Rinnone Sabrina

Laurea Magistrale in Informatica Umanistica

SEMINARIO DI CULTURA GENERALE

a.a. 2014-2015

Analisi della leggibilità di consensi informati

Sommario

1. Introduzione

La linguistica computazionale è una disciplina che si pone tra la linguistica tradizionale e la *Computer Science* e riguarda l'utilizzo di strumenti informatici per l'elaborazione del linguaggio naturale (in inglese *Natural Language Processing*, in sigla *NLP*).

Dalla sua nascita alla fine degli anni '50 e dalla sua configurazione come disciplina autonoma, ha subito una crescita esponenziale in diverse direzioni, contando oggi su numerosi gruppi di ricerca nel panorama scientifico internazionale, composti da linguisti (che producono i modelli teorici del linguaggio), psicologi (che forniscono un'analisi dei processi cognitivi umani), matematici e logici (che forniscono gli strumenti per esprimere tali modelli in modo computazionalmente trattabile) ed informatici (che sviluppano gli algoritmi atti ad implementare i modelli teorici dei fenomeni linguistici). Obiettivo comune è quello di riuscire a produrre macchine in grado di interagire con gli esseri umani utilizzando il linguaggio naturale.

Le attuali tecnologie del linguaggio permettono ai calcolatori di accedere in maniera del tutto nuova al contenuto informativo dei testi. Pur richiedendo una considerevole allocazione di risorse per lo sviluppo e il *processing*, tali tecnologie aumentano la quantità e la qualità dell'informazione estraibile dal testo rispetto all'applicazione di tecniche tradizionali e più superficiali di analisi del testo. Esse sono usate per perseguire un'ampia gamma di compiti, ad esempio la ricostruzione del profilo linguistico di generi testuali, l'estrazione di conoscenza di dominio da collezioni documentali, la valutazione del livello di leggibilità o il riconoscimento automatico della lingua madre (L1) di chi ha prodotto uno scritto in una seconda lingua (L2).

In questa relazione ho deciso di porre la mia attenzione sul tema della leggibilità. La valutazione automatica della leggibilità attraverso tecniche linguistico-computazionali rappresenta il primo passo verso la semplificazione di testi scritti. In particolare ho deciso di trattare la valutazione della leggibilità di testi di ambito medico, che può portare al miglioramento della comunicazione medico-paziente e in generale del servizio sanitario nazionale. Ho inoltre deciso di presentare il mio contributo a questo tipo di ricerca attraverso il progetto svolto durante il mio periodo di tirocinio presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa, ovvero un'indagine preliminare sulla valutazione della leggibilità di un corpus di moduli di consenso informato.

2. L'analisi automatica della leggibilità

All'interno della società dell'informazione, dove tutti dovrebbero essere in grado di accedere a tutte le informazioni disponibili, una questione centrale è il miglioramento dell'accesso alla lingua scritta.

È all'interno di questo scenario che il tema della leggibilità sta ottenendo un'importanza sempre maggiore. La leggibilità, intesa come valutazione della scorrevolezza di lettura e facilità di comprensione del significato di un testo, risulta ancora più importante nel caso in cui il destinatario di una qualunque comunicazione abbia scarse competenze linguistiche. Basti pensare a bambini e ragazzi che nella scuola dell'obbligo presentano una forma di difficoltà di

apprendimento, soggetti che vivono in situazioni di marginalità linguistico-culturale come gli immigrati in Italia, parlanti che apprendono una seconda lingua, pazienti che presentano delle disabilità.

I metodi tradizionali sino ad oggi adottati per l'analisi della leggibilità, l'**indice di Flesch** (Flesch, 1946; 1949) per la lingua inglese[1] e l'**indice GULPEASE** (Piemontese e Lucisano, 1988) per la lingua italiana[2], fanno affidamento unicamente su caratteristiche generali e formali del testo, non fornendo risultati ottimali sulla valutazione della difficoltà di lettura di un testo, soprattutto quando si analizzano testi appartenenti a determinati domini specifici.

Per questo motivo, negli ultimi anni vi è stato un progressivo affermarsi a livello internazionale dell'uso di tecnologie avanzate per il Trattamento Automatico del Linguaggio (TAL), che contribuiscono a migliorare la definizione di complessità linguistica. I nuovi indici di leggibilità sono infatti in grado di analizzare parametri di complessità linguistica più raffinati. Tali parametri spaziano tra i vari livelli di analisi linguistica e sono rintracciabili in modo automatico a partire dal risultato del processo di annotazione automatica del testo.

Per quanto riguarda la lingua italiana, il primo e al momento unico strumento avanzato basato su questo tipo di approccio, è **READ-IT**, sviluppato presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa. READ-IT è stato costruito, oltre che per misurare e valutare la leggibilità, anche per fornire un supporto alla redazione semplificata di un testo attraverso l'identificazione dei luoghi di complessità. READ-IT implementa un indice di leggibilità avanzato basato su analisi linguistica multi-livello del testo, lessicale, morfo-sintattico e sintattico, e conduce una classificazione probabilistica del testo rispetto a due classi (leggibile vs. complesso). I *corpora* rappresentativi di testi complessi e semplificati usati dal sistema appartengono allo stesso genere testuale (prosa giornalistica) e sono costituiti rispettivamente da *La Repubblica* (Rep) e *Due Parole* (2Par)[3].

I quattro modelli di analisi, corrispondenti a quattro indici di leggibilità, presentati da READ-IT sono: **Dylan BASE** (in questo modello le caratteristiche considerate sono la lunghezza della frase e delle parole); **Dylan LESSICALE** (questo modello si focalizza sulle caratteristiche lessicali del testo, costituite dalla composizione del vocabolario e la sua ricchezza lessicale); **Dylan SINTATTICO** (questo modello si basa su informazione di tipo grammaticale); **Dylan GLOBALE** (modello basato sulla combinazione dei tre modelli precedenti). Per ciascun modello, il valore ottenuto esprime il livello di difficoltà, in altre parole si riferisce alla probabilità di appartenenza alla classe dei testi di

[1] L'indice di Flesch calcola la leggibilità tenendo conto della lunghezza media delle parole, misurate in sillabe, e della lunghezza media delle frasi, misurate in parole. I valori ottenuti sono compresi in una scala da 0 a 100, dove 0 rappresenta un testo estremamente difficile e 100 un testo estremamente facile.

[2] L'indice GULPEASE calcola la leggibilità tenendo conto della lunghezza media delle parole, misurate in caratteri, e della lunghezza media delle frasi, misurate in parole. I valori ottenuti sono compresi, come per l'indice di Flesch, in una scala da 0 a 100: i lettori che hanno un'istruzione elementare leggono facilmente i testi che presentano un indice superiore ad 80, quelli che hanno un'istruzione media un indice superiore a 60 e infine quelli che hanno un'istruzione superiore un indice superiore a 40.

[3] *Due Parole* (Piemontese, 1996) è un giornale d'informazione di facile lettura i cui articoli sono scritti utilizzando in modo consapevole e sistematico criteri di scrittura controllata. Si rivolge a persone che hanno bisogno di testi informativi molto leggibili e comprensibili.

difficile leggibilità. I valori ottenuti variano su una scala che va da 0 (facile da leggere) a 100 (difficile da leggere).

2.1 Annotazione linguistica

Il prerequisito per la valutazione automatica della leggibilità è l'annotazione automatica multi-livello del testo. Il processo di annotazione avviene a partire dal monitoraggio dei *corpora* selezionati come rappresentativi e permette di identificare la struttura linguistica sottostante al testo. L'annotazione linguistica si presenta come un processo incrementale realizzato da una serie di passaggi distinti che, operando in successione, generano analisi linguistiche progressivamente più complesse per il tipo di informazione estratta dal testo.

I primi passaggi dell'annotazione linguistica consistono nella **segmentazione del testo in frasi** e nella **tokenizzazione**. La nozione di "frase" non è di facile definizione, infatti i criteri utilizzati possono variare in base al genere testuale o alla varietà della lingua. Tokenizzare un testo, invece, significa dividere le sequenze di caratteri in unità minime di analisi dette "token": parole, punteggiatura, date, numeri, sigle, ecc. I token possono essere delle entità strutturalmente complesse (es. date), ma sono comunque assunte come unità di base per i successivi livelli di elaborazione (morfologico, sintattico, ecc.). A seconda del tipo di lingua e sistema di scrittura può essere un task estremamente complesso[4].

Ai passaggi di analisi superficiale del testo seguono la **lemmatizzazione** (il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta lemma), l'**analisi morfo-sintattica** e l'**analisi sintattica**.

Nel processo di analisi morfo-sintattica, ad ogni token del testo viene associata informazione relativa alle possibili categorie grammaticali, integrata da ulteriori specificazioni morfologiche quali ad esempio genere, numero, ecc. In presenza di token morfologicamente ambigui, viene scelta la giusta interpretazione morfologica che la parola ha nel contesto specifico. Questo processo, che prende il nome di *part-of-speech tagging*, riguarda anche i token morfologicamente ambigui e, nonostante la percentuale di errori aumenti, rende il testo analizzabile ai livelli successivi[5].

Per quanto riguarda l'analisi sintattica, esistono due approcci diversi nel campo della linguistica computazionale: la rappresentazione a costituenti, basata sull'identificazione di costituenti sintattici (ad esempio sintagmi nominali, sintagmi verbali, sintagmi preposizionali, ecc.) e delle loro relazioni di incassamento gerarchico, e la **rappresentazione a dipendenze**, che fornisce una descrizione della frase in termini di dipendenze tra parole come soggetto, oggetto diretto, modificatore, ecc. Tra i due approcci, la rappresentazione a dipendenze è solitamente preferita in

[4] Nelle lingue non segmentate, dove i confini di parola non sono marcati esplicitamente nella scrittura (ad esempio cinese, giapponese, ecc.), la tokenizzazione prende il nome di *word segmentation*.

[5] Il PoS tagger utilizzato, l'unico per la lingua italiana, è quello descritto in Dell'Orletta (2009) e presenta un'accuratezza (calcolata come il rapporto tra il numero di tokens classificati correttamente e il numero totale di *tokens* analizzati) del 96,34% nell'assegnazione della giusta PoS e dei relativi tratti morfo-sintattici.

quanto presenta numerosi vantaggi in relazione a lingue caratterizzate da una certa variabilità al livello dell'ordine dei costituenti della frase, come la lingua italiana[6].

2.2 Tratti monitorati

Dal monitoraggio del profilo linguistico è possibile ricavare le caratteristiche linguistiche monitorate nella misura della leggibilità di un testo, organizzate in base alle fasi di annotazione da cui derivano.

A partire dal profilo linguistico di base i tratti monitorati sono la **lunghezza media dei periodi**, espressa in token, e la **lunghezza media delle parole**, espressa in caratteri. Periodi e parole più brevi risultano essere più comprensibili.

A partire dal profilo morfo-sintattico i tratti monitorati sono la **composizione del vocabolario**, in altre parole la percentuale di lemmi del vocabolario del testo non appartenenti al vocabolario di base del *Grande Dizionario italiano dell'uso* (De Mauro, 2000)[7], il **rapporto tipo/unità (TTR)**, la **distribuzione delle categorie morfo-sintattiche** e la **densità lessicale**. La TTR è calcolata come rapporto tra il numero di parole tipo[8] e il numero di occorrenze delle unità del vocabolario di un testo, i cui valori oscillano da 0 ad 1, dove valori vicino allo 0 indicano che il vocabolario del testo è meno vario, mentre valori vicino ad 1 caratterizzano testi particolarmente variegati dal punto di vista lessicale. Per quanto riguarda la distribuzione delle categorie morfo-sintattiche, si fa riferimento ad un sottoinsieme di categorie: sostantivi (distinguendo tra nomi comuni e propri), aggettivi, verbi e congiunzioni (distinguendo tra coordinanti e subordinanti). Una percentuale alta di aggettivi, la parte variabile del discorso che si aggiunge al nome per qualificarlo o per determinarlo meglio, corrisponde ad un alta ricchezza lessicale. Una percentuale alta di congiunzioni, la parte del discorso che serve ad unire tra loro sintagmi in una proposizione oppure proposizioni in un periodo, corrisponde ad una maggiore difficoltà dei testi. In particolare la loro ripartizione in coordinanti (che uniscono due parole o due proposizioni che non dipendono tra loro) e subordinanti (che uniscono due parole o due proposizioni che dipendono tra loro) fornisce un'indicazione approssimativa tra costruzioni paratattiche e ipotattiche dei testi e una percentuale più alta di quest'ultime rappresenta una fattore di complessità. La densità lessicale, invece, è calcolata come il rapporto tra "parole piene" (in altre parole portatrici di significato: nomi, aggettivi, verbi ed avverbi) e il numero totale delle occorrenze di parole del testo.

A partire dall'analisi sintattica sottostante al testo, i tratti che influenzano il livello di leggibilità di un testo sono diversi. Per quanto riguarda l'articolazione interna del periodo e della proposizione, i

[6] Il parser utilizzato, che rappresenta lo stato dell'arte del parsing a dipendenze per l'italiano, è *DeSR* (Attardi, Dell'Orletta et al. 2007) che ha raggiunto performance di LAS (metrica che indica la percentuale di dipendenze identificate ed etichettate correttamente) dell'83,38% - 88,67%.

[7] Il vocabolario di base è ripartito rispetto ai repertori d'uso Fondamentale (circa 2000 parole conosciute e usate da coloro che hanno almeno un'istruzione elementare), Alto uso (circa 3000 parole conosciute e usate da coloro che hanno almeno un'istruzione media), Alta disponibilità (circa 2000 parole altamente latenti, presenti all'uso che i parlanti non usano concretamente tutti i giorni, ma solo all'occorrenza).

[8] La parola tipo rappresenta la classe di tutti i *tokens* che contengono la stessa sequenza di caratteri. Due parole appartengono allo stesso tipo se sono formalmente indistinguibili a prescindere dalla posizione che occupano nel testo.

tratti monitorati sono il **numero medio di proposizioni per periodo**, costituito dal rapporto tra proposizioni e periodi, il **numero di proposizioni principali e subordinate**, il **numero medio di parole per proposizione**, costituito dal rapporto tra parole e proposizioni, e il **numero medio di dipendenti per testa verbale**. Un numero elevato di proposizioni per periodo indica un livello alto di complessità sintattica del testo, anche se questo dato non dice ancora nulla su come le diverse proposizioni si rapportino l'una con l'altra all'interno del periodo e quindi risulta interessante identificare la proporzione di proposizioni principali e subordinate. La presenza minore delle prime e maggiore delle seconde rappresenta un livello alto di complessità sintattica del testo. Anche un numero medio di parole per proposizione e di dipendenti per testa verbale elevati rappresentano un fattore di complessità. Un altro aspetto rilevante per la misura della complessità del testo riguarda i livelli di incassamento gerarchico: in presenza di più di una proposizione subordinata all'interno dello stesso periodo, è importante ricostruire quale tipo di rapporto sussista tra di esse, in altre parole se siano ricorsivamente incassate l'una all'interno dell'altra dato che solamente il numero di proposizioni subordinate non è sufficiente a definirlo. Una misura dei livelli di incassamento gerarchico all'interno della struttura sintattica è ricostruita a partire dall'**altezza massima dell'albero**, che misura la massima distanza che intercorre tra una foglia (rappresentata da parole del testo senza dipendenti) e la radice dell'albero, espressa come numero di archi (relazioni di dipendenza) attraversati nel cammino foglia-radice. Questa misura viene migliorata focalizzandosi su particolari tipi di costrutti sintattici: la ricorrenza di strutture nominali complesse costituite da una testa nominale modificata da aggettivi e/o complementi preposizionali (si prende in considerazione la **profondità media di strutture nominali complesse**, che registra la media delle profondità di strutture nominali con modificatori), e la ricorrenza di proposizioni subordinate ricorsivamente incassate (si prende in considerazione la **profondità media di "catene" di subordinazione**, che registra la ricorrenza di proposizioni subordinate ricorsivamente incassate). Valori alti corrispondono ad una significativa complessità sintattica dei testi. Infine, dato che la contiguità di elementi semanticamente e/o sintatticamente "vicini" permette una più immediata recuperabilità e accessibilità dei rapporti sussistenti tra le parole, risulta importante studiare la "lunghezza" delle relazioni di dipendenza, calcolata come la distanza in token tra la testa e il dipendente. Questo aspetto della struttura sintattica viene monitorato attraverso due parametri, ossia la **lunghezza media**, che corrisponde alla lunghezza media di tutte le relazioni di dipendenza (con esclusione dei legami riguardanti la punteggiatura) e la **media delle lunghezze massime**, che corrisponde alla media dei legami di dipendenza più lunghi per ciascuna frase. Valori alti sono rappresentativi di testi significativamente complessi.

3. L'analisi della leggibilità di testi medici

A partire dagli anni '80, l'attenzione per la qualità della comunicazione pubblica è notevolmente aumentata, specialmente in correlazione allo sviluppo di internet come mezzo di comunicazione e fonte del tutto nuova di raccolta e diffusione di documenti. Nonostante un'ampia proliferazione dei mezzi comunicativi, rimane difficile trovare regole e metodi al corretto trasferimento delle informazioni. Uno degli ambiti in cui questo fenomeno si registra con particolare frequenza e rilevanza, è la comunicazione tra medico e paziente.

La comunicazione tra medico e paziente costituisce un tassello fondamentale nella pratica medica. Attraverso una comunicazione efficace si ottengono informazioni, si fa una diagnosi, si condivide un piano di trattamento, contesti nei quali inizia la guarigione del paziente. È quindi fondamentale che tra medico e paziente si crei una relazione intensa che influenzi il livello di motivazione del paziente a star meglio aderendo al trattamento che gli viene proposto e aumentando quindi le possibilità della sua guarigione.

Da alcuni anni è nata una nuova consapevolezza dell'utilità di rendere i medici non solo dei professionisti sul campo della salute ma anche esperti della comunicazione, in quanto la relazione che si crea tra medico e paziente, i modi e i contenuti della comunicazione hanno un peso maggiore rispetto a quanto avviene in tutte le interazioni comunicative di tipo professionale. Infatti, informazioni ben precise sulle malattie e sul loro decorso, sui trattamenti medico-sanitari possibili, sulle possibilità di guarigione, sulle aspettative di vita, trasformano tutti i giorni la vita di migliaia di persone.

Il fattore che più influenza l'efficacia della comunicazione medico-paziente è il tipo di linguaggio che viene usato. Il linguaggio medico fa parte di quei linguaggi che vengono chiamati settoriali, quelle varietà di lingua utilizzate in determinati settori della vita sociale e professionale. Un linguaggio settoriale è il modo di esprimersi (parole, espressioni, termini tecnici, ecc.) proprio di un ambito specialistico, in particolare (ma non soltanto) di natura tecnica e scientifica.[9] Fra i vari linguaggi settoriali, quello medico inoltre interessa da vicino l'esperienza di tutti i parlanti, in quanto capita comunemente di imbattersi in questioni relative alla salute.

Anche all'interno di questo scenario risulta centrale quindi l'analisi della leggibilità, in questo caso di testi di dominio medico, in quanto costituisce il primo passo verso la semplificazione di testi scritti. La valutazione della leggibilità e la semplificazione, dove necessaria, dei testi scritti di dominio medico può portare al miglioramento della comunicazione medico-paziente e in generale del servizio sanitario nazionale. Il progetto che ho svolto durante il mio periodo di tirocinio si inserisce all'interno di questo ambito di ricerca.

3.1 L'analisi della leggibilità di consensi informati

Durante il mio periodo di tirocinio presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa, è stata svolta un'indagine preliminare sulla valutazione della leggibilità di un corpus di moduli di consenso informato, utilizzati prima di una procedura clinica negli ospedali del Servizio Sanitario Regionale della Toscana. Il progetto prende vita dall'idea che le informazioni relative alla salute debbano essere accessibili a tutti i membri della società, comprese le persone che hanno difficoltà di lettura a causa di un basso livello d'istruzione o perché il testo è scritto in una lingua diversa rispetto alla loro lingua madre.

In Italia il consenso informato è una forma di autorizzazione del paziente a ricevere un qualunque trattamento sanitario, medico o infermieristico, con il quale viene a conoscenza di tutte le

[9] Definizione presa dall'Enciclopedia Treccani, versione online (http://www.treccani.it/enciclopedia/).

informazioni disponibili sulla propria salute e la propria malattia. Il paziente ha il diritto/dovere di chiedere al medico tutto ciò che non è chiaro e quindi ha la possibilità di scegliere, in modo informato, se sottoporsi o meno ad una determinata terapia o esame diagnostico.

Il progetto è stato svolto in collaborazione con il Centro Gestione Rischio Clinico e Sicurezza del Paziente (Centro GRC) della regione Toscana[10], il quale ha fornito il *corpus* di consensi informati su cui poter lavorare.

Il *corpus* è composto da 584 documenti attualmente in uso presso 36 ospedali pubblici della regione Toscana e coprono 29 specialità, per un totale di 607790 *tokens*. I documenti sono inoltre organizzati in 4 macroaree (area chirurgica, area medica, area prevenzione, area servizi), cui si aggiungono altre 3 specialità (Generici, Pediatria, Riabilitazione e rieducazione funzionale). Il numero di testi e il numero di *tokens* varia tra le varie specialità. La specialità con più testi a disposizione è ORL[11], appartenente all'area chirurgica, con un totale di 134 testi e 194405 *tokens*. Le specialità con meno testi a disposizione sono invece Diabetologia (area medica) con un solo testo e 297 *tokens* e Vaccini (area prevenzione) con un solo testo e 2852 tokens.

Lo studio è stato realizzato mediante la combinazione di tecniche linguistico-computazionali e algoritmi di apprendimento automatico. Come strumento è stato usato READ-IT, di cui si è parlato nel precedente paragrafo. Rispetto ai quattro modelli di analisi della leggibilità (BASE, LESSICALE, SINTATTICO, GLOBALE), la complessità dei testi è stata monitorata rispetto al livello lessicale e sintattico. Il modello BASE, che fa affidamento unicamente su caratteristiche generali e formali del testo, quali la lunghezza delle frasi e delle parole, e il modello GLOBALE, che combina i risultati degli altri tre modelli, non forniscono risultati ottimali e non sono quindi stati presi in considerazione. In particolare si è data una maggiore rilevanza, nella prima parte del progetto, alle caratteristiche lessicali, in quanto un dominio così specifico, quale quello medico, presenta un lessico molto specializzato che richiede un discorso più approfondito.

La prima parte del progetto ha avuto come obiettivo la creazione, per ognuna delle 29 specialità, di liste contenenti le parole che non appartengono al vocabolario di base con associata la relativa frequenza. I valori di leggibilità ottenuti valutando le caratteristiche lessicali dei testi sono infatti particolarmente influenzati, oltre che dalla ricchezza lessicale, dalla composizione del vocabolario e quindi dalla percentuale di lemmi non appartenenti al vocabolario di base. Quest'ultimo, composto dai lessemi più comuni della lingua italiana, non contiene parole specifiche di ambito medico. Per esempio, il vocabolario di base non contiene parole come "farmaco", "emorragia", "rianimazione", "lesione", "prevenzione" e altre la cui presenza è spesso necessaria per esprimere il contenuto di testi come quelli del consenso informato.

[10] http://www.regione.toscana.it/centro-gestione-rischio-clinico
[11] Sigla della specialità di Otorinolaringoiatra.

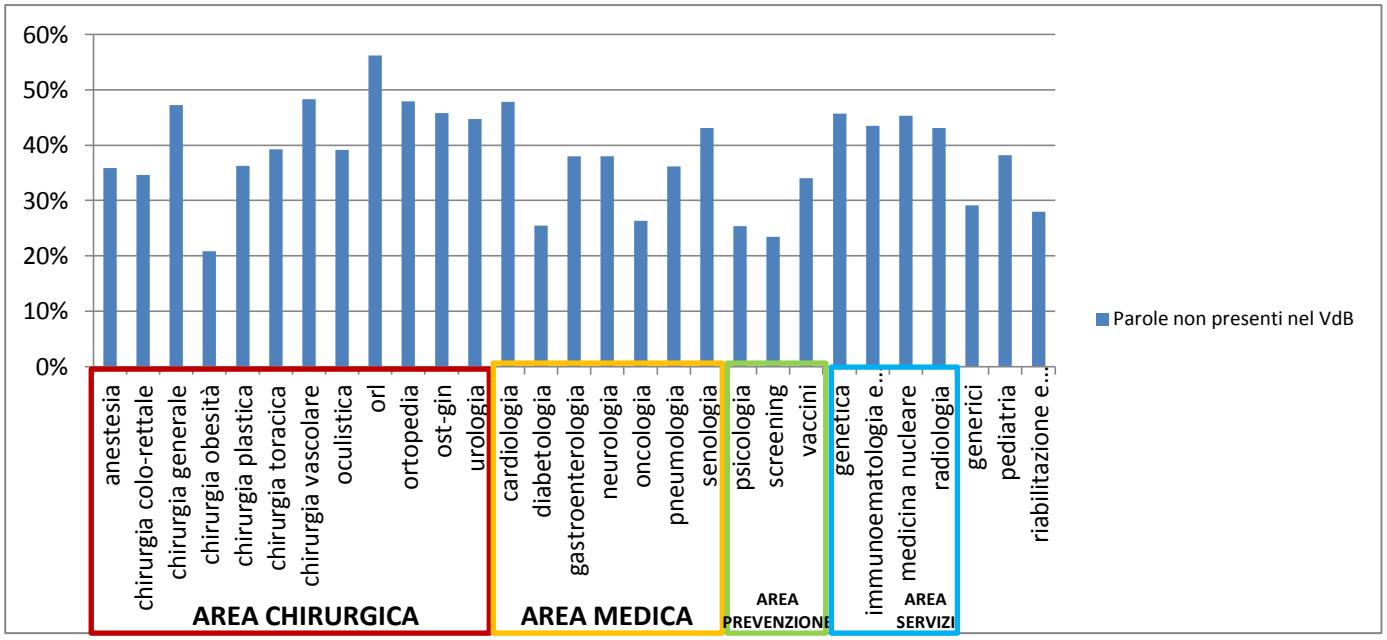


Grafico 1. Percentuale delle parole non presenti nel vocabolario di base per ogni specialità.

La specialità che contiene la più alta percentuale di parole tipo non appartenenti al VdB è ORL (56,2% su un totale di 5146 parole tipo) e, in particolare, quelle che compaiono nei testi con maggiore frequenza sono "complicanza" (557), "anestesia" (511) e "patologia" (335). Altre specialità con percentuali alte sono Chirurgia vascolare (48,30%) e Cardiologia (47,77%). La specialità che invece contiene la più bassa percentuale di parole tipo non appartenenti al VdB è Chirurgia dell'Obesità (20,85% su un totale di 1437 parole tipo) e quelle che compaiono con maggiore frequenza sono "gastrico" (60), "complicanza" (58) e "obesità" (51). Altre specialità con percentuali basse sono Screening (23,42%) e Diabetologia (25,48%).

Dopo aver creato le liste delle parole che non appartengono al vocabolario di base per tutte le specialità disponibili, l'idea è stata quella di sottoporle alle persone specializzate del team del Centro Gestione Rischio Clinico, con cui si è istaurata la collaborazione, le quali sono in grado di classificarle secondo tre livelli. Il livello 1 è costituito dalle parole di ambito medico di cui non si può fare a meno per esprimere un concetto e dalle parole che sono di facile comprensibilità nonostante non compaiono nel vocabolario di base, il quale risulta ormai abbastanza obsoleto. Il livello 2 è costituito da quelle parole di ambito medico di difficoltà media che si possono esprimere usando dei sinonimi di più facile comprensibilità. Il livello 3 è costituito da quelle parole talmente specialistiche e difficili che possono essere penalizzate. L'obiettivo di questa classificazione è la creazione, in futuro, di un vocabolario di base di ambito medico ripartito secondo i repertori d'uso **Fondamentale Medico** (livello 1), **Alto Uso Medico** (livello 2), **Specialistico Medico** (livello 3), con cui sarà possibile raffinare READ-IT e quindi migliorare la valutazione della leggibilità per i testi di un dominio specifico quale quello medico.

La seconda parte del progetto è stata invece dedicata all'analisi della leggibilità rispetto al livello lessicale e sintattico. Per ogni livello l'analisi è stata svolta rispetto alle specialità e rispetto alle aziende ospedaliere. L'obiettivo è stato quello di capire se tra le specialità vi sono alcune che

risultano essere più complesse rispetto alle altre in relazione al contenuto che esprimono, oppure se i testi appartenenti ad una determinata azienda ospedaliera risultano essere più complessi rispetto agli altri a prescindere dalla specialità.

Per quanto riguarda la valutazione della leggibilità per specialità, a livello lessicale i valori ottenuti sono molto alti (media = 0,95), invece a livello sintattico sono molto più vari (media = 0,79)

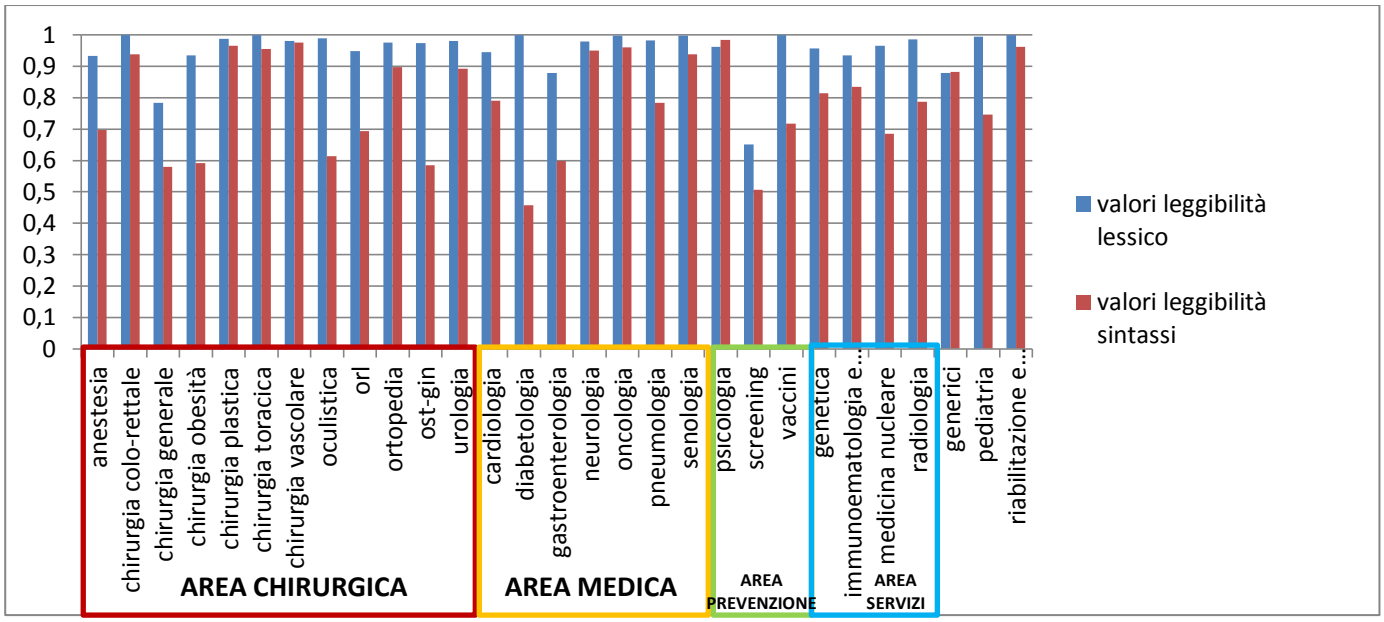


Grafico 2. Valori della leggibilità ai livelli lessicale e sintattico per ogni specialità.

A livello lessicale, considerando come tratti che influenzano la leggibilità la composizione del vocabolario e la ricchezza lessicale dei testi, i valori variano da un minimo di 0,65 (Screening) ad un massimo di 1 (Chirurgia colo-rettale, Chirurgia toracica, Diabetologia, Oncologia, Senologia, Vaccini, Riabilitazione e rieducazione funzionale). Lo Screening è un intervento sanitario che mira a mettere in evidenza la presenza di un'eventuale malattia nelle sue fasi iniziali in modo tale da poter intervenire tempestivamente con le cure più appropriate, facilitando la guarigione e riducendo la mortalità. Rispetto a specialità appartenenti all'area chirurgica o area medica, che racchiudono operazioni molto più complicate, è caratterizzato da un lessico molto più facile e quindi risulta essere la specialità meno complessa.

A livello sintattico, basandosi su informazione di tipo grammaticale come misura per la leggibilità, i valori variano da un minimo di 0,46 (Diabetologia) ad un massimo di 0,98 (Psicologia).

Dato che i valori presi in considerazione per ogni specialità sono i valori medi, di alcune di esse (Chirurgia generale, Chirurgia toracica, Ortopedia e Urologia per l'area medica, Pneumologia e Senologia per l'area medica, Psicologia per l'area prevenzione, Genetica per l'area servizi, Generici, Pediatria e Riabilitazione e rieducazione funzionale), al livello sintattico, è stata calcolata la deviazione standard, che esprime la dispersione dei dati intorno ad un indice di posizione, in questo caso appunto la media aritmetica (valore atteso).

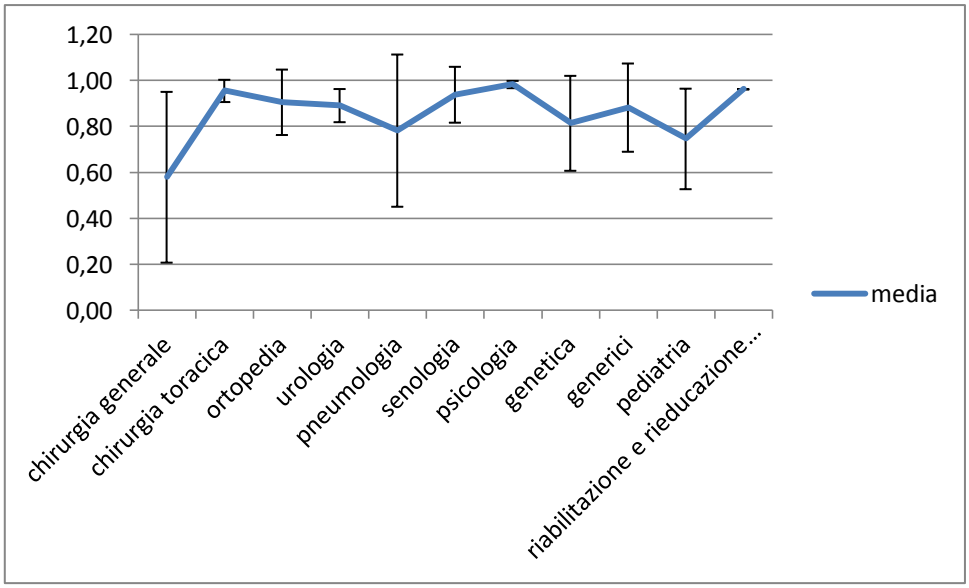
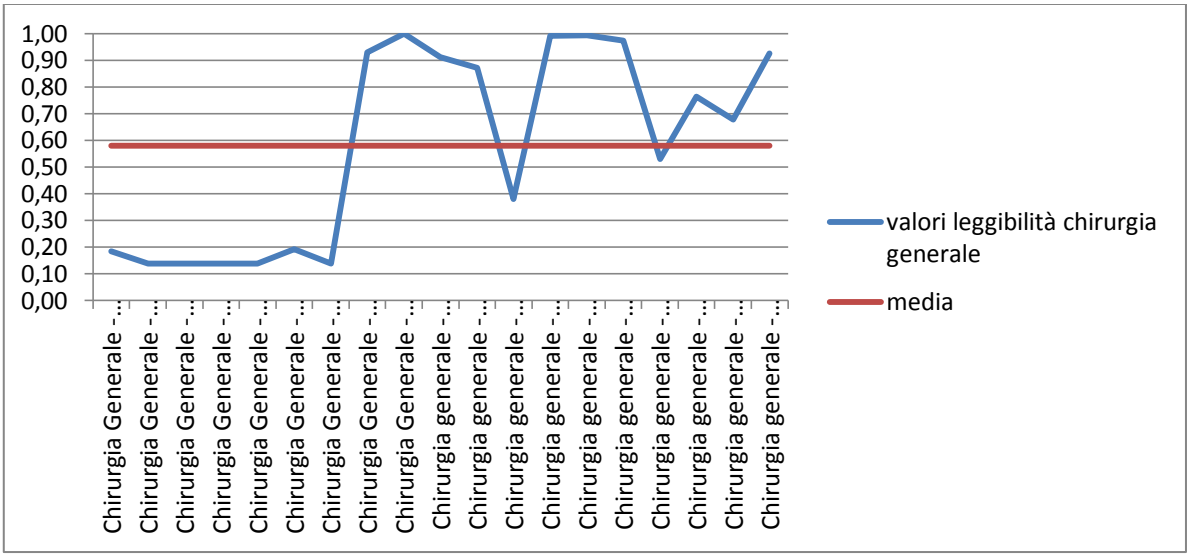
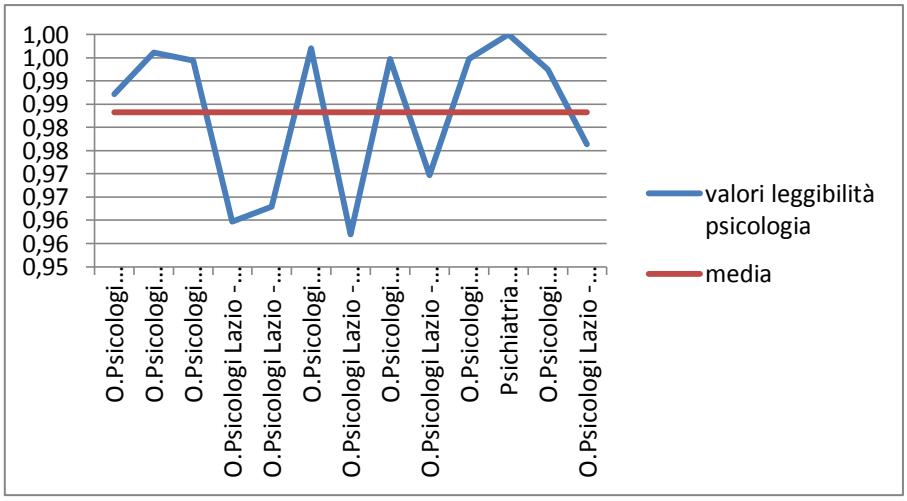


Grafico 3. Media e deviazione standard a livello sintattico per alcune specialità.

Le specialità che hanno i valori più bassi di deviazione standard sono Psicologia (0,01) e Chirurgia Toracica (0,04), invece ad avere i valori più alti sono Chirurgia Generale (0,37) e Pneumologia (0,33). Ciò vuol dire che le prime saranno caratterizzate da testi i cui valori di leggibilità sono tutti prossimi tra loro e quindi la media è un valore ragionevolmente molto preciso, al contrario delle seconde, la cui media non sarà molto accurata. Per mostrare ciò, si è deciso di studiare l'andamento interno della leggibilità per le specialità di Chirurgia generale e di Psicologia.





Grafici 4-5. Andamento interno della leggibilità a livello sintattico di Chirurgia generale e Psicologia.

La specialità di Chirurgia generale è infatti costituita da testi i cui valori di leggibilità variano da un minimo di 0,14 ad un massimo di 1 (media = 0,58). Al contrario la specialità di Psicologia è costituita da testi i cui valori di leggibilità variano da un minimo di 0,96 ad un massimo di 1 (media = 0,98).

Per quanto riguarda invece la leggibilità per azienda ospedaliera, non è stato possibile ricavare per tutti i testi a disposizione l'appartenenza ad una determinata azienda ospedaliera. Le uniche due ASL di cui è stato possibile ricavare un numero consistente di testi e su cui è stata avviata l'analisi della leggibilità sono i consensi informati dell'ASL 4 (Prato) e dell'ASL 11 (Empoli).

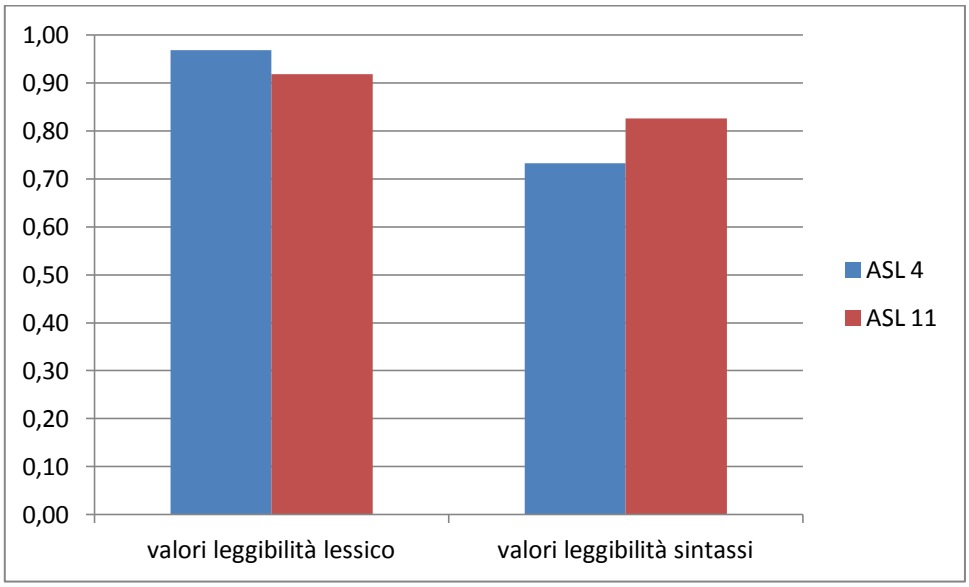


Grafico 6. Valori della leggibilità ai livelli lessicale e sintattico per ASL 4 (Prato) e ASL 11 (Empoli).

I testi di entrambe le aziende ospedaliere sono più complessi al livello lessicale rispetto al livello sintattico. In particolare i testi dell'ASL 4, al livello lessicale, risultano più complessi (0,97) rispetto a quelli dell'ASL 11 (0,92). Invece risultano meno complessi a livello sintattico (0,73 vs 0,83).

È stato poi possibile suddividere alcuni dei testi delle due ASL in specialità, così da poter valutare la leggibilità al livello sintattico sia per ASL che per specialità.

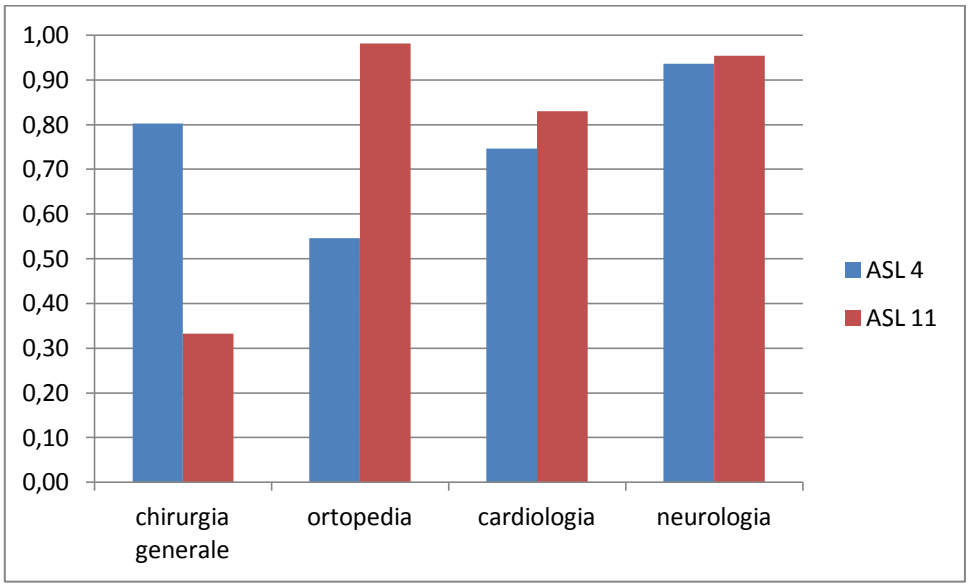


Grafico 7. Valori della leggibilità a livello sintattico per ASL e per specialità.

Le specialità che sono state ottenute per l'ASL 4 e l'ASL 11 sono Chirurgia generale, Ortopedia, Cardiologia, Neurologia. A livello sintattico la specialità che risulta essere più complessa, per entrambe le ASL, è Neurologia (0,94 - 0,95). L'altra specialità con valori abbastanza alti e molto simili per entrambe le ASL è Cardiologia (0,75 – 0,83). Invece, per le altre due specialità restanti, i valori ottenuti variano significativamente tra le due ASL. Per quanto riguarda Chirurgia generale, i testi dell'ASL 4 sono più complessi rispetto a quelli dell'ASL 11 (0,80 vs 0,33), invece per quanto riguarda Ortopedia avviene il contrario (0,55 vs 0,98).

4. Conclusioni

L'analisi automatica della leggibilità si inserisce all'interno dell'ampia gamma di compiti perseguiti dal campo della linguistica computazionale e ne costituisce uno tra i più fiorenti, specialmente quando ad essere analizzati sono testi appartenenti a domini specifici.

In questa relazione ho trattato il tema della valutazione automatica di testi scritti appartenenti al dominio medico, presentando in particolare l'analisi preliminare sulla leggibilità dei consensi informati svolta durante il mio periodo di tirocinio presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa.

Alla fine dell'analisi preliminare sulla leggibilità dei moduli di consensi informati, con la quale è stato possibile definire quali specialità e aziende ospedaliere contengono i testi più complessi a livello lessicale e sintattico, le questioni rimaste aperte sono molte. Per prima cosa sarebbe necessaria la definizione precisa della tipologia dei testi, infatti all'interno del *corpus* i testi hanno varie nomenclature, ad esempio "foglio informativo", "modulo di consenso", "modulo di dissenso", "lettera di accompagnamento", e sarebbe utile capire se esiste una relazione rispetto al

livello di leggibilità. Inoltre, sarebbe necessario definire per ogni testo l'appartenenza ad una determinata azienda ospedaliera, in modo tale da ampliare l'analisi della leggibilità a tutti i testi disponibili rispetto alle aziende ospedaliere attive nella regione Toscana.

Il passo successivo sarà la semplificazione dei testi più complessi, sia a livello lessicale che a livello sintattico, prendendo anche in considerazione il destinatario a cui tali consensi informati si rivolgono. Ad esempio, per un'ipotetica semplificazione dei testi dell'ASL 4 (Prato) si dovrà tenere in considerazione il fatto che Prato sia la città con la percentuale più alta di cittadini stranieri residenti (34.171)[12] nella regione Toscana. La semplificazione dei testi porterà quindi ad un miglioramento della comunicazione medico-paziente e del servizio sanitario nazionale.

La presentazione di questo progetto ha avuto come obiettivo la dimostrazione della validità delle tecniche linguistico-computazionali nell'analisi della leggibilità di testi scritti. Con la creazione di un vocabolario di base di ambito medico sarà inoltre possibile raffinare lo strumento READ-IT e quindi migliorare la valutazione della leggibilità per i testi di dominio medico. A mio parere, ciò su cui si dovrebbe lavorare è la creazione di risorse adeguate, promuovendo la diffusione delle fonti testuali, per sviluppare strumenti sempre migliori. Ad esempio, in questo caso, non è stato possibile lavorare sui testi appartenenti all'ASL 3 (Pistoia) in quanto protetti da password.

[12] Popolazione straniera residente al 31 dicembre 2014 in ISTAT (http://demo.istat.it/str2014/index.html).

5. Bibliografia

Dell'Orletta F., M. Wieling, A. Cimino, G. Venturi, S. Montemagni. 2014. *Assessing the Readability of Sentences: Which Corpora and Features?* In: Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, Maryland, USA.

Dell'Orletta F., S. Montemagni, G. Venturi. 2011. *READ-IT: assessing readability of Italian text with a view to text simplification*. In: Proceedings of the Second Workshop in Speech and Language Processing fot Assistive Technologies, pp. 73-83, Edinburgh, Scotland, UK.

Dell'Orletta F., S. Montemagni, G. Venturi. 2014. *Assessing document and sentence readability in less resourced languages and across textual genres*. In: Recent Adavances in Automatic Readability Assessment and Text Simplification.

Lenci A., S. Montemagni, V. Pirrelli, *Testo e computer. Elementi di linguistica computazionale,* Roma, Carocci, 2005.

Miller J.R., W. Kintsch. 1980. *Readability and recall of short prose passages: A theoretical analysis*. In: Journal of Experimental Psychology: Human Learning and Memory.

Montemagni S. 2014. *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*.

Roberti A., Belotti C., Caterino L., *Comunicazione Medico Paziente. La comunicazione come strumento di lavoro del medico*, NLP Italy – Alessio Roberti Editore, 2006.