Evoluzione della ricerca linguistica attraverso l’uso di tecnologie informatiche

Il presente articolo punta a descrivere come la spinta diretta dell'evoluzione scientifica di matrice statunitense, avvenuta in particolare nel settore delle tecnologie dell'informazione, abbia condotto la ricerca linguistica degli ultimi vent'anni verso un'espansione quasi esponenziale delle capacità di elaborazione e gestione dei dati linguistici oggetto di indagine. Tale evoluzione ha rappresentato una svolta epocale a sfavore della metodologia di indagine linguistica di tipo tradizionale, che ricorrendo necessariamente all'utilizzo di schedari cartacei, richiedeva tempi molto lunghi per la preparazione dei dati e un personale altamente qualificato composto da numerose unità.

Per contro, l'introduzione all'interno delle modalità di ricerca di sistemi informatici dotati di potenza sempre maggiore in termini di velocità di elaborazione e capacità di immagazzinamento dei dati, ha consentito nel tempo di incrementare la qualità dei risultati prodotti e di gettare le basi per la nascita di nuove discipline quali la Linguistica Computazionale e la Linguistica dei Corpora. Ovviamente tali presupposti hanno fatto sì che nascessero nuovi linguaggi di programmazione appositamente studiati per la gestione e l'elaborazione di testi in formato elettronico quali il Prolog, nato con l'obiettivo di semplificare e favorire la formalizzazione di strutture sintattiche oppure il Sed, l'Awk o il Perl, a loro volta incentrati sul concetto di "espressione regolare", ossia la formalizzazione simbolica di una determinata sequenza ricorsiva di caratteri.

Relativamente all'aspetto grafematico, è necessario altresì delineare il percorso di sviluppo tecnologico avvenuto in seno alla capacità di rappresentazione binaria dei vari simboli grafici che caratterizzano una lingua. Per rispondere alle necessità di coprire l'intera varietà dei sistemi ortografici mondiali senza essere necessariamente costretti a una preventiva traslitterazione a favore dell'alfabeto latino, nel corso degli anni si è assistito allo sviluppo di nuovi sistemi di codifica dei caratteri. In particolare, lo sviluppo di un sistema univoco quale Unicode, prodotto dell'evoluzione dei precedenti standard ASCII e ANSI (quest'ultimo utilizzato nei sistemi Microsoft Windows) consente di poter realizzare corpora in qualsivoglia lingua, comprendendo anche quelle storiche.

Partendo da tali premesse, quindi, verranno ora presentate le operazioni tipiche necessarie per la realizzazione di corpora linguistici di testi scritti. Il primo passo immediatamente successivo alla selezione dei testi in formato cartaceo consiste nella loro conversione in un formato elettronico ad opera di uno strumento di scansione ottica del testo, seguito da un programma di riconoscimento dei singoli elementi testuali.

Successivamente a questa importantissima fase di preparazione dei dati, si dovrà passare ad una seconda fase, più strettamente legata alle caratteristiche proprie delle lingue oggetto di

studio: la cosiddetta tokenizzazione. Con tale termine si intende l'operazione di trasformazione di un testo da un formato fedele alle norme ortografiche in una forma logica di rappresentazione in cui ogni elemento lessicale risulta separato da quelli circostanti, anche laddove la norma ortografica ne imponga l'unione.

Conclusa la fase di preparazione e verifica della correttezza formale dei dati testuali, sarà già possibile procedere a una prima analisi di tipo statistico mediante la preparazione di liste di frequenza. A tale proposito, è necessario menzionare la Legge di Zipf's, calcolo statistico di natura descrittiva ormai radicato nella metodologia della linguistica dei corpora. Secondo tale legge si avrà una distribuzione tale che i token con frequenza maggiore corrisponderanno principalmente a parole di natura grammaticale (articoli, congiunzioni, ecc.) oppure a nomi o verbi tra più comuni, mentre le forme caratterizzate da frequenza molto bassa (hapax legomena) rappresenteranno forme arcaiche, tecnicismi o neologismi: proprio questi ultimi rappresentano elementi di interesse primario ai fini della ricerca linguistica.

Reference

Aarts J. and Meijs W. 1990. "Theory and Practice in Corpus Linguistics." Amsterdam & Atlanta: Rodopi.

Barnbrook G. 1996. "Language and Computers. A Practical Introduction to the Computer Analysis of Language". Edinburgh: Edinburgh University Press.

Roche, E.; Schabes, Y. (eds.) 1997. "Finite-State Language Processing". Cambridge, Massachusetts. The MIT Press.

Grefenstette, Gregory. 1999. "Tokenization." In Hans van Halteren, editor, Syntactic Wordclass Tagging, pages 117-133. Kluwer, Dordrecht.

Manning C.; Schutze, H. 1999. "Foundations of Statistical Natural Language Processing". Cambridge, Massachusetts. The MIT Press.

Brennan, M. 2012. "GAWK: Effective AWK Programming: A User's Guide for GNU AWK". 4th edition, Free Software Foundation Inc.