Allegato

AVVISO PER PROGETTI DI ALTA FORMAZIONE IN AMBITO CULTURALE ATTRAVERSO L'ATTIVAZIONE DI BORSE DI STUDIO, DI BORSE DI RICERCA O ASSEGNI DI RICERCA

Bando ricerca anno 2024

TITOLO PROGETTO: Una nuova risorsa per la storia della lingua italiana: il corpus degli esempi citati nel «Grande dizionario della lingua italiana» (*GDLI*)

ACRONIMO: GDLIplus

Sintesi del progetto (abstract)

Oggi sono disponibili *online* moltissimi testi italiani di ogni epoca: le loro possibilità di interrogazione, tuttavia, sono in genere rudimentali, spesso limitate alla sola ricerca di stringhe di caratteri. Questo patrimonio documentario, poi, ha una natura frammentata e composita: non c'è un corpus testuale che possa considerarsi rappresentativo della storia della lingua italiana. Potrebbe colmare questa lacuna il corpus degli esempi citati nelle voci del *GDLI* (*corpus GDLIplus*).

Pubblicato in 21 volumi tra il 1961 e il 2002, il «Grande dizionario della lingua italiana» (*GDLI*) è il più importante dizionario storico italiano. Come tutti i vocabolari storici, il *GDLI* fonda la descrizione lessicografica delle parole sul ricchissimo corredo di citazioni esemplificative, che coprono l'intera storia dell'italiano. Grazie al lavoro di informatizzazione del *GDLI* che CNR-ILC ha già svolto con l'Accademia della Crusca, possiamo stimare che il *corpus GDLIplus* comprenda oltre due milioni e mezzo di passi, tratti da oltre 14.000 fonti (e oltre 6.000 autori), per un totale di circa 50 milioni di occorrenze.

L'italiano è rimasto a lungo una lingua "scritta": la storia dell'italiano è, di fatto, almeno fino ai *Promessi Sposi*, la storia dell'italiano letterario. Si capisce bene, dunque, come il *corpus GDLIplus* possa essere considerato a pieno titolo una risorsa formidabile per la storia della lingua italiana, utile agli studiosi così come a insegnanti e studenti, fino al cittadino navigatore di Internet.

Il progetto *GDLIplus* si propone di realizzare questa risorsa. A questo fine, sono necessari due ordini di attività.

1-Il corpus deve essere "annotato": ad ogni parola devono, cioè, essere associate informazioni linguistiche (lemma e categoria morfo-sintattica). Nonostante i recenti progressi, i metodi e le tecniche di trattamento automatico del linguaggio non sono immediatamente applicabili ai testi storici, ma necessitano di specializzazioni a vari livelli.

2-L'origine lessicografica dei testi contenuti nel corpus pone problemi specifici di gestione. La questione più macroscopica riguarda il caso in cui un medesimo passo testuale è citato più volte sotto voci diverse. L'implementazione del *corpus GDLIplus* impone la messa a punto di una strategia di gestione degli esempi ripetuti, e prima ancora la costituzione di un metodo per la loro individuazione automatica.