TRATTAMENTO AUTOMATICO DI VARIETÀ STORICHE DELL'ITALIANO (TRAVASI)

Responsabile scientifico di progetto
Simonetta Montemagni
CNR Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC)

Unità di ricerca
CNR Istituto di Linguistica Computazionale "Antonio Zampolli"; Accademia della Crusca

Responsabili delle unità di ricerca
Simonetta Montemagni (ILC); Marco Biffi (Accademia della Crusca)

Il progetto *TrAVASI* ha lo scopo di rendere fruibili a un pubblico ampio alcuni fra i contenuti digitali messi a disposizione dall'Accademia della Crusca, approntando modalità di accesso avanzate che sfruttino strategie di intelligenza linguistica. Il progetto, finanziato dalla Regione Toscana nell'ambito del Programma Operativo Regionale del Fondo Sociale Europeo 2014-20, prevede, fra gli altri obiettivi, l'alta formazione specialistica di figure professionali innovative basate sull'integrazione di conoscenze tecnologiche, competenze applicative e cultura umanistica. I lavori, che hanno avuto inizio il 1° giugno 2020, si sviluppano lungo due linee di ricerca distinte ma intercorrelate. La prima linea di ricerca riguarda la concezione e la realizzazione di sistemi di estrazione di conoscenza da dizionari storici digitalizzati, predisponendoli per l'interrogazione avanzata. Questa attività ha per oggetto il più ampio dizionario a vocazione storica dell'italiano, cioè il *Grande Dizionario della Lingua Italiana* (*GDLI*), fondato da Salvatore Battaglia: esso è uno strumento lessicografico di fondamentale importanza e si configura, inoltre, come un vasto *corpus* testuale diacronico. Il lavoro sul *GDLI* ha una duplice finalità: l'estrazione dei contenuti e la loro strutturazione in XML (Extensible Markup Language), il formato standard di rappresentazione dei dati testuali presupposto per la modalità di ricerca avanzata. La seconda linea di ricerca mira a costruire lessici computazionali e *corpora* annotati che possano aumentare l'efficacia degli strumenti di annotazione linguistica in diacronia, in diamesia e in diafasia. Essa parte dalla banca dati del *Vocabolario Dinamico dell'Italiano Moderno (VoDIM)*, che riunisce testi scritti e orali dell'italiano post-unitario e conta attualmente circa 20 milioni di parole. Il bilanciamento in diacronia e la sua notevole differenziazione diamesica e diatestuale rendono il *VoDIM* una buona palestra per la specializzazione degli strumenti di annotazione.
http://www.ilc.cnr.it/it/content/travasi

Francesca De Blasi, Manuel Favaro
CNR Istituto di Linguistica Computazionale "Antonio Zampolli"