati.

	Testi originali	**Testi semplificati**
Principali	57,90%	63,10%
Subordinate	42,10%	36,9 %

Come si nota in *tabella 7, le congiunzioni coordinanti* presenti nelle frasi semplificate sono minori rispetto a quelle originali, passando dal 69,2% al 65,4%.

Questo dato è particolarmente significativo in quanto dimostra che la semplificazione testuale è avvenuta dividendo le proposizioni coordinanti e rendendole frasi autonome. Le restanti proposizioni sono frasi subordinate introdotte dalla relativa congiunzione subordinante (quest'ultime con maggior frequenza nei testi semplificati) che non hanno aumentato il livello di frasi.

Tabella 7. confronto delle percentuali di **congiunzioni** coordinati e subordinati, nell'ordine testi originali e semplificati .

	Testi originali	**Testi**
Coordinanti	69,20%	65,40%
Subordinanti	30,80%	34,60%

Un altro dato interessante che si può vedere in *Figura 9* e *Figura 10* è la misura della *profondità dell' albero sintattico*: come prevedibile, la *media delle altezze massime* nei testi originali è di 6,042 contro il 5,311 dei testi semplificati.

Infine, le *relazioni di dipendenza* (*distanza in parole tra testa e dipendente*) contano uno scarto di circa 2,7 token (8,509 nei testi originali e 6,902 nei testi semplificati).

I risultati sulla s*truttura sintattica a dipendenze* per regole raggruppate, mostrano che la strategia di semplificazione adottata dagli esperti ha fatto ricorso di varia cancellazione (DELETE e VERBO_MENO nello specifico) senza però utilizzarle eccessivamente; presumibilmente per mantenere una struttura sintattica non troppo semplificata e incitare il lettore ad acquisire strutture grammaticali più "complesse ".

I risultati del monitoraggio sui testi orginali e semplificati per regole raggruppate dimostrano quanto le tre regole non abbiano aumentato notevolemente il livello semplicità del testo (come si evince anche dall' *Indice di Gulpease*) e, come prevedibile, dimostrano una variazione sul *Profilo Sintattico* del testo.

Come deducibile, analizzando il livello di difficoltà dei vari profili di interesse, si può notare come il *Profilo di base, lessicale* e *sintattico* nelle frasi alle quali sono state applicate le regole di "riduzione" siano diminuiti di precentualità in quanto queste tre regole di semplificazione, per loro natura, hanno eliminato parole o frasi complicate, favorendo uno snellimento della struttura sintattica a dipendenze.

I dati relativi alla diminuzione del livello di difficoltà nei singoli profili di interesse vengono avvalorati dai risultati del *Profilo Globale* che mostra una diminuzione del livello di complessità , passando dal 97,5% dei testi originali al 55,8% dei testi semplificati. L' *Indice di leggibilità Gulpease* (basato sulla combinazione di tratti di varia natura che spaziano tra aspetti lessicali e sintattici degli altri modelli) dimostra, inoltre, un aumento del livello di semplicità, passando dal 56,0 dei testi originali al 60,0 dei testi semplificati.

Di seguito riportiamo l' analisi dell' insieme delle frasi originali e semplificate alle quali non sono state applicate regole di riduzione (*Figura 11* e *12*).

Figura 11. risultato del confronto in READ-IT delle frasi *originali* alle quali *non sono* state applicate le regole *split, delete, e verbo_meno*.

indice di leggibilità	livello di difficoltà
Dylan BASE	14,9%
Dylan LESSICALE	62,8%
Dylan SINTATTICO	15,9%
Dylan GLOBALE	72,0%

indice di leggibilità	livello di semplicità
GULPEASE	61,8

[+] [–] Caratteristiche estratte dal testo

[–] Profilo di base	
Numero totale periodi:	666
Numero totale parole (token):	10926
Lunghezza media dei periodi (in token):	16,4
Lunghezza media delle parole (in caratteri):	4,7

[–] Profilo lessicale	
Composizione del vocabolario	
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	71,9%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:	
Fondamentale:	70,5%
Alto uso:	21,3%
Alta disponibilità:	8,1%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,680
Densità Lessicale:	0,583

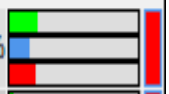
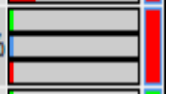
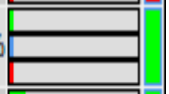
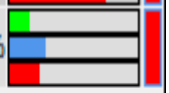
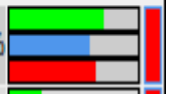
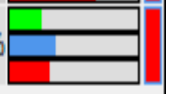
-] Profilo sintattico	
Misura" delle categorie morfo-sintattiche (%)	
Sostantivi:	18,3% 
Nomi Propri:	5,8% 
Aggettivi:	5,6% 
Verbi:	16,7% 
Congiunzioni:	5,6% 
Coordinanti:	70,5% 
Subordinanti:	29,5% 
Struttura sintattica a dipendenze	
Articolazione interna del periodo:	
Numero medio di proposizioni per periodo:	2,443 
Proposizioni principali vs subordinate (%)	
Principali:	62,8% 
Subordinate:	37,2% 
Articolazione interna della proposizione:	
Numero medio di parole per proposizione:	6,715 
Numero medio di dipendenti per testa verbale:	1,825 
"Misura" della profondità dell'albero sintattico:	
Media delle altezze massime:	5,111 
Profondità media di strutture nominali complesse:	1,211 
Profondità media di "catene" di subordinazione:	1,295 
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in arole tra testa e dipendente):	
Lunghezza media:	2,070 
Media delle lunghezze massime:	6,168 

Figura 12. risultato del confronto in READ-IT delle frasi *semplificate* alle quali *non sono* state applicate le regole *split, delete, e verbo_meno*.

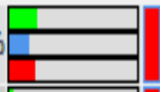
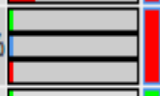
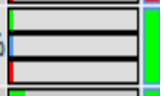
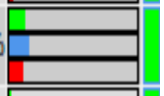
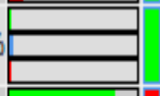
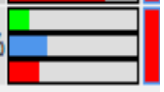
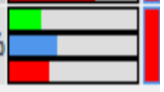
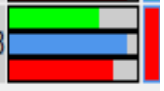
indice di leggibilità	livello di difficoltà	
Dylan BASE	15,9%	
Dylan LESSICALE	67,4%	
Dylan SINTATTICO	14,5%	
Dylan GLOBALE	68,1%	

indice di leggibilità	livello di semplicità	
GULPEASE	61,9	

[+] [–] Caratteristiche estratte dal testo

[–] Profilo di base	
Numero totale periodi:	666
Numero totale parole (token):	11187
Lunghezza media dei periodi (in token):	16,8
Lunghezza media delle parole (in caratteri):	4,7

[–] Profilo lessicale	
Composizione del vocabolario	
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	72,9%
Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:	
Fondamentale:	71,3%
Alto uso:	20,5%
Alta disponibilità:	8,3%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,700
Densità Lessicale:	0,585

[–] Profilo sintattico	
"Misura" delle categorie morfo-sintattiche (%)	
Sostantivi:	18,1% 
Nomi Propri:	5,9% 
Aggettivi:	5,7% 
Verbi:	16,9% 
Congiunzioni:	6,0% 
Coordinanti:	69,3% 
Subordinanti:	30,7% 
Struttura sintattica a dipendenze	
Articolazione interna del periodo:	
Numero medio di proposizioni per periodo:	2,541 
Proposizioni principali vs subordinate (%)	
Principali:	61,4% 
Subordinate:	38,6% 
Articolazione interna della proposizione:	
Numero medio di parole per proposizione:	6,612 
Numero medio di dipendenti per testa verbale:	1,827 
"Misura" della profondità dell'albero sintattico:	
Media delle altezze massime:	5,215 
Profondità media di strutture nominali complesse:	1,210 
Profondità media di "catene" di subordinazione:	1,313 
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente):	
Lunghezza media:	2,083 
Media delle lunghezze massime:	6,390 

Come si evince dai risultati riportati in *Figura 11* e *Figura 12,* il confronto tra gli originali e semplificati a cui non sono state applecate regole di riduzione, come prevedibile, non rivela un effetto significativo rispetto alla lunghezza media delle frasi (che anzi aumentano leggermente nei semplificati (da 16,4 a 16,8). L' effetto delle regole senza riduzione non impatta sulle caratteristiche relative alla lunghezza media dei periodi.

Invece, il processo di semplificazione si rivela efficace rispetto ad altre caratteristiche: successivamente ai dati appena elencati, si può in dettaglio analizzare la differenza di utilizzo delle *proposizioni* .

Come si nota in *Tabella 8 ,* lo scarto di entrambe le coppie di valori mostrate è equivalente: 1,4%; questo dato spiega come le *proposizioni principali* dei testi semplificati diminuiscano, aumentando leggermente le *proposizioni subordinate* (nei testi semplificati). Questo dato dimostra come le regole di inserzione presenti in questo gruppo di regole (in particolare, INSERT, VERBO _PIU' e affini) abbiano cambiato strutture lessicali e morfo-sintattiche di altro tipo. Osservando il *Profilo Lessicale*, infatti, si può notare una leggera diminuzione del *Vocabolario di Alto Uso* (21,3% nei testi originali contro I 20,5% dei testi semplificati) e un aumento altrettanto leggero del *Vocabolario di Alta disponibilità* (da 81,1% dei testi originali a 8,3% dei testi semplificati) riguardo l' uso delle parole di uso parlato.

Tabella 8. confronto delle percentuali delle *proposizioni subordinate* e *principali,* nell'ordine testi originali e semplificati.

	Testi originali	**Testi**
Principali	62,80%	61,40%
Subordinate	37,20%	38,60%

Questi dati dimostrano come i testi nei quali non sono state adottate le tre regole *split, delete e verbo meno* abbiano generato più frasi indipendenti che dipendenti. I dati sono in perfetta sintonia con quelle relative alle *congiunzioni.*
Come si evince in *Tabella 9*, infatti, le *congiunzioni coordinanti* nelle frasi originali sono minori rispetto alle frasi semplificate. Al contrario, *congiunzioni subordinanti* sono aumentate in entrambe le estrazioni: le regole differenti da *split, delete e verbo meno* hanno aumentato leggermente i valori relativi a caratteristiche di complessità sintattica, come ad esempio l' aumento delle subordinazioni .

Tabella 9. confronto delle percentuali delle *congiunzioni subordinate* e *principali,* nell'ordine testi originali e semplificati.

	Testi originali	**Testi semplificati**
Coordinanti	70,50%	69,30%
Subordinanti	29,5 %	30,7 %

Come si nota nella *Figura 11* e *Figura 12* che riportano i dati del profilo linguistico delle frasi originali e semplificate alle quali non sono state applicate le tre regole di " riduzione", gli indici di leggibilità realtivi al profilo di base, lessicale e sintattico sono aumentati nei testi semplificati.

L '*indice di leggibilità Globale* (già spiegato nel paragrafo *3.1. Analisi qualitativa delle regole di annotazione*) rivela invece una diminuzione del livello di difficoltà, passando da 72,0% dei testi originali al 68,1% dei testi semplificati (uno scarto di circa 4 punti).
Il dato appena citato risulta molto interessante in quanto dimostra che, come prevedibile, le regole di trasformazione e di inserzione hanno aumentato il livello di difficoltà del testo rispetto agli indici lessicale e sintattico di READ-IT, ma non rispetto a quello globale. Contrariamente ai dati relativi al *Profilo Globale*, l' *Indice di Gulpease* , essendo una misura tradizionale di leggibilità, non ha saputo

intercettare in maniera esaustiva il livello di semplicità delle frasi, restituendo un dato pressochè equivalente per entrambi i testi (61,8 per i testi originali e 61,9 per i testi semplificati).

Nel prossimo capitolo si analizzeranno le possibili ragioni di queste differenze nell' effetto derivante delle diverse tipologie di regole, esaminando le frasi in cui è stata utilizzata la regola SPLIT, soffermandoci sui cambiamenti grammaticali e semantici che questa regola ha avuto nei testi esaminati .

3.3 Valutazione qualitativa dello *Split*

Nel paragrafo precedente è stata descritta l' influenza che le tre regole di annotazione (*split, delete, verbo_meno*) e le regole di trasformazione, inserzione e spostamento v*erbo_piu, sogg_espl, insert, merge, spostamento, sogg_sott, sost_lex, anafora, nominalizzazione_piu, nominalizzazione_meno, pass_attivo, att_passivo*) hanno avuto sul livello di leggibilità dei testi.

Come già analizzato nel paragrafo " *3.2. Osservazioni per regole raggruppate* ", i testi originali nei quali non sono state adottate le tre regole di annotazione ottengono un livello di leggibilità globale maggiore rispetto ai testi semplificati privi dei medesimi (72,0% nei testi originali contro il 68,1% dei testi semplificati) : questo dato dimostra come il profilo di base, lessicale e sintattico nei testi privi delle regole di "rimozione" *split delete e verbo_meno* siano singolarmente aumentati, ma che globalmente non hanno incrementato il livello di difficoltà del testo. Alla stessa maniera, il *Profilo Globale* delle frasi originali e semplificate alla quale sono state applicate le tre regole di " riduzione "rivela una diminuzione del livello di difficoltà, passando dal 97,5% dei testi originali al 55,8% dei testi semplificati.

In questa sezione si andrà ad esaminare l' influenza che ha avuto una delle regole di annotazione maggiormente utilizzate secondo la frequenza distribuzionale delle regole di annotazione

(consultabile in *7. Appendice)*: lo *SPLIT* . In particolare , si esamineranno i cambiamenti grammaticali che le frasi originali hanno avuto grazie al processo di semplificazione per capire se e come questi cambiamenti sono correlati all' aumento degli indici di leggibilità.

Come già anticipato nel paragrafo "*2.2 Regole utilizzate per l' annotazione del corpus allineamento*", lo *split* è una regola di annotazione utilizzata per segnalare che una frase originale è stata divisa in due o più frasi nel testo semplificato. In particolare, questa regola è stata adottata per dividere molteplici tipologie di proposizioni : ad esempio, sono state divise delle proposizioni coordinate e subordinate eccessivamente lunghe, a favore di periodi più brevi e di proposizioni

principali. Di seguito, si riportano alcuni esempi di coordinate (con o senza configurazione coordinante):

Come si può notare dagli esempi sotto elencati, le proposizioni coordinate sono state divise e rese come proposizioni principali. Questo è stato ottenuto a livello morfosintattico tramite la rimozione della congiunzione. Inoltre si osserva che, per favorire l' aumento di proposizioni indipendenti e alleggerire le subordinazioni in un discorso, sono state introdotte inserzioni di varia natura : nel primi due casi (*Esempio 17* e *Esempio 18*), per favorire la comprensione del testo, la romozione della congiunzione coordinante ha implicato anche l' aggiunta del soggetto esplicito nella frase principale.

Esempio 17 :

frase originale: " Si trovò davanti all'improvviso Margherita vicino a un cespuglio di rosmarino, **e** stava quasi per girarsi e guardare altrove, ma notò un foglietto proprio davanti a lei. "

frase semplificata: " Sandro si trovò con Margherita vicino a un cespuglio di rosmarino**.**

Stava quasi per girarsi e guardare altrove, ma notò un foglietto proprio davanti a lei.*"*

Esempio 18 :

frase originale : " Lungo la strada c'era un gran silenzio **e** il paese pareva deserto. "

frase semplificata : " Lungo la strada c'era un gran silenzio**.** Il paese sembrava deserto. "

In *Esempio 19* c'è stata una sostituzione lessicale della congiunzione avversativa

“ *mentre* ” a favore di “*nel frattempo* ” per rendere così omogenea la restante parte della frase e mantenere la struttura informazionale della frase sul piano del discorso.

Esempio 19 :

frase originale: “Luigi l’elettricista montò dei grossi ventilatori in ogni angolo della via**, mentre** Mario l’idraulico montò dei grossi condizionatori nelle stanze più calde delle case.”

frase semplificata: “ Luigi l’elettricista portò dei grossi ventilatori in ogni angolo della via**. Nel frattempo**, Mario l’idraulico portò grossi condizionatori nelle stanze più calde delle case.*”*

In *Esempio 20* , invece, la ristrutturazione della frase è più complessa, data la presenza di una subordinata concessiva retta dalla proposizione coordinata nella frase principale e, a sua volta, modificata da altre subordinate (comparativa: *“come se fosse diversa”*; causale: *“perchè veniva dalla campagna”*). La semplificazione di una frase così complessa è stata ulteriormente possibile grazie alla rimozione di una congiunzione subordinata concessiva (*benchè)*, che regge un congiuntivo presente (*fossero)* e di un avverbio di modo (*solamente)* . In questo caso, la riformulazione della frase ha determinato l' aggiunta di un soggetto esplicito (*nomignoli*), di un complemento oggetto (*Luisa)* e di vantaggio / svantaggio (*per lei)* .

Esempio 20 :

frase originale: Alcuni dei ragazzi più scalmanati nella banda di Tonio Battaglia la chiamavano con dei nomignoli – **e benché fossero solamente** cose del tipo “topo” e “ campagnola” - **la** facevano sentire come se fosse diversa perché veniva dalla campagna, ed era difficile fare nuove amicizie.

frase semplificata: Alcuni dei ragazzi nella banda di Tonio Battaglia la chiamavano con dei nomignoli – cose del tipo "topo" e " campagnola"**.**

I nomignoli facevano sentire **Luisa** diversa, e **per lei** era difficile fare amicizia.

Come si nota in *Esempio 21,* è stata semplicemente trasformata la proposizione subordinata implicita (introdotta dal verbo *"sorridendo"*) in un frase autonoma, rimuovendo la virgola di coordinazione per asindeto e cambiando modo e tempo del verbo con un indicativo passato remoto (" *sorrise* ").

Esempio 21 :

frase originale : La Mamma ricambiò lo sguardo dei gemelli**, sorridendo,** mise il cane e il gatto nel bagagliaio, e li portò tutti a casa in macchina per finire di fare la torta di compleanno.

frase semplificata : La Mamma guardò i gemelli. **Sorrise** e mise il cane e il gatto nel bagagliaio, e li portò tutti a casa in macchina per mangiare la torta di compleanno.

In Esempio 22, la divisione è stata motivata dalla presenza di una modificazione nominale in funzione appositiva. L'apposizione nominale è stato separato dalla principale e reso come frase autonoma, cosa che ha comportato l' aggiunta della copula e lo scioglimento del costrutto nominalizzato (" *pieno di vergogna* " con " *si vergognava*").

Esempio 22 :

frase originale: Adamo Ramarri si sentì male e**, con il viso verde e pieno di vergogna,** dovette spostarsi al primo posto accanto all'autista.

frase semplificata: Adamo Ramarri si sentì male e dovette spostarsi al primo posto accanto all'autista**. La faccia di Adamo era verde e lui si vergognava.**

La medesima cosa accade in *Esempio 23,* la frase con funzione di modificatore del nome è diventata frase autonoma. (originale :*"... tutta di legno, senza pedali e assai malconcia! "* semplificato: *" Era tutta di legno, senza pedali, e molto rovinata "*).

Esempio 23 :

frase originale : All'improvviso, Ernesta notò in un angolo una strana bicicletta tutta di legno, senza pedali, e assai malconcia!

frase semplificata :All'improvviso, Ernesta notò in un angolo una strana bicicletta. **Era** tutta di legno, senza pedali, e molto rovinata!

In *Esempio 24*, invece, la divisione è stata adottata rispetto alla frase al passivo che è stata resa autonoma e modificata con l' attivo. Questo ha comportato una differente disposizione dei costrutti della frase, cambiando la sua disposizione: " ... *a Pinuccio* " , ad esempio, passa da essere complemento di termine nella frase originale a soggetto nella frase semplificata, ripetendo poi l' azione compiuta da quest'ultimo (*prese pure lui due gatti per acchiappare i topi.)*

Esempio 24 :

frase originale: La Signorina Farfalladimaggio prese cinque barboncini e cinque terrier dello Yorkshire per il suo circo, il Signor Guglielmone prese

due gatti per tenere i topi lontano dal suo negozio, e altri due gatti **furono dati a Pinuccio, il bidello della scuola, per gli stessi motivi.**

frase semplificata : La Signorina Farfalladimaggio prese cinque barboncini e cinque terrier dello Yorkshire per il suo circo, il Signor Guglielmone prese due gatti per tenere i topi lontano dal suo negozio, e altri due gatti.**Pinuccio, il bidello della scuola, prese pure lui due gatti per acchiappare i topi.**

Come si evince in *Esempio 25*, *Esempio 26* e in *Esempio 27* lo *split* ha riguardato la proposizione relativa che modifica il complemento oggetto: in *Esempio 25* si può notare come la frase originale abbia perso il grado di subordinazione relativa (*che aveva tenuto il broncio in camera sua per tutto il pomeriggio)* a favore di una divisione della frase, rendendola così indipendente dalla precedente.

Inoltre, nella proposizione principale della frase originale, il costituente con funzione di modificatore viene reso come soggetto nella frase semplificata, (frase originale: " *Erano tutti felici eccetto la Mamma"* frase semplificata: " *La Mamma era l'unica a non essere felice***.** "). Il cambiamento sintattico e lessicale della frase influisce notevolmente sul grado di comprensione e focus della informazione, perché mette ulteriormente in evidenza la contrarietà del personaggio *" la mamma"* rispetto alla famiglia. Nello specifico, la preposizione impropria "*eccetto*" è un connettivo la cui semantica potrebbe essere per gli esperti meno comprensibile per un bambino "*poor comprehender* " rispetto ad un semplice aggettivo qualificativo come " *unica* ".

Questa trasformazione però ha comportato l' aumento di inserzioni : l' aggiunta del verbo *essere* e di un avverbio di negazione (*non*) che favoriscono un aumento morfologico del testo.

Esempio 25 :

frase originale: Erano tutti felici eccetto la Mamma**, che** aveva tenuto il broncio in camera sua per tutto il pomeriggio.

frase semplificata: La Mamma era l'unica a non essere felice**.** Aveva tenuto il broncio in camera sua tutto il pomeriggio.

In *Esempio 26*, la proposizione subordinata introdotta dal preposizione *che* nella frase originale diventa proposizione principale nella frase semplificata, con l' inserzione del soggetto esplicito (*Tito).* Analogamente all' E*sempio 25* e *Esempio 28* lo *split* ha riguardato la proposizione relativa che modifica il complemento oggetto.

Esempio 26 :

frase originale: Mamma Gorilla sembrava completamente distrutta per le cure che dava al suo vivace cuccioletto Tito**, che** stava giocando vicino alle grosse sbarre di acciaio che circondavano il recinto.

Frase semplificata : Mamma Gorilla sembrava proprio distrutta per le cure che dava al suo vivace cuccioletto Tito**. Tito** stava giocando vicino alle grosse sbarre di acciaio che erano intorno alla loro area.

In *Esercizio 27 ,* proposizione relativa *che* viene separata, troncando la proposizione. Inoltre vengono date delle informazioni maggiori sul complemento oggetto (*Ventoscopio*) nella frase semplificata (*Le eliche cominciarono a girare velocemente)* grazie all' utilizzo proposizioni coordinate.

Esempio 27 :

frase originale: Raggiunta la cima della quercia, accese il Ventoscopio, **che** cominciò a misurare la forza del vento.

frase semplificata: Quando fu in cima all'albero, accese il Ventoscopio.

Le eliche cominciarono a girare velocemente e il Ventoscopio misurò la forza del vento.

In *Esempio 28,* la proposizione subordinante introdotta da *perché* è stata separata aggiungendo un modificatore *" in questo modo"* : lo *split*, quindi, ha modificato una causale a favore di una inserzione avverbile. Inoltre, il soggetto della frase semplificata diventa *" le porte bianche "* e non *" la vernice bianca "*: questo cambiamento risulta molto significativo in quanto focalizza l' attenzione del lettore al cambiamento dell' oggetto e non allo strumento che ne ha permesso il cambiamento .

Esempio 28 :

frase originale: Dopo che Luigi e Mario finirono i loro lavori, Nicola l'imbianchino pitturò di bianco tutte le porte di liquirizia, **perché** la vernice bianca rifletteva tutti i raggi solari e non permetteva al calore di entrare in casa.

frase semplificata: Dopo che Luigi e Mario finirono i loro lavori, Nicola l'imbianchino pitturò tutte le porte di liquirizia con il colore bianco**.**

In questo modo le porte bianche riflettevano i raggi del sole e così il calore non entrava in casa.

Questi esempi dimostrano che l' aumento delle proposizioni principali e la diminuzione delle proposizioni subordinante nelle frasi in cui sono state adottate le tre regole di annotazione *split, delete, verbo_meno* riportate in *Tabella 6* abbiano effetti grammaticali di vario tipo.

Inoltre, si è notevolmente osservando che la divisione di una frase in più frasi è soggetta a inserzioni(generalmente di carattere sostantivale con funzione di

soggetto, oggetto o modificatore) che mirano a aumentare la comprensione della frase per I soggetti target.

In conclusione, si può affermare quindi che, l' uso delle regole di semplificazione *split, delete, verbo_meno* accompagnate da inserzioni sintatticamente obbligatorie non hanno aumentato in maniera così rilevante la leggibilità del testo perchè si accompagnano su trasformazioni della frase talvolta complesse e sostituzioni che influenzano la complessità del testo a livello lessicale o sintattico. Le inserzioni, infatti, hanno riguardato maggiormente nomi comuni e propri , verbi con tempi e modi di facile comprensione e parole riconducibili ad ambienti familiari (come si nota nei risultati statistici di *Tabella 6 - 7- 8 - 9*).

4. CONCLUSIONE

In questo elaborato si è analizzato come le tecnologie linguistico-computazionali possano essere impiegate per favorire lo sviluppo di sistemi di semplificazione semiautomatica di testi a partire dall' annotazione di un corpus parallelo monolingua per bambini con problemi cognitivi di età compresa tra i sette e i dieci anni di età (tratto dal progetto Europeo *Terence*).

Nel dettaglio, in questo elaborato si è discussa la creazione di una risorsa costituita da due versioni allineate nella loro forma *originale* e *semplificata*, spiegando i vari metodi di allineamento delle frasi. La risorsa esemplifica una tipologia di semplificazione definita *"strutturale"*.

Nel dettaglio, la risorsa è stata annotata seguendo uno schema di semplificazione che intercetta vari fenomeni di trasformazione (*split*, *merge*, *Lexical Subtitution*, *Verbal Voice*, *Pass_attivo*, *Nominalization*, *Noun_ to_ verb*), inserimento (*Sogg_espl*, *verbo_piu*, *insert*) e rimozione (*verbo_meno*, *sogg_sott*, *delete*). Sono stati descritti, per ognuna delle regole, alcuni esempi estratti dal corpus preso in esame. Per il lavoro di annotazione delle frasi è stato usato *Brat Application Tool*, uno strumento di marcatura del testo. Una volta terminato il lavoro di annotazione si sono estratte e verificate le frequenze distribuzionali delle regole, il numero delle frasi alle quali sono state applicate e la frequenza di combinazione delle stesse. Dall' analisi delle frequenze è merso che le regole più utilizzate sono state SOST_LEX, INSERT e VERBO_ PIU e alcune di rimozione come SPLIT e DELETE . In seguito, tramite il

monitoraggio linguistico dei testi originali e semplificati condotto dal software READ-IT capace di valutare la leggibilità di un testo e di estrarne il profilo linguistico sulla base di differenti Indici di leggibilità (*Indice Base*, *Indice Lessicale*, *Indice Sintattico*, *Indice Globale*) è stato valutato l' effetto delle regole sulla complessità del testo. Dai risultati del monitoraggio, si è notato come i livelli di difficoltà dell' *Indice di Base*, *Sintattico* e *Globale* siano diminuiti nei testi semplificati grazie all' utilizzo di regole di rimozione, mentre l' *Indice di leggibilità Lessicale* è aumentato grazie all' adozione di regole di inserzione e trasformazione.

Analizzando in dettaglio le caratteristiche estratte dal *Profilo Lessicale*, si è notato che l' aumento del livello di "difficoltà" ha interessato soprattutto le parole appartenenti al *Vocabolario Fondamentale* e di parole utilizzate nella lingua parlata (*Dizionario di Alta Disponibilità*).

Inoltre, l' *Indice di leggibilità Gulpease* ha riscontrato un livello di semplicità maggiore nei testi semplificati rispetto ai quelli originali.

In seguito al confronto delle statistiche dei testi originali e semplificati sulla base dei vari Indici di leggibilità, si è passati ad un altro tipo di confronto statistico riguardante due combinazioni di regole: sono state selezionate le frasi (originali e semplificate) che contenessero, da una parte, solo le regole di annotazione *split*, *delete* e *verbo_meno* e, dall' altra, le frasi con tutte le altre regole di annotazione (v*erbo_piu, sogg_espl, insert, merge, spostamento, sogg_sott, sost_lex, anafora, nominalizzazione_piu, nominalizzazione_meno, pass_attivo, att_passivo*). Questo confronto mirava a verificare se il primo gruppo di regole (altamente usate nel testo) influissero notevolmente sul livello di

complessità del testo e se il secondo gruppo di regole avesse anch' esso un effetto sulla leggibilità del testo. Come prevedibile, le tre regole di riduzione (*split*, *delete* e *verbo_meno)* hanno favorito una diminuzione del livello di difficoltà misurata dai vari Indici di leggibilità grazie alla rimozione delle di porzioni di frasi e di snellimento delle catene di subordinazione.

Le frasi originali e semplificate alle quali è stato applicato il secondo gruppo di regole (v*erbo_piu, sogg_espl, insert, merge, spostamento, sogg_sott, sost_lex, anafora, nominalizzazione_piu, nominalizzazione_meno, pass_attivo, att_passivo*) hanno invece determinato un aumento dell' *Indice di Base* e dell' *Indice Lessicale* nelle frasi semplificate a causa soprattutto dell' adozione di regole di inserimento e trasformazione che hanno aumentato il numero totale di parole e hanno favorito un incremento della densità lessicale. Nonostante ci sia stato un aumento nel livello di difficoltà come misurato dagli Indici di leggibilità (Sintattico e Lessicale), l' *Indice Globale* ha riscontrato una diminuzione del livello di difficoltà nelle frasi semplificate.

Al contrario dell' *Indice di leggibilità Gulpease* che, essendo una misura tradizionale di leggibilità, ha riportato un livello di semplicità quasi equivalente in entrambi in corpora. Da questi risultati si è quindi rilevato che il secondo gruppo di estrazione contenente regole di inserimento e trasformazione ha influito sul livello di leggibilità globale del testo e che, a differenza di una misura tradizionale come quella di Gulpease, READ-IT è riuscito a intercettare questo cambiamento.

Il risultato dell' intero lavoro dimostra l' importanza di creare risorse specifiche per la semplificazione del testo annotate con varie tipologie di regole di trasformazione, come prerequisito per condurre analisi linguistico-computazionali sulla complessità del testo e sulla possibilità di autorizzare alcuni interventi di seplificazione del testo condotti da esperti. Inoltre, il lavoro svolto dimostra come gli strumenti per l' analisi automatica della complessità dei testi siano, a differenza delle misure tradizionali di leggibilità, utili per verificare in maniera esaustiva l' effetto della semplificazione a partire da un corpus annotato con diverse tipologie di trasformazione.

5. BIBLIOGRAFIA

Lenci A., Montemagni S., Pirelli V. *Testo e computer – elementi di linguistica computazionale.* Roma, Carocci, 2005.

Dayley B. Python – *Codice e comandi essenziali.* Piacenza, Pearson, 2007.

Bird S., Klein E., Loper E. *Natural Language processing with Python*. A cura di Livio Mondini, Sebastopol, O'Reilly, 2009.

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

Dell'Orletta F., Montemagni S., Venturi G. *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, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

De Mauro T. *Il dizionario della lingua italiana.* Torino, Paravia, 2000.

Pietro L., Piemontese M. E. *"GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana"*, «Scuola e città», 3, 31, marzo 1988, La Nuova Italia.

Bott S., Saggion H.: *Text simplification resources for Spanish.* Language Resources and Evaluation 48(1): 93-120 (2014)

Dell'Orletta F. *"Ensemble system for Part-of-Speech tagging"*. In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009)

Attardi G., Dell'Orletta F. "*Reverse Revision and Linear Tree Combination for Dependency Parsing"*. In: NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies (Boulder, Colorado, June2009). Proceedings, pp. 261 – 264. Association for Computational Linguistics, 2009.
Attardi G., Dell'Orletta F., Simi M., Turian J. *"Accurate Dependency Parsing with a Stacked Multilayer Perceptron"*. In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009)

Arfè, B., Oakhill, J., Pianta, E., Alrifai, M.: *Story simplification user guide*, Technical report D .2.2, TERENCE Project (2012)

Dell'Orletta F., Montemagni S., Venturi G. *"*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, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.

Brunato D., Dell'Orletta F., Venturi G., Montemagni S. (2014) *"Defining an annotation scheme with a view to automatic text simplification"*. In Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it), 9-10 December, Pisa, Italy.

Allen D. *" A study of the role of relative clauses in the simplification of news texts of learners of English "* Accepted 14 April 2009, Available online 23 October 2009

6. SITI WEB

ProgettoTerence:
http://terenceproject.eu/web/guest/home

Wikipedia, *voce Quadro comune europeo di riferimento per la conoscenza dellelingue*
http://it.wikipedia.org/wiki/Quadro_comune_europeo_di_riferimento_per_la_conoscenza_delle_lingue (visitato il 16 Aprile 2014)

Corpus analizzati in questo elaborato:

Ernesta Sparalesta Esploratrice di Monica Massaro;
http://www.terenceproject.eu/repository/bookit/bookit.html

Le avventure di Sofia e Benedetto di Adel Varzegi;
http://www.terenceproject.eu/repository/bookit/bookit.html

Muoversi di Suzanna Drew- Edwards ;

http://www.terenceproject.eu/repository/bookit/bookit.html

Ugo Scellino Giramondo di Monica Massaro;

http://www.terenceproject.eu/repository/bookit/bookit.html

Un'estate da ricordare di Nykki Irvin;

http://www.terenceproject.eu/repository/bookit/bookit.html

Demo e tool utilizzati per l' analisi dei corpora :

Lingua (http://linguistic-annotation-tool.italianlp.it/)

Read-it(http://www.italianlp.it/wp-content/uploads/2014/06/Demo-Documentation.pdf)

Monitor-it (http://monitor-it.italianlp.it/ *)*

7. APPENDICE

Principalmente la selezione delle frasi e la selezione di eventuali porzioni di testo, prima dell'effettivo processo di analisi del testo, è avvenuta tramite il linguaggio di programmazione Python ; questo linguaggio processa il testo e ne estrae alcune distribuzioni statistiche, come ad esempio la distribuzione delle regole di annotazione.

Esempio di programma scritto per estrarre la distribuzione statistica delle regole nel corpus

Uno dei programmi a livello distribuzionale delle regole di semplificazione è il programma *estraiStatistiche.py*. Come abbiamo visto nel cap.2 viene fornito un output riguardante le frequenze delle regole in ordine decrescente, il numero di frasi alle quali sono state applicate le regole e la frequenza di combinazione in ordine decrescente:

```
import sys

import codecs

if len(sys.argv)<2:
        print"python programma_estraiRegole.py
        nomefile.out"
        exit()

def Ordina(dict):
```

```
        return      sorted(dict.items(),key=lambda x: x[1], reverse=True)

def main(file1):
    fileRegole=codecs.open(file1, "r", "utf8") Frasi={}
    Frasi_freq={
    }
    REGOLE={
    }
    REGOLE_freq={}
    for l in fileRegole:
    if l =="":
            break
        lS=l.strip().split("\t")
        frase=tuple(lS[1:])
        if    not(frase    in    Frasi_freq):
            Frasi_freq[frase]=0.0
        Frasi_freq[frase]+=1.0
        stringa=""
        for (x,y) in enumerate(frase):
            if not(x%2==1):
                if   not(y   in   REGOLE):
                    REGOLE_freq[y]=0
                    .0 REGOLE[y]=0.0
                REGOLE[y]+=1.0
                stringa+=y+" "
            else:
                REGOLE_freq[frase[x-1]]+=float(y)
        stringa=stringa.strip()
        if    not(stringa    in    Frasi):
            Frasi[stringa]=0.0
        Frasi[stringa]+=1.0
```

```
REGOLE_ord=Ordina(REGOLE)
REGOLE_freq_ord=Ordina(REGOLE_freq)
Frasi_ord=Ordina(Frasi)
Frasi_freq_ord=Ordina(Frasi_freq)


print "RISULTATI:"
print
print "1) Frequenza di applicazione delle REGOLE:"
for elem in          REGOLE_freq_ord:
        stringa=elem[0]+"\t"+str(elem[1])     print
        stringa
print
print     "2)    numero    di    frasi    alle    quali    sono
           state applicate le varie REGOLE:"
for elem in          REGOLE_ord:
        stringa=elem[0]+"\t"+str(elem[1])     print
        stringa
print
print "3) Frequenza di combinazione di regole (senza   frequenza) nelle
            frasi:"
for elem in          Frasi_ord:
        stringa=elem[0]+"\t"+str(elem[1])     print
        stringa
        print
print       "4)      Frequenza di combinazione di
                     regole (compresa la frequenza) nelle frasi:"
for elem in          Frasi_freq_ord:
        stringa1=""
        for      x      in      elem[0]:
              stringa1+=str(x)+" "
        stringa=stringa1.strip() +"\t"+str(elem[1]) print stringa
```

```
main(sys.argv[1])
```

Output:

Frequenza di applicazione delle REGOLE:

SOST_LEX 891.0

DELETE 534.0

INSERT 329.0

SPOSTAMENTO 182.0

VERBO_PIU 128.0

TRATTI_VERBO 110.0

SOGG_ESPL 52.0

SPLIT
41.0
VERBO_MENO
25.0
Nominalizzazione- 22.0

Numero di frasi alle quali sono state applicate le varie REGOLE:

SOST_LEX 525.0

DELETE 353.0

INSERT 251.0

SPOSTAMENTO 157.0

VERBO_PIU 117.0

TRATTI_VERBO 98.0

SOGG_ESPL 49.0

SPLIT 37.0
VERBO_MENO
24.0
Nominalizzazione- 22.0

Frequenza di combinazione di regole secondo il numero delle frasi :

SOST_LEX 144.0

DELETE SOST_LEX 67.0

DELETE INSERT SOST_LEX
43.0 DELETE 36.0
INSERT SOST_LEX 19.0

INSERT 18.0

DELETE SOST_LEX SPOSTAMENTO 15.0

SOST_LEX SPOSTAMENTO 15.0

DELETE INSERT SOST_LEX VERBO_PIU 13.0
DELETE INSERT 12.0
DELETE INSERT SOST_LEX SPOSTAMENTO 9.0

8. RINGRAZIAMENTI

Un riconoscimento speciale va alle persone del laboratorio Italian Natural Language Processing Lab (ItaliaNLP Lab) dell'Istituto di Linguistica Computazionale "A. Zampolli" ed in particolar modo la Dott.ssa Dominique Brunato e al Prof. Felice Dell' Orletta per il supporto che mi hanno dato durante le diverse fasi del mio lavoro.

Ringrazio anche il Prof. Alessandro Lenci e tutti i familiari a me più vicini che mi hanno costantemente incoraggiato a compiere gli approfondimenti necessari alla realizzazione del mio lavoro.