Università di Pisa
Dipartimento di Filologia, Letteratura e Linguistica
Corso di Laurea Magistrale in Informatica Umanistica

Analisi della leggibilità dei consensi informati: un approccio linguistico-computazionale

Sabrina Rinnone

Relatore
Prof.ssa Simonetta Montemagni

Correlatore
Dott.ssa Giulia Venturi

Anno Accademico 2015/2016

Indice dei contenuti

Elenco delle figure

Elenco delle tabelle

Introduzione

Questa tesi affronta il tema della valutazione automatica della leggibilità come passo preliminare alla semplificazione dei testi scritti. In particolare la tesi si concentra sulla valutazione automatica della leggibilità di testi appartenenti al dominio medico, un dominio cruciale in quanto le informazioni che riguardano la salute dovrebbero essere accessibili a tutti i membri della società indipendentemente dallo status socio-culturale e dal livello di competenza linguistica.

Le misure tradizionali per valutare la leggibilità si concentrano solamente su caratteristiche generali e formali del testo. Invece attraverso l'uso delle attuali tecnologie del linguaggio è possibile monitorare un'ampia varietà di fattori linguistici che influenzano la leggibilità di un testo. Nel dominio medico queste potenzialità non sono state però indagate a fondo per valutare la leggibilità dei testi medici e per supportare i professionisti sanitari alla semplificazione, dove necessaria, dei documenti destinati al pubblico. Per questo motivo, a partire dai metodi oggi usati e dallo strumento generale attualmente utilizzato nella valutazione automatica della leggibilità e nella semplificazione dei testi per la lingua italiana, in questa tesi è presentata una metodologia di analisi automatica della leggibilità di moduli di consenso informato. L'obiettivo è quello di verificare se questi metodi di nuova generazione possano essere specializzati rispetto al dominio medico e in una prospettiva interlinguistica.

Il **capitolo 1** introduce il tema della valutazione automatica della leggibilità. Sono elencate le metriche principali utilizzate distinguendo quelle tradizionali da quelle che si basano sull'uso di tecnologie avanzate per il Trattamento Automatico del Linguaggio (TAL). Sono poi presentati i metodi principali adottati per la valutazione automatica della leggibilità di testi appartenenti al dominio medico per l'inglese, lo svedese e l'italiano.

Il **capitolo 2** introduce la metodologia di analisi automatica della leggibilità di un testo e illustra gli strumenti di annotazione linguistica automatica e di valutazione della leggibilità utilizzati in questa tesi. Essi sono applicati ad un corpus di moduli di consenso informato attualmente in uso presso gli ospedali e le aziende sanitarie locali della regione Toscana, fornito dal Centro Gestione Rischio Clinico e Sicurezza del Paziente (Centro GRC). L'analisi preliminare del corpus è stata svolta durante il periodo di tirocinio presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa e ampliata in questo lavoro di tesi. Dai risultati dell'annotazione linguistica automatica è ricostruito il profilo linguistico del corpus grazie a corpora rappresentativi di altre varietà linguistiche, con lo scopo di identificare l'insieme dei tratti linguistici che caratterizzano i consensi informati. La valutazione automatica della leggibilità è condotta con READ-IT, primo strumento di valutazione automatica della leggibilità per la lingua italiana. L'analisi della leggibilità dei documenti ha permesso di mettere alla luce differenze e similarità dei consensi informati che riguardano diverse specialità e che sono stati rilasciati da determinate aziende sanitarie locali. Un'attenzione particolare è posta sull'analisi sul lessico del corpus caratterizzato da parole specifiche dell'ambito medico che non sono solitamente usate nel linguaggio comune, alcune delle quali indispensabili per esprimere il contenuto dei consensi informati.

Nel **capitolo 3** si affronta il tema della leggibilità di testi medici da una prospettiva interlinguistica presentando un caso di studio su due lingue tipologicamente distanti: il basco e l'italiano. La scelta della lingua basca deriva dall'esperienza di tirocinio svolto presso l'IXA Group, gruppo di ricerca dell'Università dei Paesi Baschi

(UPV/EHU) che lavora nel campo della Linguistica Computazionale e del Trattamento Automatico del Linguaggio. A partire dal confronto dei principi su cui si basano gli strumenti per la valutazione automatica della leggibilità, vale a dire *ErreXail* e READ-IT, e la definizione delle caratteristiche linguistiche in comune prese in considerazione, è condotto un confronto linguistico di corpora comparabili di consensi informati in italiano e basco alla ricerca di fenomeni di complessità linguistica comuni alle due lingue.

Prendendo spunto dai risultati della valutazione della leggibilità, nel **capitolo 4** si presenta una metodologia di semplificazione semi-automatica di un modulo di consenso informato all'interno di un programma per la procreazione assistita. Il lavoro nasce dalla collaborazione dell'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa con il Centro Demetra di Firenze, centro che opera nel campo della Procreazione Medicalmente Assistita (PMA) e convenzionato con il Sistema Sanitario Nazionale. I risultati dell'analisi della leggibilità sono confrontati rispetto alla risposta di pazienti che si sono sottoposti al trattamento attraverso la somministrazione di un questionario anonimo per monitorare il loro grado di soddisfazione, dovuto anche alla chiarezza del consenso informato. La metodologia di semplificazione riguarda sia il livello lessicale e sia il livello sintattico ed è applicata ad uno dei paragrafi del consenso informato.

Nel **capitolo 5** si discutono i risultati ottenuti e si propongono possibili sviluppi futuri.

1

La leggibilità di testi di dominio medico: stato dell'arte

Questo capitolo introduce il tema della valutazione automatica della leggibilità come passo preliminare alla semplificazione di testi scritti. In particolare si concentra sulla leggibilità di testi appartenenti ad un dominio specifico, quello medico.

Nel paragrafo 4.2 sono presentate le metriche principali utilizzate nella valutazione della leggibilità. Nel paragrafo 1.2 si descrivono i metodi principali adottati per i testi appartenenti al dominio medico. Infine nei paragrafi 1.3, 1.4 e 1.5 sono trattati i principali approcci presenti nello stato dell'arte per la lingua inglese, svedese e italiana.

1.1 La leggibilità

All'interno della società dell'informazione, dove tutti dovrebbero essere in grado di accedere alla vastissima quantità di informazioni disponibili, migliorare l'accessibilità alla lingua scritta è una questione centrale che non può essere trascurata.

È il caso delle informazioni amministrative, le quali dovrebbero essere accessibili a tutti i membri della società, anche agli individui che presentano scarse competenze linguistiche dovute ad un livello basso di educazione o a specifiche disabilità linguistiche, o perché i testi non sono scritti nella loro lingua madre. Un altro dominio cruciale è rappresentato dalle informazioni inerenti la salute, che dovrebbero essere accessibili ad un gruppo più largo ed eterogeneo di cittadini. La comprensibilità è anche una questione importante per accedere alle informazioni sul web, come indicato nelle *Web Content Accessibility Guidelines* (WCAG), proposte dal *Web Accessibility Initiative* del W3C[1], che contengono un'ampia gamma di metodi studiati per la creazione di siti usufruibili da ogni tipologia di utente. Nell'area educativa, invece, la valutazione della leggibilità rende possibile la personalizzazione dei materiali didattici tenendo conto delle competenze linguistiche dello studente.

La leggibilità è un concetto legato all'aspetto linguistico di un testo ed esprime la probabilità che esso risulti più o meno accessibile al lettore finale, data la presenza di parametri che vengono identificati dalla letteratura specialistica come spie di complessità a vari livelli di descrizione linguistica.

Le formule tradizionali per valutare la leggibilità avanzate a partire dagli anni '80 si focalizzano su un insieme di caratteristiche generali e formali del testo che vengono assunte come approssimazioni dei fattori linguistici in gioco nella valutazione della leggibilità.

Due misure tradizionali per valutare la leggibilità sono il *Flesch-Kincaid Reading Ease* e il *Flech-Kincaid Grade Level* [Flesch, 1949] che tengono conto della lunghezza media delle parole, misurate in sillabe, e della lunghezza media delle frasi, misurate in parole. Anche se queste due metriche usano le stesse caratteristiche, hanno fattori di peso differenti: i risultati ottenuti sono inversamente correlati in quanto un testo con un punteggio alto con la prima, dovrà averne uno basso con la seconda.

I valori ottenuti con il *Flesch-Kincaid Reading Ease* sono compresi in una scala da 0 a 100, dove 0 rappresenta un testo estremamente difficile, con frasi lunghe

[1] http://www.w3.org/TR/WCAG20/

mediamente 37 parole e parole composte mediamente da più di due sillabe, e 100 un testo estremamente facile, con frasi lunghe mediamente 12 parole o meno e parole composte mediamente da meno di due sillabe. La formula specifica per calcolare questo valore è:

$$Reading\ Ease = 206,835 - 1,015\left(\frac{totale\ parole}{totale\ frasi}\right) - 84,6\left(\frac{totale\ sillabe}{totale\ parole}\right)$$

La formula è stata testata su alcuni testi tipici ottenendo per esempio un valore di 92, quindi molto facile, su testi di genere fumettistico e un valore di 6, quindi molto difficile, per l'*Internal Revenue Code*, il codice tributario statunitense. La formula si basa sugli studi che Flesch ha condotto sulla psicologia umana, in particolare su come lavora la mente dell'uomo [Flesch, 1981].

Il *Flech-Kincaid Grade Level* è invece un indice che associa la leggibilità al numero di anni di istruzione necessari per comprendere un testo, rendendo più facile per gli insegnanti, i genitori e altri giudicare il livello di leggibilità dei diversi libri e testi. La formula specifica è:

$$Grade\ Level = 0,39 - \left(\frac{totale\ parole}{totale\ frasi}\right) + 11,8\left(\frac{totale\ sillabe}{totale\ parole}\right) - 15,59$$

La formula *Flesch-Kincaid Reading Ease*, creata e testata per la lingua inglese, è stata poi adattata per l'italiano con la scala Flesch-Vacca [Franchina and Vacca, 1986]:

$$Flesch\text{-}Vacca\ scale = 206 - (0,65 * sillabe_{nelle\_prime\_100\_parole}) - \left(\frac{parole_{in\_ogni\_frase}}{totale\ frasi}\right)$$

Un'altra misura tradizionale per la valutazione della leggibilità, tarata direttamente sulla lingua italiana, è data dall'indice Gulpease

[Lucisano and Piemontese, 1988], nato da un gruppo di linguisti dell'Università La Sapienza di Roma che nel 1987 si è riunito attorno a Tullio de Mauro per costituire il GULP (Gruppo Universitario Linguistico Pedagogico):

$$Gulpease = 89 - \left(\frac{totale\ caratteri * 100}{totale\ parole}\right) + 3 * \left(\frac{totale\ frasi * 100}{totale\ parole}\right)$$

I valori che si ottengono da questa formula sono, come per l'indice di Flesch, compresi in una scala che va 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. Uno dei principali vantaggi dell'indice Gulpease è dato dal calcolo della lunghezza delle parole in lettere e non più in sillabe, prestandosi bene ad essere automatizzato.

Un'altra misura di leggibilità usata per stimare gli anni di educazione necessari per capire un testo è l'indice SMOG [McLaughlin, 1969], acronimo di *Simple Measure of Gobbledygook* ("misurazione semplificata del linguaggio incomprensibile"), il quale è ampiamente utilizzato per il controllo di messaggi relativi alla salute. Si calcola prendendo a campione 30 frasi - di cui 10 all'inizio, 10 al centro e 10 alla fine del testo - e calcolando il rapporto tra parole lunghe (composte da almeno tre sillabe) e frasi, secondo determinati parametri numerici:

$$SMOG = 1,0430 * \sqrt{numero\ di\ polisillabi * \frac{30}{numero\ di\ frasi}} + 3,1291$$

Questi indici per il calcolo della leggibilità fanno parte dei cosidetti indici di leggibilità di "prima generazione" sviluppati a partire dalla seconda metà del Novecento grazie all'uso di strumenti oggettivi di misurazione della leggibilità. Tutti i metodi tradizionali per l'analisi della leggibilità sono facili da calcolare e usare ma presentano degli svantaggi. Per esempio, l'uso della lunghezza della frase come misura della

complessità sintattica presuppone che una frase più lunga è grammaticalmente più complessa rispetto ad una più corta, ma in realtà non è sempre così. Il conteggio delle sillabe delle parole viene usato partendo dal presupposto che le parole più frequenti hanno più probabilità di avere un numero minore di sillabe rispetto a quelle meno frequenti (un'associazione che è legata alla legge di Zipf, [Zipf, 1935]), ma la lunghezza della parola non necessariamente riflette la sua difficoltà.

Un'evoluzione delle metriche tradizionali della leggibilità è rappresentata dalle indagini avviate per la valutazione della difficoltà lessicale di un testo. Un primo passo è rappresentato dalle formule che fanno riferimento al vocabolario, come la formula Dale-Chall [Chall and Dale, 1995] e Lexile [Stenner, 1996] che usano una combinazione tra la frequenza con cui ricorrono le parole e la lunghezza media delle frasi in parole. In particolare la prima fornisce la percentuale delle parole che non sono presenti nella lista di 3000 parole "facili". Ma queste formule, nonostante rappresentino un passo in avanti nel valutare la leggibilità grazie alla disponibilità di dizionari di frequenza e corpora di riferimento, risultano ancora insoddisfacenti per quando riguarda la struttura della frase. Infatti in tante situazioni risulta necessario l'utilizzo di parole specifiche, e quindi considerate "difficili", per trasmettere il contenuto previsto e un ruolo centrale nella valutazione della leggibilità è quindi costituito dalla struttura della frase.

Con i progressi conseguiti nel settore del Trattamento Automatico del Linguaggio (TAL), negli ultimi anni sono andati affermandosi indici avanzati per la valutazione della leggibilità che rendono possibile monitorare una varietà più ampia di fattori linguistici che influenzano la leggibilità di un testo [Collins-Thompson, 2014]. Questo contesto ha favorito la nascita degli indici di leggibilità cosiddetti di "seconda generazione" che sono 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 di annotazione automatica del testo.

Esistono due diversi tipi di approccio alla valutazione automatica della leggibilità

di un testo [Dell'Orletta et al., 2014a]:

- **Valutazione della leggibilità come compito di classificazione**: assegnazione del documento analizzato ad una specifica classe di leggibilità.
- **Valutazione della leggibilità come compito di ranking**: assegnazione del documento analizzato ad una posizione all'interno di una scala di leggibilità.

La valutazione della leggibilità come compito di classificazione è l'approccio più utilizzato (per esempio in [Petersen and Ostendorf, 2009], [Aluisio et al., 2010], [Feng et al., 2010]) ma il suo problema principale è rappresentato dal fatto che richiede dati di addestramento che possono non essere disponibili, specialmente per un dominio specifico. Invece la valutazione della leggibilità come compito di ranking rappresenta un'alternativa valida al precedente in quanto richiede solamente dati di addestramento rispetto due livelli di leggibilità (facile-difficile). Questo approccio è utilizzato per esempio da [Inui et al., 2001], [Tanaka-Ishii et al., 2010], [Ma et al., 2012].

Gli approcci per la valutazione automatica della leggibilità si differenziano tra loro anche rispetto al tipo di caratteristiche linguistiche prese in considerazione (ad esempio lessicali, sintattici, semantiche). Alternative interessanti alle formule statiche che fanno riferimento al vocabolario prima citate sono state avanzate da [Si and Callan, 2001] che usano modelli del linguaggio ad unigrammi combinati alla lunghezza della frase per catturare informazioni del contenuto da pagine web scientifiche, o da [Collins-Thompson and Callan, 2004] che hanno adottato un modello del linguaggio simile per predire la difficoltà di lettura di brevi passaggi e documenti dal web. Invece il ruolo delle caratteristiche sintattiche è stato indagato per esempio da [Schwarm and Ostendorf, 2005] e [Heilman et al., 2007]: in questi studi la struttura sintattica è definita attraverso una combinazione di caratteristiche dai modelli del linguaggio a n-grammi (trigrammi, bigrammi e unigrammi) e dagli alberi sintattici (altezza dell'albero sintattico, numero di frasi nominali, frasi verbali e clausole subordinate) con caratteristiche più tradizionali.

Un altro fattore importante che determina la tipologia di caratteristiche da considerare è dato dai destinatari del testo sotto valutazione: la facilità di lettura non dipende solo dalle proprietà intrinseche del testo ma è anche influenzata dal pubblico di lettori previsto: per esempio gli studi di [Schwarm and Ostendorf, 2005], [Heilman et al., 2007] e [François and Fairon, 2012] si rivolgono a studenti stranieri o [Feng et al., 2009] si concentrano su persone con lieve disabilità intellettuale.

La maggior parte della ricerca si è focalizzata sulla valutazione della leggibilità a livello di documento rispetto che a livello di frase. Tuttavia, le metriche di valutazione della leggibilità a livello di frase risultano importanti per scopi applicativi specifici, come nel caso in cui il compito perseguito è la semplificazione automatica del testo.

1.2 La leggibilità di testi di dominio medico

Un dominio cruciale è rappresentato dalle informazioni inerenti la salute che dovrebbero essere accessibili a tutti i membri della società. I contenuti della comunicazione in ambito medico sono di grande importanza in quanto 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. Anche la comunicazione tra medico e paziente costituisce un tassello fondamentale nella pratica medica e ciò a cui si punta è rendere i medici non solo dei professionisti sul campo ma anche esperti della comunicazione.

La comunicazione inefficiente tra medico e paziente è uno dei fattori che porta inevitabilmente al verificarsi di incidenti ed errori. Con l'obiettivo di prevenire il verificarsi di un errore e, qualora questo accada, contenerne le conseguenze, i servizi sanitari si occupano della progettazione di specifici modelli di controllo del rischio clinico. Recenti evidenze dalla ricerca sostenuta dall'Organizzazione Mondiale della Sanità (OMS) suggeriscono che gli eventi avversi sono una delle principali cause di disabilità e morte in tutto il mondo, specialmente tra le persone che vivono in paesi

a basso e medio reddito. Sono stati stimati in tutto il mondo 421 milioni di ricoveri ogni anno e circa 42,7 milioni di eventi avversi, con più del 50% di eventi prevenibili [Jha et al., 2013]. In Italia l'incidenza media di eventi avversi in un campione rappresentativo di pazienti ricoverati in cinque grandi ospedali italiani ubicati al Nord, al Centro e al Sud del Paese è stata determinata del 5,2%, per un totale di eventi prevenibili a livello nazionale pari al 56,7% [Tartaglia et al., 2012]. Per questi motivi, la comunità medica ha sempre mostrato un forte interesse per il miglioramento delle informazioni inerenti la salute in termini di qualità e comprensibilità dei documenti.

Nel dominio medico sono stati proposti meno approcci basati sull'uso di tecniche di NLP (*Natural Language Processing*) per la valutazione della leggibilità di testi relativi alla salute.

Un approccio alla valutazione della leggibilità di testi medico come compito di classificazione è riportato in [Kauchak et al., 2014], in cui gli autori presentano un classificatore addestrato su un corpus parallelo costituito da coppie di frasi originali/semplificate dall'*English Wikipedia*[2] e dal *Simple English Wikipedia*[3] in grado di predire la difficoltà dei testi relativi alla salute.

Invece un approccio alla valutazione della leggibilità come compito di ranking è presentato per esempio da [Kim et al., 2007], i quali hanno sviluppato un punteggio di distanza calcolato in base a come le caratteristiche del testo di un documento differiscono da quelle di un campione facile, costituito da testi medici raccolti da diverse risorse online.

Uno degli scopi applicativi specifici per il quale viene affrontata la valutazione della leggibilità nel dominio medico è la semplificazione dei testi medici, condotta attraverso la creazione di metodi che possono aiutare a rendere i documenti relativi alla salute più comprensibili. Questi metodi fanno riferimento a caratteristiche lessicali e sintattiche del testo.

[2]http://en.wikipedia.org
[3]http://simple.wikipedia.org

Grazie al ruolo centrale delle caratteristiche lessicali nel determinare la leggibilità dei testi inerenti la salute, la semplificazione lessicale costituisce il livello più esplorato nella semplificazione del testo. Sono stati elaborati diversi approcci per rendere questi testi più comprensibili, riducendo la difficoltà del vocabolario. Anche se con alcune differenze, tutti gli approcci si basano sull'identificazione delle parole difficili e la loro sostituzione con sinonimi più facili. Per questo scopo sono stati usate diverse risorse, sia specifiche del dominio, come l'*Unified Medical Language System* (UMLS) [Bodenreider, 2004] e l'*Open Access and Collaborative Consumer Health Vocabulary* (OAC CHV)[4], o generali come i sinonimi e iperonimi di WordNet [Miller, 1995] o il conteggio delle frequenze delle parole nel Google Web Corpus [Brants and Franz, 2006].

Per quanto riguarda le caratteristiche sintattiche, è interessante notare che quelle prese in considerazione tipicamente non vanno oltre la distribuzione delle parti delle parti del discorso (PoS) e/o sintagmi nominali. Nessuno dei metodi specifici del dominio proposti finora fa uso di caratteristiche sintattiche che possono essere estratte dall'output di un parser sintattico.

1.3 Gli approcci per la lingua inglese

Per quanto riguarda la lingua inglese, sono stati sviluppati più metodi per la valutazione della leggibilità e la semplificazione dei testi medici rispetto alle altre lingue. Questo è dovuto principalmente alla maggiore disponibilità di risorse sfruttabili per la creazione di un sistema capace di semplificare materiali di informazione relativa alla salute.

1.3.1 L'approccio di Leroy

Un approccio molto interessante è quello presentato in [Leroy et al., 2012]. Gli autori hanno sviluppato un algoritmo semi-automatico per la semplificazione di te-

[4]http://consumerhealthvocab.chpc.utah.edu/CHVwiki/

sti di ambito medico, in grado di poter aiutare gli scrittori a produrre documenti facilmente comprensibili.

La prima fase del lavoro è stata costituita dall'identificazione automatica delle caratteristiche linguistiche rappresentative di un testo difficile attraverso strumenti di elaborazione del linguaggio naturale. La seconda fase invece è stata costituita dalla presentazione e dalla valutazione di alternative più semplici che possono essere usate come sostituzione al testo difficile.

Per identificare le caratteristiche linguistiche che differenziano i testi facili da quelli difficili, sono stati confrontati corpora composti da testi facili e difficili. Sono stati condotti due tipi di analisi: lessicale e grammaticale. L'analisi lessicale ha avuto il compito di definire quanto risultano difficili le parole per qualsiasi tipo di lettore utilizzando la metrica della "familiarità del termine", stimata usando le frequenze delle parole nel Google Web Corpus. Quest'ultima rappresenta una stima della conoscenza che un lettore inesperto ha con un dato termine. L'analisi grammaticale invece ha identificato differenze significative nell'occorrenza di parti del discorso differenti: i testi semplici contengono più parole funzionali, avverbi e verbi, al contrario dei testi difficili che contengono più nomi e aggettivi [Leroy and Endicott, 2012].

L'algoritmo per la semplificazione del testo basato sulla familiarità del termine è costituito da tre passaggi:

1. Identificazione automatica delle parole difficili
2. Recupero automatico e ordinamento delle alternative più facili
3. Scelta di una delle alternative per sostituire la parola o il segmento del testo difficile

L'identificazione automatica delle parole difficili si basa sui dati raccolti dalle frequenze degli unigrammi all'interno del Google Web Corpus. È stata scelta come soglia per distinguere le parole difficili da quelle facili la frequenza di 15.377.914, la

frequenza della 5.000^{esima} parola più comune del corpus. Sulla base di questa soglia, circa l'85% di tutte le parole in inglese sono state classificate come facili.

Per ogni parola etichettata come difficile viene poi creata e ordinata una lista di potenziali sostituzioni. Le sostituzioni possono essere singole parole o segmenti di testo e sono recuperati da fonti diverse: sinonimi e iperonimi da *WordNet* 2.1 [Miller et al., 1990], tipi semantici dall' *Unified Medical Language System* (UMLS) [Bodenreider, 2004], definizioni dall'*English Wiktionary*[5] e dal *Simple English Wiktionary*[6]. Le alternative sono limitate basandosi sulla PoS della parola difficile nel testo originale e sulla familiarità del termine, questo per assicurarsi che la parola difficile non sia sostituita da un'alternativa altrettanto difficile.

Le possibili sostituzioni estratte da *WordNet* sono costituite da sinonimi ed iperonimi della parola etichettata come difficile. Sono selezionati i sinonimi e gli iperonimi diretti con il valore di familiarità più alto per ogni synset della PoS della parola difficile. In particolare gli iperonimi vengono aggiunti alla lista delle possibili sostituzioni in questo modo: "<parola difficile>, a kind of <iperonimo> "; "<parola difficile>, kinds of <iperonimo> "; o "<parola difficile>, a way of <iperonimo>".

Dall'UMLS sono estratti invece i tipi semantici che soddisfano il requisito della familiarità e vengono aggiunti alla lista come: "<parola difficile>, a kind of <tipo semantico>".

Infine sono estratte le definizioni da *WordNet*, dall'*English* e *Simple English Wiktionary*. In particolare queste ultime due risorse sono usate per evitare che l'algoritmo sostituisca solamente i termini medici. Inoltre, le definizioni vengono troncate alla prima occorrenza di una virgola, o di "for example", "as in", "expecially" o "such as". Come nei casi precedenti vengono selezionate le definizioni che soddisfano il requisito della familiarità. Le definizioni dal *Simple English Wiktionary* sono aggiunte senza nessuna modifica, invece quelle da *WordNet* e dall'*English Wiktionary* sono modificate in una delle seguenti forme: "<parola difficile> is a

[5]http://en.wiktionary.org
[6]http://simple.wiktionary.org

<definizione>"; "When something is <parola difficile>, it is a <definizione>"; o "Something that is <parola difficile>, is <definizione>".

L'algoritmo è stato testato su materiale medico disponibile on-line della biblioteca medica di San Diego, Università della California (UCSD), misurando il suo impatto su due tipi differenti di difficoltà: *percepita* ed *effettiva*. La prima misura il livello di difficoltà di un testo a prima vista e la seconda il livello di comprensibilità concreta e la memorizzazione delle informazioni contenute nel testo.

Lo studio è stato condotto su un'ampia gamma di utenti online, reclutati da Amazon Mechanical Turk (AMT)[7], un servizio internet di crowdsourcing in cui i partecipanti svolgono obiettivi conosciuti come HIT (Human Intelligence Tasks) in cambio di un piccolo pagamento.

La prima parte dello studio si è concentrata sulla misurazione della difficoltà percepita del testo. Sono state selezionate 12 frasi, le quali sono state processate dall'algoritmo e semplificate da un bibliotecario medico che ha scelto una tra le alternative fornite. Per un confronto con i lavori esistenti, è stata aggiunta allo studio un'altra variabile, il *Flesch-Kincaid grade level* [Flesch, 1949]. Sono state selezionate dodici frasi abbastanza diverse tra loro, sei con un alto e sei con un basso *Flesch-Kincaid grade level*. Ai partecipanti sono state mostrate le due versioni di tutte le frasi (originale e semplificata) ed è stato chiesto loro di giudicarle su una scala Likert, dove un punteggio di 1 rappresenta una frase molto facile e un punteggio di 5 una frase molto difficile. Le frasi originali e le frasi semplificate hanno ricevuto una punteggio medio complessivo rispettivamente di 2,56 e 2,44. Nonostante i risultati non mostrino dei grandi cambiamenti, le differenze maggiori si riscontrano per le frasi che hanno un *Flesch-Kincaid readability grade level* alto.

La seconda parte dello studio si è invece concentrata sulla misurazione della difficoltà effettiva del testo, per valutarne sia la comprensibilità e sia la memorizzazione delle informazioni. Sono stati scelti due documenti, uno sulla policitemia vera e uno sul pemfigo, entrambi paragonabili a livello di lunghezza e difficoltà. Come nel caso

[7]http://www.mturk.com/mturk/welcome

precedente, ciascun testo è stato semplificato da un bibliotecario medico utilizzando l'algoritmo di semplificazione.

In particolare, per misurare la comprensibilità è stato chiesto ai partecipanti di rispondere a domande a risposta multipla con quattro possibili risposte affiancate al testo e ognuna delle frasi ha ricevuto un punteggio di 1 se la risposta è corretta e 0 se incorretta. Invece per misurare la memorizzazione dell'informazione è stato chiesto ai partecipanti di rispondere a 30 domande senza il testo: le risposte che potevano dare erano "Vero", "Falso" o "Non lo so" e in questo caso è stato assegnato un punteggio di 1 se la risposta è corretta, -1 se incorretta e 0 per "Non lo so". Infine i punteggi finali sono stati riformulati e presentati su una scala da 0 a 100. Nella valutazione della comprensione delle informazioni la percentuale di risposte corrette per i testi originali e semplificati date dai partecipanti è stata rispettivamente del 51% e 58%. Infine, per quanto riguarda la memorizzazione dell'informazione i partecipanti hanno raggiunto il 73% di risposte corrette con il documento semplificato e 71% con quello originale.

I risultati mostrano che la difficoltà percepita dei testi semplificati è minore rispetto a quella dei testi originali. Anche i risultati che riguardano la comprensibilità e la memorizzazione delle informazioni migliorano con il passaggio dai testi originali a quelli semplificati. Quindi l'algoritmo basato sulla familiarità del termine presentato dagli autori dimostra di essere in grado di supportare gli scrittori nella semplificazione del testo.

1.3.2 L'approccio di Kandula

Un altro strumento per la semplificazione di documenti medici è presentato in [Kandula et al., 2010], il quale si rivolge sia alla difficoltà semantica e sia a quella sintattica.

Questo strumento di semplificazione estende quello descritto nell'articolo [Zeng-Treitler et al., 2007], in grado di identificare i termini difficili e sostituirli con

sinonimi o spiegandoli usando termini correlati più facili, riportando risultati corretti nel 68% dei casi.

L'identificazione dei termini difficili è anche in questo caso la prima fase nella semplificazione del testo. Gli autori hanno proposto una tecnica basata sulla frequenza per stimare la difficoltà dei termini, partendo dall'osservazione che termini che occorrono più frequentemente in fonti biomediche mirate a lettori inesperti, come *Reuters News*[8] o su *MedlinePlus*[9], tendono ad essere più facili [Zeng et al., 2005]. Un altro approccio proposto dagli autori stimava la difficoltà del termine in base ai suoi contesti d'uso [Zeng-Treitler et al., 2008]. Una combinazione di queste due tecniche è stata usata per ottenere un singolo punteggio nell'intervallo tra 0 (molto difficile) ed 1 (molto facile), il punteggio di familiarità. Come in [Leroy et al., 2012], tutti i termini con un punteggio di familiarità minore rispetto a una data soglia sono stati considerati dfficili e necessitano quindi della semplificazione.

I termini difficili sono stati semplificati attraverso la sostituzione con un sinonimo più facile, estratto dall'*Open Access and Collaborative Consumer Health Vocabulary* (OAC CHV). Nel caso in cui non siano stati trovati sinonimi adatti, sono state generate delle spiegazioni, delle frasi che rendono il termine più comprensibile. La spiegazione è differente dalla definizione: la definizione si concentra sulla descrizione semantica precisa e completa di un termine e può quindi usare termini difficili ed essere abbastanza lunga, mentre la spiegazione usa termini più facili ed è molto più breve. Per generare automaticamente le spiegazioni, lo strumento estraeva un insieme di parole gerarchicamente e/o semanticamente correlate al termine originale, selezionava tra quelle che soddisfano il criterio della soglia del punteggio di familiarità la più facile e creava una frase corta per descrivere la relazione tra quest'ultima e la parola difficile. Per generare l'insieme delle parole relazionate sono stati estratti i sinonimi e le relazioni gerarchiche definite nell'*Unified Medical Language System* (UMLS) [Bodenreider, 2004]. Una revisione ha però mostrato che il 32% delle spie-

[8]http://www.reuters.com/news
[9]http://www.nlm.nih.gov/medlineplus

gazioni ricavate erano inutili o incorrette, a causa dell'uso di relazioni gerarchiche non applicabili.

Il nuovo strumento presentato usa un set più ampio di relazioni per la generazione della spiegazione. Sono state analizzate manualmente le spiegazioni contenute in un set di 150 documenti relativi al diabete e identificate le relazioni chiave usate per spiegare 5 gruppi semantici comuni di concetti relativi alla salute ("disease name", "anatomical structure", "device", "procedure" e "medication"), in quanto i tipi di relazione tra il termine difficile e i termini della spiegazione più facile sono molto spesso dipendenti dal tipo semantico del termine difficile. Per esempio è molto comune spiegare una malattia con i sintomi comuni osservati o facendo riferimento alla specifica parte del corpo affetta. Quindi è stato modificato l'algoritmo per la generazione della spiegazione in modo che prenda in considerazione il tipo semantico del termine difficile.

Il nuovo strumento comprende anche una componente sintattica. Gli autori hanno focalizzato la loro attenzione sull'identificazione e la semplificazione delle frasi complesse e composte, per esempio quelle che sono costituite da una clausola dipendente e una indipendente.

La semplificazione sintattica è portata avanti al livello della frase: gli autori hanno assunto che frasi lunghe più di 10 parole richiedano la semplificazione sintattica e sono state processate attraverso una serie di moduli. Alla fine della semplificazione la frase originale può restare invariata o essere sostituita da due o più frasi più corte e quindi presumibilmente più facili.

I moduli usati sono il Part-of-Speech tagger, il semplificatore grammaticale e il validatore dell'output. Con l'analisi da parte del Part-of-Speech Tagger, le frasi costituite da più di 10 parole vengono tokenizzate e ad ogni token viene associata la corretta PoS: è stato usato il software open-source OpenNLP[10]. Il semplificatore grammaticale, basato su quello proposto in [Siddharthan, 2006], spezza la frase lunga in due o più frasi più corte applicando un set di regole trasformazionali. In-

[10]Disponibile su http://opennlp.apache.org/

fine il validatore ha il compito di controllare l'output per evitare semplificazioni sgrammaticate o frammentate.

Questo strumento può essere potenzialmente usato per semplificare qualsiasi tipo di testo medico. In questo lavoro è stato testato su due set di documenti medici: 40 cartelle cliniche elettroniche e 40 articoli di riviste biomediche disponibili su *PubMed*, database della letteratura scientifica biomedica. Per questi due set di documenti, la leggibilità dei testi originali e semplificati è stata assegnata usando le formule di leggibilità tradizionali, il *Flesch kincaid Grade Level* [Flesch, 1949] e il *Simple of Measure of Gobbledygook* (SMOG) [McLaughlin, 1969]. Dato che entrambe queste misure sono limitate a caratteristiche formali per la valutazione della leggibilità, ne sono state riportate anche altre due: la difficoltà semantica e la coesione. La difficoltà semantica è calcolata come la media del punteggio di familiarità di tutti i termini del documento. La coesione è invece calcolata come il rapporto tra il numero di parole alla loro forma radice che si sovrappongono tra frasi adiacenti e il numero totale di termini distinti.

I risultati hanno mostrato che la leggibilità delle cartelle cliniche elettroniche è bassa a causa di un uso esteso di termini medici e abbreviazioni, frasi corte e sgrammaticate, e una bassa coesione. Invece gli articoli delle riviste sono generalmente più coesi ma usano termini difficili e frasi eccessivamente lunghe con strutture sintattiche complesse.

Dopo la semplificazione gli articoli delle riviste biomediche hanno mostrato un miglioramento nella leggibilità in tutte e 4 le misure. I valori di FKGL e SMOG sono scesi marginalmente (rispettivamente da 15,83 a 15,58 e da 17,27 a 17,08), il numero medio di sillabe per parola è diminuito (da 1,83 a 1,80) ma il numero medio di parole per frase è leggermente aumentato (da 25,23 a 25,52) a causa dell'aggiunta delle spiegazioni, invece il punteggio di familiarità e la coesione sono migliorate (rispettivamente da 0,75 a 0,77 e da 0,66 a 0,71). I risultati sulle cartelle cliniche elettroniche semplificate hanno mostrato miglioramenti nel punteggio della familiarità coerenti con quelli dei risultati precedenti (da 0,73 a 0,76) ma pur sempre una

bassa coesione nonostante un aumento marginale (da 0,09 a 0,10), i valori di FKGL e SMOG invece sono aumentati (rispettivamente da 8,31 a 9,23 e da 11,31 a 12,27), il numero medio di parole per frase è aumentato significativamente (da 10,62 e 13,30) mentre il numero medio di sillabe per parole è rimasto quasi invariato (da 1,67 a 1,66).

In aggiunta a questi 80 documenti, gli autori hanno condotto un'analisi dettagliata della semplificazione su altre 9 cartelle cliniche elettroniche che sono state usate per testare la versione precedente dello strumento. È stato eseguito un cloze test[11], su 8 dei 9 documenti usando 4 revisori. Ogni quinta parola delle prime 250 parole di ogni documento è stata sostituita da uno spazio vuoto che doveva essere riempito dai revisori. Ad ogni revisore sono stati assegnati 8 documenti (4 prima e 4 dopo la semplificazione) e ogni documento è stato rivisto da due revisori. Assumendo che una percentuale alta di risposte indichi la presenza di un testo più semplice, i risultati hanno mostrato che il punteggio medio è aumentato da 35,8% nei testi originali a 43,6% nei testi semplificati.

Nonostante i risultati mostrino dei miglioramenti, restano comunque abbastanza scarsi. Questo è dovuto principalmente alla presenza di poche istanze di semplificazione sintattica rispetto al numero di spiegazioni che invece sono state aggiunte, le quali hanno portato necessariamente all'aumento della lunghezza media della frase e quindi anche ai valori di FKGL e SMOG. Aggiungere spiegazioni come frasi distinte può diminuire questi ultimi due valori ma non sempre ciò può essere appropriato, specialmente in certi contesti. Per quanto riguarda i risultati del cloze test, nonostante un miglioramento significativo i risultati non si trovano ancora nell'intervallo del punteggio desiderabile (almeno 50%-60%).

[11]Il cloze test è una prova di accertamento della capacità di comprensione di testi scritti e consiste nella ricostruzione di un brano tramite il reinserimento di alcune parole precedentemente cancellate. Originariamente proposto per valutare il grado di "difficoltà" di testi scritti in inglese, il cloze test è ora utilizzato per accertare abilità di comprensione nello studio delle lingue straniere.

1.3.3 L'approccio di Peng

In [Peng et al., 2012] gli autori propongono un approccio alternativo per rilevare ed estrarre informazione da frasi complesse presenti negli abstract dei testi biomedici, in modo che possano essere facilmente processati da applicazioni di text mining. Molte di queste applicazioni rilevano le informazioni nel testo basandosi su alcuni modelli comuni affidabili, ma la progettazione di regole per tutte le variazioni sintattiche è quasi impossibile per il fatto che le costruzioni della frase e gli stili di scrittura variano considerevolmente da un autore all'altro e da una pubblicazione all'altra.

Gli autori, invece di abbinare tutte le possibili variazioni del testo, propongono per prima cosa la semplificazione delle frasi complesse per poi tentare di farli corrispondere ai modelli ordinari. È stato quindi creato iSimp, un semplificatore che ha l'obiettivo di ridurre la complessità sintattica delle frasi, attraverso il rilevamento di diversi costrutti della frase e la loro trasformazione in un formato che è facilmente accessibile agli strumenti di text mining. Esistono diversi tipi di semplificazione che riguardano sia obiettivi *machine-oriented* che *human-oriented*. In questo caso l'applicazione della semplificazione del testo si inserisce nel secondo filone e rappresenta quindi un passo di pre-elaborazione per migliorare l'efficienza di altri obiettivi di NLP.

I costrutti sintattici per la semplificazione trattati sono:

- **Coordinazione**: strutture sintattiche complesse che collegano insieme più elementi in una proposizione o più proposizioni nel periodo attraverso l'uso di congiunzioni coordinative ("and", "or" e "but").

- **Clausole relative**: clausole che modificano le frasi nominali. Se ne distinguono due tipi, quelle complete introdotte da pronomi relativi ("which", "who" e "that") e ridotte che iniziano con un gerundio/participio passato e non hanno il soggetto esplicito.

- **Apposizioni**: costrutti di due frasi nominali una vicina all'altra, tipicamente separate da una virgola e che si riferiscono alle stesse entità.

La metodologia proposta è costituita da tre fasi: pre-elaborazione della frase originale, rilevamento dei costrutti per la semplificazione e la generazione delle frasi semplificate.

Nella pre-elaborazione della frase originale, è usato un PoS tagger per determinare la categoria linguistica corrispondente per ogni parola della frase (nomi, verbi, aggettivi, preposizioni, ecc.). Dopo è usato un parser superficiale per rilevare i chunk, in particolare ne vengono trattati tre tipi: frasi nominali (NP), gruppi verbali (VG) e frasi preposizionali (PP). Le altre parole non incluse in questi tre tipi sono marcati con "altro" (O).

Nella seconda fase, per il rilevamento di ogni costrutto per la semplificazione, chiamato C, viene costruita una macchina a stati finiti M_c. Per ogni macchina sono possibili tre risultati: 1) *"successo"* quando rileva il costrutto; 2) *"fallimento"* quando non rileva il costrutto; e 3) *"in attesa"* quando la macchina sospetta che un'altra struttura è annidata all'interno del costrutto e quindi deve essere rilevata per prima.

Infine il semplificatore genera frasi separate per ogni costrutto rilevato. Per la coordinazione, la frase originale può essere divisa in più frasi, ciascuna delle quali contiene un sintagma o una proposizione. Anche se le frasi semplificate non sembrano necessariamente più semplici, la divisione rende meno complessi i compiti di estrazione dell'informazione. Una frase contenente una clausola relativa può essere semplificata generando due frasi, una contenente la clausola relativa e un'altra che combina la frase nominale con la clausola relativa. Le apposizioni possono essere semplificate allo stesso modo delle clausole relative.

Il semplificatore è stato valutato usando un corpus annotato manualmente, costituito da 100 abstract (per un totale di 954 frasi) estratti da *Medline* che contengono le parole "protein" e "gene" nel titolo. I risultati hanno mostrato che iSimp è in

grado di riconoscere i tre tipi di costrutti sintattici presi in considerazione con una media di F-measure tra l'86,5% e il 92,7%.

1.4 Gli approcci per la lingua svedese

Un'eccezione per le metriche di leggibilità per i testi medici non tarate per la lingua inglese è rappresentata dallo svedese. Così come per l'inglese, gli algoritmi di semplificazione di testi medici svedesi sono stati concepiti facendo riferimento a metodi di sostituzione delle parole difficili con un sinonimo più facile, oppure nel rilevamento automatico di parole e abbreviazioni non presenti nel vocabolario o sulla divisione delle parole composte e sulla correzione dell'ortografia.

1.4.1 L'approccio di Kvist

Il passo preliminare per lo sviluppo di uno strumento di semplificazione di testi medici è presentato in [Kvist and Velupillai, 2013], in cui è presentata un'analisi quantitativa e qualitativa di un corpus contenente una collezione di rapporti radiologici.

I rapporti radiologici analizzati dagli autori fanno parte dello Stockholm EPR Corpus [Dalianis et al., 2012], un ampio corpus di cartelle cliniche elettroniche di più di 600.000 pazienti nell'area metropolitana di Stoccolma durante gli anni 2006-2010. Sono stati estratti i rapporti degli esami eseguiti durante gli anni 2009-2010 presso l'Ospedale Universitario Karolinska, per un totale di 434.427 documenti contenenti sia il testo del rinvio e sia il risultato radiologico.

Per l'analisi statistica quantitativa del corpus, gli autori hanno estratto, attraverso l'uso della libreria NLTK [Bird et al., 2009], un numero di categorie differenti: frasi, parole, bigrammi e trigrammi. Inoltre hanno estratto tutti i nomi, i verbi e gli aggettivi attraverso l'uso del Part-of-Speech tagger descritto in [Stagger, 2012], allenato per lo svedese. Per ogni categoria è stata infine calcolata la frequenza dei tipi e dei token.

I risultati hanno mostrato che la somma delle prime 100 frasi rappresenta il 7.8% del totale delle frasi nel corpus. Tuttavia, il vocabolario è ricorrente, infatti la somma delle prime 100 parole più frequenti rappresenta il 35% del totale delle parole usate nel corpus, e il 16% di tutti i bigrammi sono stati trovati tra i primi 100 più frequenti. Con riferimento alle tre classi di parole, le proporzioni sono ancora più alte. I risultati hanno mostrato che il 53% dei nomi, il 79% dei verbi e il 74% degli aggettivi si trovano rispettivamente tra i primi 100 più frequenti.

È stato osservato che in media un rapporto radiologico consiste di 5 frasi (min=1, max=66), la cui lunghezza media è di 12 parole. Molte frasi non sono complete, infatti tra le prime 100 solo 23 contengono sia il soggetto e sia il predicato. Inoltre, 7 di esse sono composte da una singola parola e 30 da due parole. Tra le frasi più lunghe, molte sono frasi standard di carattere amministrativo.

Infine è stata effettuata un'analisi qualitativa sui 100 elementi più frequenti per ogni categoria. La struttura dei rapporti radiologici è solitamente coerente: prima un'intestazione o una frase che descrive la procedura o il metodo usato e quale parte del corpo viene esaminata, e a seguire una descrizione di ciò che è stato visto nella figura radiologica durante l'esame, un'interpretazione dei risultati e una diagnosi. L'informazione amministrativa, come le date e i nomi dei medici in carica, si trovano solitamente alla fine del documento.

Le prime 100 frasi più frequenti contengono principalmente informazioni riguardo a risultati, parti del corpo, procedure e informazioni amministrative. In particolare parole e frasi amministrative sono presenti in 20 tra le prime 100 parole e in 16 tra le prime 100 frasi più frequenti. Inoltre in almeno metà delle prime 100 frasi sono menzionate parti del corpo.

Le parole straniere sono prevalentemente latine, greche ed inglesi, usate per indicare parti del corpo, posizioni e procedure. Ben 18 parole tra le prime 100 più frequenti sono abbreviazioni, di cui 7 sono comuni (per esempio tel=telefono e cm=centimetro), 10 sono specifiche del dominio (per esempio iv=intravenoso) e una ambigua (ca=cancro o circa).

Per quanto riguarda invece le classi di parole, comunemente i nomi vengono usati per descrivere risultati, parti del corpo e informazioni amministrative, gli aggettivi per descrizioni dei risultati (come posizioni e misure) e i verbi per i risultati (in particolare tra i primi 100 verbi più frequenti, 70 sono attivi e 30 passivi).

Con l'analisi quantitativa e qualitativa di questo corpus è quindi possibile capire meglio il contenuto dei rapporti radiologici e utilizzarla nello sviluppo di uno strumento di semplificazione in grado di creare testi medici più comprensibili per i pazienti.

1.4.2 L'approccio di Abrahamsson

In [Abrahamsson et al., 2014], gli autori presentano un algoritmo di semplificazione di testi medici basato sul metodo di sostituzione delle parole difficili con un sinonimo più facile, ma adattato alla natura composta della lingua svedese: la difficoltà di una parola non è assegnata misurando solamente la sua frequenza all'interno di un corpus generale, ma anche quella delle sottostringhe di cui è composta.

La semplificazione è stata studiata su testi di riviste mediche. In particolare, è stato usato un sottoinsieme della rivista Läkartidningen, la Rivista dell'Associazione Medica Svedese [Kokkinakis, 2012], costituita da 10.000 frasi selezionate casualmente da numeri pubblicati nel 1996. Come risorsa lessicale è stata usata la versione svedese di MeSH, vocabolario ideato con l'obiettivo di indicizzare la letteratura in ambito biomedico.

Per le statistiche di frequenza delle parole è stato usato il corpus *Parole*, contenente circa 19 milioni di token. Per ogni parola in Läkartidningen, l'algoritmo controlla se su MeSH è presente un sinonimo e se quest'ultimo ha una frequenza maggiore rispetto all'originale all'interno del corpus *Parole*. Se le due condizioni sono verificate, avviene la sostituzione. Nello svedese generale molte parole mediche occorrono raramente, ma la maggior parte di esse, essendo tipicamente descrittive, consistono di composti di parole usate nel linguaggio giornaliero, presenti frequentemente nel corpus. Per gestire questi casi, l'algoritmo è stato modificato basandosi

sull'idea che una parola composta da elementi che occorrono nel linguaggio standard è più facile da capire rispetto ad una parola rara. Quindi l'algoritmo, quando sia la parola originale e sia il sinonimo non occorrono nel corpus, conduce una ricerca per le sottostringhe di cui sono composte. La parola originale è sostituita dal sinonimo nel caso in cui consiste di un ampio numero di sottostringhe presenti in *Parole* rispetto a quelle della parola originale. Inoltre, per assicurarsi che le sottostringhe siano parole rilevanti, devono consistere di almeno quattro caratteri.

Per la valutazione dell'effetto della sostituzione sono state usate due metriche, sia sul testo originale e sia su quello modificato: LIX (misura di leggibilità) è la metrica standard usata per misurare la leggibilità dei testi svedesi e OVIX (indice di variazione delle parole) è la misura della varianza lessicale [Falkenjack et al., 2013]. Valori di LIX tra 25-30 e di OVIX tra 60-69 sono stati riscontrati in testi facili.

Secondo la metrica LIX, dopo la semplificazione il testo medico diventa leggermente più difficile (da 50 a 51) ma secondo la metrica OVIX diventa leggermente più facile (da 87,2 a 86,9). Dato che la metrica LIX prende il considerazione il numero di parole lunghe nel testo (più di 6 caratteri), una spiegazione plausibile per il suo aumento è che le parole corte derivate dal greco e dal latino siano state sostituite da parole composte più lunghe e che le abbreviazioni e gli acronimi siano stati estesi. Questo ha portato anche all'aumento dei tipi delle parole, riflettendosi su una diminuizione del valore di OVIX.

Per ottenere risultati da metodi non automatici, è stata condotta anche una piccola valutazione manuale della correttezza e della leggibilità percepita. Un sottoinsieme delle frasi selezionato casualmente nel quale almeno un termine è stato sostituito, è stato classificato da un medico in base a tre categorie: 1) il significato originale è stato mantenuto dopo la sostituzione; 2) il significato originale è stato alterato leggermente dopo la sostituzione, e 3) il significato originale è stato alterato significativamente dopo la sostituzione. Per la valutazione della leggibilità percepita, le frasi appartenenti alla prima categoria sono state poi ulteriormente classificate da altri due valutatori, laureati in discipline non biologiche: 1) entrambe le frasi

sono facili/difficili da capire, o 2) una delle due frasi è più facile da capire rispetto all'altra. Nel secondo caso il valutatore doveva indicare quale delle due frasi era più facile.

I risultati hanno mostrato che il significato semantico originale è stato leggermente alterato in almeno un terzo delle frasi, indicando che il set di sinonimi di MeSH ha bisogno di essere adattato per il compito di semplificazione del testo. Tre tipi di potenziali problemi sono: 1) sinonimi distanti, 2) i termini non sono sempre scritti in maniera appropriata ma hanno bisogno di essere trasformati in un altro formato prima di essere utilizzati, e 3) anche se le abbreviazioni sono state ampliate nella forma corretta, molte di esse all'interno del dominio medico posseggono più significati diversi.

1.4.3 L'approccio di Grygonit

Un altro approccio per la lingua svedese è presentato in [Grigonyte et al., 2014], nel quale gli autori presentano uno strumento per la semplificazione lessicale testato su un dataset di cartelle cliniche elettroniche (EHR).

Una cartella clinica elettronica contiene documentazione sistematica della storia medica di un singolo paziente durante il tempo, inserita e organizzata da operatori sanitari con lo scopo di consentire una cura sicura e consapevole. Essa è costituita sia da parti strutturate (come dettagli riguardo il paziente, risultati di laboratorio, codici diagnostici, ecc.) e parti non strutturate (nel formato di testo libero). Linguisticamente, le cartelle cliniche rappresentano un dominio altamente specializzato, costituite da frasi telegrafiche che coinvolgono parole dislocate o mancanti, abbreviazioni abbondanti ed errori di ortografia.

L'approccio proposto è costituito da tre fasi: in primo luogo tutte le abbreviazioni conosciute sono riconosciute e marcate; in secondo luogo vengono controllate le parole sconosciute per vedere se sono dei composti; infine per le parole sconosciute rimaste vengono fatte delle correzioni dipendenti dal contesto.

Per il rilevamento delle abbreviazioni è stato usato lo Swedish Cinical Terminology Matcher (SCATM), basato su SCAN [Isenius et al., 2012], il quale usa diversi lessici di termini medici, abbreviazioni, parole e nomi svedesi. Per la divisione dei composti, è stata usata una collezione di risorse lessicali generali e specifiche del dominio medico come la terminologia SNOMED CT svedese. Infine il rilevamento degli errori di ortografia è stato basato sulle pratiche usate nei correttori ortografici per le lingue indo-europee: similarità fonetica combinata con un metodo per misurare la vicinanza tra le stringhe.

L'approccio è stato testato su un dataset composto da un sottoinsieme casuale di 100 note giornaliere di diverse cartelle cliniche nel periodo 2009-2010 presenti all'interno dello Stockholm EPR Corpus [Dalianis et al., 2012]: un medico ha annotato le abbreviazioni, le parole composte e gli errori di ortografia.

I risultati hanno mostrato una precisione del 91,1% nel rilevamento delle abbreviazioni, dell'85,5% nella divisione delle parole composte e dell'83,87% nella correzione degli errori dell'ortografia.

1.5 Gli approcci per la lingua italiana

Le iniziative sviluppate finora per quanto riguarda la lingua italiana sono basate sulle formule tradizionali della leggibilità. Questo è il caso del progetto ETHIC (*Evaluation Tool of Health Information for Consumer*) [Cocchi et al., 2014], finalizzato allo sviluppo di uno strumento efficace per bibliotecari medici e professionisti dell'informazione relativa alla sanità per valutare la qualità dei documenti prodotti e per supportarli nella preparazione di testi di qualità migliore, adatti e comprensibili dai pazienti e dai consumatori in generale.

ETHIC permette di valutare testi di informazione relativa alla sanità (opuscoli, depliant, ecc.) e consiste in una lista di controllo e un manuale di istruzioni. Inoltre incorpora strumenti per la valutazione della leggibilità del testo, della comprensione lessicale e degli elementi non testuali come le tabelle.

La lista di controllo consiste di 24 elementi raggruppati in 5 sezioni: trasparenza, idoneità, caratteristiche del documento, caratteristiche linguistiche e testuali, caratteristiche grafiche. Include anche le valutazioni della leggibilità del testo ricorrendo all'indice Gulpease [Lucisano and Piemontese, 1988] e della comprensione lessicale, basata sulla distribuzione delle parole all'interno del Vocabolario di Base per la lingua italiana [De Mauro, 2000]. Il manuale di istruzioni mostra come eseguire la valutazione, spiega come assegnare il punteggio corretto ad ogni elemento preso in considerazione e contiene esempi pratici.

Lo strumento dà la possibilità di assegnare 3 punteggi differenti ad ogni singolo elemento: 2 punti se il documento possiede parzialmente la caratteristica presa in considerazione, 1 punto se la possiede parzialmente e 0 se invece non è presente. Quando, a causa della peculiarità del documento, non è possibile applicare ad esso uno o più elementi, il valutatore ha la possibilità di assegnare N/A ("Non Applicabile"), senza condizionare il punteggio finale, il quale consiste nella percentuale calcolata sul punteggio massimo ottenibile secondo il numero di elementi applicati al documento su cui si porta avanti la valutazione. Il metodo del punteggio rende la lista di controllo facilmente adattabile alla valutazione di diversi tipi di materiale di informazioni relative alla sanità e permette di confrontare tra loro diversi documenti (considerando il punteggio totale o sezione per sezione).

ETHIC non è stato ancora sottoposto ad una procedura di convalida per dimostrare la sua efficacia, la quale permetterà in futuro di individuare possibili errori nella lista di controllo. In particolare i passaggi essenziali per la valutazione dello strumento possono essere un pre-test su un piccolo campione di documenti, la valutazione della validità del contenuto da un gruppo di esperti e infine la sua applicazione su un ampio campione di testi.

Un altro caso di studio che tratta linguaggi differenti includendo anche l'italiano è riportato in [Terranova et al., 2012], il cui scopo è stato di valutare e migliorare la qualità e la leggibilità dei moduli di consenso informato usati in cardiologia.

Sono stati valutati i moduli attualmente in uso, precedentemente scritti in ita-

liano e in inglese, di 7 esami di imaging comuni, secondo le raccomandazioni delle società scientifiche. Per ogni testo è stato anche sviluppato un modulo rivisto secondo gli standard di riferimento, tra cui le linee guida del Federal Plain Language[12]. Per quanto riguarda i punteggi di leggibilità, ogni testo (originale e rivisto) è stato valutato facendo riferimento all'indice *Flesch-Kincaid grade level* [Flesch, 1949] e all'indice Gulpease [Lucisano and Piemontese, 1988].

Il progetto presentato dagli autori è costituito da due fasi: in primo luogo è stata effettuata un'analisi su un campione di volantini attuali sviluppati per scopi di consenso informato sulla base degli standard di riferimento e in secondo luogo sono stati creati dei moduli di consenso informato rivisti convalidati da un gruppo di esperti, costituito da un medico legale, un avvocato, un esperto di linguistica computazionale, medici, un esperto di comunicazione e un membro dell'organizzazione di difesa del paziente. Anche se la valutazione della leggibilità è stata condotta facendo riferimento alle metriche tradizionali per garantire la comparabilità dei risultati tra le lingue in esame, la novità principale di questo approccio è che la riscrittura dei moduli di consenso informato è stata condotta attraverso READ-IT [Dell'Orletta et al., 2011], il primo strumento per la valutazione automatica della leggibilità per l'italiano.

All'analisi qualitativa, i moduli di consenso informato originali erano complessi e organizzati male, scritti in gergo e comprendevano contenuti incompleti (non erano presenti informazioni riguardo alle opzioni di trattamento, il rischio delle radiazioni a lungo termine e dosi). Inoltre non erano correttamente evidenziate le probabilità di esito dell'esame.

All'analisi quantitativa, i punteggi di leggibilità erano particolarmente bassi per tutti i tipi di moduli di consenso e sono sostanzialmente migliorati nei documenti rivisti (per quelli inglesi il *Flesch-Kincaid grade level* passa da 10,2 a 6,5 e per quelli italiani l'indice Gulpease passa da 45,7 a 84,09). Le versioni riviste includono anche una discussione esplicita delle opzioni di trattamento e i loro relativi rischi e benefici, danni potenziali che potrebbero derivare dal non sottoporsi alla procedura,

[12]http://www.plainlanguage.gov/howto/guidelines/bigdoc/fullbigdoc.pdf

una spiegazione sui rischi delle radiazioni ionizzanti previsti a lungo termine (a riguardo sono stati aggiunte una tabella e una figura) e una linea da compilare dopo l'esame che riporta la dose effettiva (non quella teorica, la dose di riferimento attesa) erogata al paziente.

Infine, un altro caso di studio volto ad indagare e a migliorare l'accessibilità dei moduli di consenso informato italiani attraverso tecniche di NLP è presentato in [Venturi et al., 2015]. La metodologia di analisi automatica della leggibilità proposta è illustrata ed estesa nel lavoro di tesi presentato nel capitolo 2.

2

Metodi, strumenti e risultati della valutazione della leggibilità dei consensi informati

In questo capitolo è introdotta la metodologia di analisi automatica della leggibilità di un testo basata su strumenti di annotazione linguistica automatica. Questa metodologia è stata applicata ad un corpus di moduli di consenso informato in uso presso gli ospedali e le aziende sanitarie locali della regione Toscana.

Nel paragrafo 2.1 è presentata la catena di analisi linguistica automatica per l'italiano, prerequisito per la valutazione automatica della leggibilità. Nel paragrafo 2.2 si descrive il corpus analizzato e le caratteristiche linguistiche dei documenti di cui è composto. Nel paragrafo 2.3 sono presentati i risultati della valutazione automatica della leggibilità del corpus ottenuti usando lo strumento READ-IT, primo strumento di valutazione della leggibilità per la lingua italiana. Infine, nel paragrafo 2.4 è posta l'attenzione sul lessico del corpus, in quanto caratterizzato da parole specifiche dell'ambito medico.

2.1 I sistemi di analisi linguistica per l'italiano

Le tecnologie linguistico-computazionali permettono di accedere al contenuto informativo dei testi attraverso l'individuazione della struttura linguistica sottostante e la sua rappresentazione esplicita [Montemagni, 2013]. L'identificazione della struttura linguistica del testo 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: segmentazione del testo in frasi e tokenizzazione, ovvero la divisione delle sequenze di caratteri in unità minime di analisi dette "token" (parole, punteggiatura, date, numeri, sigle, ecc.); lemmatizzazione del testo "tokenizzato" e analisi morfo-sintattica; analisi della struttura sintattica della frase in termini di relazioni di dipendenza tra parole come soggetto, oggetto diretto, modificatore ecc

Un testo arricchito con informazioni di questo tipo diventa il punto di partenza per ulteriori elaborazioni automatiche, in particolare per l'identificazione di parametri che possono essere sfruttati nel compito di ricostruzione del profilo linguistico di un testo.

Lo stato dell'arte nei compiti di annotazione linguistica è rappresentato da sistemi basati su algoritmi di apprendimento automatico supervisionato. Il compito di annotazione linguistica viene modellato come un compito di classificazione probabilistica: ad ogni passo di computazione il sistema sceglie l'annotazione più probabile data la parola in input, i suoi tratti descrittivi, il contesto e le annotazioni linguistiche già identificate. A partire da un corpus di addestramento, annotato con informazione morfo-sintattica e sintattica, viene costruito un modello probabilistico per l'annotazione linguistica del testo.

I moduli di annotazione linguistica usati in questa tesi rappresentano lo stato dell'arte per la lingua italiana.

Per quanto riguarda l'annotazione morfo-sintattica, il PoS tagging descritto in [Dell'Orletta, 2009] presenta un'accuratezza (calcolata come il rapporto tra il nu-

mero di token classificati correttamente e il numero totale di token analizzati) del 96,34% nell'assegnazione della giusta PoS e dei relativi tratti morfo-sintattici.

Per quanto riguarda invece l'analisi a dipendenze, il parser utilizzato è DeSR [Attardi et al., 2009]. L'accuratezza viene calcolata rispetto alle metriche tipicamente usate per misurare la performarce dei parser a dipendenze:

- **LAS** (*Labelled Attachment Score*): proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda sia la testa sintattica sia il ruolo svolto in relazione ad essa, con un'accuratezza dell'83,38%.

- **UAS** (*Unlabelled Attachment Score*): proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione della testa sintattica, con un'accuratezza dell'87,71%.

Questi valori di accuratezza sono quelli standard ottenuti nell'addestramento su corpora giornalistici.

Nonostante i risultati dell'annotazione linguistica automatica non siano perfetti e includano inevitabilmente un margine di errore che varia a seconda del livello e del tipo di informazione linguistica presa in considerazione forniscono indicazioni affidabili per la descrizione delle principali caratteristiche linguistiche di un testo. Queste caratteristiche sono descritte in dettaglio nel paragrafo 2.2.2.

I risultati dell'annotazione linguistica possono essere usati in diversi compiti, tra cui il calcolo automatico della leggibilità. Il primo strumento avanzato per la valutazione automatica della leggibilità per la lingua italiana basato su metodi e strumenti di TAL è READ-IT, che viene presentato nel paragrafo 2.3.1.

2.2 Il corpus di consensi informati

In questo paragrafo viene presentato il corpus che è stato usato in un caso di studio volto ad indagare e migliorare l'accessibilità dei moduli di consenso informato sulla base di tecniche avanzate di NLP. In particolare nel paragrafo 2.2.2 viene

presentata la ricostruzione del profilo linguistico del corpus preso in analisi per individuare le caratteristiche linguistiche principali che caratterizzano il linguaggio usato nei moduli di consenso informato (paragrafo 2.2.2).

Basandosi sui risultati del processo di ricostruzione del profilo linguistico è stata condotta un'analisi preliminare del corpus durante il periodo di tirocinio presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa.

Il modulo di consenso informato è un documento che riporta informazioni dettagliate su un trattamento/intervento sanitario, medico o infermieristico proposto per una specifica malattia/condizione, con il quale il paziente viene informato sulle modalità di esecuzione, i benefici, gli effetti collaterali, i rischi prevedibili e l'esistenza di valide alternative terapeutiche [Cordasco, 2013]. Lo scopo principale del modulo di consenso informato è quello di supportare la comunicazione e la relazione medico-paziente, facilitando un'espressione volontaria, informata e consapevole della volontà del paziente.

Il lavoro di tesi è stato svolto in collaborazione con il Centro Gestione Rischio Clinico e Sicurezza del Paziente (Centro GRC) della regione Toscana[13]. In particolare, dal 2010 il Centro GRC lavora ad un programma di comunicazione e compensazione degli eventi avversi, al fine di migliorare l'efficienza della gestione dei sinistri. Grazie a questo programma, l'efficienza è significativamente migliorata, con un risparmio di circa 5 milioni di euro e una riduzione di 5 mesi nei tempi di chiusura dei reclami. Tuttavia, il numero di reclami è stabile e sono ancora presenti casi relativi ad una comunicazione medico-paziente inefficiente, spesso legata alla mancanza di consensi informati adeguati.

L'approccio linguistico-computazionale per la valutazione della leggibilità del corpus di moduli di consenso informati qui analizzati si presenta come punto di partenza per il miglioramento della comunicazione medico-paziente, che sarà perseguito attraverso la progettazione e lo sviluppo di uno strumento indirizzato agli operatori sanitari che includa funzionalità avanzate per la valutazione della qualità

[13]http://www.regione.toscana.it/centro-gestione-rischio-clinico

dei documenti scritti e che li supporti, dove necessario, nella loro semplificazione [Venturi et al., 2015].

2.2.1 Composizione del corpus

Specialità	Nº documenti	Nº parole (tokens)
Anestesia	20	21.065
Chirurgia colo-rettale	2	1.997
Chirurgia dell'obesità	3	8.091
Chirurgia generale	19	11.588
Chirurgia plastica	4	3.550
Chirurgia toracica	9	5.608
Chirurgia vascolare	16	22.739
Oculistica	7	10.496
Otorinolaringoiatria	134	194.421
Ortopedia	44	76.712
Ostetricia e ginecologia	35	31.243
Urologia	17	19.576
Totale: Area Chirurgica	**310**	**407.086**
Cardiologia	54	39.887
Diabetologia	1	297
Gastroenterologia	9	9.856
Neurologia	8	5.199
Oncologia	3	1.692
Pneumologia	4	3.220
Senologia	17	20.455
Totale: Area Medica	**96**	**80.309**
Psicologia	13	11.651
Screening	8	2.007
Vaccini	1	2.852
Totale: Area Prevenzione	**22**	**16.510**
Genetica	11	6.416
Immunoematologia e trasfusionale	43	45.962
Medicina nucleare	29	18.045
Radiologia	24	17.358
Totale: Area Servizi	**107**	**87.781**
Generici	**33**	**8.928**
Pediatria	**13**	**6.092**
Riabilitazione e rieducazione funzionale	**2**	**674**

Tabella 2.1: Statistiche generali del corpus di consensi informati

Il corpus è composto da 583 documenti, per un totale di 607.380 token, costituiti da moduli di consenso informato e varie tipologie di informativa medico-paziente

attualmente in uso nelle 16 aziende ospedaliere del Servizio Sanitario della Toscana, vale a dire 4 ospedali universitari e 12 aziende sanitarie locali. I documenti coprono 29 specialità e sono organizzati in 4 macro-aree (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 token varia tra le varie specialità: quella con più testi a disposizione è otorinolaringoiatria (area chirurgica) con un totale di 134 testi e 194.421 token, invece quelle con meno testi a disposizione sono diabetologia (area medica) con un solo testo e 297 token e vaccini (area prevenzione) con un solo testo e 2.852 token.

2.2.2 Le caratteristiche linguistiche del corpus

I risultati dell'annotazione linguistica automatica se appropriatamente esplorati possono fornire indicazioni affidabili per la descrizione delle principali caratteristiche linguistiche dei moduli di consenso informato. Sulla base di queste caratteristiche è ricostruito il profilo linguistico.

Con lo scopo quindi di tracciare il profilo linguistico dei moduli di consenso informato contenuti nel corpus preso in esame, è stata condotta una metodologia di analisi finalizzata a descriverne le caratteristiche lessicali, morfo-sintattiche e sintattiche sulla base del modo in cui alcuni significativi tratti linguistici si distribuiscono nei testi. Queste caratteristiche sono state estratte automaticamente dal corpus dei moduli di consenso informato usando il Part-of-Speech tagger e il parser sintattico descritti nel paragrafo 2.1.

La fase di ricostruzione del profilo linguistico si è concentrata sul confronto della lingua di un corpus di dominio specifico (in questo caso medico) con quella di corpora rappresentativi dei generi testuali tradizionali, vale a dire giornalismo, prosa scientifica, materiali didattici e narrativa [Dell'Orletta et al., 2013], con lo scopo finale di individuare le caratteristiche linguistiche principali che caratterizzano il linguaggio usato nei moduli di consenso informato.

Le caratteristiche linguistiche monitorate per la ricostruzione del profilo linguistico possono essere organizzate in quattro categorie principali in base alle fasi di annotazione da cui sono state estratte, vale a dire tokenizzazione, lemmatizzazione, PoS tagging e analisi delle dipendenze:

- **Caratteristiche del testo grezzo**: comprendono le caratteristiche utilizzate nelle metriche tradizionali per la misurazione della leggibilità, come la lunghezza della frase (calcolata come la media delle parole per frase) e la lunghezza delle parole (calcolata come la media dei caratteri per parola).

- **Caratteristiche lessicali**: comprendono la composizione interna del vocabolario del testo rispetto al vocabolario di base (VdB) del *Grande Dizionario italiano dell'uso* [De Mauro, 2000] e la Type Token Ratio (TTR), usata per esprimere la ricchezza lessicale di un testo.

  - **VdB**: include una lista di circa 7000 parole altamente familiari ai parlanti nativi dell'italiano. Sono calcolate due caratterististiche diverse:

    * La percentuale di tutte le parole uniche (tipi) presenti nel VdB.
    * La distribuzione interna delle parole presenti nel VdB ripartite 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)

  - **TTR**: calcolata come rapporto tra il numero di parole tipo e il numero di occorrenze delle unità del vocabolario di un testo. I valori ottenuti 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. Dato che questa

misura è particolarmente sensibile alla lunghezza del testo, viene calcolata sulle prime 100 parole.

- **Caratteristiche morfo-sintattiche**: comprendono la distribuzione della categorie morfo-sintattiche, la densità lessicale e il rapporto nomi/verbi.
  - **Distribuzione delle categorie morfo-sintattiche**: si fa riferimento ad un sottoinsieme di categorie, vale a dire nomi, verbi, aggettivi e congiunzioni (distinguendo tra coordinanti e subordinanti).
  - **Densità lessicale**: calcolata come rapporto tra "parole piene" (vale a dire portatrici di significato: nomi, aggettivi, verbi e avverbi) e il numero totale delle occorrenze di parole del testo.
  - **Distribuzione dei modi e dei tempi dei verbi**: è una caratteristica nuova e specifica del linguaggio che sfrutta il potere predittivo della ricca morfologia verbale italiana.
- **Caratteristiche sintattiche**: comprendono le caratteristiche relative alla struttura sintattica di ogni periodo del corpus:
  - **Caratteristiche della subordinazione**: comprendono il numero medio di frasi per periodo, la proporzione di principali e subordinate e la lunghezza media delle catene subordinanti
  - **Profondità dell'albero sintattico**: riguarda i livelli di incassamento gerarchico nell'albero sintattico di un periodo. Comprende tre diverse caratteristiche:
    - **Altezza massima dell'albero**: calcolata come 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.

    - **Lunghezza media delle "catene" preposizionali**: calcolata come la profondità (o, in altri termini, pesantezza) media delle catene di dipendenza a testa nominale. Con questo dato si va a misurare l'incidenza di strutture nominali complesse contraddistinte dalla presenza di modificatori (aggettivali, nominali, preposizionali).
    - **Distribuzione delle "catene" di dipendenza a testa nominale per profondità**: calcolata come percentuale e numero di occorrenze di sequenze gerarchiche e ricorsive di modificatori di teste nominali lunghe 1, 2, ecc.

- **Caratteristiche dei predicati verbali**: comprendono aspetti differenti del comportamento dei predicati verbali, vale a dire la percentuale di radici verbali rispetto a tutte le radici delle frasi che occorrono nel testo e la loro arità.
    - **Arità dei predicati verbali**: calcolata come il numero medio di dipendenti istanziati che condividono la stessa testa verbale (coprendo sia argomenti e sia modificatori).
    - **Distribuzione delle teste verbali per numero di dipendenti istanziati**: calcolata come percentuale e numero di occorrenze di verbi con un numero di dipendenti istanziati uguale a 0, 1, 2, ecc.
- **Lunghezza delle relazioni di dipendenza**: calcolata come la distanza in token tra la testa e il dipendente in una proposizione.

La (Tabella 2.2) riporta la selezione delle caratteristiche linguistiche che caratterizzano maggiormente il corpus di consensi informati (*ConInf*) rispetto a 4 corpora che rappresentano diversi generi testuali: giornalismo (*Gior*), prosa scientifica (*ProSc*), materiali didattici (*Didat*) e narrativa (*Narr*).

Per quanto riguarda le caratteristiche di base, il corpus è caratterizzato da frasi corte (media = 16,06) e da parole lunghe (media = 6,75) rispetto agli altri corpora.

Caratteristiche	***ConInf***	***Gior***	***ProSc***	***Didat***	***Narr***
Lunghezza media delle frasi	16,06	22,9	27,19	28,15	17,99
Lunghezza media delle parole	6,75	5,09	5,57	5	4,91
% di lemmi (tipi) nel VdB	57,24	70,95	58,54	74,05	70,77
TTR (prime 100 parole)	0,72	0,63	0,66	0,69	0,71
Distribuzioni delle POS:					
- nomi	28,51%	28,29%	28,53%	23,25%	23,63%
- verbi	11,83%	13,3%	10,67%	13,87%	15,2%
- aggettivi	9,26%	6,17%	8,8%	7,9%	6,28%
- preposizioni	16,19%	15,87%	16,77%	14,73%	12,31%
- avverbi	3,6%	4,18%	4,09%	5,84%	5,64
- congiunzioni	4,29%	3,65%	3,69%	4,75%	4,48%
- coordinanti	82,54%	78,4%%	83,79%	78,29%	70,2%
- subordinanti	17,69%	21,6%	16,21%	21,71%	29,8%
Densità lessicale	0,59	0,56	0,58	0,56	0,57
Rapporto nomi/verbi	2,41	2,13	2,67	1,68	1,55
Numero medio di frasi per periodo	1,3	2,46	2,41	3,07	2,3
Proporzione di frasi principali e subordinate:					
- frasi principali	74,7%	70,55%	72,26%	67,01%	66,53%
- frasi subordinate	25,3%	29,3%	27,47%	32,23%	33,23%
Lunghezza media delle catene subordinanti	1,02	1,09	0,96	1,09	1,14
Media delle altezze massime degli alberi	4,86	5,91	6,74	6,57	4,57
Lunghezza media delle catene preposizionali	1,31	1,3	1,38	1,22	1,17
Distribuzione delle catene di dipendenza a testa nominale per profondità:					
- 1 complemento incassato	74,25%	75,97%	69,77%	78,97%	78,78%
- 2 complementi incassati	21%	19,26%	22,66%	17,93%	15,29%
- $\geq$3 complementi incassati	4,73%	4,62%	7,05%	2,34%	2,43%
Media dei dipendenti per testa verbale	1,84	2,15	2,04	1,96	1,8
Distribuzione delle teste verbali per numero di dipendenti istanziati:					
- arità 0	10,88%	5,35%	6,28%	6,99%	8,58%
- arità 1	32,71%	28,03%	29,23%	29,13%	31,74%
- arità 2	30,53%	31,84%	33,34%	34,59%	31,8%
- arità $\geq$3	26,12%	34,47%	30,88%	28,53%	24,15%
Media della lunghezza massima delle relazioni di dipendenza (esclusa la punteggiatura)	6,43	9,11	10,37	10,91	7,26
Numero di token per clausola	11,29	9,98	11,63	8,76	7,36
Percentuale di radici verbali con soggetto esplicito	57%	69,6%	76,6%	66,9%	48,79%

Tabella 2.2: Le caratteristiche linguistiche del corpus di consensi informati rispetto ai 4 corpora: giornalismo, prosa, scientifica, materiali didattici e narrativa.

Queste caratteristiche sono quelle monitorate nelle metriche tradizionali della leggibilità, come l'indice Gulpease [Lucisano and Piemontese, 1988], secondo le quali parole lunghe sono meno comprensibili e frasi lunghe sono grammaticalmente più complesse. Questi risultati possono quindi suggerire una tendenza alla scrittura di consensi informati caratterizzati da un vocabolario inevitabilmente complesso ma in costruzioni sintattiche più facili. Inoltre le parole lunghe sono in genere più informative.

Questo risultato è confermato anche dai valori delle caratteristiche lessicali, infatti i consensi informati contengono una percentuale abbastanza bassa di lemmi che appartengono al VdB (57,24%) e quindi un gran numero di parole specifiche del dominio medico che non sono solitamente usate nel linguaggio comune. Un valore molto simile è rintracciato anche nei testi scientifici (58,54%) e questo conferma che i testi informativi sono caratterizzati dall'uso massiccio di termini tecnici abbastanza lunghi. In aggiunta, il corpus dei consensi informati mostra il valore più alto di TTR (0,72), calcolato sui primi 100 token di ogni testo, ed è quindi caratterizzato dalla più grande varietà lessicale. Quest'ultima rappresenta un indice di grande complessità: i moduli di consenso informato sono costituiti da un numero elevato di concetti nuovi nonostante dovrebbero essere dei documenti che ripetono sempre le stesse cose.

Per quanto riguarda le caratteristiche morfo-sintattiche, in particolare la distribuzione delle parti del discorso (PoS), i consensi informati sono caratterizzati da un'alta percentuale di nomi, aggettivi e preposizioni (rispettivamente 28,51%, 9,26% e 9,26%) e una percentuale bassa di verbi (11,83%) e avverbi (3,60%). Per quanto riguarda invece la percentuale di congiunzioni, è interessante vedere la distribuzione delle due sotto-categorie di cui è composta: coordinanti e subordinanti. Infatti se per la categoria generale non si nota una differenza significativa tra i diversi generi testuali, i moduli di consenso informato presentano una percentuale più alta di congiunzioni coordinanti (82,54%) rispetto a quelle subordinanti (17,69%), valori nuovamente molto simili a quelli riscontrati nei testi scientifici (83,79% e 16,21%) ma

significativamente differenti rispetto agli altri corpora, specialmente quello dei testi narrativi (70,2% e 29,8%). Assumendo la percentuale di congiunzioni subordinanti come indicatore della proporzione di costruzioni ipotattiche all'interno del testo, si può affermare che i moduli di consenso informato fanno un uso limitato di questo tipo di costruzioni: ciò è associato alla presenza di costruzioni sintattiche più facili, caratterizzate da una presenza minore di frasi subordinate.

A partire dalla distribuzione delle singole categorie morfo-sintattiche, è possibile mettere in relazione la frequenza di ricorrenza di certe categorie grammaticali rispetto ad altre: per esempio misurare la densità lessicale e il rapporto tra nomi e verbi. L'alta ricorrenza di nomi e aggettivi nei moduli di consenso informato ma di contro l'alta ricorrenza di verbi e avverbi negli altri generi testuali, porta a valori di densità lessicale molto simili (da 0,56 a 0,59). Invece l'alta ricorrenza di nomi e bassa di verbi nel corpus *ConInf* dà luogo ad un rapporto nomi/verbi abbastanza alto (2,41). Anche in questo caso un valore molto alto è riscontrato nei testi scientifici (2,67) e ciò dimostra che questi tipi di testi sono più informativi rispetto agli altri, al contrario dei testi narrativi in cui vi è una ricorrenza alta di pronomi e verbi e bassa di nomi che da luogo ad un rapporto nomi/verbi più basso (1,55).

Questo tipo di informazione rappresenta un passo in avanti rispetto all'analisi della distribuzione delle categorie morfo-sintattiche considerate al di fuori del contesto di ricorrenza, ma non è tuttavia sufficiente a ricavare indicazioni precise in merito alla struttura sintattica complessiva sottostante al testo. È dalla caratteristiche strutturali dell'albero sintattico che si studiano le dipendenze presenti tra le parole indicanti relazioni grammaticali.

Per quanto riguarda le caratteristiche sintattiche, un dato elementare ma significativo è dato dal numero di frasi per periodo, calcolato a partire dal numero di teste verbali rispetto al numero di periodi: si nota che il corpus preso in analisi è caratterizzato da una maggiore proporzione di periodi monoclausali, con una media di 1,30 clausole per periodo, al contrario degli altri generi testuali caratterizzati da valori più alti. Questo dato però non dice nulla su come le diverse clausole si

rapportino l'una con l'altra all'interno del periodo ed è quindi possibile procedere in questa indagine identificando la proporzione di frasi principali e subordinate che è ricostruita a partire dal rapporto tra le radici verbali (corrispondenti alle frasi principali) e le clausole argomentali (ovvero sottocategorizzate dal verbo reggente) e quelle con valore temporale, causale, locativo, ecc. La percentuale più bassa di frasi subordinate si riscontra in *ConInf* (25,3%), mentre i valori più alti sono presenti nei corpora della didattica (32,23%) e della narrativa (33,23%). Inoltre i moduli di consenso informato sono caratterizzati da frasi subordinate corte (*Lunghezza media delle catene subordinanti* = 1,02), al contrario di quelle dei testi narrativi che sono in media un po' più lunghe (1,14).

Un altro aspetto rilevante riguarda i livelli di incassamento gerarchico: in presenza di più di una clausola subordinata all'interno dello stesso periodo, diventa cruciale ricostruire quale tipo di rapporto sussista tra di esse, ovvero se siano ricorsivamente incassate l'una all'interno dell'altra. Una prima e approssimativa misura dei livelli di incassamento gerarchico all'interno della struttura sintattica è data dall'altezza massima dell'albero: nei consensi informati sono registrati valori più bassi rispetto agli altri generi testuali (media = 4,86). Questa misura può essere raffinata focalizzandosi su particolari tipi di costrutti sintattici, per esempio la ricorrenza media di strutture nominali complesse. A questo livello la caratteristica monitorata è la profondità media delle catene di dipendenza a testa nominale (*Lunghezza media delle catene preposizionali*) e quindi la loro presenza all'interno di strutture nominali complesse che includono modificatori aggettivali, nominali e preposizionali: i valori più alti si riscontrano nei testi scientifici (1,38) e a seguire nei moduli di consenso informato (1,31). Questo si riflette anche al livello della distribuzione delle catene di dipendenza a testa nominale per profondità: si osserva una percentuale bassa di sequenze corte (1 complemento incassato = 74,25%) e alta di sequenze lunghe (2 complementi incassati = 21% e $\geq$3 complementi incassati = 4,73%). Questi valori sono simili a quelli riscontrati nei testi giornalistici ma molto distanti da quelli dei materiali didattici e della narrativa, caratterizzati da percentuali alti di sequenze

corte (rispettivamente 1 complemento incassato = 78,97% e = 78,78%) e bassa di sequenze lunghe (rispettivamente 2 complementi incassati = 17,93% e = 15,29% e ≥3 complementi incassati = 2,34% e 2,43%). Quindi a questo livello, i consensi informati risultano essere più complessi rispetto agli altri generi testuali (con l'eccezione dei testi scientifici), essendo caratterizzati dalla presenza di sequenze profonde di strutture nominali complesse.

Un'altra caratteristica monitorata è l'arità delle teste verbali, in modo tale da ricostruire il numero di tutti i dipendenti associati ad essa: in questo caso i valori più alti si riscontrano nei testi giornalistici (2,15), invece i moduli di consenso informato sono caratterizzati da uno dei valori più bassi (1,84) insieme ai testi narrativi (1,8). Questo dato diventa ancora più significativo ricostruendo la distribuzione delle teste verbali per numero di dipendenti istanziati: in questo caso i diversi generi testuali presentano un andamento analogo e le uniche differenze si riscontrano tra i moduli di consenso informato e i testi giornalistici: i primi sono caratterizzati da una percentuale di verbi con arità 0 e 1 maggiore rispetto ai secondi (rispettivamente 10,88% - 5,35% e 32,71% - 28,03%) e di contro più bassa di verbi con arità ≥3 (rispettivamente 26,12% e 34,47%). Una differenza significativa si riscontra nella lunghezza delle relazioni di dipendenza: la media delle lunghezze massime per frase delle relazioni di dipendenza (esclusa la punteggiatura) per i consensi informati presenta il valore più basso rispetto a tutti gli altri generi (6,43). Al livello delle caratteristiche tipicamente associate alla complessità *strutturale*, il corpus dei moduli di consenso informato quindi risulta essere meno complesso rispetto agli altri generi testuali.

Un'altra caratteristica interessante dei consensi informati è rappresentata dal numero di token per clausola (11,29), valore simile a quello riscontrato nei testi scientifici (11,29) ma significativamente diverso da quelli degli altri generi, ad esempio la narrativa (7,36). Infine si riscontra la percentuale più bassa di radici verbali con il soggetto esplicito (57%). Facendo riferimento a queste due caratteristiche, *ConInf* invece risulta essere il più complesso.

Le caratteristiche linguistiche del corpus organizzato in macro-aree

La ricostruzione del profilo linguistico è stata condotta anche internamente al corpus dei moduli di consenso informato con lo scopo finale di individuare le caratteristiche linguistiche principali che differenziano i testi organizzati per macro-aree, considerando anche le tre specialità non appartenenti a nessuna di esse, vale a dire generici, pediatria e riabilitazione e rieducazione funzionale.

La (Tabella 2.3) riporta la selezione delle caratteristiche linguistiche che caratterizzano maggiormente il corpus di consensi informati organizzato per macro-aree, vale a dire area chirurgica (*ArChir*), area medica (*ArMed*), area prevenzione (*ArPrev*), area servizi (*ArServ*), generici (*Gen*), pediatria (*Ped*) e riabilitazione e rieducazione funzionale (*Riab*).

Per quanto riguarda le caratteristiche di base, la macro-area caratterizzata da frasi più lunghe è l'area prevenzione (20,37) ma anche da parole più corte (6,03). La specialità dei generici invece è quella caratterizzata da frasi più corte (10,45) ma da parole più lunghe (9,93).

Per quanto riguarda le caratteristiche lessicali, la percentuale di lemmi appartenenti al VdB è abbastanza simile per tutte le macro-aree: la maggior parte dei valori cadono nell'intervallo 56-58%, con la sola eccezione dell'area prevenzione che presenta una percentuale del 63,19% ed è quindi caratterizzata dall'uso maggiore di termini medici. D'altro canto quest'ultima macro-area presenta il valore più basso di TTR (0,70) e questo vuol dire che nonostante il maggiore utilizzo di termini non comuni, questi sono ripetuti più volte rispetto ai documenti delle altre macro-aree: quella con la più grande varietà lessicale è pediatria (0,8).

A livello morfo-sintattico non si notano grandi differenze nella distribuzione delle parti del discorso (PoS): come è stato visto nella fase precedente, i consensi informati sono caratterizzati da una alta percentuale di nomi, aggettivi e preposizioni e bassa di verbi e avverbi. I valori di densità lessicale sono quindi simili, nell'intervallo tra 0,57 (area prevenzione) e 0,61 (pediatria). I testi pediatrici sono caratterizzati da

Caratteristiche	***ArChir***	***ArMed***	***ArPrev***	***ArSer***	***Gen***	***Ped***	***Riab***
Lunghezza media delle frasi	15,27	19,29	20,37	16,59	10,45	14,06	10,81
Lunghezza media delle parole	6,6	6,31	6,03	6,66	9,93	7,04	7,84
% di lemmi (tipi) nel VdB	56,5%	56,8%	63,19%	58,37%	57,44%	57,93%	56,67%
TTR (prime 100 parole)	0,71	0,73	0,70	0,75	0,72	0,8	0,79
Distribuzioni delle POS:							
- nomi	27,79%	28,66%	28,79%	29,49%	30,94%	29,77%	29,03%
- verbi	11,63%	11,74%	12,46%	12,39%	11,95%	10,93%	12,92%
- aggettivi	9,76%	9,32%	8,72%	8,47%	7,26%	9,45%	9,92%
- preposizioni	16,12%	16,33%	17,98%	16,34%	15,44%	14,49%	16,86%
- avverbi	3,9%	3,55%	3,46%	3,47%	2,5%	3,87%	3,4%
- congiunzioni	4,44%	4,11%	4,46%	4,2%	3,9%	4,06%	3,65%
- coordinanti	81,82%	83,27%	76,41%	81,73%	92,53%	79,45%	94,13%
- subordinanti	18,18%	16,73%	23,59%	18,27%	7,47%	20,55%	5,87%
Densità lessicale	0,60	0,59	0,57	0,59	0,57	0,61	0,59
Rapporto nomi/verbi	2,39	2,44	2,31	2,38	2,59	2,72	2,25
Numero medio di frasi per periodo	1,19	1,74	1,93	1,58	0,98	1,32	1,23
Proporzione di frasi principali e subordinate:							
- frasi principali	76,52%	73,06%	67,1%	74,53%	66,44%	78,06%	77,08%
- frasi subordinate	23,48%	26,94%	32,9%	25,47%	33,56%	21,94%	22,92%
Lunghezza media delle catene subordinanti	1,03	1,02	0,98	1,1	0,74	0,81	1
Media delle altezze massime degli alberi	4,56	5,57	6,58	5,13	3,61	4,27	4,14
Lunghezza media delle catene preposizionali	1,29	1,36	1,39	1,35	1,22	1,3	1,31
Distribuzione delle catene di dipendenza a testa nominale nominale per profondità:							
- 1 complemento incassato	75,6%	71,33%	70,27%	71,58%	81,26%	74,51%	75,15%
- 2 complementi incassati	20,14%	22,72%	22,51%	23,15%	15,74%	21,64%	20,68%
- $\geq$3 complementi incassati	4,25%	5,95%	7,22%	5,27%	3%	3,86%	4,17%
Media dei dipendenti per testa verbale	1,82	1,91	1,86	1,93	1,64	1,83	1,63
Distribuzione delle teste verbali per numero di dipendenti istanziati:							
- arità 0	12%	9,1%	6,6%	9,14%	14,79%	8,05%	14,44%
- arità 1	32,11%	32,84%	37,1%	31,81%	36,31%	35,99%	37,78%
- arità 2	30,94%	29,54%	34,89%	29,24%	30,57%	31,43%	28,33%
- arità $\geq$3	24,95%	28,52%	21,41%	29,81%	18,32%	24,54%	8,47%
Media della lunghezza massima delle relazioni di dipendenza (esclusa la punteggiatura)	6,15	7,41	7,71	6,67	4,88	5,82	5,45
Numero di token per clausola	11,44	11,38	9,77	11,04	11,58	11,26	10,36
Percentuale di radici verbali con soggetto esplicito	57,58%	61,52%	71,5%	60,85%	18,46%	53,9%	39,29%

Tabella 2.3: Le caratteristiche linguistiche del corpus di consensi informati organizzato per macro-aree.

una delle percentuali più alte di nomi (29,77%) e più bassa di verbi (10,93%): da ciò segue il rapporto nomi/verbi più alto (2,72). Invece quello più basso si riscontra nella specialità di riabilitazione e rieducazione funzionale (2,25). La percentuale di congiunzioni cade nell'intervallo tra 3,9% (generici) e 4,46% (area prevenzione). Per quanto riguarda invece la distribuzione delle due sotto-categorie di cui è composta, la macro-area che si discosta maggiormente dalle altre è la specialità di riabilitazione e rieducazione funzionale, con una percentuale di congiunzioni coordinanti del 94,13% e di congiunzioni subordinanti del 5,87% . In questo caso le differenze sono significative, infatti il valore più basso di congiunzioni coordinanti è 76,41% e più alto di congiunzioni subordinanti 23,59% (area prevenzione).

Per quanto riguarda le caratteristiche sintattiche e in particolare il numero di frasi per periodo, la macro-area caratterizzata da una maggiore proporzione di periodi monoclausali è la specialità dei generici, con una media di 0,98 clausole per periodo. In questo caso si contrappongono l'area medica e l'area prevenzione, i cui testi contengono una proporzione maggiore di periodi costituiti da più di una clausola (rispettivamente si registra una media di 1,74 e di 1,93). A livello della proporzione di frasi principali e subordinate, la macro-area che possiede la percentuale maggiore delle prime e minore delle seconde è pediatria (rispettivamente 78,06% e 21,94%), ma con una media della lunghezza delle catene subordinanti tra i più alti (1). Invece quella che possiede una percentuale più alta di frasi subordinate è l'area prevenzione (32,9%) ma che sono in media le più corte se confrontate con i testi delle altre macro-aree (0,98).

Per quanto riguarda invece i livelli di incassamento gerarchico all'interno della struttura sintattica, si riscontrano differenze significative nella media delle altezze massime degli alberi: i valori variano da 3,61 (generici) a 6,58 (area prevenzione). Focalizzandosi sulla ricorrenza media di strutture nominali complesse, anche in questo caso il valore più basso si riscontra per la specialità dei generici (1,22) e il più alto per l'area prevenzione (1,29). Questo risultato si riflette anche al livello della distribuzione delle catene di dipendenza a testa nominale per profondità: la macro-area

generici è caratterizzata da una percentuale alta di sequenze corte (1 complemento incassato = 81,26%) e bassa di sequenze lunghe (2 complementi incassati = 15,74% e ≥3 complementi incassati = 3%), al contrario dell'area prevenzione che ha invece la percentuale più alta di sequenze lunghe (≥3 complementi incassati = 7,22%).

Un'altra caratteristica monitorata è il numero medio di dipendenti per testa verbale: in questo caso i valori più alti si riscontrano per i testi appartenenti all'area servizi e all'area medica (rispettivamente 1,93 e 1,91). I testi della specialità dei generici hanno nuovamente uno dei valori più bassi (1,64) insieme a quelli della specialità di riabilitazione e rieducazione funzionale (1,63). Ricostruendo la distribuzione delle teste verbali per numero di dipendenti istanziati, i testi dell'area servizi sono caratterizzati dalla percentuale più alta di verbi con arità ≥3 (29,81%) cui si contrappone quella della specialità di riabilitazione e rieducazione funzionale, caratterizzata dal valore più basso (8,47%), cui segue la specialità dei generici (18,32%). Una differenza significativa si riscontra anche nella lunghezza delle relazioni d dipendenza: la media delle lunghezza massima per frase delle relazioni di dipendenza (esclusa la punteggiatura) presenta nuovamente il valore più basso per i testi delle specialità dei generici e di riabilitazione e rieducazione funzionale (rispettivamente 4,88 e 5,45). Invece il valore più alto è per l'area prevenzione (7,71).

Un'altra caratteristica interessante è rappresentata dal numero di token per clausola: la differenza più significativa si riscontra nei documenti dell'area prevenzione (9,77) e in quelli della specialità di riabilitazione e rieducazione funzionale (10,36), mentre per le altre macro-aree i valori cadono nell'intervallo tra 11,04 (area servizi) e 11,58 (generici). Infine si riscontra la percentuale più bassa di radici verbali con il soggetto esplicito per la specialità dei generici (18,46%), risultato significativamente differente rispetto a quello delle altre macro-aree, i cui valori cadono nell'intervallo tra 39,29% (riabilitazione e rieducazione funzionale) e 71,5% (area prevenzione).

Alla luce dei risultati ottenuti durante il monitoraggio linguistico del corpus dei moduli di consenso informato organizzato per macro-aree, quelli più significativi riguardano il livello sintattico. In particolare è interessante notare l'opposizione tra i

moduli appartenenti alla specialità dei generici e all'area prevenzione: al livello della struttura dell'albero sintattico (media delle altezze massime degli alberi, ricorrenza di strutture nominali complesse, numero medio di dipendenti per testa verbale) la prima è caratterizzata da valori associabili ad una maggiore semplicità rispetto alla seconda, la quale risulta essere la più difficile. Invece per le altre caratteristiche (numero di token per clausola, percentuale di radici verbali con soggetto esplicito) la situazione è inversa: la specialità dei generici risulta essere la più difficile, mentre l'area prevenzione la più facile.

2.3 La valutazione della leggibilità

La valutazione della leggibilità dei moduli di consenso informato è stata effettuata usando lo strumento READ-IT (paragrafo 2.3.1), primo strumento di valutazione della leggibilità per la lingua italiana. La scelta di utilizzare uno strumento addestrato su corpora giornalistici è stata necessaria data la mancanza di risorse specifiche del dominio. Nonostante ciò i risultati ottenuti sono in grado di rappresentare un quadro generale della complessità dei consensi informati.

Come discusso nel capitolo 1, la valutazione della leggibilità nel dominio medico in genere non va oltre la distribuzione delle parti del discorso (PoS) e/o sintagmi nominali: in questo caso la novità sta nell'uso di risultati ottenuti dal testo annotato sintatticamente (vale a dire a dipendenze), rendendo così possibile monitorare una varietà più ampia di fattori che influenzano la leggibilità di un testo.

In questa prima parte dello studio ci si è focalizzati sui risultati degli esperimenti effettuati al livello del documento. La valutazione della leggibilità è stata effettuata rispetto alle specialità mediche (paragrafo 2.3.2) e rispetto le aziende sanitarie locali (paragrafo 2.3.3).

2.3.1 READ-IT: strumento di analisi della leggibilità per la lingua italiana

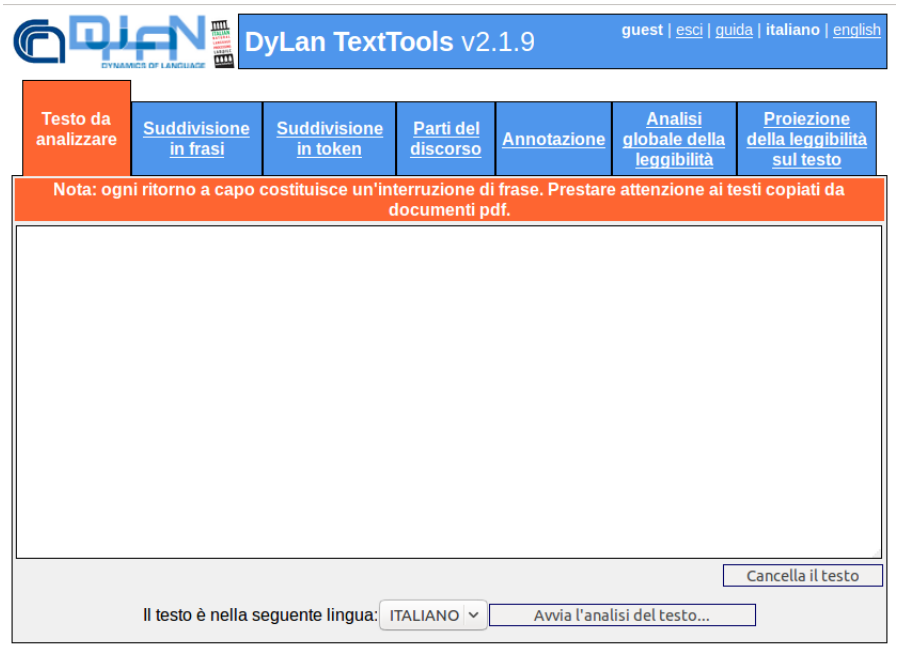


Figura 2.1: Demo online di READ-IT.

READ-IT[14], sviluppato presso l'Istituto di Linguistica Computazionale (ILC) "A. Zampolli" del CNR di Pisa, rappresenta il primo strumento di valutazione della leggibilità per la lingua italiana [Dell'Orletta et al., 2011]. 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. Nella Figura 2.1 è riportata la pagina iniziale della demo online di READ-IT.

READ-IT lavora sui testi sintatticamente analizzati e assegna ad essi dei punteggi che ne quantificano la leggibilità. READ-IT affronta la questione del calcolo della leggibilità come un problema di classificazione, quindi assegna il documento analizzato ad una specifica classe di leggibilità. Il sistema è basato su Support Vector Machines che usa LIBSVM [Chang and Lin, 2011] il quale, dato un set di caratte-

[14]http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest

ristiche e un training corpus, crea un modello statistico usando le caratteristiche estratte dal corpus.

Il training corpus è formato da due corpora che appartengono allo stesso genere testuale (prosa giornalistica): *La Repubblica* (Rep) e *Due Parole* (2Par), quest'ultimo costituito da articoli scritti per un pubblico di adulti con deficit cognitivo o caratterizzato da un basso livello di alfabetizzazione. Gli articoli di 2Par sono stati scritti da linguisti italiani esperti nella semplificazione del testo usando criteri di linguaggio controllato sia a livello lessicale che sintattico [Piemontese, 1996]. Ci sono diverse motivazioni per cui sono stati selezionati questi due corpora. Sul lato pratico 2Par è l'unico corpus disponibile di testi semplificati indirizzato appositamente a persone con un basso livello di alfabetizzazione e quindi rappresenta l'unica opzione possibile sul fronte dei testi semplificati. Per la selezione dell'altro corpus, si è optato per testi appartenenti allo stesso genere testuale, vale a dire prosa giornalistica, con lo scopo di evitare interferenze dovute alla variazione del genere testuale nella misura della leggibilità del testo. Questo è confermato dal fatto che i due corpora mostrano un comportamento simile rispetto a diversi parametri che sono indicativi delle differenze tra i generi testuali: per esempio la densità lessicale, il rapporto nomi/verbi, la percentuale delle radici verbali, ecc. Dall'altro lato, i due corpora si differenziano significativamente rispetto alla distribuzione di caratteristiche tipicamente associate alla complessità del testo: per esempio la composizione del vocabolario usato o, dal punto di vista sintattico, la profondità dell'albero parsato, la ripartizione di frasi principali e subordinate, ecc.

I corpora Rep/2Par ricordano inoltre quelli usati in altri studi sulla leggibilità, come l'*Encyclopedia Britannica* e *Britannica Elementary* [Barzilay and Elhadad, 2003], ma con una principale differenza: mentre i corpora inglesi consistono di coppie di testi originali/semplificati, vale a dire "corpora monolingue paralleli", quelli italiani si presentano come "corpora monolingue comparabili", senza nessuna coppia completa/semplificata dello stesso testo. La comparabilità è garantita dall'inclusione di testi appartenenti allo stesso genere testuale.

Dato un testo nuovo in analisi, READ-IT lo assegna alla classe dei testi facili o difficili da leggere.

La valutazione globale della leggibilità del testo avviene quindi rispetto a due classi (semplice vs. complesso).

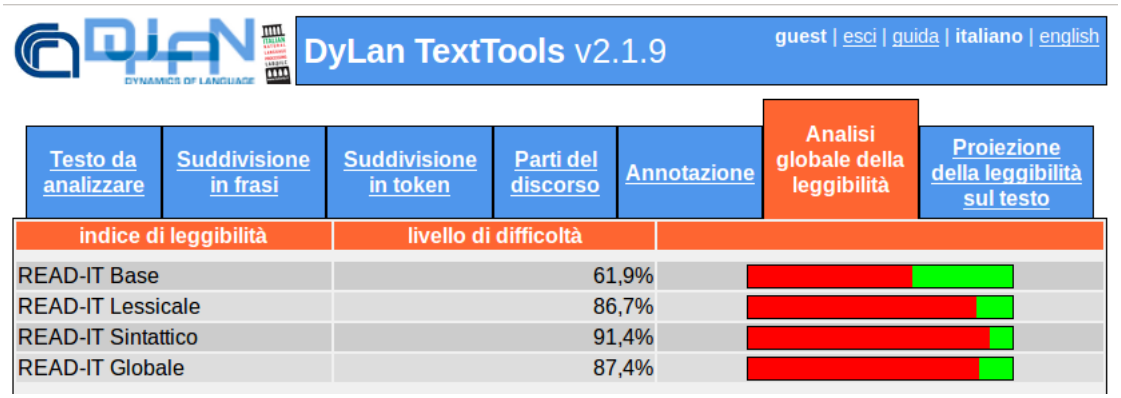


Figura 2.2: Analisi globale della leggibilità in READ-IT.

READ-IT valuta la leggibilità di un testo sulla base di diversi "gruppi" delle caratteristiche linguistiche descritte nella fase di ricostruzione del profilo linguistico (paragrafo 2.2.2) (Figura 2.2):

- **READ-IT Base**: in questo modello le caratteristiche considerate sono quelle tipicamente usate nelle misure tradizionali della leggibilità di un testo, vale a dire 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.

- **READ-IT Lessicale**: questo modello si focalizza sulle caratteristiche lessicali del testo, costituite dalla composizione del vocabolario e la sua ricchezza lessicale.

- **READ-IT Sintattico**: questo modello si basa su informazione di tipo grammaticale, vale a dire sulla combinazione di tratti morfo-sintattici e sintattici ricavati dai corrispondenti livelli di analisi linguistica.

- **READ-IT Globale**: questo modello si basa 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 difficile leggibilità. I valori ottenuti variano su una scala che va da 0 (facile da leggere) a 100 (difficile da leggere).

Figura 2.3: Proiezione della leggibilità sul testo in READ-IT.

La valutazione al livello della singola frase, invece, riveste un ruolo importante quando la valutazione della leggibilità è finalizzata alla semplificazione del testo. READ-IT consente di identificare i periodi che necessitano di revisione e il tipo di difficoltà (base, lessicale, sintattica e globale) riscontrata in relazione ad essi.

Come si può osservare in Figura 2.3, per ciascun periodo il livello di difficoltà è rappresentato cromaticamente mediante colori che vanno dal verde (testo leggibile) al rosso (testo particolarmente difficile). Quando la difficoltà si situa al livello lessicale viene richiesta la revisione delle parole usate nel testo, quando invece si situa al livello sintattico, l'utente è invitato a riformulare il periodo facendo uso di strutture grammaticali più semplici (ad esempio, evitando l'uso ripetuto di strutture subordinate).

2.3.2 La valutazione della leggibilità per specialità medica

La prima parte del lavoro di tesi è stata la valutazione della leggibilità per specialità medica, con lo scopo di capire se tra di esse alcune risultano essere più complesse rispetto alle altre in relazione al contenuto che esprimono. Rispetto ai quattro modelli di analisi della leggibilità, la complessità dei testi è stata monitorata rispetto ai modelli BASE, LESSICALE e SINTATTICO. Come è stato visto nel paragrafo 2.3.1, per ciascun modello il valore ottenuto si riferisce alla probabilità di appartenenza alla classe dei testi di difficile leggibilità. I valori ottenuti variano su una scala che va da 0 (facile da leggere) a 100 (difficile da leggere).

La (Tabella 2.4) riporta i risultati ottenuti all'interno dell'intero corpus, per tutte le 29 specialità mediche. Inoltre è stato calcolato un punteggio per ognuna delle quattro macro-aree (area chirurgica, area medica, area prevenzione e area servizi), come la media dei punteggi registrati per ogni specialità. L'intero corpus è caratterizzato da un livello basso di leggibilità, anche se con differenze significative tra i diversi modelli e tra le macro-aree.

Per quanto riguarda i risultati ottenuti dal modello BASE, il quale fa riferimento alle caratteristiche del testo grezzo come la lunghezza delle frasi e delle parole, la specialità che risulta essere più facile è diabetologia (23,05), mentre quella più difficile è chirurgia toracica (94,98). Altre specialità caratterizzate da valori bassi sono otorinolaringoiatria (25,14) e vaccini (33,72), mentre altre caratterizzate da valori alti sono chirurgia toracica (94,98), chirurgia plastica (88,95), urologia (85,40), senologia (85,09) e psicologia (80,44). I valori per le restanti specialità oscillano invece da 41,12 (gastroenterologia) e 75,18 (chirurgia colo-rettale). Considerando i valori medi ottenuti per i testi organizzati in macro-aree, quella che risulta essere più facile è la specialità di pediatria (51,59) e quella più difficile riabilitazione e rieducazione funzionale (63,84). Questo modello non risulta comunque affidabile per catturare la difficoltà dei moduli di consenso informato in quanto, come si è visto nella fase del monitoraggio linguistico rispetto ai corpora generali, questi testi

Specialità	READ-IT		
	BASE	**LESSICALE**	**SINTATTICO**
Anestesia	50	93,37	69,62
Chirurgia colo-rettale	75,18	100	93,81
Chirurgia dell'obesità	51,63	93,42	59,20
Chirurgia generale	43,03	78,29	58
Chirurgia plastica	88,95	98,72	96,51
Chirurgia toracica	94.98	99,94	95,55
Chirurgia vascolare	88,64	98,13	97,62
Oculistica	49,21	98,89	61,29
Otorinolaringoiatria	25,14	94,90	69,42
Ortopedia	50,54	97,58	89,66
Ostetricia e ginecologia	60,37	97,31	58,52
Urologia	85,40	98,08	89,16
Totale: Area Chirurgica	**63,59**	**95,72**	**78,19**
Cardiologia	66,20	94,50	78,99
Diabetologia	23,05	100	45,68
Gastroenterologia	41,12	87,90	59,82
Neurologia	69,44	97,96	94,98
Oncologia	46,34	99,73	96,07
Pneumologia	49,57	98,18	78,27
Senologia	85,09	99,68	93,88
Totale: Area Medica	**54,26**	**96,85**	**78,24**
Psicologia	80,44	96,25	98,32
Screening	53,13	65,14	50,60
Vaccini	33,72	100	71,76
Totale: Area Prevenzione	**55,76**	**87,13**	**73,56**
Genetica	56,26	95,65	81,45
Immunoematologia e trasfusionale	56,84	93,39	83,47
Medicina nucleare	52,62	96,56	68,48
Radiologia	63,78	98,61	78,68
Totale: Area Servizi	**57,38**	**96,05**	**78,02**
Generici	**51,59**	**87,81**	**88,27**
Pediatria	**49,84**	**99,46**	**74,67**
Riabilitazione e rieducazione funzionale	**63,84**	**99,99**	**96,25**

Tabella 2.4: Risultati della valutazione della leggibilità secondo i modelli BASE, LESSICALE e SINTATTICO dei documenti organizzati in specialità.

sono caratterizzati da frasi abbastanza corte, una caratteristica tipica dei testi facili. Questo è evidente per esempio nel confronto tra l'area medica e l'area prevenzione: la prima risulta essere più facile (54,26) rispetto alla seconda (55,26) nonostante il fatto che i risultati ottenuti negli altri due modelli mostrino il contrario. Infatti al livello LESSICALE e al livello SINTATTICO, i testi dell'area medica risultano essere molto più difficili rispetto a quelli dell'area prevenzione (rispettivamente 96,85 e 78,24 contro 87,13 e 73,56).

Rispetto al modello LESSICALE e SINTATTICO, si riscontrano forti differenze tra le macro-aree e le 29 specialità. In particolare, tutte le specialità risultano essere più difficili al livello lessicale (media = 95,15) rispetto a quello sintattico (media = 78,55), con valori molto più vari.

Considerando i risultati ottenuti dal modello LESSICALE, il quale fa riferimento alla composizione del vocabolario e la sua ricchezza lessicale, i valori variano da un minimo di 65,14 (screening) ad un massimo di 100 (chirurgia colo-rettale, diabetologia e vaccini). 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. In particolare, i testi di screening all'interno del corpus contengono informative sull'esame mammografico. Rispetto a specialità appartenenti alle altre macro-aree, che contengono moduli di consenso informato su operazioni più complicate, è caratterizzato da un lessico molto più facile. Questa specialità fa parte dell'area prevenzione, che risulta essere la macro-area più facile (87,13). Considerando sempre la suddivisione in macro-aree, la più difficile è invece la specialità di riabilitazione e rieducazione funzionale (99,99).

Considerando i risultati ottenuti dal modello SINTATTICO, il quale fa riferimento a informazione di tipo sintattico, i valori variano da un minimo di 45,68 (diabetologia) ad un massimo di 98,32 (psicologia). Un'altra specialità ad avere un valore basso rispetto alle altre è lo screening (50,60). Per quanto riguarda la suddivisione in macro-aree, anche in questo caso quella più facile è l'area prevenzio-

ne (73,56), invece quella più difficile è la specialità di riabilitazione e rieducazione funzionale (96,25).

Alla luce dei risultati ottenuti con i modelli LESSICALE e SINTATTICO, è interessante notare l'opposizione tra diverse specialità che risultano essere più difficili rispetto al primo ma più facili rispetto al secondo. Escludendo la specialità di screening che risulta essere la più facile tra le specialità rispetto ad entrambi i modelli (rispettivamente 65,14 e 50,60), per esempio diabetologia e ostetricia e ginecologia sono tra le più difficili a livello lessicale (rispettivamente 100 e 97,31) ma tra le più facili a livello sintattico (rispettivamente 45,68 e 58,52). Altre specialità risultano essere molto difficili rispetto a entrambi i modelli, per esempio chirurgia vascolare (rispettivamente 98,13 e 97,62) e oncologia (rispettivamente 99,73 e 96,07). Seguendo la suddivisione per macro-aree, la differenza maggiore tra modello LESSICALE e SINTATTICO si ha per l'area medica (rispettivamente 96,85 e 78,24), mentre la specialità di riabilitazione e rieducazione funzionale risulta essere molto difficile ad entrambi i livelli (rispettivamente 99,99 e 96,25).

Come è stato presentato durante la fase di monitoraggio linguistico, la tipologia di caratteristiche che contribuiscono a questi risultati riguardano le caratteristiche locali dell'albero sintattico a dipendenza, considerate dalla letteratura come indici della complessità del linguaggio, invece di quelle *strutturali*. Questa consapevolezza può essere usata nella costruzione di una versione specifica del dominio del modello SINTATTICO della leggibilità all'interno di READ-IT.

2.3.3 La valutazione della leggibilità per azienda sanitaria

La seconda parte del lavoro di tesi è stata la valutazione della leggibilità per azienda sanitaria locale, con lo scopo di capire se i testi appartenenti ad una determinata azienda risultano essere più complessi rispetto agli altri a prescindere dalla specialità a cui appartengono. Non per tutti i testi all'interno del corpus è stato possibile ricavare l'azienda sanitaria da cui sono stati rilasciati e per questo motivo l'analisi è stata condotta su 4 delle 12 ASL esistenti nella regione Toscana, quelle

che contengono un numero più consistente di testi. Gli identificativi delle aziende sanitarie sono stati anonimizzati e per questo motivo si parla di ASL A, B, C e D. Rispetto ai quattro modelli di analisi, la complessità dei testi è stata monitorata rispetto ai modelli BASE, LESSICALE e SINTATTICO.

Specialità	**N° documenti**	**N° token**	READ-IT		
			BASE	**LESSICALE**	**SINTATTICO**
ASL A	207	153.957	58,40	96,88	73,18
ASL B	67	59.696	57,40	91,88	82,54
ASL C	28	30.781	68,13	95	84,5
ASL D	16	7.317	64,69	85,11	88,99
Totale	**318**	**251.751**	**62,16**	**92,22**	**82,30**

Tabella 2.5: Statistiche generali e risultati della valutazione della leggibilità secondo i modelli BASE, LESSICALE e SINTATTICO dei testi organizzati per aziende sanitarie locali.

La (Tabella 2.5) riporta le statistiche generali e i risultati della valutazione della leggibilità dei testi organizzati per azienda sanitaria. Il sottocorpus è composto da 318 documenti, per un totale di 251.751 token. Il numero di testi e il numero di token varia tra le quattro ASL prese in esame: quelle con più testi a disposizione sono la A con un totale di 207 testi e 153.957 token e la B con un totale di 67 testi e 59.969 token, invece quella con meno testi a disposizione è la D con 16 testi e 7.317 token.

Per quanto riguarda i risultati della valutazione della leggibilità, anche in questo caso il modello BASE non risulta essere affidabile: a questo livello le ASL risultano essere più facili rispetto agli altri due. In particolare a questo livello l'ASL più facile è la B (57,40) e la più difficile è la C (68,13). Inoltre i testi delle ASL sono più complessi al livello lessicale rispetto a quello sintattico. In particolare l'ASL A è la più difficile al primo livello (96,88) ma la più facile al secondo (73,18). Al livello lessicale l'ASL più facile è invece la D (85,11).

Per le due ASL che contengono il numero maggiore di moduli di consenso informato, vale a dire la A e la B, i testi sono stati inoltre suddivisi in specialità, in modo da poter valutare la leggibilità sia per ASL che per specialità. Le specialità

che le due ASL condividono sono quattro: chirurgia generale, ortopedia, cardiologia e neurologia. La valutazione della leggibilità è stata monitorata a livello lessicale e sintattico.

Specialità	ASL A		ASL B	
	LESSICALE	**SINTATTICO**	**LESSICALE**	**SINTATTICO**
Chirurgia generale	99,65	80,24	54,53	33,28
Ortopedia	90,02	54,57	99,74	98,13
Cardiologia	99,87	74,63	97,2	82,96
Neurologia	99,6	93,64	97,41	95,43

Tabella 2.6: Risultati della valutazione della leggibilità secondo i modelli LESSICALE e SINTATTICO dei documenti organizzati per specialità all'interno delle ASL.

Come si vede nella (Tabella 2.6), a livello lessicale i valori sono significativamente alti per tutte le specialità all'interno delle due aziende sanitarie e l'unica eccezione è data da chirurgia generale per l'ASL B (54,53), a cui si contrappone il secondo valore più alto tra le specialità dell'ASL A (99,65). Sempre a questo livello, per l'ASL B la specialità ad avere il valore più alto è ortopedia (99,74), che si contrappone invece al valore più basso tra le specialità dell'ASL A (90,02).

Al livello sintattico i valori sono più vari e la specialità più complessa per entrambe le ASL è neurologia (93,64 e 95,45). È interessante notare come la specialità a risultare più leggibile a questo livello è chirurgia generale (33,28) dell'ASL B, la stessa che risulta più facile a livello lessicale. Invece la specialità di chirurgia generale dell'ASL A è la seconda meno leggibile tra tutte (80,24). L'unica specialità dell'ASL A che risulta essere più facile rispetto a quelle dell'ASL B è ortopedia (rispettivamente 54,57 e 98,13).

2.4 Il lessico dei moduli di consenso informato

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. Il rapporto tra lingua comune e i diversi linguaggi settoriali è stato ampiamente

dibattuto nella letteratura italiana [Cavagnoli, 2007] e internazionale [Cabré, 1999]. Se in passato si mirava a definire un confine che li separava nettamente, oggigiorno si è andata affermando l'idea che i due tipi di linguaggio rappresentano i due poli di un continuum che si estende dalla lingua comune ai linguaggi settoriali e caratterizzato da una gamma di livelli intermedi.

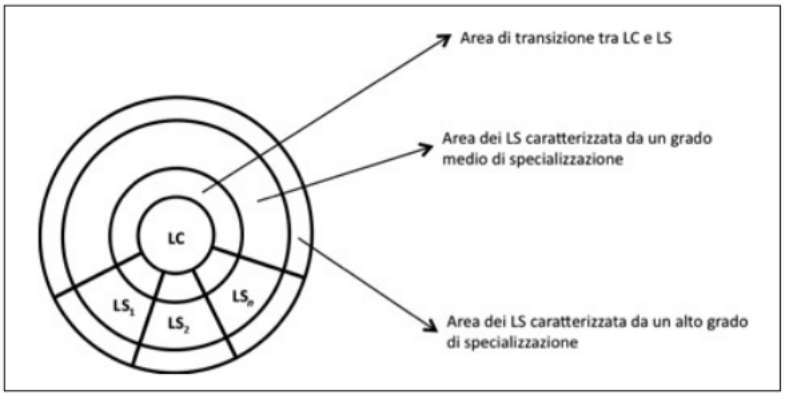


Figura 2.4: Rapporto tra lingua comune e linguaggi settoriali.

La Figura 4.3 riporta il grafico proposto da [Rondeau, 1983] che visualizza il complesso rapporto che lega lingua comune (LC) e linguaggi settoriali (LS) di cui si distinguono diverse varietà a seconda del grado di specializzazione. Ciascun settore corrispondente a un linguaggio settoriale presenta variazioni di tipo verticale, corrispondenti a diversi livelli di comunicazione, che vanno da uno altamente specialistico e specializzato comprensibile dagli esperti di dominio, a quelli più vicini all'utente comune comprensibili dagli esperti e non. A seconda dei diversi livelli di comunicazione, la terminologia usata per convogliare gli stessi contenuti varia in modo significativo, presentando insiemi di terminologia specialistica e lessico comune.

A partire dalle considerazioni riguardo il linguaggio medico discusse nel capitolo 1 e dato che i moduli di consenso informato risultano essere particolarmente difficili dal punto di vista lessicale a causa della presenza di un gran numero di parole specifiche del dominio medico che non sono solitamente usate nel linguaggio comune, si è deciso di approfondire lo studio del lessico e quindi della terminologia usata del corpus.

L'identificazione e il recupero da una collezione documentale di termini che sono ritenuti significativi rispetto al dominio al quale i documenti si riferiscono vengono usati per la creazione di vocabolari controllati e thesauri[15]. Un vocabolario controllato può essere costruito manualmente da esperti del dominio, oppure essere acquisito in modo semi-automatico facendo uso di metodi e tecniche di NLP. Nel secondo caso, il metodo più elementare è quello di scartare dal vocabolario sottostante a una collezione di testi, assunta come rappresentativa delle conoscenze relative a uno specifico settore del sapere, le cosiddette *stop-words* (ad esempio le parole semanticamente vuote come articoli, preposizioni, pronomi, ecc.): in questo caso però non tutte le parole ottenute sono rilevanti rispetto al dominio. Uno dei principali problemi di estrazione automatica di terminologia specialistica da corpora di dominio riguarda quindi la difficoltà di distinguere tra termini del dominio (linguaggio settoriale) e non-termini (lessico comune) e la soluzione ottimale è rappresentata dal ricorso a tecniche avanzate di filtraggio dei termini rilevanti [Montemagni, 2008].

Per affrontare questo problema si è deciso di estrarre, per ognuna delle 29 specialità, le liste dei termini 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, 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.

Come si vede nella (Tabella 2.7), la specialità che contiene la più alta percentuale di parole tipo non appartenenti al VdB è otorinolaringoiatria (56,2% su un totale di 5146 parole tipo) e, in particolare, quelle che compaiono nei testi con mag-

[15]Nel caso del thesaurus, il vocabolario controllato si arricchisce di relazioni semantiche tra i termini identificati.

Specialità	N°tipi	N° tipi non presenti nel VdB
Anestesia	1.934	694 (35,83%)
Chirurgia colo-rettale	624	216 (34,62%)
Chirurgia dell'obesità	1.437	587 (20,85%)
Chirurgia generale	1.492	705 (47,25%)
Chirurgia plastica	414	150 (36,23%)
Chirurgia toracica	648	254 (39,2%)
Chirurgia vascolare	2.348	1.134 (48,3%)
Oculistica	1.649	645 (39,11%)
Otorinolaringoiatria	5.146	2.892 (56,2%)
Ortopedia	3.905	1.869 (47,86%)
Ostetricia e ginecologia	3.046	1.394 (45,76%)
Urologia	1.624	727 (44,77%)
Totale: Area Chirurgica	**24.267**	**11.267 (46,43%)**
Cardiologia	3.427	1.637 (47,77%)
Diabetologia	157	40 (25,48%)
Gastroenterologia	1.240	471 (37,98%)
Neurologia	1.189	452 (38,02%)
Oncologia	558	147 (26,34%)
Pneumologia	896	324 (36,16%)
Senologia	1.865	804 (43,11%)
Totale: Area Medica	**9.332**	**3.875 (41,52%)**
Psicologia	745	189 (25,37%)
Screening	470	104 (23,42%)
Vaccini	752	256 (34,04%)
Totale: Area Prevenzione	**1.967**	**549 (27,91%)**
Genetica	1.128	515 (45,66%)
Immunoematologia e trasfusionale	2.454	1.068 (43,52%)
Medicina nucleare	2.049	929 (45,34%)
Radiologia	2.202	949 (43,1%)
Totale: Area Servizi	**7.833**	**3.461 (44,18%)**
Generici	**1.224**	**356 (29,08%)**
Pediatria	**1.398**	**534 (38,2%)**
Riabilitazione e rieducazione funzionale	**290**	**81 (27,93%)**

Tabella 2.7: Percentuale dei lemmi non appartenenti al vocabolario di base dei documenti organizzati per specialità.

giore 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 percentuale più bassa è chirurgia dell’obesità (20,85% su un totale di 1437 parole tipo). Altre specialità caratterizzate da percentuali basse sono screening (23,42%) e psicologia (25,37%). Le specialità di otorinolaringoiatria e di chirurgia vascolare appartengono all’area chirurgica, che risulta essere la macro-area più difficile (46,43% su un totale di 24.267 parole tipo) mentre quelle di screening e psicologia appartengono all’area prevenzione che risulta essere la macro-area più facile (27,91% su un totale di 549 parole tipo).

La difficoltà significativa registrata al livello lessicale suggerisce che lo strumento READ-IT, addestrato su testi del genere giornalistico, necessita di essere specializzato al livello del vocabolario usato, il quale dovrebbe contenere anche una selezione dei termini di dominio da usare nei moduli di consenso informato per non penalizzare il livello di leggibilità. L’idea è stata quella di sottoporre le liste dei lemmi non appartenenti al vocabolario di base a due esperti in valutazione della qualità dell’assistenza sanitaria, in modo da poterli organizzare in tre classi diverse: 1)“parole specifiche del dominio”, cioè parole di ambito medico di cui non si può fare a meno per esprimere un concetto (ad esempio “infezione” o “articolazione”); 2) “parole tecniche”, cioè parole che sono specifiche del dominio ma che dovrebbero essere integrate con glosse (ad esempio “patologia” o “ipertensione”); e 3)“tecnicismi”, cioè parole che vanno penalizzate e che necessitano di essere sostituite con un sinonimo più semplice in modo da essere comprensibili al lettore inesperto (ad esempio “monitorizzare” o “escissione”). Lo scopo finale è la creazione di un vocabolario di base in ambito medico ripartito secondo i repertori d’uso Fondamentale Medico (classe 1), Alto Uso Medico (classe 2) e Specialistico Medico (classe 3), con cui sarà possibile raffinare lo strumento READ-IT e quindi migliorare la valutazione della leggibilità per i testi di dominio medico. Inoltre il vocabolario ottenuto potrà essere usato nel compito di semplificazione del testo: analizzando i termini appartenenti alle classi 2 e 3 e quindi integrandoli con le glosse e sostituendoli con i sinonimi,

sarà possibile semplificare i testi in modo da renderli più comprensibili ai lettori inesperti, rendendo possibile il passaggio del linguaggio settoriale usato nei moduli di consenso informato dal livello di alto grado di specializzazione ad uno di grado medio.

3

Verso l'analisi della leggibilità in diverse lingue

In questo capitolo è illustrata una metodologia di valutazione della leggibilità valida per più lingue a partire dall'indagine di quali sono le caratteristiche linguistiche del testo correlate con gli elementi di complessità del testo. In particolare è presentato un caso di studio su due lingue tipologicamente lontane: il basco e l'italiano.

Nel paragrafo 3.1, dopo una breve caratterizzazione delle due lingue prese in esame, è riportato il confronto dei principi su cui si basano gli strumenti per la valutazione della leggibilità per le due lingue. Nel paragrafo 3.2 è invece presentato un confronto linguistico alla ricerca di fenomeni di complessità linguistica comuni alle due lingue.

3.1 Un caso di studio: basco e italiano a confronto

In questo caso di studio volto a definire le basi per una metodologia di valutazione della leggibilità valida per più lingue si è posta l'attenzione su due lingue

tipologicamente lontane: il basco e l'italiano.

La scelta della lingua basca nasce dalla collaborazione tra l'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa e l'IXA Group, gruppo di ricerca dell'Università dei Paesi Baschi (UPV/EHU) che lavora nel campo della Linguistica Computazionale e del Trattamento Automatico del Linguaggio.

3.1.1 Breve caratterizzazione delle due lingue

Il basco è una lingua pre-indoeuropea isolata parlata attualmente nei Paesi baschi, cioè nel nord della Spagna (nel Paese Basco spagnolo: regione della Navarra e della Comunità autonoma dei Paesi Baschi) e nell'estremo sud-ovest della Francia (nel Paese Basco francese: dipartimento dei Pirenei Atlantici) e ha uno status diverso a seconda della provincia in cui è parlato. Oltre il basco standard si contano 5 dialetti più 2 quasi persi e un altro documentato. Viene parlato come madrelingua dal 20,3% dei baschi (547.747 di 2.648.998). Considerando anche i parlanti del basco come seconda lingua, è conosciuto da 714.136 persone.

L'italiano e il basco si presentano come due lingue tipologicamente lontane: una comparazione interlinguistica è stata presentata in [Atorino, 2011].

In generale un confronto tra il basco e l'italiano implica innanzitutto un riconoscimento della loro diversità tipologica. Secondo la classificazione tipologica tradizionale, le due lingue appartengono a gruppi differenti.

Dal punto di vista morfologico, il basco appartiene al gruppo delle lingue agglutinanti, mentre l'italiano fa parte del gruppo delle lingue flessive. Le caratteristiche fondamentali che definiscono una lingua agglutinante sono: a) la tendenza di ogni affisso a rappresentare una categoria grammaticale e b) l'invariabilità dei morfemi con i quali si rappresenta una categoria. Di conseguenza, le lingue agglutinanti sono facilmente segmentabili nei morfemi che costituiscono le parole e le parole constano di vari morfemi. In contrapposizione alla morfologia agglutinante, le lingue flessive presentano proprietà inverse: a) abbondanza di morfemi che rappresentano più di una categoria grammaticale e b) tendenza alla variabilità dei morfemi con cui si

rappresenta ogni categoria. Di conseguenza la segmentazione delle parole in morfemi è più difficile e le parole di solito contengono meno morfemi rispetto alle lingue agglutinanti.

Dal punto di vista sintattico, la distinzione tipologica più importante è quella che classifica le lingue del mondo secondo l'ordine dei principali costituenti della frase: il soggetto, il verbo e i complementi. Anche secondo questo parametro, il basco e l'italiano appartengono a due tipi di lingue distinti: il basco è SOV e l'italiano SVO. Quindi le lingue hanno la tendenza a situare il nucleo sempre nella stessa posizione rispetto ai complementi, indipendentemente dalla categoria lessicale del nucleo: lasciando da parte la posizione del soggetto, le lingue VO hanno la testa a sinistra del complemento e quello OV hanno il testa a destra.

La nozione di testa non è legata esclusivamente alla struttura sintattica: anche nella struttura delle parole c'è un elemento che funziona come testa, nel senso che è il morfema che determina la categoria e buona parte delle proprietà della parola complessa. Nelle lingue del mondo c'è una chiara tendenza alla suffissazione, indipendentemente dall'ordine sintattico che manifesta la lingua: sia il basco che le lingue romanze utilizzano la suffissazione come processo morfologico altamente produttivo. Il processo di formazione delle parole attraverso la prefissazione invece sembra essere molto più legato all'ordine sintattico e quindi alla posizione della testa: le lingue con testa finale nella sintassi sono meno propense alla prefissazione. Quindi, una delle caratteristiche della morfologia lessicale basca è l'assenza quasi totale di prefissi, perfettamente coerente con l'ordine sintattico, trattandosi di una lingua con la testa in posizione finale.

3.1.2 Confronto tra gli strumenti di analisi della leggibilità per il basco e l'italiano

In [Gonzalez-Dios et al., 2014] è presentato lo strumento di valutazione della leggibilità per la lingua basca, *ErreXail*.

Prima della creazione di *ErreXail*, per calcolare la complessità dei testi baschi non sono state usate metriche specifiche. L'unica eccezione è data da un sistema per la valutazione della saggistica, *Idazlanen Autoebaluaziorako Sistema* (IAS) [Aldabe et al., 2012], il quale include metriche simili a quelle usate nella valutazione della leggibilità. IAS analizza i testi baschi con l'obiettivo di migliorarli attraverso l'uso di diversi criteri come la correzione del numero della clausole all'interno delle frasi, dei tipi di frasi e parole, del numero di lemmi, ecc. Lo strumento proposto intende aggiungere ad IAS la capacità di valutare la complessità dei testi aggiungendo nuove caratteristiche linguistiche.

Come READ-IT è stato sviluppato in vista della semplificazione del testo, *ErreXail* basco costituisce il modulo di pre-elaborazione per un sistema di semplificazione del testo.

Una caratteristica che *ErreXail* condivide con READ-IT riguarda l'approccio alla questione del calcolo della leggibilità come un problema di classificazione: anche lo strumento basco assegna il documento analizzato ad una specifica classe di leggibilità (facile vs difficile). Inoltre entrambi i sistemi sono basati su Support Vector Machine, ma usano algoritmi diversi: come è stato visto nel paragrafo 2.3.1, quello italiano usa LIBSVM [Chang and Lin, 2011], mentre quello basco SMO [Platt, 1998]. Quindi anche il sistema basco, dato un set di parametri e un training corpus, crea un modello statistico usando le caratteristiche linguistiche del corpus.

Al contrario di READ-IT, allenato su corpora appartenenti al genere della prosa giornalistica, il training corpus di *ErreXail* è formato da due corpora che appartengono al genere della prosa scientifica. Il primo corpus, inteso come quello difficile, è composto da 200 testi estratti dall'*Elhuyar aldizkaria*[16], una rivista mensile scientifica. Il secondo corpus, inteso come quello facile, è invece composto da 200 testi estratti da *ZerNola*[17], un sito web creato per la diffusione della scienza ai bambini fino a 12 anni.

[16] http://aldizkaria.elhuyar.org/
[17] http://www.zernola.net/

Entrambi gli strumenti si basano su una metodologia di valutazione della leggibilità che tiene conto di un'analisi linguistica multi-livello. Al contrario gli strumenti di analisi linguistica automatica e gli schemi di annotazione sono diversi.

I vari livelli di analisi linguistica per il basco sono esplorati usando diversi strumenti:

1. Analisi morfo-sintattica con *Morpheus* [Alegria et al., 2002]
2. Lemmatizzazione e identificazione delle funzioni sintattiche con *Eustagger* [Aduriz et al., 2003]
3. Identificazione degli elementi multi-parola [Alegria et al., 2004a]
4. Riconoscimento e classificazione delle entità nominate con *Eihera* [Alegria et al., 2004b]
5. Riconoscimento di strutture linguistiche parziali con *Ixati* [Aduriz et al., 2004]
6. Determinazione dei confini della frase e della clausola con *MuGak* [Aranzabe et al., 2013]
7. Identificazione dell'apposizione [Gonzalez-Dios et al., 2013]

Questa pre-elaborazione è stata quindi necessaria per effettuare l'analisi di diversi gruppi di caratteristiche linguistiche: globali, lessicali, morfologiche, morfo-sintattiche, sintattiche e pragmatiche per un totale di 94 caratteristiche:

- **Caratteristiche globali**: prendono in considerazione il testo nel suo complesso e sono basate sulle formule tradizionali della leggibilità, come la lunghezza media della frase e delle parole.
- **Caratteristiche lessicali**: sono basate sui lemmi e comprendono la distribuzione di tutte le parti del discorso (PoS) e tipi differenti di abbreviazioni e simboli, focalizzandosi su particolari tipi di sostantivi e verbi per un totale di 39 caratteristiche.

- **Caratteristiche morfologiche**: analizzano le diverse forme che i lemmi possono assumere, per un totale di 24 caratteristiche.

- **Caratteristiche morfo-sintattiche**: sono basate sul riconoscimento di strutture linguistiche parziali (chunk) e sull'identificazione delle apposizioni, per un totale di 5 caratteristiche. Contrariamente alle caratteristiche presentate finora, quelle morfo-sintattiche prendono in considerazione principalmente sequenze di parole.

- **Caratteristiche sintattiche**: considerano la media e i tipi delle clausole subordinate per un totale di 10 caratteristiche.

- **Caratteristiche pragmatiche**: sono costituite dagli elementi di coesione, ad esempio i diversi tipi di congiunzione, per un totale di 12 caratteristiche.

Per quanto riguarda le caratteristiche linguistiche prese in considerazione nella valutazione della leggibilità, una delle principali differenze riscontrate tra *ErreXail* e READ-IT riguarda le caratteristiche sintattiche: lo strumento basco non include caratteristiche sintattiche più profonde usate invece dallo strumento italiano che, come discusso nel capitolo 2 , fa affidamento su caratteristiche ricavabili a partire da un'analisi sintattica completa della frase.

Le caratteristiche linguistiche per cui si è trovata una corrispondenza tra il basco e l'italiano sono 21. Oltre le caratteristiche linguistiche di base, vale a dire la lunghezza media delle frasi e delle parole, le altre caratteristiche in comune sono principalmente morfo-sintattiche: la distribuzione delle parti del discorso (PoS), i modi e i tempi dei verbi e la densità lessicale. L'unica caratteristica lessicale riscontrata per entrambe le lingue è la Type Token Ratio (TTR). A livello sintattico invece le caratteristiche in comune sono rappresentate dal numero medio di clausole per frase e la distribuzione dei verbi per numero di dipendenti istanziati. Altre caratteristiche sono molto simili ma anche troppo generali o specifiche per le due lingue.

3.2 Confronto linguistico di corpora di consensi informati in italiano e basco

In questo paragrafo è presentato un confronto tra le caratteristiche linguistiche di corpora non paralleli ma comparabili di moduli di consenso informato in italiano e basco. Il confronto è stato condotto rispetto a corpora di lingua standard, alla ricerca di caratteristiche linguistiche correlate alla complessità del testo. Dato che queste caratteristiche sono quelle utilizzate dagli strumenti di analisi automatica della leggibilità, questo tipo di confronto pone le basi per una futura metodologia di valutazione della leggibilità valida per più lingue.

In particolare nel paragrafo 3.2.1 è presentato il corpus di moduli di consenso informato in basco che è stato preparato per il confronto e quindi i risultati ottenuti (paragrafo 3.2.2). Infine nel paragrafo 3.2.3 sono elencati alcuni esempi interessanti estratti dai documenti in basco.

3.2.1 Il corpus di consensi informati baschi

Specialità	Nº documenti	Nº parole (tokens)
Cardiologia	2	761
Gastroenterologia	1	345
Oculistica	1	304
Ostetricia e ginecologia	1	439
Otorinolaringoiatria	4	1.473
Pediatria	1	408
Vaccini	1	391
Totale	**11**	**4.121**

Tabella 3.1: Statistiche generali dei consensi informati baschi

Per effettuare il confronto delle caratteristiche linguistiche dei moduli di consenso informato in basco e in italiano è stato necessario per prima cosa la raccolta di un corpus: 11 documenti bilingui paralleli spagnolo - basco, di cui è stata estratta solo la parte in basco per un totale di 4.121 token, organizzati in specialità per renderlo comparabile al corpus italiano che si è presentato nel capitolo 2. Questi moduli di

consenso informato baschi sono stati preparati dal centro sanitario basco Osakindetza[18], configurato come un organismo autonomo per fornire assistenza sanitaria completa ai cittadini.

I documenti sono stati scritti seguendo una guida pratica per la redazione dei moduli di consenso informato, [Marijuan et al., 1998]. Nella scrittura dei documenti i professionisti hanno seguito una struttura fissa:

1. Identificazione
    - Ospedale - Servizio
    - Medico responsabile
    - Paziente
2. Tipo di procedura
    - Descrizione della procedura
    - Benefici / vantaggi attesi
    - Rischi (più rari e gravi / più frequenti e più o meno gravi)
    - Disagi ed effetti collaterali a breve, medio e lungo termine
    - Possibili alternative
    - Disponibilità dei professionista a chiarire dubbi e/o fornire ulteriori informazioni
    - Espressione della libertà di scelta o riconsiderazione della decisione presa
3. Firme
    - Medico responsabile
    - Paziente

[18]http://www.osakidetza.euskadi.eus/

- Nel caso di impossibilità del paziente: tutor, responsabile legale o persona designata dal paziente

Oltre all'attenzione al contenuto, la guida alla scrittura dei consensi informati descrive anche tre parametri da utilizzare per una revisione della comprensibilità del testo, assumendo che un documento corretto nel suo contenuto ma difficile nella comprensione è inadeguato, allo stesso modo in cui un documento facile da comprendere ma poco veritiero è eticamente e legalmente inaccettabile. I tre parametri sono il lettore esterno, il formato, gli indici di leggibilità.

I documenti dovrebbero essere valutati da un lettore esterno e inesperto che dovrebbe dare la propria opinione sulla comprensibilità del testo. Il lettore esterno può aiutare a migliorare il testo dicendo se il testo informa senza spaventare e se dispone di informazioni sufficienti per aiutare il paziente a prendere le decisioni appropriate o se, al contrario, è causa di confusione.

Invece il formato del documento facilita od ostacola la lettura. Per questo motivo Osakidetza ha inviato a tutti i suoi centri un tipo di formato che consente di unificare, per quanto possibile, i moduli di consenso informato, con l'obiettivo di raggiungere una certa omogeneità.

Infine, l'ultimo metodo per verificare se i documenti sono comprensibili ad un lettore medio è rappresentato dallo studio della leggibilità. Un metodo usato per valutare la leggibilità di un testo in spagnolo è "Grammatik" del programma Word Perfect [Word Perfect, 1994], disponibile ai professionisti nei Comitati Etici degli ospedali. Il programma permette di confrontare diverse metriche rispetto ad altri testi noti: numero delle parole, numero delle frasi, indice di Flesch-Kincaid e indice della complessità sintattica della frase. In particolare quest'ultimo realizza un'analisi della struttura sintattica per ognuna delle frasi, assumendo che un testo con più frasi subordinate sia più difficile: il valore 0 indica un testo facile e 100 un testo difficile. Con gli ultimi due indici è possibile calcolare l'indice Legin [Lorda et al., 1996], il cui valore oscilla tra 0 e 200, dove 0 indica un testo difficile e 200 un testo facile.

$$LEGIN = (100 + \textit{Indice di Flesch-Kincaid} - \textit{Indice della complessità sintattica})$$

Affinché il documento sia leggibile da un pubblico generale, gli autori assumono che deve avere un indice di Flesch-Kincaid superiore a 10, un indice di complessità sintattica delle frasi minore di 14 e soprattutto un valore LEGIN maggiore di 70. Quando un testo ha valori di Flesch-Kincaid inferiori a 10 è necessario rivedere la dimensione delle parole e delle frasi. Se il testo ha una complessità sintattica delle frasi superiore a 40 è necessario rivedere il numero di frasi composte, coordinate e subordinate, e cercare di semplificarle. Il testo andrebbe modificato fin quando si raggiunge un punteggio accettabile.

Questi indici facilitano solo la valutazione in modo obiettivo della leggibilità di un documento, cioè se sono facili da leggere dal pubblico inesperto, e fanno affidamento unicamente su caratteristiche generali e formali del testo che sono facili da calcolare e usare nonostante un primo passo verso l'analisi della struttura sintattica. Non fornendo quindi risultati ottimali sulla valutazione della difficoltà di lettura di un testo, si è deciso di analizzare il corpus con *ErreXail*, lo strumento per la valutazione della leggibilità di testi baschi descritto nel paragrafo 3.1.2, con il quale è stato possibile valutare le caratteristiche linguistiche principali che caratterizzano i moduli di consenso informato in basco.

3.2.2 Risultati del confronto

Nel confronto tra le caratteristiche linguistiche comuni degli strumenti per la lingua basca e italiana per la valutazione della leggibilità di moduli di consenso informato nelle due lingue sono state analizzate le specialità che nel corpus basco contengono più testi, vale a dire cardiologia e otorinolaringoiatria. I testi appartenenti alle due specialità sono stati confrontati con quelli equivalenti all'interno del corpus italiano presentato nel capitolo 2.

Come è stato visto nel paragrafo precedente i documenti in basco sono stati scritti seguendo dei parametri per la revisione della comprensibilità. Questo dato di fatto rende il confronto più difficile poiché la scrittura dei consensi informati italiani non è stata accompagnata da controlli di questo tipo.

	Cardiologia	**Otorinolaringoiatria**			
	Cateterismo	Adenoidectomia	Otoplastica	Tonsillectomia	Exeresi
Testi in basco	2	1	1	1	1
Testi in italiano	1	2	1	2	3

Tabella 3.2: Composizione dei testi in basco e in italiano presi in analisi

La (Tabella 3.2) riporta la composizione dei testi in basco e in italiano presi in analisi. I documenti trattano lo stesso argomento: la specialità di cardiologia contiene moduli di consenso informato sull'operazione di cateterismo cardiaco (2 per il basco e 1 per l'italiano), mentre quella di otorinolaringoiatria sulle operazioni di adenoidectomia (1 per il basco e 2 per l'italiano), otoplastica (1 per il basco e 1 per l'italiano), tonsillectomia (1 per il basco e 2 per l'italiano) ed exeresi di cisti del dotto tireoglosso (1 per il basco e 3 per l'italiano).

Cardiologia						
Caratteristiche	**Basco**			**Italiano**		
	ConInf	***Gior***	***ProSc***	***ConInf***	***Gior***	***ProSc***
Lunghezza media delle frasi	19,2	15,92	21,42	15,29	22,9	27,19
Lunghezza media delle parole	4,37	4,64	5,26	6,12	5,09	5,57
TTR	0,23	0,32	0,33	0,71	0,63	0,66
Distribuzione delle PoS:						
- nomi	39,1%	40,57%	42,8%	27,8%	28,29%	28,53%
- verbi	10,87%	12,82%	14,07%	11,1%	13,3%	10,67%
- aggettivi	9,31%	4,86%	4,8%	4,98%	6,17%	8,8%
- avverbi	3,68%	3,71%	4,2%	4,98%	4,18%	4,09%
- congiunzioni	4,72%	2,43%	3,2%	3,74%	3,65%	3,69%
Densità lessicale	0,80	0,81	0,85	0,57	0,56	0,58
Numero medio di clausole per frase	2,98	2,22	3,12	1,5	2,37	2,41
Distribuzione delle teste verbali per numero di dipendenti istanziati:						
- arità 1	21,75%	16,57%	19,6%	40,48%	28,03%	29,23%
- arità 2	12,85%	27,29%	21,6%	23,81%	31,84%	33,34%
- arità 3	0,52%	2,43%	0,74%	19,05%	20,47%	20,59%
Modi verbali:						
- indicativo	33,18%	30,14%	25,8%	21,43%	31,13%	27,09%
Tempi verbali:						
- presente	39,32%	26,86%	28%	28,95%	55,63%	54,33%
- passato	0%	18,57%	11,4%	1,75%	2,85%	1,34%
- futuro	3,69%	2,43%	1,14%	36,84%	3,57%	0,97%

Tabella 3.3: Selezione delle caratteristiche linguistiche comuni che caratterizzano maggiormente i consensi informati in basco e in italiano: specialità di cardiologia.

Otorinolaringoiatria						
Caratteristiche	**Basco**			**Italiano**		
	ConInf	***Gior***	***ProSc***	***ConInf***	***Gior***	***ProSc***
Lunghezza media delle frasi	24,14	15,92	21,42	12,87	22,9	27,19
Lunghezza media delle parole	5,13	4,64	5,26	7,02	5,09	5,57
TTR	0,37	0,32	0,33	0,69	0,63	0,66
Distribuzione delle PoS:						
- nomi	41,77%	40,57%	42,8%	25,06%	28,29%	28,53%
- verbi	14,56%	12,82%	14,07%	10,73%	13,3%	10,67%
- aggettivi	9,58%	4,86%	4,8%	5,99%	6,17%	8,8%
- avverbi	3,5%	3,71%	4,2%	4,57%	4,18%	4,09%
- congiunzioni	4,47%	2,43%	3,2%	3,75%	3,65%	3,69%
Densità lessicale	0,9	0,81	0,85	0,59	0,56	0,58
Numero medio di clausole per frase	4,22	2,22	3,12	1,22	2,37	2,41
Distribuzione delle teste verbali per numero di dipendenti istanziati:						
- arità 1	27,82%	16,57%	19,6%	35,73%	28,03%	29,23%
- arità 2	9,28%	27,29%	21,6%	31,31%	31,84%	33,34%
- arità 3	0,76%	2,43%	0,74%	13,53%	20,47%	20,59%
Modi verbali:						
- indicativo	31,04%	30,14%	25,8%	17,79%	31,13%	27,09%
Tempi verbali:						
- presente	39,41%	26,86%	28%	43,5%	55,63%	54,33%
- passato	0,66%	18,57%	11,4%	1,22%	2,85%	1,34%
- futuro	0,65%	2,43%	1,14%	10,3%	3,57%	0,97%

Tabella 3.4: Selezione delle caratteristiche linguistiche comuni che caratterizzano maggiormente i consensi informati in basco e in italiano: specialità di otorinolaringoiatria.

La (Tabella 3.3) e la (Tabella 3.4) riportano una selezione delle caratteristiche linguistiche comuni che caratterizzano maggiormente il corpus basco e il corpus italiano per le due specialità prese in analisi: cardiologia e otorinolaringoiatria. Per ognuna delle caratteristiche sono riportati i valori riscontrati nei moduli di consenso informato (*ConInf*) e quelli dei corpus rappresentanti generi testuali tradizionali, vale a dire prosa giornalistica (*Gior*) e scientifica (*ProSc*). Il confronto con quest'ultimi ha lo scopo di individuare le caratteristiche linguistiche correlate alla complessità che caratterizzano i moduli di consenso informato rispetto a varietà linguistiche standard. In particolare per l'italiano sono stati utilizzati gli stessi corpora che sono stati descritti nel paragrafo 2.2.2. Invece per il basco, il primo corpus è composto da

7 testi nel dominio giornalistico (per un totale di 1036 token) e il secondo da 5 testi nel dominio di divulgazione scientifica sulla medicina (per un totale di 1113 token), diversi da quelli su cui è stato addestrato *ErreXail*.

Per quanto riguarda le caratteristiche di base, vale a dire il numero medio di parole per frase e di caratteri per parola, i risultati mostrano che i moduli di consenso informato in basco, al contrario di quelli italiani, sono costituiti da frasi più lunghe (19,2 contro 15,29 per cardiologia e 24,14 contro 12,87 per otorinolaringoiatria) ma da parole più corte (4,37 contro 6,12 per la prima e 5,13 contro 7,02 per la seconda). Queste due caratteristiche sono solamente descrittive, in quanto questi tipi di valori vanno confrontati con corpora rappresentanti diverse varietà della stessa lingua. Per quanto riguarda la lunghezza media delle frasi, i due valori per le due specialità in basco sono abbastanza differenti ma si avvicinano più a quello riscontrato nei testi della prosa scientifica (21,42) che a quello nei testi giornalistici (15,92). Per quanto riguarda invece la lunghezza media delle parole, i testi della specialità di cardiologia sono caratterizzati dal valore più basso in confronto agli altri due corpora (rispettivamente 4,64 e 5,26), mentre quelli di otorinolaringoiatria presentono un valore più vicino a quello del corpus di prosa scientifica (5,13).

Al livello lessicale, la caratteristica lessicale presa in analisi è la TTR e in questo caso i moduli di consenso informato baschi presentano valori più bassi (0,23 e 0,37) rispetto a quelli italiani (0,71 e 0,69) e quindi una minore varietà lessicale. A questo livello quindi i testi baschi risultano essere meno complessi in quanto sono composti da un numero inferiore di concetti, risultato che ci si aspetta per documenti che dovrebbero ripetere sempre le stesse cose: i valori riscontrati nei testi giornalistici e scientifici sono maggiori (rispettivamente 0,32 e 0,33).

Per quanto riguarda la distribuzione delle parti del discorso (PoS), i moduli di consenso informato baschi sono caratterizzati da una percentuale più alta di nomi (39,1% e 41,77%) e di aggettivi (rispettivamente 9,31% e 9,58%) rispetto a quelli italiani (rispettivamente 27,8% e 25,06% per i nomi e 4,98% e 5,99% per gli aggettivi). La percentuale di verbi è simile per cardiologia (10,87% contro 11,1%) ma

molto più diversa per otorinolaringoiatria in cui il basco presenta una percentuale più alta (14,56% contro 10,73%). Prendendo in considerazione le altre PoS, i moduli di consenso baschi presentano una percentuale più bassa di avverbi (3,68% e 3,5% contro 4,98% e 4,57%) e più alta di congiunzioni (4,72% e 4,47% contro 3,74% e 3,75%). Confrontando i valori dei moduli di consenso informato baschi con quelli dei corpora di altri generi testuali, le differenze maggiori si riscontrano per gli aggettivi e le congiunzioni: i testi giornalistici e scientifici presentano percentuali molto più basse (per gli aggettivi rispettivamente 4,86% e 4,8%, per le congiunzioni 2,43% e 3,2%). Infine, sia per i moduli di consenso basco e sia per quelli italiani si sono riscontrate percentuali molto basse di nomi propri, verbi modali, pronomi e abbreviazioni, che non sono state inserite nelle tabelle.

A partire dalla distribuzione delle singole categorie morfo-sintattiche, è stata misurata la densità lessicale: le percentuali più alte di nomi, verbi ed aggettivi nei testi baschi danno luogo a valori più alti (0,8 e 0,9 contro 0,57 e 0,56).

Per quanto riguarda le caratteristiche sintattiche, una prima caratteristica monitorata è data dal numero di clausole per frase: i moduli baschi presentano una media significativamente più alta rispetto a quelli italiani (rispettivamente 2,98 e 4,22 contro 1,5 e 1,22). Quindi se a livello lessicale i testi baschi risultano essere più facili rispetto a quelli italiani, questo primo risultato mostra che la loro maggiore complessità risiede a livello sintattico.

Un'altra caratteristica sintattica presa in analisi è data dalla distribuzione delle teste verbali per numero di dipendenti istanziati: si nota che i testi in italiano hanno un numero significativamente più alto di teste verbali mono-argomentali rispetto a quelli in in basco (rispettivamente 40,48% e 35,73% contro 21,75% e 27,82%). Ma è interessante notare che i testi baschi presentano anche una percentuale minore di verbi con valenza 2 e 3 (rispettivamente 12,85% - 0,52% e 9,28% - 0,76% contro 23,81% - 19,05% e 31,31% - 13,53%), essendo quindi caratterizzati da verbi con valenza $\geq$3. In questo caso questi tipi di testi si differenziano significativamente dal genere giornalistico, i cui testi sono caratterizzati da una percentuale di verbi

mono-argomentali minore (16,57%) ma maggiore di verbi con valenza 2 e 3 (27,29% e 2,43%). Anche i testi scientifici sono caratterizzati da una percentuale dei primi minore (19,6%), ma per quanto riguarda i secondi, quella di verbi con valenza 2 è maggiore (21,6%) ma quelli di valenza 3 presentano un valore simile (0,74%).

Infine, per quanto riguarda i modi e i tempi verbali, i moduli di consenso informato baschi presentano la percentuale più alta di verbi all'indicativo (33,18% e 31,04%), rispetto sia ai testi giornalistici e scientifici (rispettivamente 30,14% e 25,8%) della stessa lingua e sia rispetto ai documenti italiani (21,43% e 17,79%). Invece, a proposito dei tempi verbali si nota per entrambe le lingue una percentuale maggiore di verbi al presente (rispettivamente 39,32% e 39,41% contro 28,95% e 43,25%) e minore di verbi al passato (rispettivamente 0% e 0,66% contro 1,75% e 1,22%). Ciò che è interessante notare è che i testi giornalistici e scientifici in basco presentano percentuali alte di quest'ultimi (rispettivamente 18,57% e 11,4%), al contrario di quelli italiani (rispettivamente 2,85% e 1,34%). In questo caso ciò che distingue i corpora presi in analisi è la distribuzione di verbi al futuro: i testi baschi presentano una percentuale minima (3,69% e 0,65%) al contrario di quelli italiani, caratterizzati dall'occorrenza più alti di questo tipo di verbi (36,84% e 10,3%).

Dato che i corpora sono composti da moduli di consenso informato delle due lingue che trattano gli stessi argomenti, sono state confrontate le caratteristiche linguistiche per ognuno dei documenti, in modo da individuare le maggiori differenze.

Caratteristiche	***Catet***	
	B	**I**
TTR	0,23	0,71
Distribuzione delle PoS:		
- nomi	39,1%	27,8%
- verbi	10,87%	11,1%
- aggettivi	9,31%	4,98%
- avverbi	3,68%	4,98%
- congiunzioni	4,72%	3,74%
Densità lessicale	0,80	0,57
Media di clausole per frase	2,98	1,5
Distribuzione delle teste verbali per numero di dipendenti istanziati:		
- arità 1	21,75%	40,48%
- arità 2	12,85%	23,81%
- arità 3	0,52%	19,05%

Tabella 3.5: Selezione delle caratteristiche linguistiche comuni che caratterizzano maggiormente i testi della specialità di cardiologia.

Caratteristiche	***Aden***		***Otop***		***Tons***		***Exer***	
	B	**I**	**B**	**I**	**B**	**I**	**B**	**I**
TTR	0,34	0,70	0,39	0,7	0,38	0,68	0,36	0,66
Distribuzione delle PoS:								
- nomi	42,07%	24,79%	41,05%	25,08%	42,04%	25,51%	41,91%	24,87%
- verbi	14,94%	10,53%	14,94%	10,8%	14,59%	10,82%	13,78%	10,78%
- aggettivi:	9,57%	4,32%	11,29%	6,99%	9,24%	5,82%	8,22%	6,83%
- avverbi	3,78%	4,32%	4,96%	4,29%	1,27%	4,75%	3,98%	4,90%
- congiunzioni	4,53%	3,13%	3,31%	3,68%	4,46%	4,47%	5,57%	3,72%
Densità lessicale	0,91	0,60	0,93	0,59	0,87	0,58	0,87	0,59
Medie di clausole per frase	3,89	1,31	4,43	0,99	4,67	1,32	3,88	1,25
Distribuzione delle teste verbali per numero di dipendenti istanziati:								
- arità 1	31,2%	38,29%	27,34%	36,72%	29,36%	31,31%	23,39%	36,61%
- arità 2	7,09%	30,51%	10,16%	36,72%	11,01%	29,11%	8,87%	28,91%
- arità 3	0,71%	15,03%	2,34%	7,81%	0%	16,87%	0%	14,41%

Tabella 3.6: Selezione delle caratteristiche linguistiche comuni che caratterizzano maggiormente i testi della specialità di otorinolaringoiatria.

La (Tabella 3.5) e la (Tabella 3.6) riportano una selezione delle caratteristiche linguistiche comuni che caratterizzano maggiormente i testi delle specialità di cardiologia e otorinolaringoiatra, vale a dire cateterismo cardiaco (*Catet*), adenoidectomia (*Aden*), otoplastica (*Otop*), tonsillectomia (*Tons*) ed exeresi (*Exer*) per la lingua basca (*B*) e italiana (*I*).

Per quanto riguarda le caratteristiche lessicali, i valori di TTR per i testi baschi oscillano da un minimo di 0,23 (cateterismo cardiaco) ad un massimo di 0,39 (otoplastica), valori estremamente diversi rispetto a quelli nei testi italiani che oscillano da un minimo di 0,66 (exeresi) ad un massimo di 0,71 (cateterismo cardiaco): la differenza maggiore di riscontra quindi nel testo *Catet*. L'altra caratteristica lessicale monitorata è la densità lessicale: i valori dei testi baschi oscillano da un minimo di 0,80 (*Catet*) ad un massimo di 0,93 (*Otopl*) mentre quelli italiani da 0,57 (*Cater*) a 0,60 (*Aden*): in questo caso la differenza più significativa si registra per il documento sull'otoplastica (0,93 contro 0,59). Il fatto che *Otopl* sia caratterizzato dal valore più alto di densità lessicale è confermato dalla distribuzione delle parti del discorso (PoS), infatti anche se è caratterizzato da percentuali simili di nomi e verbi simili agli altri documenti appartenenti alla specialità di otorinolaringoiatria (rispettivamente 41,05% e 14,94%), registra quella più alta di aggettivi (11,29%) e avverbi (4,96%) e più bassa di congiunzioni (3,31%). Tra i testi baschi, quello che si discosta maggiormente è *Catet*, caratterizzato dalla percentuale minore di nomi (39,1%) e verbi (10,87%).

Per quanto riguarda le caratteristiche sintattiche, in particolare il numero di clausole per frase, i testi baschi della specialità di otorinolaringoiatria risultano essere più complessi rispetto a quello della specialità di cardiologia in quanto nei primi si riscontrano valori nell'intervallo 3,88 - 4,67 cui si oppone la media di 2,98 per il testo sul cateterismo cardiaco. La differenza maggiore tra i testi baschi ed italiani si riscontra per *Otop* (rispettivamente 4,43 e 0,99). L'altra caratteristica presa in considerata è la distribuzione delle teste verbali per numero di dipendenti istanziati: in questo caso *Catet* è il documento che ha la percentuale più bassa di teste verbali mono-argomentali (21,75%), cui si contrappongono i valori ottenuti per i testi della specialità di otorinolaringoiatria che presentano percentuali che oscillano da 23,39% (*Exer*) a 31,2% (*Aden*). Ma è interessante notare che invece il testo sul cateterismo cardiaco italiano è quello che ha la percentuale più alta di verbi con valenza 1 (40,48%). Il testo sull'otoplastica basco invece è quello che presenta la

percentuale più alta di verbi con valenza 3 (2,34%) che si contrappone al valore più alto riscontrato tra i testi italiani (7,81%).

3.2.3 Alcuni esempi

Prendendo in considerazione il corpus dei moduli di consenso informato basco sono state notate alcune caratteristiche interessanti.

I testi sono stati scritti nella lingua standard, anche se molti di essi presentano un mix dei dialetti parlati in Guipúzcoa e in Biscaglia dovuto al fatto che il centro sanitario Osakindetza che li ha prodotti si trova in una città al confine tra queste due province, Arrasate. Quindi molto probabilmente il traduttore è nativo di questa zona ed è stato influenzato dal proprio linguaggio. Un esempio è presente nel testo sull'otoplastica:

(1) *Interbentzioaren ostean, beharrezkoa da bendaje bat* **eroatea***; eta egun batzuen buruan kentzen da.*

"Dopo l'intervento, è necessario **portare** una benda; e viene rimossa in pochi giorni."

Il testo sulla tonsillectomia è inoltre caratterizzato da un linguaggio molto informale:

(2) **Oso oso** *gutxitan behar izaten da anestesia orokorra.*

"**Molto raramente** richiede l'anestesia totale."

Alcuni testi presentano al loro interno un mix di stile personale e impersonale, anche all'interno della stessa frase. Alcuni esempi sono presenti nel testo sul cateterismo cardiaco:

(3) a. *Izterrondoko arteria bat* **ziztatuko dizute,** *anestesia lokala erabiliz.*

"**Le faranno una puntura** in un'arteria dell'inguine, utilizzando l'anestesia locale"

b. *Ostean, bihotzeraino iritsiko den hodi estu bat* **sartuko da** *bertatik (kateterra).*

'Dopo, da quella parte **verrà messo** un canale stretto (il catetere) che arriverà fino al cuore."

c. *Hori egin ondoren, likido bat* **injektatzen da** *(kontrastea), eta honek, X Izpien bidez, zure bihotzaren funtzionamendua eta zure bihotzeko arterien egoera* **ikusten lagunduko digu.**

"Dopo aver fatto quello, **viene iniettato** un liquido (il contrasto), e questo, mediante raggi X, **ci aiuterà a vedere** lo stato del funzionamento del suo cuore e delle arterie del suo cuore."

In molti testi invece sono presenti delle informazioni complementari tra parentesi. Un esempio è presente nel testo riguardo l'operazione di tonsillectomia:

(4) *Amigdalektomiari esker,* **amigdalen hipertrofiak eragindako arazoak** *gutxituko dira* **(arnasteko eta irensteko),** *eta, halakorik balego, infekzio-foku bat ere desagerraraziko du.*

"Grazie alla tonsillectomia, i **problemi causati per l'ipertrofia delle tonsille** diminuiranno **(per respirare e per ingerire)**, ed eventualmente, farà scomparire anche la fonte dell'infezione."

Inoltre sono presenti anche delle informazioni complementari tra parentesi composte da spiegazioni usate nel linguaggio comune e informale. Un esempio è presente nel testo sull'otoplastica:

(5) *Aldaketa estetiko hori egiteko arrazoirik ohikoena da belarria aurrerantz bereizia izatea* **(belarri askeak, "hegal itxurakoak")**

"La ragione più comune per questo cambiamento è l'estetica dell'orecchio **(orecchio libero, "a forma di ali di farfalla")**"

4

Verso la semplificazione dei consensi informati

Uno degli scopi applicativi della valutazione automatica della leggibilità è rappresentata dalla semplificazione dei testi scritti. In questo capitolo sono presentati i primi risultati di una metodologia di semplificazione semi-automatica sperimentata nella riscrittura guidata di un consenso informato per il programma di fecondazione in vitro e iniezione intracitoplastica degli spermatozoi (ICSI).

Nel paragrafo 4.1 è presentato il modulo di consenso informato su cui sono stati condotti i primi passi di semplificazione. Nel paragrafo 4.2 sono riportati i risultati della valutazione automatica della leggibilità del consenso. Nel paragrafo 4.2.1 è descritto il questionario utilizzato per confrontare i valori della leggibilità del consenso rispetto alla risposta di pazienti reali che si sono sottoposti al programma di fecondazione in vitro. Infine nel paragrafo 4.3 è presentato l'approccio alla semplificazione seguito e discussa quella condotta per il paragrafo 5 del consenso.

4.1 Un caso di studio: il consenso informato per il programma di fecondazione in vitro

Il caso di studio presentato in questo lavoro di tesi finalizzato a fornire un esempio di semplificazione semi-automatica di testi appartenenti al dominio medico nasce dalla collaborazione dell'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa con il Centro Demetra di Firenze[19], centro che opera nel campo della Procreazione Medicalmente Assistita (PMA) e convenzionato con il Sistema Sanitario Nazionale per i cicli di trattamento. Il centro nasce come associazione di professionisti che si dedicano all'infertilità e alla medicina della riproduzione, con l'obiettivo di offrire alle coppie un'assistenza completa, dalla diagnosi alla terapia.

L'obiettivo della collaborazione è il supporto al miglioramento della comunicazione medico-paziente a partire dalla valutazione della leggibilità e dalla sua eventuale semplificazione del modulo di consenso informato per il programma di fecondazione in vitro e iniezione intracitoplastica degli spermatozoi (ICSI) distribuito dal centro.

Il documento preso in analisi è costituito da un'introduzione, 18 paragrafi e un'autocertificazione finale. La maggior parte dei paragrafi contengono articoli di legge che non possono essere sostituiti o modificati in quanto devono essere presenti a norma di legge. Per questo motivo la valutazione automatica della leggibilità è stata ristretta all'analisi di quattro paragrafi su cui era possibile effettuare la semplificazione: il paragrafo 5 che contiene una descrizione della tecnica riassumendone ogni fase della sua applicazione, il paragrafo 13 che riporta la spiegazione delle modifiche della legge riguardante i limiti dell'applicazione della tecnica sugli embrioni, il paragrafo 14 contenente la certificazione della possibilità di crionconservazione dei gameti maschili e femminili e infine il paragrafo 16 che riporta le modalità e le condizioni di crioconservazione degli embrioni.

[19] http://www.centrodemetra.com

4.2 La valutazione della leggibilità del consenso informato

	N° parole (tokens)	READ-IT		
		BASE	**LESSICALE**	**SINTATTICO**
Testo totale	**7.021**	**85,2**	**58,1**	**100**
Paragrafo 5	767	97,7	91,1	100
Paragrafo 13	422	94,5	19	100
Paragrafo 14	99	84,2	0,2	100
Paragrafo 15	257	98	55,6	100

Tabella 4.1: Risultati della valutazione della leggibilità secondo i modelli BASE, LESSICALE e SINTATTICO del consenso informato e dei paragrafi presi in analisi.

La valutazione della leggibilità è stata effettuata utilizzando lo strumento READ-IT, facendo riferimento in particolare ai modelli BASE, LESSICALE e SINTATTICO.

La (Tabella 4.1) riporta i risultati della valutazione della leggibilità ottenuti per il testo totale e per i paragrafi 5, 13, 14 e 15 del consenso informato. Il documento è composto in totale da 7.021 token ed è caratterizzato da valori di leggibilità differenti per quanto riguarda i modelli. Differenze significative si registrano anche nel confronto dei valori ottenuti per i paragrafi.

Rispetto al modello BASE, il quale fa riferimento alle caratteristiche del testo grezzo come la lunghezza delle frasi e delle parole, il testo totale presenta un valore abbastanza alto (85,2). I valori ottenuti per i paragrafi cadono invece nell'intervallo tra 84,2 (paragrafo 14) a 97,7 (paragrafo 5). Per quanto riguarda invece gli altri due modelli di READ-IT, è interessante notare che il documento è più difficile a livello sintattico che a quello lessicale.

Considerando i risultati ottenuti dal modello LESSICALE, il quale fa riferimento alla composizione del vocabolario e la sua ricchezza lessicale, il valore per il consenso informato totale è di 58,4 ma si notano differenze significative per quanto riguarda la suddivisione in paragrafi: a questo livello il 5 risulta essere particolarmente

difficile (91,1), a cui segue il 15 (55,6), il 13 (19) e infine il 14 che risulta essere particolarmente facile (0,2).

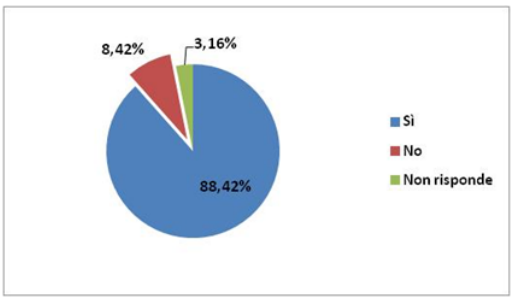
Invece i risultati ottenuti dal modello SINTATTICO, il quale fa riferimento a informazione di tipo grammaticale, non presentano variazioni: tutti i valori sono = 100 e quindi a questo livello il testo è estremamente complesso.

4.2.1 Il questionario per la valutazione della soddisfazione dei pazienti

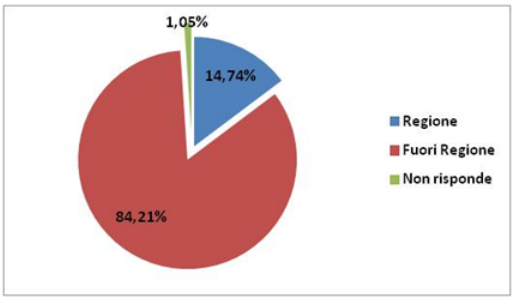
La novità dell'approccio proposto è costituita dal confronto tra la valutazione della leggibilità rispetto alla risposta di un gruppo di pazienti che si sono recati al Centro Demetra per seguire la procedura di procreazione assistita ottenuti in seguito alla somministrazione di un questionario anonimo finalizzato a monitorare il loro grado di soddisfazione. Il questionario contiene una serie di domande chiuse, divise in tre sezioni differenti.

La prima sezione del questionario è costituita da domande sul paziente. La seconda sezione riporta invece la valutazione di diversi fattori che vanno dall'assistenza, la cortesia e la preparazione del personale medico e tecnico alla chiarezza delle spiegazioni e dei paragrafi del consenso informato. Le possibili risposte che i pazienti possono dare sono: "molto soddisfatto", "soddisfatto", "poco soddisfatto" e "insoddisfatto". Infine, il questionario contiene una terza sezione per la raccolta di commenti liberi, volto a cogliere suggerimenti dagli intervistati.

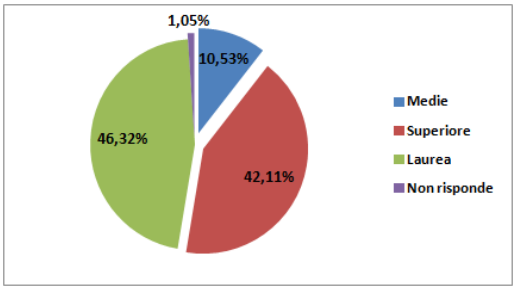
In totale sono stati erogati 95 questionari tra la metà di settembre e la metà di ottobre del 2016. Ai fini di questo studio sono state studiate le caratteristiche dei pazienti (prima sezione) e le domande che riguardano la chiarezza dei paragrafi del consenso informato considerati (all'interno della seconda sezione), con lo scopo di confrontarli con i risultati della valutazione automatica della leggibilità.



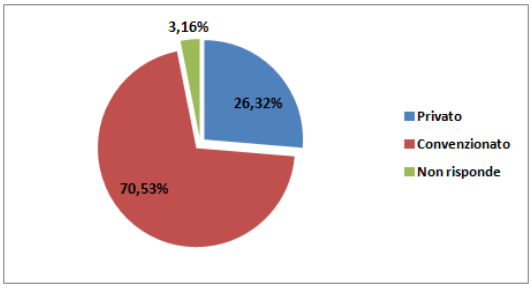
(a) Madrelingua italiana.



(b) Canale di accesso.



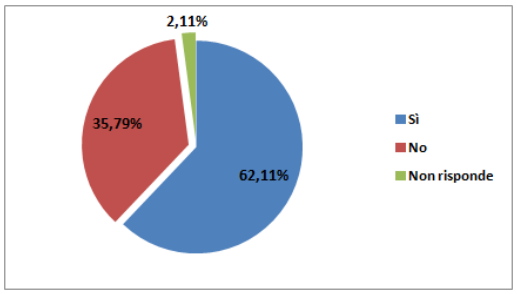
(c) Titolo di studio



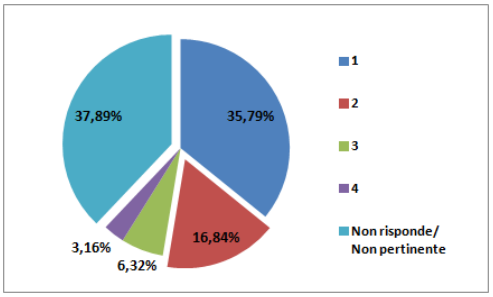
(d) Tipologia di prestazione.

Figura 4.1: Caratteristiche dei pazienti.

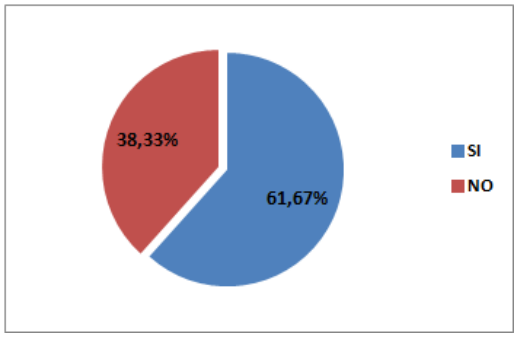
La (Figura 4.1) riporta le caratteristiche del campione di pazienti a cui è stato sottoposto il questionario. Tra le 95 persone a cui è stato somministrato il questionario, l'88,42% è madrelingua italiana e il 14,74% proviene da una regione diversa dalla Toscana. Per quanto riguarda il titolo di studio posseduto, il 46,32% dei pazienti possiede la laurea e il 42,11% il diploma superiore. In questo caso è interessante notare che solo il 10,53% degli intervistati possiede un diploma di scuola media. Inoltre la maggiorparte dei pazienti possiede una convenzione per il trattamento (70,53%).



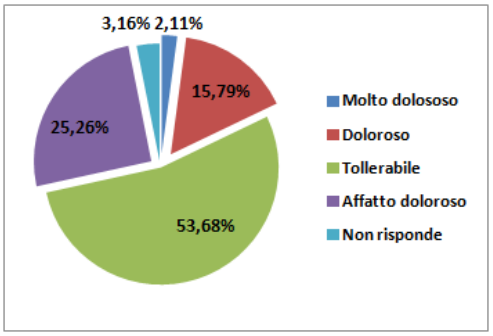
(a) Sottomissione ad altri cicli.



(b) Numero di cicli



(c) Stesso centro.



(d) Percezione dell'intervento di aspirazione ovarica.

Figura 4.2: Informazioni sul trattamento di fecondazione assistita dei pazienti.

Facendo riferimento alle caratteristiche dei pazienti in relazione alla procedura da effettuare (Figura 4.2), il 62,11% si è già sottomesso ad altri cicli, mentre per il 35,79% è la prima volta. In particolare per quanto riguarda i primi, il 37,89% si è sottoposto ad un ciclo, il 16,84% a due, il 6,32% a tre e 3,16% a quattro. In questo caso il 35,79% ha deciso di non rispondere considerando la domanda non pertinente. Inoltre il 61,67% si è sottomesso agli altri cicli sempre al Centro Demetra. Per quanto riguarda invece la percezione del dolore per l'intervento di aspirazione ovarica, le percentuali più alte si riscontrano per i pazienti che lo hanno trovato tollerabile e affatto doloroso (rispettivamente 53,68% e 25,26%), invece il 15,79% e il 2,11% affermano sia stato doloroso e molto doloroso.

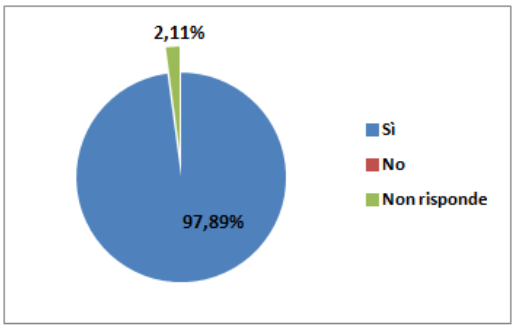


Figura 4.3: Raccomandazione del centro ad altri pazienti.

Infine, una valutazione molto importante riguarda la percentuale di pazienti che raccomanda il centro ad altre persone: in questo il 97,89% ha dato una risposta positiva.

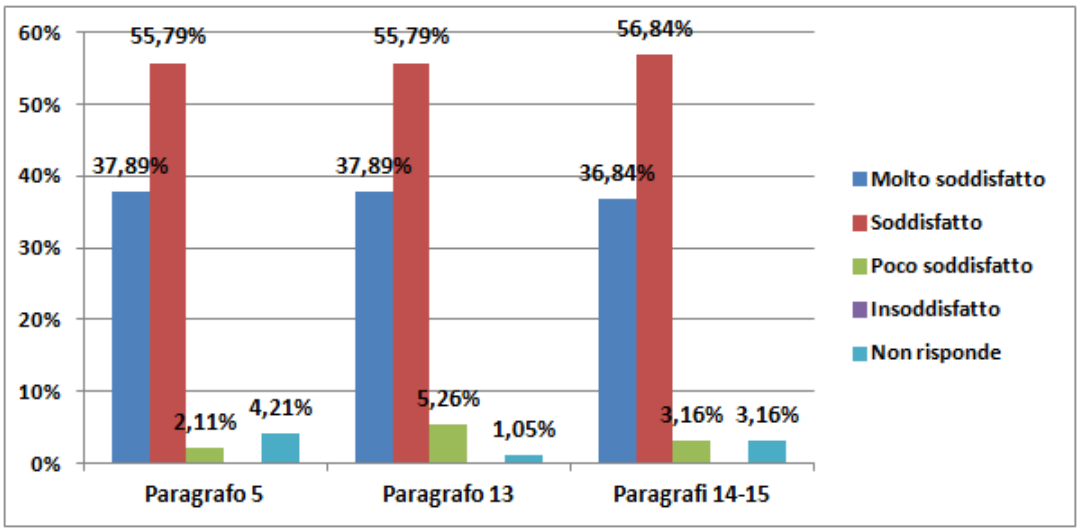


Figura 4.4: Chiarezza dei paragrafi del consenso informato.

Invece per quanto riguarda la percezione della chiarezza dei paragrafi 5, 13 e 14-15, i risultati ottenuti sono mostrati nella Figura 4.4. La maggior parte dei pazienti risulta essere "soddisfatto": le percentuali sono rispettivamente di 55,79% per i paragrafi 5 e 13 e di 56,84% per i 14-15. Le percentuali di pazienti "molto soddisfatti" sono invece minori (rispettivamente 37,89% per i primi due e 36,84%

per i terzi). Per quanto riguarda invece le percentuali di pazienti "poco soddisfatti" si notano delle differenze un po' più significative: in questo caso è il paragrafo 13 che risulta essere meno chiaro (5,26%) al contrario del 5 che presenta invece la percentuale più bassa (2,11%). Nessuno dei 95 partecipanti al questionario risulta essere invece "non soddisfatto". Infine è interessante notare che anche se il paragrafo 5 è quello con la percentuale più bassa di persone "poco soddisfatte" è anche quello con la percentuale più alta di pazienti che non hanno voluto esprimere il loro parere (4,21%). Considerando i risultati avuti con il paragrafo 13, la situazione è opposta: la percentuale di pazienti "poco soddisfatti" è più alta e solo l'1,05% ha preferito non rispondere.

Confrontando questi risultati con quelli ottenuti dalla valutazione della leggibilità descritta nel paragrafo 4.2, si nota che il paragrafo 5 è quello che ha i valori più alti per tutti i modelli presi in considerazione da READ-IT ma che ha la percentuale più bassa di pazienti poco soddisfatti. Quindi anche se per la valutazione automatica della leggibilità risulta essere il più difficile, la percezione dei pazienti è molto più positiva, al contrario invece del paragrafo 13 che ha la percentuale più alta di pazienti poco soddisfatti ma che ha un valore molto basso per il modello lessicale (19) e quindi risulta essere più facile. Invece i paragrafi 14-15 hanno la percentuale più alta di pazienti soddisfatti e risultano essere più facili a livello lessicale (in particolare il 14 presenta un valore di 0,2).

Questo confronto mostra quindi che sebbene i risultati della valutazione di leggibilità siano bassi i giudizi della soddisfazione della chiarezza dei paragrafi del consenso informato sono molto alti. Questo può essere letto alla luce del fatto che i pazienti a cui è stato somministrato il questionario hanno un livello di istruzione medio-alto mentre invece lo strumento READ-IT, come si è visto nel capitolo 2, si rivolge ad un pubblico caratterizzato da un basso livello di alfabetizzazione.

In futuro il questionario sarà nuovamente somministrato ad un altro gruppo di pazienti simili per valutare la loro soddisfazione dopo la semplificazione del consenso informato.

4.3 La semplificazione del consenso informato

Come si è visto, la valutazione automatica della leggibilità rappresenta il primo passo verso la semplificazione di testi scritti. La maggior parte della ricerca si è focalizzata sulla valutazione della leggibilità a livello di documento rispetto che a livello di frase. Tuttavia, valutare la leggibilità anche a questo livello risulta importante quando il compito perseguito è la semplificazione del testo. Lo scopo della semplificazione del testo è di ridurre la complessità lessicale e sintattica preservando il significato originale del testo.

Nel paragrafo 4.3.1 è presentato l'approccio seguito per la semplificazione e nel paragrafo 4.3.2 quella condotta per il paragrafo 5 del consenso informato.

4.3.1 Approccio alla semplificazione

In seguito al ruolo centrale delle caratteristiche lessicale nel determinare la leggibilità dei testi inerenti la salute, la semplificazione lessicale costituisce il livello più esplorato. Sono stati elaborati diversi approcci per rendere questi testi più comprensibili, riducendo in particolare la difficoltà del vocabolario. A livello sintattico invece, l'approccio principale perseguito è costituito dall'identificazione e la semplificazione per esempio delle frasi che sono costituite da una clausola dipendente e indipendente (capitolo 1).

Per la semplificazione del modulo di consenso informato, sono state seguite tre fasi distinte:

1. Valutazione automatica della leggibilità a livello di frase
2. Semplificazione lessicale guidata dall'esperto di dominio
3. Semplificazione sintattica sulla base di regole linguistiche

Tramite READ-IT è stato possibile valutare la leggibilità a livello della frase del modulo di consenso informato e identificare il tipo di difficoltà (base, lessicale, sintattica e globale) presente. Per ciascun periodo il livello di difficoltà è rappresentato

cromaticamente mediante colori che vanno dal verde (testo leggibile) al rosso (testo particolarmente difficile). Ad ogni singola frase sono inoltre associati dei valori che variano, come nel caso della valutazione della leggibilità a livello di documento, su una scala che va da 0 (facile da leggere) a 100 (difficile da leggere).

In base al tipo di difficoltà riscontrata, sono stati seguiti due metodi diversi. Quando la frase presa in analisi risulta più complessa a livello lessicale è stata effettuata una revisione delle parole usate nel testo guidata dall'esperto del dominio. Quando invece risulta più complessa a livello sintattico, si è deciso di riformulare la frase facendo uso di strutture grammaticali più semplici. Queste due fasi sono descritte in dettaglio nei paragrafi che seguono.

Il lessico del consenso informato

Si è deciso di approfondire, come è stato fatto per il corpus descritto nel capitolo 2, lo studio del lessico e quindi della terminologia usata nel modulo di consenso informato sulla procreazione assistita. Quindi è stata estratta la lista dei termini che non appartengono al vocabolario di base con associata la relativa frequenza.

	N° tipi	**N° tipi non presenti nel VdB**
Testo totale	1472	571 (38,79%)

Tabella 4.2: Percentuale dei lemmi non appartenenti al vocabolario di base del consenso informato sulla procreazione assistita.

Come si vede nella (Tabella 4.1), il modulo di consenso informato contiene una percentuale del 38,79% di parole tipo non appartenenti al vocabolario di base e quelle che ricorrono con più frequenza sono "embrione" (51), "ovociti" (28) e "procreazione" (27).

A differenza dell'approccio messo a punto per l'analisi del lessico del corpus di consensi informati descritto nel capitolo 2, non sono stati analizzati solo termini singoli ma anche termini polirematici, costituiti da sequenze di più parole. I termini polirematici sono stati estratti utilizzando *Text-to-Knowledge* (T2K), una piattaforma software progettata e sviluppata congiuntamente dall'Istituto di Linguistica

Computazionale "A. Zampolli" (ILC) del CNR di Pisa e dal Dipartimento di Linguistica dell'Università di Pisa, che si propone di offrire una batteria integrata di strumenti avanzati di analisi linguistica del testo, analisi statistica e apprendimento automatico del linguaggio, destinati a offrire una rappresentazione accurata del contenuto di una base documentale non strutturata [Dell'Orletta et al., 2014b]. T2K trasforma le conoscenze implicitamente codificate all'interno di un corpus di testi in conoscenza esplicitamente strutturata: il risultato finale di questo processo interpretativo spazia dall'acquisizione di conoscenze lessicali e terminologiche complesse alla loro organizzazione in strutture proto-concettuali.

La lista ottenuta analizzando il testo completo del modulo di consenso informato è stata filtrata eliminando i termini polirematici che hanno un termine singolo contenuto nel vocabolario fondamentale. I termini polirematici che ricorrono con più frequenza sono "stimolazione ovarica", "prelievo chirurgico" e "monitoraggio ecografico" (4).

Le liste dei lemmi e delle parole polirematiche non appartenenti al vocabolario di base sono state sottoposte alla validazione di un esperto in ostetricia e ginecologia, in modo da seguire la stessa classificazione che è stata presentata nel paragrafo 2.4. In questo caso non è stata portata avanti solamente la classificazione ma è stata aggiunta anche la possibile sostituzione: 1)"parole specifiche del dominio", cioè parole di ambito medico di cui non si può fare a meno per esprimere un concetto (ad esempio "embrione" e "prelievo chirurgico"); 2)"parole tecniche", cioè parole che sono specifiche del dominio ma che dovrebbero essere integrate con glosse (ad esempio "tromboembolici" > "fenomeni che possono portare ad una trombosi e conseguentemente a una embolia" o "stimolazione ovarica" > "stimolazione farmacologica dell'ovaia per ottenere un numero maggiore di ovuli di quelli prodotti naturalmente ogni mese dalla donna"); e 3) "tecnicismi", cioè parole che necessitano di essere sostituite con un sinonimo (o quasi) più semplice in modo da essere comprensibili al lettore inesperto (ad esempio "prelievo ovocitario" > "prelievo degli ovuli").

La classificazione dei termini non appartenenti al VdB del consenso informato

sulla procreazione assistita, e in particolare i sinonimi e le glosse aggiunti a quelli appartenenti alle classi 2 e 3, è stata usata nel compito di semplificazione lessicale.

Le regole di semplificazione

Le regole di semplificazione utilizzate nell'approccio qui proposto sono contenute nello schema definito da [Brunato et al., 2015]:

- **Divisione**: consiste nel segmentare una frase in più parti in modo da avere delle frasi più brevi per esprimere lo stesso concetto, evitando quindi al lettore di comprendere frasi con dipendenze molto lunghe come ad esempio clausole coordinanti (introdotte da congiunzioni coordinanti, due punti o punti e virgola), clausole subordinanti, apposizioni e clausole avverbiali. Questa regola però non può essere applicata in ogni caso in quanto molto spesso alcune subordinate possono fornire informazioni che devono ricorrere insieme alla principale.

- **Fusione**: è l'opposto dello *split* e consiste nell'unione di due o più frasi in un'unica frase semplificata. Adottare questa trasformazione è meno probabile in quanto crea frasi semanticamente dense che sono più difficili da processare. L'esperto può decidere di applicarla per verificare se le frasi fuse presentano un pattern regolare di caratteristiche linguistiche che possono essere catturate automaticamente.

- **Riordinamento**: consiste nel cambiare l'ordine di alcune parti della frase, stando attenti ad intervenire anche a livello lessicale e sintattico in quanto si potrebbero ottenere risultati agrammaticali.

- **Inserimento**: consiste nell'aggiunta di informazioni utili a comprendere meglio una frase in quanto non sempre una frase semplice è quella che risulta essere più breve rispetto alla sua originale. A volte risulta però difficile definire cosa possa essere necessario inserire per rendere la frase più semplice. Si

distinguono due tipi di inserimento, vale a dire uno per i verbi e un altro per i soggetti (che sono sottintesi).

- **Cancellazione**: consiste nell'eliminazione di parole ridondanti o poco utili alla comprensione del concetto espresso nella frase. Come l'operazione di *insert*, definire cosa cosa sia possibile eliminare risulta molto difficile. Si può assistere all'eliminazione di avverbi o aggettivi, verbi o soggetti.

- **Trasformazione**: comprende sei diverse tipologie di trasformazione che una frase può subire per diventare più comprensibile. Queste operazioni intervengono a livello lessicale, morfo-sintattico e sintattico, dando anche luogo a fenomeni di sovrapposizione.

  - **Sostituzione lessicale (a livello di parola)**: sostituzione di una parola singola con un'altra parola (o più di una) che è solitamente un sinonimo più comune o un termine meno specifico.
  - **Sostituzione lessicale (a livello di parola polirematica)**: sostituzione di una parola polirematica con una parola singola o un'altra parola polirematica.
  - **Sostituzione dell'anafora**: sostituzione di un pronome con il suo antecedente lessicale.
  - **Nome in verbo**: sostituzione di una nominalizzazione con un verbo
  - **Verbo in nome**: sostituzione di un verbo con una nominalizzazione
  - **Voce del verbo**: sostituzione di una frase passiva ad una attiva e viceversa

4.3.2 La semplificazione del paragrafo 5

Seguendo l'approccio proposto è stata condotta la semplificazione del paragrafo 5 del modulo di consenso informato.

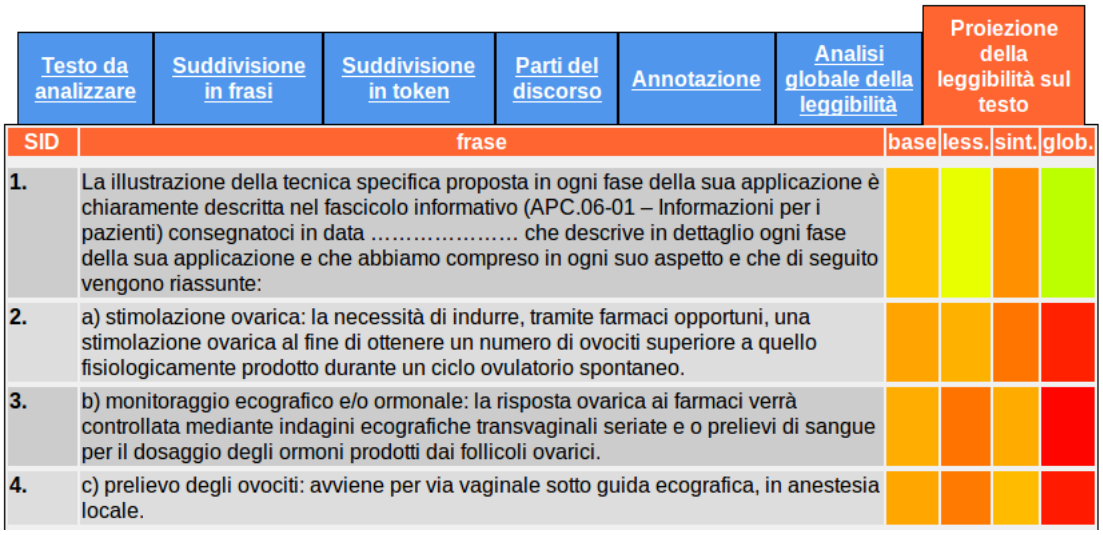


Figura 4.5: Valutazione della leggibilità a livello di frase della prima parte del paragrafo 5.

Nella (Figura 4.5) è riportata come esempio la prima parte del paragrafo 5 del modulo di consenso informato. Le prime due frasi risultano essere più facili a livello lessicale (rispettivamente 45,4 e 65,2) e più complesse a livello sintattico (rispettivamente 71,8 e 77,1), mentre per le successive due frasi la situazione è inversa (rispettivamente 77,4 e 76,1 a livello lessicale e 66,5 e 63,4 a quello sintattico).

Vengono qui riportate le frasi semplificate dell'esempio in (Figura 4.5):

(6) O: **La illustrazione della tecnica specifica proposta in ogni fase della sua applicazione** è chiaramente descritta nel fascicolo informativo (APC.06-01 – Informazioni per i pazienti) **consegnatoci** in data ..... **che** descrive in dettaglio ogni fase **della sua applicazione e che** abbiamo compreso in ogni suo aspetto e che **di seguito** vengono riassunte:

S: **Ogni fase dell'applicazione della tecnica** è chiaramente descritta nel fascicolo informativo (APC.06-01 – Informazioni per i pazienti) **che ci è stato consegnato** in data ..... **Il fascicolo** descrive in dettaglio ogni fase **dell'applicazione della tecnica che** abbiamo compreso in ogni suo aspetto e che **qui** vengono riassunte:

(7) O: a) stimolazione **ovarica**: la necessità di indurre**, tramite farmaci opportuni, una stimolazione ovarica al fine di** ottenere un numero **di ovociti superiore a quello fisiologicamente prodotto durante un ciclo ovulatorio spontaneo.**

S: a) stimolazione **dell'ovaia**: la necessità di indurre **una stimolazione farmacologica dell'ovaia per** ottenere un numero **maggiore di ovuli di quelli prodotti naturalmente ogni mese dalla donna.**

(8) O: b) monitoraggio ecografico e/o ormonale: la risposta **ovarica** ai farmaci verrà controllata mediante indagini ecografiche transvaginali **seriate** e/o prelievi di sangue per **il dosaggio degli** ormoni prodotti dai follicoli **ovarici.**

S: b) monitoraggio ecografico e/o ormonale: la risposta **dell'ovaia** ai farmaci verrà controllata mediante indagini ecografiche transvaginali **in serie** e o prelievi di sangue per **dosare gli** ormoni prodotti dai follicoli **dell'ovaia**.

(9) O: c) prelievo degli **ovociti**: avviene per via vaginale sotto guida ecografica**,** in anestesia locale.

S: c) prelievo degli **ovuli**: avviene per via vaginale sotto guida ecografica **e** in anestesia locale.

Per valutare l'effetto della semplificazione nei valori di leggibilità, nella (Tabella 4.3) è riportato un confronto tra i valori ottenuti per le frasi originali e quelli ottenuti dopo la semplificazione secondo i modelli LESSICALE e SINTATTICO di READ-IT.

Frasi originali	READ-IT		**Frasi semplificate**	READ-IT	
	LESS	**SINT**		**LESS**	**SINT**
1ª frase	45,4	71,8	1ª frase	50	68,4
			2ª frase	43,6	45,9
2ª frase	65,2	77,1	3ª frase	53,6	64,9
3ª frase	77,4	66,5	4ª frase	71,4	63,3
4ª frase	76,1	63,4	5ª frase	74,1	62

Tabella 4.3: Confronto tra i risultati della valutazione della leggibilità a livello di frase secondo i modelli LESSICALE e SINTATTICO della prima parte del paragrafo 5 originale e semplificata.

La prima frase è stata segmentata in due parti, in modo da ottenere due frasi più brevi per esprimere lo stesso concetto. Inoltre sono state eliminate parole poco utili alla comprensione del concetto espresso ed è stato cambiato il loro ordine. In questo caso non essendoci stati casi di sostituzioni lessicali o aggiunte di glosse in quanto non sono presenti parole specifiche del dominio medico, i risultati più interessanti riguardano il livello sintattico: si passa da un valore di 71,8 per la frase originale a 68,4 e 45,9 per le due frasi semplificate.

Nella seconda frase si è optato invece per due sostituzioni lessicali, una a livello di parola singola e una a livello di parola polirematica, attraverso il supporto del vocabolario medico contenente i termini singoli e polirematici classificati presentato nel paragrafo precedente. Per quanto riguarda la prima, il termine "ovarica" è stato sostituito con un sinonimo rappresenta il suo referente fisico. Per quanto riguarda la seconda, tramite la glossa presente nel vocabolario medico presente alla voce "stimolazione ovarica" è stato possibile spiegare con termini più facili questo trattamento. I valori di leggibilità migliorano sia a livello lessicale e sia sintattico: per quanto riguarda il primo si passa da un valore di 65,2 a 53,6 e per quanto riguarda il secondo da 77,1 a 64,9.

Anche nella terza frase sono state effettuate tre sostituzioni lessicali, tutte a livello di parola singola. Inoltre si è optato alla trasformazione di un nome in verbo, vale a dire "dosaggio" > "dosare". In questo caso si ha un miglioramento del valore

di leggibilità a livello lessicale che passa da 77,4 a 71,3 e a livello sintattico da 66,5 a 63,3.

Infine nella quarta frase, oltre la sostituzione lessicale a livello di parola di "ovociti", si è deciso di usare nella frase semplificata la congiunzione copulativa "e". In questo caso si ha un leggero miglioramento nei due livelli esaminati: a livello lessicale si passa da un valore di 76,1 a 74,1 e a quello sintattico da 63,4 a 62.

Infine l'effetto della semplificazione è stato valutato anche prendendo in considerazione il paragrafo 5 nella sua globalità (Tabella 4.4).

Paragrafo 5	READ-IT		
	BASE	**LESSICALE**	**SINTATTICO**
Originale	97,7	91,1	100
Semplificato	90,1	59,7	99,7

Tabella 4.4: Confronto tra i risultati della valutazione della leggibilità secondo i modelli BASE, LESSICALE e SINTATTICO del paragrafo 5 originale e semplificato.

I risultati mostrano un miglioramento dei valori di leggibilità, specialmente per quanto riguarda i modelli BASE e LESSICALE.

Per quanto riguarda il modello BASE, il quale fa riferimento alle caratteristiche del testo grezzo come la lunghezza delle frasi e delle parole, il valore ottenuto passa da 97,7 nel paragrafo originale a 90,1 in quello semplificato. Questo risultato è dovuto principalmente al fatto che la regola di semplificazione maggiormente usata è stata quella dello *split*: segmentare più frasi in più parti ha portato alla presenza di frasi più brevi che vanno ad influenzare il risultato a questo livello.

Il miglioramento più significativo che si è ottenuto riguarda il modello LESSICALE, il quale fa riferimento alla composizione del vocabolario e la sua ricchezza lessicale: il valore passa da 91,1 nel paragrafo originale a 59,7 in quello semplificato. Questo risultato è dovuto alle sostituzioni dei termini medici con sinonimi più semplici ed alle aggiunte delle glosse, entrambi contenuti all'interno del vocabolario medico. Quest'ultimo dimostra quindi di essere in grado di supportare la semplificazione lessicale del testo.

Per quanto riguarda invece il modello SINTATTICO, il quale fa riferimento a informazione di tipo grammaticale, non si ha un miglioramento significativo: il valore passa da 100 a 99,7. A questo livello il documento analizzato globalmente risulta essere particolarmente complesso anche dopo la semplificazione. Questo risultato suggerisce che il livello sintattico deve essere ulteriormente esplorato.

5
Conclusioni

In questo lavoro di tesi è stato presentato un approccio linguistico-computazionale per la valutazione della leggibilità di testi appartenenti al dominio medico, in particolare di moduli di consenso informato, come passo preliminare alla semplificazione. L'obiettivo è stato quello di valutare come i metodi e gli strumenti attualmente disponibili possano essere specializzati rispetto al dominio medico e in una prospettiva interlinguistica.

È stata presentata una metodologia per il calcolo della leggibilità di un testo basata su strumenti di annotazione linguistica automatica. La novità dell'approccio proposto rispetto allo stato dell'arte, in cui la valutazione della leggibilità nel dominio medico non va oltre la distribuzione delle parti del discorso (PoS) e/o sintagmi nominali, è costituita dall'uso di risultati ottenuti dal testo annotato sintatticamente (vale a dire a dipendenza) ed è stata quindi monitorata una varietà più ampia di fattori che influenzano la leggibilità di un testo. Nonostante il metodo di ricostruzione del profilo linguistico e calcolo della leggibilità adottato in questo lavoro di tesi sia stato progettato per l'analisi di testi rappresentativi della lingua comune, gli esperimenti condotti sul corpus di consensi informati hanno mostrato come la meto-

dologia seguita riesca tuttavia a definire quali sono le caratteristiche linguistiche che caratterizzano i documenti sotto valutazione, mettendone in luce gli specifici luoghi di complessità. I punteggi di leggibilità ottenuti dallo strumento READ-IT hanno mostrato che i consensi informati sono più complessi al livello lessicale che a quello sintattico dato che contengono una percentuale alta di parole non appartenenti al vocabolario di base. La difficoltà registrata al livello lessicale suggerisce quindi che lo strumento READ-IT, addestrato su testi del genere giornalistico, necessita di essere specializzato al livello di vocabolario usato, il quale dovrebbe contenere una selezione dei termini di dominio da usare nei moduli di consenso informato per non penalizzare il livello di leggibilità. Per questo motivo è stata posta un'attenzione particolare sull'analisi del lessico del corpus che ha portato alla creazione di un vocabolario medico con il quale sarà possibile specializzare lo strumento READ-IT. La creazione del vocabolario medico rappresenta un significativo passo in avanti rispetto allo stato dell'arte in quanto finora per la lingua italiana non esistono lessici specialistici che sono gia disponibili invece per altre lingue.

La valutazione della leggibilità del corpus preso in esame ha mostrato differenze significative per le differenti macro-aree e specialità, per esempio i testi dell'area prevenzione si sono rivelati i più facili rispetto alle altre macro-aree e al suo interno i testi di screening si sono rivelati i più facili rispetto alle altre specialità. In futuro questa analisi potrà essere ampliata con la definizione precisa della tipologia dei testi, infatti all'interno del corpus i testi hanno diverse 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. Per quanto riguarda l'analisi dei testi organizzati per aziende sanitarie locali di provenienza, è stato possibile avviare un'analisi preliminare solo su 4 tra le 12 esistesti in Toscana. Definendo per ogni testo l'appartenenza ad una determina azienda sanitaria in futuro sarà possibile ampliare questa analisi a tutte quelle operanti in Toscana e approfondirla rispetto a tutte le specialità.

Il tema della valutazione automatica della leggibilità è stato trattato anche da

una prospettiva interlinguistica. A questo scopo è stato presentato un caso di studio su due lingue tipologicamente lontane, vale a dire il basco e l'italiano. A partire dal confronto dei principi su cui si basano gli strumenti per la valutazione della leggibilità, vale a dire *ErreXail* e READ-IT, e la definizione delle caratteristiche in comune prese in considerazione, è stato condotto un confronto linguistico di corpora comparabili di consensi informati in italiano e basco alla ricerca di fenomeni di complessità linguistica comuni alle due lingue. I risultati hanno mostrato che i testi baschi sono più complessi al livello sintattico al contrario di quelli italiani che invece lo sono a livello lessicale. In futuro questi risultati andranno indagati più a fondo dato che i documenti baschi sono stati scritti seguendo dei parametri per la revisione della comprensibilità, rendendo questo confronto più difficile. Dato che le caratteristiche prese in considerazione sono quelle utilizzate dagli strumenti di analisi automatica della leggibilità, questo tipo di confronto pone le basi per una futura metodologia di valutazione della leggibilità valida per più lingue.

A partire dai risultati della valutazione della leggibilità, è stata messa a punto una metodologia di semplificazione semi-automatica di un consenso informato all'interno di un programma per la fecondazione assistita. I risultati della valutazione della leggibilità sono stati confrontati rispetto alle risposte di 95 pazienti che si sono sottoposti al trattamento attraverso la somministrazione di un questionario anonimo per monitorare il loro grado di soddisfazione, dovuto anche alla chiarezza del consenso. I risultati hanno mostrato che nonostante i risultati della leggibilità abbiano mostrato che il consenso è particolarmente complesso, la percezione della sua chiarezza è positiva. Questi risultati possono essere letti alla luce del fatto che i pazienti che hanno compilato il questionario hanno un livello di istruzione medio-alto ma necessitano di essere indagati ulteriormente. Rispetto ai metodi presenti nello stato dell'arte che si focalizzano prevalentemente alla difficoltà lessicale dei testi medici, la metodologia proposta in questa tesi riguarda anche la difficoltà sintattica. I risultati ottenuti nell'applicazione della metodologia ad un paragrafo del consenso mostrano un miglioramento dei valori di leggibilità, specialmente a livello lessicale e quindi

l'approccio proposto dimostra di essere in grado di supportare la semplificazione del testo. Al momento il paragrafo semplificato è sottoposto alla valutazione degli esperti del dominio il cui parere sarà fondamentale per impostare sviluppi futuri della metodologia, in modo da poterla applicare anche agli altri paragrafi del consenso. A completamento del lavoro sarà somministrato nuovamente il questionario per valutare l'effettivo miglioramento della percezione della chiarezza del consenso informato.

Bibliografia

[Abrahamsson et al., 2014] Abrahamsson, E., Forni, T., and Skeppstedt (2014). Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014, EACL Workshop)*, pages 57–65.

[Aduriz et al., 2003] Aduriz, I., Aldezabal, I., Alegria, I., Arriola, J. M., de Ilarraza, A. D., Ezeiza, N., and Gojenola, K. (2003). Finite State Applications for Basque. In *EACL 2003 Workshop on Finite-State Methods in Natural Language Processing*, pages 3–11.

[Aduriz et al., 2004] Aduriz, I., Aranzabe, M. J., Arriola, J. M., de Ilarraza, A. D., Gojenola, K., Oronoz, M., and Uria, L. (2004). A Cascaded Syntactic Analyser for Basque. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 124–134. Springer.

[Aldabe et al., 2012] Aldabe, I., Maritxalar, M., Perez De Viñaspre, O., and Larraitz, U. (2012). Automatic Exercise Generation in an Essay Scoring System. In *Proceedings of the 20th International Conference on Computers in Education*, pages 671–673.

[Alegria et al., 2004a] Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004a). Representation and Treatment of Multiword Expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55. Association for Computational Linguistics.

[Alegria et al., 2002] Alegria, I., Aranzabe, M. J., Ezeiza, A., Ezeiza, N., and Urizar, R. (2002). Robustness and customisation in an analyser/lemmatiser for Basque.

In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6. Las Palmas de Gran Canaria, May.

[Alegria et al., 2004b] Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2004b). Design and Development of a Named Entity Recognizer for an Agglutinative Language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition.*

[Aluisio et al., 2010] Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

[Aranzabe et al., 2013] Aranzabe, M. J., de Ilarraza, A. D., and Gonzalez-Dios, I. (2013). Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del lenguaje natural*, 50:61–68.

[Atorino, 2011] Atorino, M. T. (2011). Comparazione interlinguistica: italiano e basco. *Un mundo, muchas miradas*, (2).

[Attardi et al., 2009] Attardi, G., Dell'Orletta, F., Simi, M., and Turian, J. (2009). Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Evalita 2009.*

[Barzilay and Elhadad, 2003] Barzilay, R. and Elhadad, N. (2003). Sentence Alignment for Monolingual Comparable Corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

[Bird et al., 2009] Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

[Bodenreider, 2004] Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

[Brants and Franz, 2006] Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1.

[Brunato et al., 2015] Brunato, D., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2015). Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of the 9th Linguistic Annotation Workshop (LAW'15)*. Denver, Colorado, USA.

[Cabré, 1999] Cabré, M. T. (1999). *Terminology: Theory, Methods, and Applications*. John Benjamins. Amsterdam.

[Cavagnoli, 2007] Cavagnoli, S. (2007). *La comunicazione specialistica*. Carocci. Roma.

[Chall and Dale, 1995] Chall, J. S. and Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookile Books, Cambridge, MA.

[Chang and Lin, 2011] Chang, C. C. and Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. Software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[Cocchi et al., 2014] Cocchi, M., Mazzocut, C., Cipolat Mis, C., Truccolo, I., Cervi, E., Iori, R., and Orlandini, D. (2014). ETHIC. Evaluation Tool of Health Information for Consumers. Development, features and validation. *Divided we fall, united we inform. Building alliances for a new European cooperation, 14th EAHIL Annual Conference*. Roma (Italy), 11-13 June.

[Collins-Thompson, 2014] Collins-Thompson, K. (2014). Computational Assessment of Text Readability: A Survey of Current and Future Research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

[Collins-Thompson and Callan, 2004] Collins-Thompson, K. and Callan, J. (2004). A Language Modeling Approach to Predicting Reading Difficulty.

[Cordasco, 2013] Cordasco, K. M. (2013). Obtaining Informed Consent From Patients: Brief Update Review.

[Dalianis et al., 2012] Dalianis, H., Hassel, M., Henriksson, A., and Skeppstedt, M. (2012). Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In *Swedish Technology Conference*, pages 17–18.

[De Mauro, 2000] De Mauro, T. (2000). *Il dizionario della lingua italiana*. Torino, Paravia.

[Dell'Orletta, 2009] Dell'Orletta, F. (2009). Ensamble system for Part-of-Speech tagging. In *Proceedings of Evalita '09, Evaluation of NLP and Speech Tools for Italian*. Reggio Emilia, December.

[Dell'Orletta et al., 2011] Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 73–83. Edinburgh, UK.

[Dell'Orletta et al., 2013] Dell'Orletta, F., Montemagni, S., and Venturi, G. (2013). Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An exploratory Study on Italian Fictional Prose. In *Proceedings of Recent Advances in Natural Language Processing*, pages 189–197. Hissar, Bulgaria, 7-13 September 2013.

[Dell'Orletta et al., 2014a] Dell'Orletta, F., Montemagni, S., and Venturi, G. (2014a). Assessing Document and Sentence Readability in Less Resourced Languages and across Textual Genres. *Recent Advanced in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company:163–193.

[Dell'Orletta et al., 2014b] Dell'Orletta, F., Venturi, G., Cimino, A., and Montemagni, S. (2014b). T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2062–2070. 26-31 May, Reykjavik, Iceland.

[Falkenjack et al., 2013] Falkenjack, J., Heimann Mühlenbock, K., and Jönsson, A. (2013). Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODAL-IDA 2013)*, pages 27–40.

[Feng et al., 2009] Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics.

[Feng et al., 2010] Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics.

[Flesch, 1949] Flesch, R. (1949). *The Art of Readable Writing*. Harper & Row.

[Flesch, 1981] Flesch, R. (1981). *How to Write Plain English*. Barnes & Noble.

[Franchina and Vacca, 1986] Franchina, V. and Vacca, R. (1986). Adaptation of Flesch readability index on a bilingual text written by the same both in Italian and English languages. *Linguaggi*, 3:229–237.

[François and Fairon, 2012] François, T. and Fairon, C. (2012). An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computa-*

tional Natural Language Learning, pages 466–477. Association for Computational Linguistics.

[Gonzalez-Dios et al., 2014] Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., and Salaberri, H. (2014). Simple or Complex? Assessing the readability of Basque Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344. Dublin, Ireland, August 23-29.

[Gonzalez-Dios et al., 2013] Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., and Soraluze, A. (2013). Detecting Apposition for Text Simplification in Basque. In *Computational Linguistics and Intelligence Text Processing*, pages 513–524. Springer.

[Grigonyte et al., 2014] Grigonyte, G., Kvist, M., Velupillai, S., and Wirén, M. (2014). Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014, EACL Workshop)*, pages 74–83.

[Heilman et al., 2007] Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of NAACL HLT*, pages 460–467. sn.

[Inui et al., 2001] Inui, K., Yamamoto, S., and Inui, H. (2001). Corpus-based acquisition of sentence readability ranking models for deaf people. In *NLPRS*, pages 159–166.

[Isenius et al., 2012] Isenius, N., Velupillai, S., and Kvist, M. (2012). Initial Results in the Development of SCAN: a Swedish Clinical Abbreviation Normalizer. In *Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of*

Methods, Applications, and Resources for eHealth Document Analysis - CLEF eHealth2012. Rome, Italy, September.

[Jha et al., 2013] Jha, A. K., Larizgoitia, I., Audera-Lopez, C., Prasopa-Plaizier, N., Waters, H., and Bates, D. (2013). The global burden of unsafe medical care: analytic modelling of observational studies. *BMJ Quality and Safety*, 22(10):809–815.

[Kandula et al., 2010] Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010). A Semantic and Syntactic Text Simplification Tool for Health Content. In *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 366–370. United States.

[Kauchak et al., 2014] Kauchak, D., Mouradi, O., Pentoney, C., and Leroy, G. (2014). Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 2616–2625. IEEE.

[Kim et al., 2007] Kim, H., Goryachev, S., Rosemblat, G., Browne, A. C., Keselman, A., and Zeng-Treitler, Q. (2007). Beyond Surface Characteristics: A New Health Text-Specific Readability Measurement. In *AMIA*.

[Kokkinakis, 2012] Kokkinakis, D. (2012). The Journal of the Swedish Medical Association - a Corpus Resource for Biomedical Text Mining in Swedish. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey.*

[Kvist and Velupillai, 2013] Kvist, M. and Velupillai, S. (2013). Professional Language in Swedish Radiology Reports - Characterization for Patient-Adapted Text Simplification. In *Proceedings of the Scandinavian Conference on Health Informatics 2013*, pages 55–59.

[Leroy and Endicott, 2012] Leroy, G. and Endicott, J. E. (2012). Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754. ACM.

[Leroy et al., 2012] Leroy, G., Endicott, J. E., Mouradi, O., Kauchak, D., and Just, M. L. (2012). Improving Perceived and Actual Text Difficulty for Health Information Consumers using Semi-Automated Methods. In *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 522–531.

[Lorda et al., 1996] Lorda, S. P., Cantalejo, I. M. B., and Concheiro Carro, L. (1996). Legibilidad de los formularios escritos de consentimiento informado. *Medicina clínica*, 107(14):524–529.

[Lucisano and Piemontese, 1988] Lucisano, P. and Piemontese, M. E. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città*, 3:57–68.

[Ma et al., 2012] Ma, Y., Fosler-Lussier, E., and Lofthus, R. (2012). Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 548–552. Association for Computational Linguistics.

[Marijuan et al., 1998] Marijuan, M., Lejona, B., de Heredia, J. L., Arcelay, A., Martinez, S., Hernanz, M., Urkola, B., Hernando, A., and Gabaldón, O. (1998). *Guía práctica para la elaboración de documentos de información y consentimiento.* Osakidetza.

[McLaughlin, 1969] McLaughlin, G. H. (1969). SMOG Grading: A New Readability Formula. *Journal of Reading*, 12(8):639–646.

[Miller, 1995] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

[Montemagni, 2008] Montemagni, S. (2008). Estrazione Terminologica Automatica e Indicizzazione: Scenari Applicativi, Problemi e Possibili Soluzioni. In *Documenti Digitali, Guarasci, R*, pages 241–284. ITER.

[Montemagni, 2013] Montemagni, S. (2013). Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *In Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, 42(1):145–172.

[Peng et al., 2012] Peng, Y., Tudor, C. O., Torii, M., Wu, C. H., and Vijay-Shanker, K. (2012). iSimp: A Sentence Simplification System for Biomedical Text. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM2012)*, pages 211–216.

[Petersen and Ostendorf, 2009] Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

[Piemontese, 1996] Piemontese, M. E. (1996). *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnodid.

[Platt, 1998] Platt, J. C. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

[Rondeau, 1983] Rondeau, G. (1983). *Introduction à la terminologie*. Gaëtan Morin éditeur. Québec.

[Schwarm and Ostendorf, 2005] Schwarm, S. E. and Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

[Si and Callan, 2001] Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM.

[Siddharthan, 2006] Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Springer*, Volume 4:77–109.

[Stagger, 2012] Stagger, Ö. R. (2012). A modern POS tagger for Swedish. In *Proceedings of the Fourth Swedish Language Technology Conference*.

[Stenner, 1996] Stenner, A. J. (1996). Measuring Reading Comprehension with the Lexile Framework. *Paper presented at the California Comparability Symposium*.

[Tanaka-Ishii et al., 2010] Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

[Tartaglia et al., 2012] Tartaglia, R., Albolino, S., Bellandi, T., Bianchini, E., Biggeri, G., Fabbro, G., Bevilacqua, L., Dell'Erba, A., Privitera, G., and Sommella, L. (2012). Eventi avversi e conseguenze prevenibili: studio retrospettivo in cinque grandi ospedali italiani. *Epidemiologia & Prevenzione*, 36(3-4):151–161.

[Terranova et al., 2012] Terranova, G., Ferro, M., Carpeggiani, C., Recchia, L., Braga, L., Semelka, R., and Picano, E. (2012). Low Quality and Lack of Clarity of Current Informed Consent Form in Cardiology - How to Improve Them. *Journal of the American College of Cardiology (JACC): Cardiovascular Imaging*, Elsevier inc., vol.5(6):649–655.

[Venturi et al., 2015] Venturi, G., Bellandi, T., Dell'Orletta, F., and Montemagni, S. (2015). NLP-Based Readability Assessment of Health-Related Texts: a Case

Study on Italian Informed Consent Forms. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 131–141. Lisbon, Portugal, 17 september.

[Word Perfect, 1994] Word Perfect (1994). Word perfect v: 6.1: Guía para el usuario.

[Zeng et al., 2005] Zeng, Q., Kim, E., Crowell, J., and Tse, T. (2005). A Text Corpora-Based Estimation of the Familiarity of Health Terminology. *Springer*, Volume 3745:184–192.

[Zeng-Treitler et al., 2007] Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., and Roseandale, D. (2007). Making Texts in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. In *AMIA Annual Symposium proceedings / AMIA Symposium*, pages 846–850.

[Zeng-Treitler et al., 2008] Zeng-Treitler, Q., Goryachev, S., Tse, T., Keselman, A., and Boxwala, A. (2008). Estimating Consumer Familiarity with Health Terminology: A Context-based Approach. *J Am Med Informatics Assoc. 2008*, 15(3):349–356.

[Zipf, 1935] Zipf, G. K. (1935). The psycho-biology of language. Boston.