

La disponibilità crescente di corpora di grandi dimensioni permette di avere a disposizione una quantità di dati prima inimmaginabile per l'indagine lessicale e la pratica lessicografica. La linguistica computazionale è il complemento essenziale a tale disponibilità di dati perché consente di estrarre informazione utile all'analisi linguistica e al tempo stesso permette di arricchire i testi con livelli di annotazione linguistica in grado di moltiplicarne le potenzialità di analisi. La piena consapevolezza dei metodi computazionali di rappresentazione ed estrazione dell'informazione lessicale è, dunque, una parte ineludibile del bagaglio di competenze del linguista e del lessicografo di nuova generazione. In questo corso, verranno presentati, in maniera chiara ed accessibile a tutti, i principali metodi computazionali per la ricerca semantico-lessicale e per la costruzione di risorse lessicografiche attraverso la ricerca e l'analisi di dati estratti da corpora testuali. Esistono oggi molti strumenti e ambienti per l'esplorazione dei corpora, di fronte ai quali però spesso il linguista si pone solo come mero utente passivo. È invece essenziale che il linguista comprenda a pieno anche il "dietro le quinte" di questi ambienti informatici. L'obiettivo del corso è infatti di formare un linguista come "fruitore attivo" delle nuove tecniche computazionali, in grado di acquisire una consapevolezza critica delle loro potenzialità.

Nel corso verranno affrontati i seguenti temi:

1. metodi per l'annotazione linguistica dei corpora (es. annotazione morfosintattica, lemmatizzazione, analisi sintattica);
3. metodi statistici per l'estrazione di informazione linguistica dai testi: es. collocazioni, espressioni polirematiche, strutture argomentali dei verbi, ecc.
3. induzione di classi paradigmatiche sulla base delle proprietà distribuzionali dei termini lessicali.

Il corso è integrato da un **laboratorio** nel quale gli studenti avranno la possibilità di applicare operativamente le nozioni di statistica lessicale all'analisi di dati estratti da corpora per la compilazione di entrate lessicali. Il laboratorio verrà realizzato in collaborazione con Francesca Masini e in coordinamento con il laboratorio di Lessicografia Computazionale da lei tenuto.

Riferimenti bibliografici

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

*Alessandro Lenci (alessandro.lenci@ling.unipi.it) è ricercatore presso il Dipartimento di Linguistica "T. Bolelli" dell'Università di Pisa. Si occupa prevalentemente di linguistica computazionale e di semantica lessicale.