Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico

Felice dell'Orletta, Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR

Se da una lato le tecnologie linguistico-computazionali svolgono un ruolo ormai indiscusso per l'accesso al contenuto testuale, sia esso rappresentato dalla conoscenza specifica di un dominio oppure dalla conoscenza linguistica sottostante (es. collocazioni, strutture argomentali, relazioni semantico-lessicali tra parole, ecc.), ciò non appare scontato quando si vada a considerare il loro ruolo nella valutazione della competenza linguistica di apprendenti. La presente comunicazione intende indagare questo interrogativo, in particolare se e in che misura le tecnologie linguistico-computazionali possano costituire un valido ausilio nella valutazione della competenza linguistica italiana di studenti in ambito scolastico.

L'intuizione di partenza riguardante il "potere diagnostico" delle tecnologie linguistico-computazionali in questo compito specifico affonda le sue radici in un filone di studi avviato a livello internazionale negli ultimi cinque anni e all'interno del quale analisi linguistiche generate da strumenti di trattamento automatico del linguaggio sono usate per: monitorare lo sviluppo della sintassi nel linguaggio infantile (Sagae *et al.* 2005; Lu 2008); identificare deficit cognitivi attraverso misure di complessità sintattica (Roark *et al.* 2007); misurare la leggibilità di testi per studenti di L1 e L2 (Heilman *et al.* 2007; Collins-Thompson 2005); monitorare la capacità di lettura come componente centrale della competenza linguistica (Schwarm *et al.* 2005; Petersen *et al.* 2009). Sulla scia di questi studi, esperimenti preliminari sono stati anche condotti per il monitoraggio della lingua italiana nelle sue varietà diamesiche, diafasiche e diastratiche (Montemagni 2010). I risultati promettenti raggiunti nell'ambito di queste ricerche mostrano a nostro avviso che le tecnologie linguistico-computazionali cominciano a essere mature per poter essere sfruttate anche in contesti applicativi di monitoraggio della competenza linguistica.

La comunicazione intende esplorare le potenzialità di tali tecnologie nel monitorare la competenza linguistica degli apprendenti la lingua italiana (primariamente come L1, ma anche come L2) a partire dall'analisi automatica delle loro produzioni scritte. In particolare, si focalizzerà sull'identificazione di parametri che spaziano tra diversi livelli di descrizione linguistica che possano essere di aiuto nel ricostruire il profilo linguistico di chi ha prodotto il testo oggetto dell'analisi. A partire da una piattaforma ormai consolidata e ampiamente sperimentata di metodi e strumenti per il trattamento automatico dell'italiano sviluppati presso l'Istituto di Linguistica

Computazionale del CNR, la comunicazione illustrerà i risultati conseguiti nella scelta e definizione di strumenti di rilevazione di tipo quantitativo riguardanti il grado di competenza linguistica italiana sottostante a produzioni scritte. Tali parametri riguarderanno aspetti della competenza lessicale, morfo-sintattica e sintattica, ad esempio: la ricchezza lessicale e la tipologia del vocabolario usato (definita in relazione ai repertori del Vocabolario di Base) nel testo; la densità lessicale definita come rapporto tra parole piene e parole funzionali; la "complessità" delle strutture sintattiche prodotte, concernente aspetti quali il rapporto tra clausole e periodi, tra complementi e clausole, tra clausole principali e subordinate, così come altre peculiarità quali il livello di saturazione delle valenze verbali, la "profondità" delle catene di dipendenza sintattica, la distribuzione dei tempi e modi verbali. Nella tipologia di parametri proposti, l'aspetto di maggiore novità riguarda quelli basati su "microprelievi" effettuati sul testo annotato al livello morfo-sintattico e a dipendenze che, per quanto includa un inevitabile margine di errore, se appropriatamente esplorato rende possibile l'indagine di aspetti della struttura linguistica altrimenti difficilmente investigabili.

Riferimenti bibliografici

Collins-Thompson, K., Callan, J. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*. Vol. 56, No. 13, 1448-1462.

Heilman, M. , Collins-Thompson, K., Callan, J., 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.

Lu, Xiaofei 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1): 3-28.

Montemagni, S. 2010. Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana. Presentazione tenuta nell'ambito della Giornata di Studio "Lo stato della lingua. Il CNR e l'italiano nel terzo millennio", Roma, 8 marzo 2010, Consiglio Nazionale delle Ricerche - Dipartimento Identità Culturale.

Petersen, S.E., Ostendorf, M. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23: 89-106.

Roark, B., Mitchell, M., and Hollingshead, K. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proc. ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP'07)*, pp. 1-8.

K. Sagae, A. Lavie, and B. MacWhinney. 2005. Automatic measurement of syntactic development in child langugage. In *Proceedings of the 43rd Annual Meeting of the ACL*.

Schwarm, S., and Ostendorf, M. 2005. Reading level assessment using support vector machines and statistical language models. In *Proc. ACL*, pp. 523-530.