La valutazione automatica della leggibilità.

I tools messi in campo per iSLe da ILC (Isitituto di Linguistica computazionale – CNR Pisa)

Tra le funzionalità innovative messe a disposizione dalla piattaforma editoriale iSLe, è prevista la possibilità per l'autore di poter controllare automaticamente il grado di leggibilità del proprio testo in funzione del lettore di riferimento.

La leggibilità (traduzione del termine inglese “readability”) 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. Ad esempio, dal punto di vista lessicale, l'uso massiccio di termini tecnici propri di varietà specialistiche della lingua, può tradursi in una maggior difficoltà di comprensione del testo da parte del lettore medio. Una costruzione sintattica caratterizzata da una subordinazione di grado elevato o da una maggior distanza tra gli elementi grammaticalmente dipendenti (ad esempio, il soggetto e il verbo della frase) è invece annoverata tra i fattori di complessità sul piano sintattico.

Sebbene gli studi sulla leggibilità e il suo trattamento automatico abbiano una lunga tradizione (le prime “formule” di leggibilità per la lingua inglese risalgono alla prima metà del secolo scorso), per molto tempo parametri di questo tipo non sono stati considerati, data la difficoltà ad essere intercettati in maniera affidabile dagli strumenti allora in uso. Al contrario, gli indici automatici proposti (tra cui il noto Gulpease per la lingua italiana) si sono limitati a fornire un punteggio di leggibilità basato su caratteristiche del testo molto superficiali, tipicamente la lunghezza della parola e della frase.

È solo negli ultimi anni che la ricerca in questo settore ha compiuto notevoli progressi, grazie all'impulso dato dallo sviluppo concomitante delle tecnologie basate sul Trattamento Automatico del Linguaggio, che hanno consentito di rendere computabili uno spettro di parametri linguistici più ampio e sofisticato. Questo contesto ha favorito la nascita di sistemi di leggibilità cosiddetti di “seconda generazione”, nei quali la valutazione della leggibilità viene affrontata come un compito di classificazione probabilistica basato su algoritmi di apprendimento supervisionato. Ciò significa che, a partire da un corpus di addestramento accuratamente selezionato come rappresentativo dei livelli di leggibilità da considerare (ad esempio livello “semplice” e livello “complesso” in una classificazione binaria), il sistema “impara” a riconoscere le caratteristiche linguistiche che permettono di discriminare ciascun livello e che sono estratte automaticamente dal risultato dell'annotazione linguistica del testo. Seguendo questo approccio, il punteggio di leggibilità di un nuovo testo sarà calcolato sulla base della maggior somiglianza del suo profilo linguistico ad uno dei livelli di leggibilità definiti in fase di addestramento.

Per quanto riguarda la lingua italiana, il primo strumento per la valutazione automatica della leggibilità dei testi fondato su questi presupposti è rappresentato da READ-IT, realizzato dall'Istituto di Linguistica Computazionale “Antonio Zampolli” del CNR di Pisa. READ-IT calcola la leggibilità dei testi sulla base di un'analisi sofisticata della struttura linguistica sottostante al testo e articolata su diversi livelli di descrizione linguistica. Esso, inoltre, fornisce una predizione della leggibilità non solo per l'intero testo ma anche in relazione alla frase, un aspetto del tutto innovativo nel panorama nazionale e internazionale della ricerca del settore. Proprio la capacità di intercettare i “luoghi di complessità” all'interno delle singole frasi, qualificandone la natura (in particolare, lessicale o sintattica), permette a READ-IT di proporsi come un

ausilio alla riscrittura semplificata del testo, della quale possono beneficiare categorie di studenti con profili atipici (ad esempio, apprendenti l'italiano come seconda lingua o studenti affetti da lievi disturbi del linguaggio).

Nell'ambito del progetto ISLE, le funzionalità offerte da READ-IT sono state specializzate per adattarsi alle specificità del contesto didattico, sia dal punto di vista dei materiali di riferimento, sia rispetto ai potenziali fruitori. Ciò ha richiesto uno studio preliminare delle caratteristiche linguistiche del genere testuale da trattare, finalizzato ad adattare i modelli di leggibilità di partenza alle peculiarità lessicali e sintattiche dei "linguaggi" caratterizzanti le diverse discipline oggetto di sperimentazione. In particolare, la personalizzazione ha comportato l'identificazione di una lista di termini di dominio, propri di ciascun lessico disciplinare, che sono stati considerati parte del "vocabolario fondamentale" dei testi di ambito didattico, allo scopo di evitare inopportune penalizzazioni del lessico di dominio sulla valutazione della leggibilità dei testi appartenenti a questa tipologia testuale. È stata inoltre sperimentata la possibilità di modellare i livelli di leggibilità previsti dalla versione base di READ-IT in relazione al profilo linguistico estratto automaticamente a partire da corpora rappresentativi delle principali classi scolastiche in cui si articola il ciclo formativo.

Per maggiori approfondimenti sulle metodologie finalizzate alla valutazione automatica della leggibilità dei testi e sullo strumento READ-IT, si rimanda a:

- Dell'Orletta F., Wieling M., Cimino A., Venturi G., Montemagni S. (2014) "*Assessing the Readability of Sentences: Which Corpora and Features?*". In Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014), 26 June, Baltimore, Maryland, USA.
- Dell'Orletta F., Montemagni S., Venturi G. "*READ-IT: assessing readability of Italian texts with a view to text simplification*". In: SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (Edimburgo, UK, 30 Luglio 2011). Proceedings, pp. 73 – 83. Association for Computational Linguistics Stroudsburg, PA, USA, 2011.