# ANNOTAZIONE DI UN TESTO A SUPPORTO DELL'ELABORAZIONE DI UN SISTEMA DI SEMPLIFICAZIONE AUTOMATICA DEL TESTO

a cura di Guglielmo Lischi

Seminario di Cultura Digitale 2014 - 2015

## 0. Introduzione

Pensando come possa evolversi l'italiano parlato del terzo millennio, è lecito chiedersi se e in quale misura le tecnologie linguistico-computazionali possano essere di aiuto nel monitoraggio della lingua italiana (Montemagni, 2013), ma prima salta alla mente un'altra domanda: che cosa è un "monitoraggio" della lingua? Il monitoraggio di una lingua, o meglio, di un testo, comprende un insieme arbitrario di estrazioni statistiche riguardanti la sua struttura linguistica a vari livelli, compiute con tecnologie linguistico-computazionali (algoritmi sviluppati con i linguaggi di programmazione, trasposizione di un testo dalla versione cartacea alla versione digitale, ecc.). L'identificazione della struttura linguistica di un testo avviene tipicamente in modo incrementale, attraverso analisi linguistiche a livelli di complessità crescente; partendo da una "segmentazione" di un testo potremmo considerare in un primo luogo di dividere il testo per frasi, considerando come elemento delimitatore di frase il punto, ma, nel monitoraggio della lingua, le unità di base del testo in formato digitale sono i tokens, una famiglia eterogenea che raggruppa, oltre alle parole ortografiche, anche numeri, sigle, segni di punteggiatura, e altri elementi del vastissimo inventario testuale (Lenci, 2005). «L'intuizione di partenza riguardante il "potere diagnostico" delle tecnologie linguistico-computazionali in compiti di monitoraggio linguistico trova conferma in un recente filone di studi avviato a livello internazionale all'interno del quale le analisi linguistiche generate da strumenti di trattamento automatico del linguaggio sono usate, ad esempio, per misurare la leggibilità di testi (Montemagni, 2013, p. 145)», nonché per supportare la semplificazione semiautomatica degli stessi[1]. In questa direzione di ricerca si colloca questo elaborato, che prende spunto da un lavoro svolto da ben quattro poli di ricerca[2]. Esso è consistito nella ricostruzione della prima e della seconda guerra mondiale tramite l'analisi linguistico-computazionale dei bollettini di guerra, e dello svolgimento delle operazioni, nello studio delle strategie di propaganda, nella comparazione delle due guerre mondiali per tipologia (guerra di posizione e guerra di movimento) e per differenti governi (liberale e fascista), allo scopo di studiare il cambiamento della lingua italiana

1 Vedi ad esempio i lavori di (Saggion, 2014) per lo Spagnolo.

2 Coordinamento:
- Alessandro Lenci (Università di Pisa, CoLing Lab)
- Simonetta Montemagni (ILC-CNR, ItaliaNLP Lab)

Analisi linguistico-computazionali:
- ILC-CNR, CoPhi Lab » Federico Boschetti, Paolo Picchi
- ILC-CNR, ItaliaNLP Lab » Andrea Cimino, Felice dell'Orletta, Giulia Venturi
- Università di Pisa, CoLing Lab » Gianluca Lebani, Lucia Passaro
- Informatica Umanistica » Giacomo Corsini, Michele Mallia, Federica Semplici

Consulenza storica:
- Nicola Labanca (Università di Siena)

Software di ricerca e interfaccia grafica:
- Stefano Dei Rossi (WebSoup)

durante quegli anni. Questo elaborato, seguendo quasi le stesse metodologie applicative e lo stesso tipo di percorso d'analisi (v. fig. 1), si è proposto di analizzare come le tecnologie linguistico computazionali possano essere impiegate per favorire lo sviluppo di sistemi di semplificazione semiautomatica del testo.

Figura 1. ricerche a confronto, da Seminario di Cultura Digitale (Pisa, 22 ottobre 2014)

Il punto di partenza di questa ricerca è stata l'annotazione[3] di un corpus, costituito in modo tale da rappresentare una risorsa esemplificativa di un tipo di semplificazione che si può definire "intuitiva". Infatti, caratteristica di questo corpus, è di essere costituito da due versioni allineate: una contenente dei testi nella loro forma "originale", e l'altra gli stessi testi in una versione riadattata da alcune insegnanti per diverse categorie di persone (principalmente studenti stranieri con una competenza limitata di italiano, inseriti in ogni ordine e grado scolastico). Dopo aver portato a termine l'opera di annotazione, sono state eseguite diverse analisi linguistico-computazionali finalizzate ad intercettare gli interventi di semplificazione degli insegnanti: attraverso tali analisi è stato evidenziato l'effetto di ogni regola, o combinazioni di regole, per la semplificazione del testo, e si è mostrato quali regole si sono rivelate più efficaci per lo studio e la comparazione dei due corpora (corpora plurale di corpus che indica un insieme di testi).

## 1. Introduzione ai metodi di semplificazione del testo

La semplificazione manuale del testo può seguire due approcci: l'approccio intuitivo e l'approccio strutturale. L'approccio strutturale segue delle regole definite a priori da esperti e concepite per uno specifico destinatario, come ad esempio i bambini con difficoltà di comprensione del testo. Queste regole sono potenzialmente sfruttabili da sistemi linguistico-computazionali. È questo il metodo seguito dal progetto europeo Terence finalizzato alla pianificazione, allo sviluppo, e alla valutazione di un sistema adattivo di apprendimento per *poor comprehenders* sia per la lingua italiana che per la lingua inglese[4].

In questo contesto, le costruzioni che sono state più frequentemente semplificate nei testi per bambini sono quelle relative alle voci passive, alle proposizioni relative ed ipotetiche, dal momento che ricerche psicolinguistiche sulla comprensione attraverso la lettura hanno evidenziato che la comprensione di un testo è più legata alla coerenza e alla relazione fra gli elementi del testo che semplicemente alla somma delle caratteristiche linguistiche delle parole o delle frasi individuali nel testo. Inoltre è stato evidenziato che durante la lettura i bambini sono guidati 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 quel testo un dato che gli è particolarmente familiare o di sua appartenenza, e, di conseguenza, apprenda le relazioni semantiche nei testi. Il processo della semplificazione del testo tende a conservare quanto più possibile della struttura linguistica e testuale

3 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.

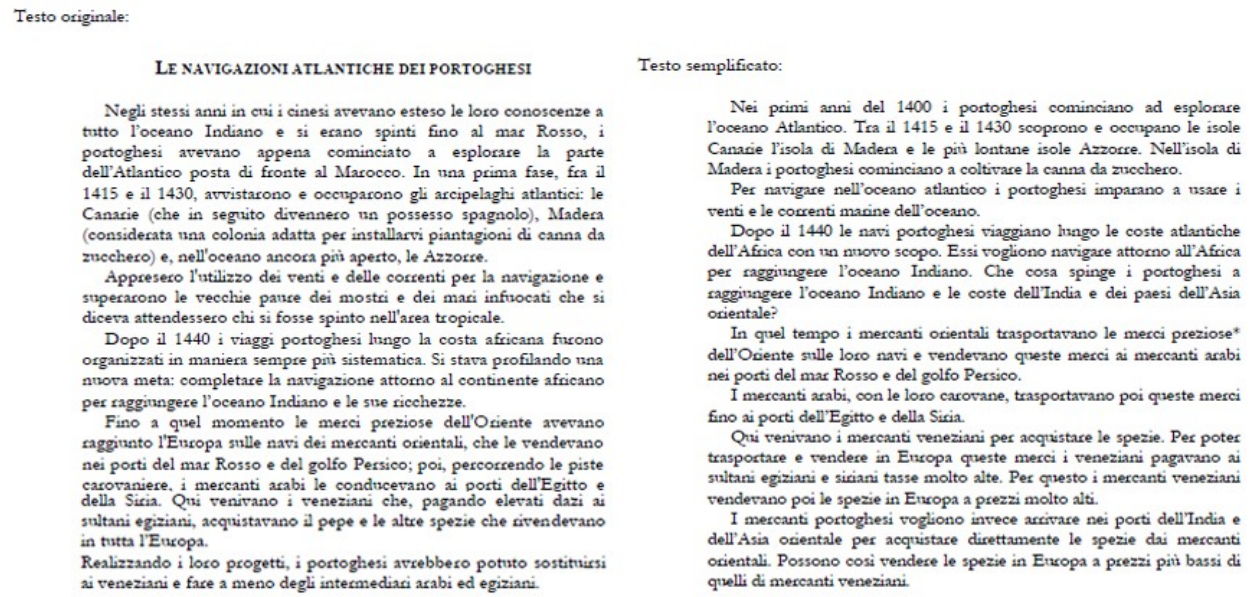
4 http://terenceproject.eu/web/guest/home

della storia autentica. Invero anche 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, questo sistema di semplificazione offre ai lettori livelli graduali di difficoltà, accostandosi progressivamente alla difficoltà che i lettori incontrano nel testo di partenza. Ma a tutti i livelli, l'attenzione è posta sulla struttura globale e sulla coerenza del testo, cosicché anche la versione più semplice del testo conservi quanto più possibile la struttura narrativa e lo stile della storia originale.
L'oggetto di trattazione di questo elaborato è invece la semplificazione intuitiva, ovvero un tipo di semplificazione del testo che è normalmente raggiunta dalle insegnanti servendosi della loro conoscenza del contesto scolastico e delle abilità linguistiche dei propri studenti. Nel presentare i testi utilizzati per studiare questo tipo di approccio, verranno presentate le strategie di interventi sul testo che caratterizzano il lavoro di queste insegnanti.

## 2. Descrizione dei corpora e metodologie utilizzate dagli insegnanti per la semplificazione dei testi

Come anticipato nel capitolo precedente, vediamo adesso in che modo diversi insegnanti hanno, intuitivamente, apportato una versione semplificata di diversi testi, poi confluiti nel corpus qui analizzato, spaziando tra molteplici tipologie di argomento e di narrazione. I testi trattati sono 24 per ciascun sotto-corpus parallelo allineato. Tipicamente un corpus parallelo allineato comprende testi nella loro lingua originale definita come L1, 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'unità tipica di allineamento è la frase:

Figura 2. Un esempio di passaggio dal testo originale al semplificato, da: http://riviste.unimi.it/index.php/promoitals/article/view/832/1073 (Giugno 2015)



Testo originale:

**LE NAVIGAZIONI ATLANTICHE DEI PORTOGHESI**

Negli stessi anni in cui i cinesi avevano esteso le loro conoscenze a tutto l'oceano Indiano e si erano spinti fino al mar Rosso, i portoghesi avevano appena cominciato a esplorare la parte dell'Atlantico posta di fronte al Marocco. In una prima fase, fra il 1415 e il 1430, avvistarono e occuparono gli arcipelaghi atlantici: le Canarie (che in seguito divennero un possesso spagnolo), Madera (considerata una colonia adatta per installarvi piantagioni di canna da zucchero) e, nell'oceano ancora più aperto, le Azzorre.

Appresero l'utilizzo dei venti e delle correnti per la navigazione e superarono le vecchie paure dei mostri e dei mari infuocati che si diceva attendessero chi si fosse spinto nell'area tropicale.

Dopo il 1440 i viaggi portoghesi lungo la costa africana furono organizzati in maniera sempre più sistematica. Si stava profilando una nuova meta: completare la navigazione attorno al continente africano per raggiungere l'oceano Indiano e le sue ricchezze.

Fino a quel momento le merci preziose dell'Oriente avevano raggiunto l'Europa sulle navi dei mercanti orientali, che le vendevano nei porti del mar Rosso e del golfo Persico; poi, percorrendo le piste carovaniere, i mercanti arabi le conducevano ai porti dell'Egitto e della Siria. Qui venivano i veneziani che, pagando elevati dazi ai sultani egiziani, acquistavano il pepe e le altre spezie che rivendevano in tutta l'Europa.

Realizzando i loro progetti, i portoghesi avrebbero potuto sostituirsi ai veneziani e fare a meno degli intermediari arabi ed egiziani.

Testo semplificato:

Nei primi anni del 1400 i portoghesi cominciano ad esplorare l'oceano Atlantico. Tra il 1415 e il 1430 scoprono e occupano le isole Canarie l'isola di Madera e le più lontane isole Azzorre. Nell'isola di Madera i portoghesi cominciano a coltivare la canna da zucchero.

Per navigare nell'oceano atlantico i portoghesi imparano a usare i venti e le correnti marine dell'oceano.

Dopo il 1440 le navi portoghesi viaggiano lungo le coste atlantiche dell'Africa con un nuovo scopo. Essi vogliono navigare attorno all'Africa per raggiungere l'oceano Indiano. Che cosa spinge i portoghesi a raggiungere l'oceano Indiano e le coste dell'India e dei paesi dell'Asia orientale?

In quel tempo i mercanti orientali trasportavano le merci preziose* dell'Oriente sulle loro navi e vendevano queste merci ai mercanti arabi nei porti del mar Rosso e del golfo Persico.

I mercanti arabi, con le loro carovane, trasportavano poi queste merci fino ai porti dell'Egitto e della Siria.

Qui venivano i mercanti veneziani per acquistare le spezie. Per poter trasportare e vendere in Europa queste merci i veneziani pagavano ai sultani egiziani e siriani tasse molto alte. Per questo i mercanti veneziani vendevano poi le spezie in Europa a prezzi molto alti.

I mercanti portoghesi vogliono invece arrivare nei porti dell'India e dell'Asia orientale per acquistare direttamente le spezie dai mercanti orientali. Possono così vendere le spezie in Europa a prezzi più bassi di quelli di mercanti veneziani.

Ad ogni frase del testo originale, corrisponde una frase semplificata (del testo semplificato). Le tipologie di alunni ai quali i testi sono stati proposti sono:

- alunni della scuola secondaria di primo grado (scuola media);
- studenti stranieri non aventi conoscenza della lingua italiana e nozioni cognitive e culturali;
- studenti di italiano L2 livello B1[5], di età compresa fra i 16 ed i 17 anni;
- studenti di quarta primaria e prima secondaria di primo grado;
- classe di alunni multietnici del terzo anno della scuola secondaria di primo grado e nel primo anno della scuola secondaria di secondo grado, aventi differente padronanza della lingua;
- alunni al 5° anno della scuola primaria;
- studenti stranieri inseriti nel biennio della scuola secondaria di II grado, provenienti da diverse aree geografiche;
- alunni del primo anno del liceo.

Uno dei passi applicativi da parte degli insegnanti (per quanto riguarda gli alunni della scuola

5 Secondo il *Quadro comune europeo di riferimento per la conoscenza delle lingue* (QCER), in inglese *Common European Framework of Reference for Languages* (CEFR), un sistema descrittivo impiegato per definire le abilità conseguite da chi studia una lingua straniera europea designante il livello di un insegnamento linguistico, il livello *B1* è considerato il livello di "soglia" o "intermedio". L'individuo che possiede questo livello di conoscenza della lingua comprende i punti chiave di argomenti quotidiani che riguardano la scuola, il tempo libero e la famiglia; inoltre sa muoversi con disinvoltura in situazioni che possono verificarsi se si trova in viaggio nel paese di cui parla la lingua. È in grado di produrre un testo semplice relativo ad argomenti che siano familiari o di interesse personale. È in grado di esprimere esperienze ed avvenimenti, sogni, speranze e ambizioni e di spiegare brevemente le ragioni delle sue opinioni e dei suoi progetti.

media) è stato quello di fornire agli alunni un messaggio immediato riguardante il contesto introduttivo. Il modo per facilitare la lettura e fornire un'idea globale riguardante l'argomento è quello di far visionare preventivamente agli alunni alcuni film. Ad esempio, tra i sotto-corpora, ne emergono due: uno riguardante Anna Frank ed uno riguardante il mito di Pangu. Per avvicinare gli studenti ai rispettivi contesti storici (la seconda guerra mondiale e la persecuzione ebraica per quanto concerne Anna Frank, e il mondo leggendario per quanto riguarda il mito di Pangu), i docenti hanno suggerito la visione del film *La vita è Bella* diretto da Roberto Benigni, l'ascolto della colonna sonora del film *Parla con lei* diretto da Pedro Almodovar per il quadro storico relativo ad Anna Frank, e la visione del Film *Hercules* per l'ambito del mito di Pangu.

Inoltre gli insegnanti hanno fornito un italiano utile per questo tipo di studenti stranieri che non hanno nozioni cognitive e culturali, ed una scarsa conoscenza della lingua italiana. I docenti quindi si propongono di offrire un testo ad alta comprensibilità, non un testo banale e riduttivo o un "surrogato" estremamente ridotto di ciò che veniva spiegato nel testo originale, ma un testo capace di essere comprensibile e di semplice approccio.

Uno dei settori grammaticali sottoposti alla semplificazione testuale, riguarda la sintassi, costituita principalmente da frasi brevi, una struttura della frase secondo l'ordine SVO (Soggetto - Verbo - Oggetto), l'uso dei verbi nei modi finiti e nella forma attiva, l'uso esplicito dei soggetti, l'omissione delle forme impersonali e delle subordinazioni superiori al primo grado. I testi ad alta comprensibilità creati dai docenti non sostituiscono però il libro di testo, ma lo affiancano, favorendo l'attenzione dell'allievo e insegnandogli tecniche di studio che non stimolino solo la memoria, ma anche la comprensione delle informazioni e dei concetti. Nonostante la resa ottimale del testo semplificato, gli alunni dovranno sempre tenere il testo originale accanto a quello semplificato per evitare che si fossilizzino su un livello linguistico basso. I sotto-corpora semplificati quindi avranno un alto grado di comprensibilità per quanto riguarda il lessico, la morfosintassi e la sintassi. Inerentemente a questi ambiti linguistici osserviamo su che cosa si sono concentrati maggiormente i docenti:

1. Lessico:
    - Per la sostituzione delle parole complicate sono state adottate due risorse di riferimento:

VdB[6] (vocabolario di base) e LIP[7] (Lessico di frequenza dell'italiano parlato); la sostituzione è stata realizzata mediante l'uso di sinonimi più vicini alla lingua comune e di parafrasi esplicative;

- uso molto ridotto delle nominalizzazioni.

2. Morfosintassi:
   - Verbale: passaggio da passato prossimo a presente storico.
3. Sintassi:
   - riduzione della lunghezza della frase in caso di periodo ricco di subordinate;
   - esplicitazione delle proposizioni implicite;
   - splitting[8] in più frasi;
   - passaggio da ordine marcato ad ordine non marcato (SVO);
   - preferenza di utilizzo della paratassi[9];
   - vengono evitate le espressioni idiomatiche.

Una volta che gli insegnanti hanno individuato i fattori determinanti per la semplificazione del testo, si sono posti degli obiettivi riguardo agli alunni. Non bisogna dimenticare il focus primario: favorire la comprensione di un testo. La lista seguente riporta gli scopi finali che gli insegnanti vorrebbero che gli alunni raggiungessero:

- fare previsioni e ipotesi, che vengono poi ridefinite nel corso della lettura;

---

6 Il vocabolario di base (VdB) della lingua italiana è stato creato da Tullio De Mauro. Comprende circa 7.000 parole, quelle che hanno la maggiore frequenza statistica nella nostra lingua. Sono quelle che più usiamo, che più ci sono familiari. Esso è diviso in:

1. Vocabolario fondamentale, composto da 1.991 parole. Sono le più usate in assoluto 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).

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

7 Il Lessico di frequenza dell'italiano parlato, curato anch'esso da Tullio de Mauro insieme a Mancini, Vedovelli e Voghera, è tratto da un corpus di circa 500.000 parole grafiche, trascrizioni di registrazioni effettuate a Milano, Firenze, Roma e Napoli, pari a quasi 57 ore di parlato. Le tipologie del parlato rappresentate sono dialoghi faccia a faccia. I lemmi sono consultabili secondo frequenza e secondo ordine alfabetico; vi è anche una lista di frequenza dei fonosimboli (ogni manifestazione fonica non riconducibile alle strutture fonematiche e morfematiche proprie di una data lingua, ad esempio varie forme esclamative o espressive quali *uffa, uh, mah, bah*, talora non ben rappresentabili con segni grafici tradizionali come la *m* prolungata a bocca chiusa per indicare dubbio, incredulità, ecc.) e delle polirematiche (le parole polirematiche, dette anche semplicemente polirematiche sono elementi lessicali, formati da più di una parola, che hanno una particolare coesione strutturale e semantica interna e possono appartenere a varie categorie lessicali, per esempio *anima gemella, carta di credito, acqua e sapone, portare avanti, dare una mano, a fior di pelle, a furia di, ecc*). Il volume del lessico è corredato da due dischetti che contengono le trascrizioni di tutti i testi del corpus, permettendo dunque a chi voglia svolgere ulteriori analisi l'accesso diretto ai materiali.

8 Lo "splitting" è la divisione di una frase in due o più frasi; considerando come delimitatore di frase il punto.

9 La paratassi è un modo di costruire il periodo basata sulla coordinazione di frasi indipendenti per mezzo di congiunzioni (ad esempio vado e torno) o per semplice accostamento (mangiate, bevete, fate come a casa vostra).

- collegare le informazioni che vengono presentate nel testo;
- sviluppare il lessico (con attenzione anche all'uso figurativo dello stesso) ed avere la capacità di analizzarlo;
- sviluppare la conoscenza di alcune strutture della lingua italiana;
- sviluppare la capacità di ascolto;
- sviluppare la capacità di confronto interculturale;
- conoscere le origini e la struttura dell'argomento;
- manipolare testi semplici;
- analisi del lessico: (i nomi delle piante e dei frutti, i nomi dei personaggi fantastici, ecc.).
- comprendere il testo;
- riprodurre un testo analogo a livello orale e scritto.

## 3. Metodi linguistico-computazionali per l'analisi e il riconoscimento delle regole di semplificazione utilizzate

Gran parte del lavoro svolto per la stesura di questo elaborato, è consistito nell'annotazione delle regole di semplificazione, mettendo a confronto i corpora allineati (originale – semplificato). In questo capitolo conclusivo verranno di seguito: 1. spiegate il tipo di regole usate e quella con la maggiore frequenza di utilizzo 2. analizzati i risultati della semplificazione dal punto di visto qualitativo. In questo contesto verrà introdotto uno strumento d'analisi per la valutazione automatica della leggibilità: READ-IT[10] (Dell'Orletta et al, 2011).

La tabella 2 elenca tutte le regole utilizzate per l'annotazione del corpus allineato, le quali specificano quale regola dedita alla semplificazione testuale è stata utilizzata; queste regole sono state redatte dal laboratorio di ricerca Italian Natural Language Processing Lab[11] dell'Istituto di Linguistica Computazionale "Antonio Zampolli" all'interno del Centro Nazionale delle ricerche di Pisa, ed equivalgono precisamente a tags[12] nella annotazione XML[13].

Sono stati evidenziati tags che trasformano una parola o una porzione di testo, regole che inseriscono un elemento mancante nella frase (come una parola, un'altra frase, o parte di essa), oppure tags dediti a marcare la rimozione e la cancellazione di una parola o di una parte intera della frase o del testo.

10 www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest
11 www.italianlp.it
12 Da ora in avanti considereremo equivalenti regola e tag.
13 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.

Tabella 2. regole utilizzate per marcare la semplificazione testuale

| | |
|---|---|
| **<split>** | Da inserire per segnalare che una parte della frase originale (es. proposizione coordinata) è stata resa come frase autonoma nella versione semplificata. |
| **<merge>** | Da inserire per segnalare la frase (o le frasi) autonoma(e) nella versione originale che sono state unite in una singola frase nella versione semplificata. |
| **<spostamento>** | Da aggiungere per segnalare uno spostamento di parti della frase (es. una frase subordinata che nell'originale precede la principale mentre nel semplificato segue la principale). |
| **<sost_lex>** | Da aggiungere per segnalare una sostituzione lessicale (es. uso di un sinonimo) dall'originale al semplificato. Questo tag possiede l'attributo "forma" che indica il sostituto (può essere una o più parole). |
| **<anafora>** | Da aggiungere per segnalare i casi in cui un pronome è stato sostituito da un sintagma nominale lessicale. |
| **<tratti_verbo>** | Da aggiungere per segnalare i casi in cui il verbo è stato mantenuto ma sono cambiati alcuni dei suoi tratti (tempo, modo, persona, es. dal passato remoto al presente). Anche in questo caso, indica i tratti come attributi del tag (modo, tempo e persona). |
| **<att_passivo>** | Indica il cambiamento della diatesi verbale (da attivo a passivo). Tag da marcare sul verbo. |
| **<pass_attivo>** | Indica il cambiamento della diatesi verbale (da passivo ad attivo). Anch'esso da marcare sul verbo. |
| **<nominalizzazione_piu>** | Da inserire nel caso in cui un verbo, nella versione semplificata diventa un sostantivo. |
| **<nominalizzazione_meno>** | Da inserire per segnalare lo “scioglimento” di una nominalizzazione o di una perifrasi nominale, trasformata nella corrispondente struttura verbale. Sia *<nominalizzazione_piu>* che *<nominalizzazione_meno>*, possiedono l'attributo "forma". |
| **<sogg_espl>** | Da aggiungere per segnalare i casi in cui nella frase originale c'è un soggetto sottinteso (esplicitato nell'attributo "sog") che è stato esplicitato nella frase semplificata. |
| **<verbo_piu>** | Da aggiungere per segnalare i casi in cui nella |

| | |
|---|---|
| | frase originale manca un verbo che è stato inserito nella frase semplificata. Può avere gli attributi “tempo”, “modo”, “persona”. |
| **<insert>** | Da inserire per segnalare altri tipi di inserimento (parole che non sono soggetto o verbo) oppure sequenze di più parole. Possiede l'attributo "forma". |
| **<verbo_meno>** | Da aggiungere per segnalare i casi in cui un verbo nella frase originale è stato eliminato nella frase semplificata. |
| **<sogg_sott>** | Da aggiungere per segnalare i casi in cui nella frase originale c'è un soggetto esplicito che è stato sottinteso nella frase semplificata. |
| **<delete>** | Da aggiungere per segnalare una frase originale (o una parte di frase) completamente rimossa nella versione semplificata. |

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

Il lavoro di annotazione delle frasi è stato lungo e laborioso. A lavoro ultimato è stato molto probabile che alcuni tags non fossero nella corretta annotazione XML. La prima azione da svolgere è stata la validazione del documento grazie all'editor Xml Copy Editor. Da ricordare che per portare a termine delle corrette analisi linguistico-computazionali, i corpora paralleli, previa annotazione, dovevano essere prima spezzati per separare i testi in due documenti differenti, contenenti uno le frasi originali, l'altro le frasi semplificate.

Utilizzando il linguaggio di programmazione Python, è stato rintracciato subito il tag con la maggiore frequenza, ovvero <*sost_lex*> applicato ben 495 volte su 206 frasi. Questo altissima frequenza indica come gli insegnanti abbiano "abusato" di molteplici sostituzioni lessicali: fra queste, la trasformazione di alcuni aggettivi, come per esempio “bei” che diventa “felici”, o dei bigrammi[14] all'interno della frase “Un tempo la terra era vuota e senza abitanti” che diventa “Una volta la terra era vuota e solitaria”. Un altro tag molto ricorrente è <*tratti_verbo*>. Andando a sbirciare tra i valori degli attributi di questo tag possiamo vedere che, nella maggior parte dei casi, frequentemente il tempo dei verbi è stato cambiato in presente; per esempio la frase “alle tre qualcuno suonò alla porta.” diventa “alle tre una persona suona alla porta.”

Ora valutiamo l'effetto delle regole di semplificazione rispetto alla leggibilità del testo.

A tale scopo è stato utilizzato un tool chiamato READ-IT, un'applicazione web capace di valutare la leggibilità di un testo e di estrarne il profilo linguistico. Con questo tool possiamo verificare quanto

14 Sequenze formate da due parole consecutive.

un testo sia leggibile, e verificare se le semplificazioni apportate dalle insegnanti siano effettivamente riuscite. L'output del programma si articola in due sezioni distinte dedicate a:

- La valutazione della leggibilità del documento effettuata da diversi modelli di analisi basati su diversi tipi di informazione, che potremmo vedere come diversi indici di leggibilità;
- la ricostruzione del profilo linguistico del testo, condotta in relazione a un sottoinsieme dei parametri utilizzati dal programma per la valutazione della sua leggibilità, articolati secondo il livello di descrizione linguistica di appartenenza. Questa seconda sezione è tesa a fornire elementi di analisi utili a comprendere i risultati riportati nella prima sezione: si tratta di informazioni utili per il linguista e il linguista computazionale che permettono di monitorare il funzionamento del sistema ed eventualmente correggerlo.

Il tool sfrutta una catena di analisi linguistica in grado di analizzare il testo fino alla sintassi ed utilizza le caratteristiche linguistiche ricavate da quest'analisi automatica per assegnare quattro livelli di leggibilità.

La valutazione globale della leggibilità del testo viene condotta sulla base di diverse configurazioni di caratteristiche del testo che producono quattro modelli di leggibilità:

- Dylan 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 (Piemontese, Lucisano, 1988), un indice specificamente concepito per la lingua italiana, 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.
- Dylan LESSICALE: questo modello si focalizza sulle caratteristiche lessicali del testo, costituite dalla composizione del vocabolario e dalla sua ricchezza lessicale.
- Dylan SINTATTICO: questo modello si basa su un'informazione di tipo grammaticale, ovvero sulla combinazione di tratti morfo-sintattici e sintattici desunti dai corrispondenti

livelli di analisi linguistica.

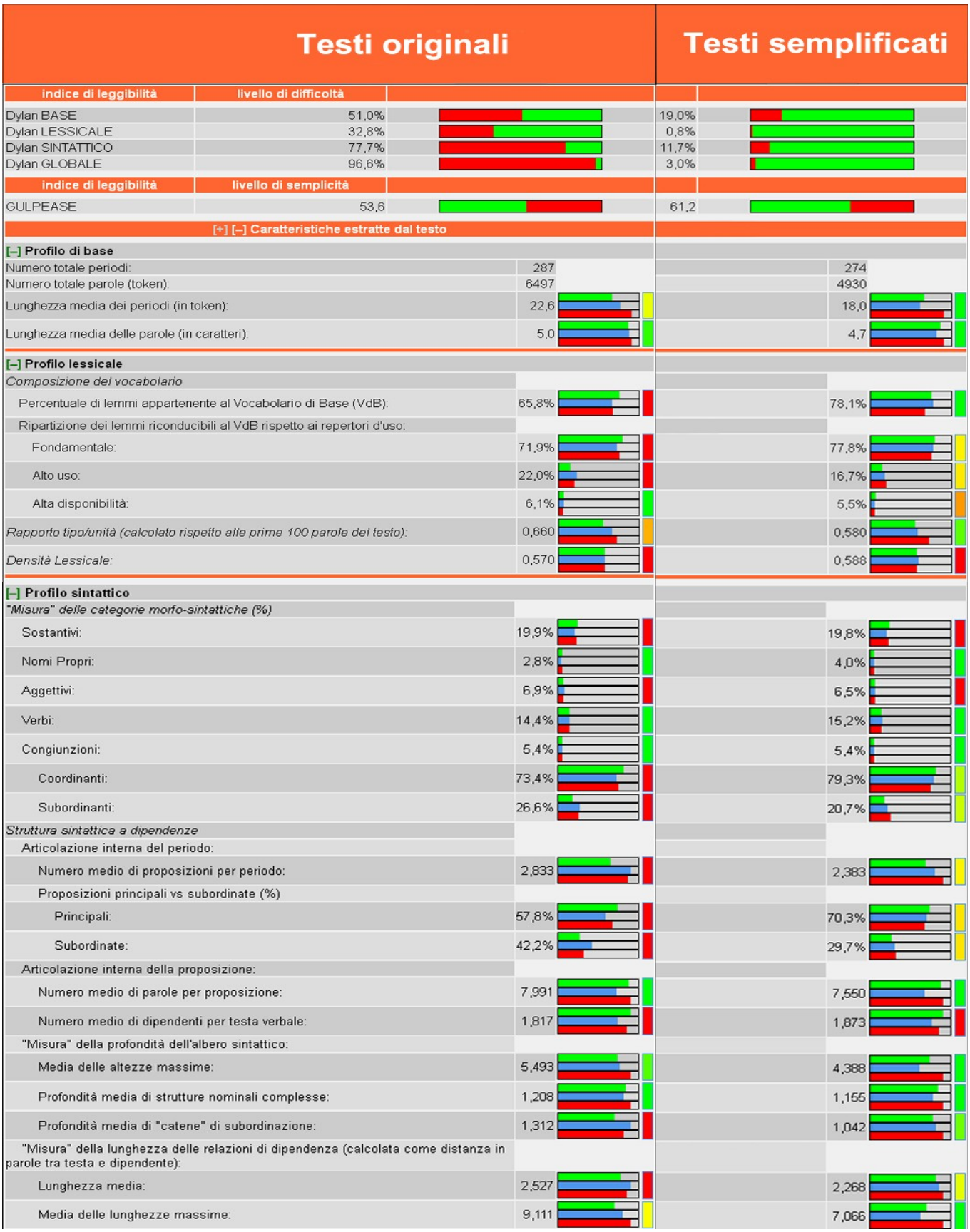
- Dylan GLOBALE: si tratta di un modello basato sulla combinazione di tratti di varia natura, che spaziano dalle caratteristiche generali del testo del modello Dylan BASE a quelle lessicali e sintattiche degli altri due modelli.

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: la barra a fianco esprime visivamente questo valore, dove il rosso rappresenta la probabilità di appartenenza alla classe dei testi difficili e il verde a quelli di facile lettura.

Infine, nel modello sintattico, vengono considerate proprietà che caratterizzano l'albero sintattico di una frase come ad esempio la media delle altezze massime, la profondità media di strutture nominali complesse (cioè il numero di modificatori che dipendono da un nome testa della dipendenza) e la profondità media di catene di subordinazione.

Tornando all'analisi della semplificazione, la prima operazione che è stata svolta è il confronto tra tutti i testi originali e i semplificati, per verificare se effettivamente gli indici in READ-IT intercettano gli interventi di semplificazione.

Figura 3. output di READ-IT riguardante il confronto tra tutti i testi originali e semplificati, da http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest (Maggio 2015)



| | Testi originali | | Testi semplificati | |
|---|---|---|---|---|
| **indice di leggibilità** | **livello di difficoltà** | | | |
| Dylan BASE | 51,0% | | 19,0% | |
| Dylan LESSICALE | 32,8% | | 0,8% | |
| Dylan SINTATTICO | 77,7% | | 11,7% | |
| Dylan GLOBALE | 96,6% | | 3,0% | |
| **indice di leggibilità** | **livello di semplicità** | | | |
| GULPEASE | 53,6 | | 61,2 | |
| **[+] [–] Caratteristiche estratte dal testo** | | | | |
| **[–] Profilo di base** | | | | |
| Numero totale periodi: | | 287 | | 274 |
| Numero totale parole (token): | | 6497 | | 4930 |
| Lunghezza media dei periodi (in token): | | 22,6 | | 18,0 |
| Lunghezza media delle parole (in caratteri): | | 5,0 | | 4,7 |
| **[–] Profilo lessicale** | | | | |
| *Composizione del vocabolario* | | | | |
| Percentuale di lemmi appartenente al Vocabolario di Base (VdB): | | 65,8% | | 78,1% |
| Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso: | | | | |
| Fondamentale: | | 71,9% | | 77,8% |
| Alto uso: | | 22,0% | | 16,7% |
| Alta disponibilità: | | 6,1% | | 5,5% |
| *Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):* | | 0,660 | | 0,580 |
| *Densità Lessicale:* | | 0,570 | | 0,588 |
| **[–] Profilo sintattico** | | | | |
| *"Misura" delle categorie morfo-sintattiche (%)* | | | | |
| Sostantivi: | | 19,9% | | 19,8% |
| Nomi Propri: | | 2,8% | | 4,0% |
| Aggettivi: | | 6,9% | | 6,5% |
| Verbi: | | 14,4% | | 15,2% |
| Congiunzioni: | | 5,4% | | 5,4% |
| Coordinanti: | | 73,4% | | 79,3% |
| Subordinanti: | | 26,6% | | 20,7% |
| *Struttura sintattica a dipendenze* | | | | |
| Articolazione interna del periodo: | | | | |
| Numero medio di proposizioni per periodo: | | 2,833 | | 2,383 |
| Proposizioni principali vs subordinate (%) | | | | |
| Principali: | | 57,8% | | 70,3% |
| Subordinate: | | 42,2% | | 29,7% |
| Articolazione interna della proposizione: | | | | |
| Numero medio di parole per proposizione: | | 7,991 | | 7,550 |
| Numero medio di dipendenti per testa verbale: | | 1,817 | | 1,873 |
| "Misura" della profondità dell'albero sintattico: | | | | |
| Media delle altezze massime: | | 5,493 | | 4,388 |
| Profondità media di strutture nominali complesse: | | 1,208 | | 1,155 |
| Profondità media di "catene" di subordinazione: | | 1,312 | | 1,042 |
| "Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente): | | | | |
| Lunghezza media: | | 2,527 | | 2,268 |
| Media delle lunghezze massime: | | 9,111 | | 7,066 |

Ciò che si osserva è fondamentalmente la diminuzione dei valori, in molteplici casi, nella versione semplificata.

Tabella 3. Percentuali Dylan, nell'ordine testi originali e semplificati

| | | |
|---|---|---|
| Dylan BASE | 51,0% | 19,0% |
| Dylan LESSICALE | 32,8% | 0,8% |
| Dylan SINTATTICO | 77,7% | 11,7% |
| Dylan GLOBALE | 96,6% | 3,0% |
| GULPEASE | 53,6 | 61,2 |

L'indice più significativo di questa tabella è il Dylan lessicale, il quale riesce ad approssimare quasi a 0 la sua percentuale, motivo per cui il livello di difficoltà di lettura decrementa maggiormente. Non da trascurare anche il Dylan Globale che mostra per i testi semplificati una percentuale di ben 3,0% per i testi semplificati, contro 96,6% dei testi originali. Si può dedurre inoltre a chi siano rivolti tutti i testi semplificati, grazie all'indice di Gulpease: l'indice di 53,6 implica che i testi originali siano rivolti a ragazzi compresi tra le scuole medie e le scuole superiori, mentre 61,2 implica che il target delle insegnanti sia complessivamente la scuola media. Si può osservare nel profilo di base come il numero dei periodi si abbassi rispetto ai testi originali (274 contro 287), oppure come il numero di tokens diminuisca di ben 1567 tokens nella versione semplificata. Ad influire ulteriormente sono anche la media delle dimensioni dei periodi e dei tokens i quali diminuiscono sempre nel testo semplificato rispetto al testo originale (periodi: 18,0 contro 22,6; tokens: 4,7 contro 5,0). Nel profilo lessicale si possono osservare alcuni elementi relativi al vocabolario di base e la ripartizione di diversi lemmi[15]. Complessivamente si può osservare come la percentuale dei lemmi del vocabolario di base sia superiore nei testi semplificati (78,1% contro 65,8 nei testi originali); i lemmi ad uso fondamentale sono il 77,8% nella versione semplificata e 71,9% nella versione originale; questo significa che la ricorrenza delle parole “fondamentali” all'interno dei testi in formato originale è minore rispetto alla versione semplificata. Il rapporto tipo/unità (*Type/Token Ratio* abbreviata come TTR) è un indice di ricchezza lessicale, calcolato come il numero di parole tipo, o vocabolario[16], diviso il numero di tutte le parole del testo. Il quoziente è sempre compreso tra 0 e 1; se si approssima a 0 significa che il testo non è molto vario lessicalmente, mentre se si approssima a 1 significa che è molto vario; se il quoziente equivale a uno (caso rarissimo ma non impossibile), significa che il testo non presenta parole ripetute. Nei testi

15 In linguistica, e in particolare in morfologia, il lemma costituisce la forma canonica di una parola. Il rapporto fra lemmi e parole è particolarmente importante nelle lingue dotate di un ricco paradigma flessivo delle parole. Tipicamente il lemma è la parola di ricerca del dizionario.

16 Il vocabolario di un testo, è l'insieme di tutte le parole contate una sola volta, ovvero l'insieme delle parole “tipo”; da differenziare con la definizione di parole “unità” che sono tutte le parole del testo.

analizzati si può riscontrare una TTR per i testi originali di 0,660, e di 0,580 per i testi semplificati; questo indica che gli insegnanti autori delle semplificazioni hanno volutamente ripetuto alcune parole, all'interno dei testi semplificati, proprio per semplificarne la lettura. Non a caso troviamo a destra un valore diminuito di ben 0,08, che per questi range di punteggi è un valore molto significativo. La densità lessicale è un indice che caratterizza variazioni di registro linguistico e viene calcolata come il rapporto tra il numero totale di occorrenze nel testo di sostantivi, verbi, avverbi, aggettivi, e il numero totale di parole nel testo, ad esclusione dei segni di punteggiatura (Dell'Orletta, 2012-2013). In questo caso si riscontrano dei valori al limite dell'equivalenza: 0,570 per i testi originali, e 0,588 per i testi semplificati; questo indica una leggera variazione del registro linguistico di 0,018 nei testi semplificati. Plausibile dato che le opzioni di semplificazione variano in base alla persona che adotta le semplificazioni.

Nel profilo sintattico, e in un primo luogo rispetto alle categorie morfosintattiche, possiamo osservare gli aumenti e le diminuzioni dei valori relativi agli elementi del testo più importanti tra cui i nomi propri che nei corpora semplificati hanno una percentuale di 4,0% e negli originali di 2,8%; ciò implica che molti nomi propri vengono ripetuti nella versione semplificata per renderli più salienti al lettore: un indice di semplicità da non trascurare. Un'analisi simile si può fare nei confronti delle congiunzioni coordinanti che hanno uno scarto del 5,9% (79,3% per i testi semplificati e 73,4 per quanto concerne i testi originali). Nelle articolazioni dei periodi possiamo trovare un'ascesa della percentuale di utilizzo delle proposizioni principali nei corpora semplificati di 12,5 (70,3 contro 57,8 negli originali), ed un abbassamento delle proposizioni subordinate ugualmente di 12,5 nei corpora semplificati. Le misure coincidono perfettamente perché quelle che in un primo momento erano proposizioni subordinate, dopo diventano proposizioni principali.

All'interno dell'articolazione delle proposizione si può trovare un abbassamento di 0,441 del numero medio di parole per proposizione (7,991 nei testi originali e 7,550 nei testi semplificati).

La media delle profondità degli alberi sintattici[17] nei corpora originali si abbassa di ben 1,105 (5,493 nei corpora originali e 4,388 nei corpora semplificati); la leggibilità di una frase con alberi più corti è molto più semplice rispetto ad una frase lunga e dotata di molte articolazioni; infatti la misura media delle catene di subordinazione, nei testi semplificati, cala di ben 0,27 (1,312 per i testi originali e 1,042 per i testi semplificati).

Si può osservare anche un abbassamento della media delle lunghezze massime delle relazioni di dipendenza, calcolata come distanza in parole tra la testa (verbo della proposizione principale) e l'ultima parola della dipendente: l'abbassamento è di 2,045 (9,111 per i corpora originai e 7,066 per i

17 L'albero sintattico in un'annotazione sintattica a dipendenze rappresenta il numero di archi che intercorrono tra una foglia (rappresentata da parole del testo senza dipendenti) e la radice (root) dell'albero.

corpora semplificati).

In questo lavoro è stato affrontato un tema molto attuale e ancora poco esplorato in linguistica computazionale: la semplificazione automatica del testo. È stata sottolineata l'importanza di creare una risorsa adeguata al tipo di compito, che, più nel dettaglio, ha affrontato l'aspetto della semplificazione “intuitiva”. Per questo è stato costituito un corpus, che abbiamo qui definito come corpus “parallelo monolingue”. Va sottolineata la difficoltà di reperire testi totalmente allineati (ovvero a livello di singole frasi), dal momento che la produzione di un testo semplificato si inserisce in un contesto di attività più ampio, che include vari interventi previsti dagli insegnanti per facilitare la comprensione in favore di specifici destinatari (studenti con un livello di conoscenza dell'italiano limitato).

Una volta costituito un corpus composto da un numero di testi significativo, la fase successiva è stata l'annotazione, il cui obiettivo è stato quello di intercettare i tipi di semplificazione, sia lessicale che sintattica, attraverso delle regole appositamente predisposte. Successivamente per valutare l'incidenza di ciascuna di queste regole è stato sviluppato un programma che ha consentito di identificare le regole maggiormente produttive: come abbiamo visto, si tratta principalmente di quelle che intercettano cambiamenti a livello del lessico. Per valutare l'effetto delle regole di semplificazione anche da un punto di vista più qualitativo, è stato introdotto nell'analisi il software READ-IT. Questo tool misura la leggibilità del testo sulla base di complesse configurazioni di caratteristiche linguistiche estratte in maniera automatica. Come abbiamo visto dalla prima estrazione, che compara i due corpora nella totalità, READ-IT ha attribuito un punteggio di leggibilità superiore ai testi semplificati. Questo testimonia l'importanza di un monitoraggio linguistico del testo a livello di complessità crescente, partendo prima dall'analisi distribuzionale (ovvero il calcolo della frequenza di applicazione delle regole) e terminando con l'analisi qualitativa grazie al tool READ-IT che ci consente di osservare dei parametri articolati secondo il livello di descrizione linguistica di appartenenza.

**5. Bibliografia**


Bibliografia Primaria

Lenci, Alessandro, Simonetta Montemagni, Vito Pirelli. *Testo e computer – elementi di linguistica computazionale.* Roma, Carocci, 2005.

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

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

Simonetta Montemagni. *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, Felice, Simonetta Montemagni, Giulia Venturi. *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.

Tullio De Mauro. *Il dizionario della lingua italiana.* Torino, Paravia, 2000. Lucisano, Pietro, Maria Emanuela Piemontese. "*GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana*", «*Scuola e città»*, 3, 31, marzo 1988, La Nuova Italia.

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

Marina Tassara, *La lingua per studiare: una rassegna bibliografica,* Italiano LinguaDue, n. 2. 2010

Monografie

Maria Ferrari, Elisa Maggi, Franca Marchesi. 2008. *Antologia ITALIANO L2 – Testi d'autore*

*facilitati e semplificati per classi plurilingue*. Bergamo, Sestante.

Tiziano Franzi, Simonetta Damele. *A ciascuno il suo.* A cura di Gabriella Candia, Torino, Loescher, 2010.

Alessandro Lenci, *et al*. *Memorie di Guerra Un progetto di linguistica computazionale per le Digital Humanities*. Seminario di Cultura Digitale, Pisa, 22 ottobre 2014.

Siti web

Wikipedia, voce *Quadro comune europeo di riferimento per la conoscenza delle lingue* http://it.wikipedia.org/wiki/Quadro_comune_europeo_di_riferimento_per_la_conoscenza_delle_lingue (visitato il 16 Aprile 2015)

*Progetto Terence*:
http://terenceproject.eu/web/guest/home (visitato il 29 Maggio 2015)

Siti web per la creazione del corpus parallelo allineato

*Capire per studiare 3*:
http://www.google.it/url?
sa=t&rct=j&q=&esrc=s&source=web&cd=1&sqi=2&ved=0CCAQFjAA&url=http%3A%2F
%2Fnuke.istitutocomprensivoroncalli.it%2FLinkClick.aspx%3Ffileticket%3DgYWntxgeATE
%253D%26tabid%3D507%26mid
%3D1649&ei=p4uyU43_BYrf4QS3u4DYCA&usg=AFQjCNGegjrdr_mkn-f8f_pF6gV-
3WgCRQ&sig2=m6aeVfhw0cEhrJuDHETbcQ&bvm=bv.69837884,bs.1,d.ZGU (visitato il 9 maggio 2015)

*Percorsi di apprendimento per gli stranieri nella scuola italiana:*
http://www.researchgate.net/publication/40783189_Percorsi_di_apprendimento_per_
gli_stranieri_nella_scuola_italiana (visitato il 13 Aprile 2015)

*Approccio alla lingua italiana per allievi stranieri:*
http://www.retetrevisointegrazionealunnistranieri.it/download/laboratoriorete.pdf (visitato il 16

Aprile 2015)

*Il mito:*
http://www.scuolavicospinea.it/docenti/RISM/public/gruppo%2013.pdf (visitato il 15 Aprile 2015)

*Io sono così:*
http://www.scuolavicospinea.it/docenti/RISM/public/gruppo%208.pdf (visitato il 17 Aprile 2015)

*Il pinocchio di Collodi:*
http://www.scuolavicospinea.it/docenti/RISM/public/gruppo%208.pdf (visitato il 28 Aprile 2015)