L'EMPORIO DELLA LINGUA
Parlando, scrivendo, dubitando

La nostra lingua: parliamo di strumenti (2ª parte).

Lavorando con il DVD del "Primo Tesoro della Lingua Letteraria Italiana del Novecento" si esplorano, ricordiamolo, 100 opere letterarie degli ultimi 60 anni, dal "Tempo di uccidere" di Ennio Flanano (1947) a "Caos calmo" di Sandro Veronesi (2006) per un totale di 8.076.576 forme riconducibili ai già citati 94.254 lemmi.
" Non tutti i lemmi del tesoro sono parole in senso usuale, vocaboli. Come potrà vedersi, nel lemmario del tesoro sono catalogati, categorizzati e contestualizzati molti materiali di varia natura, che si possono trovare nei testi, dai simboli ideografici ai numeri e alle date ecc. Più precisamente tra i 94.254 lemmi figurano 26.542 "nomi propri"[1]. […] Vi sono inoltre 49 simboli, 102 sigle, 132 abbreviazioni e 2 acronimi che possono trovarsi anche nel GRADIT. Vi sono poi 1.225 cifre e aggettivi numerali ordinali o cardinali (espressi nei testi ora in parole-numero ora in cifre arabe o romane), 14 ideogrammi e 25 tra prefissi, confissi e suffissi. Infine abbiamo categorizzato come "altro" un ulteriore insieme eterogeneo: 229 parole inventate, come se ne trovano in Buzzati (*affioriccio, ganolsi* ecc.), giocose, come nel *baco del calo del malo, beco del chelo del melo* ecc. evocato da Natalia Ginzburg. Ancora: abbiamo lemmatizzato come "locuzioni" senz'altra aggiunta (da non confondere con le locuzioni polirematiche verbali, sostantivali ecc. di cui diremo oltre), e cercato di sciogliere alcune (59) concrezioni per lo più dialettali, tipo *figliesfaccimm* o altre anche più complesse care all'inventiva napoletana e affioranti specialmente in Via Gemito di Domenico Starnone. In complesso, occorre sottrarre alla cifra complessiva di 94.254 lemmi tali 28.379 lemmi per circoscrivere quei lemmi che corrispondano a vocaboli italiani o dialettali o d'altre lingue straniere e che risultano quindi 65.875. Nel lemmario i dialettalismi sono 3.688 e coprono lo 0,24% delle occorrenze complessive. […] Più folta è la quantità di lemmi marcati come esotismi. Sono nel complesso 8.803, occorrono ciascuno abbastanza raramente e coprono nel complesso lo 0,43% del totale delle occorrenze. Tra i quasi novemila lemmi marcati come esotismi spiccano i latinismi, 1.983 […] 29.701 occorrenze sono occupate da altri esotismi: dominano i francesismi con 11.647 occorrenze […], alcune centinaia di ispanismi e altri vocaboli di lingue meno rappresentate. Come nel GRADIT, in questo Primo tesoro le polirematiche (o lessemi complessi) sono incassate sotto il lemma del primo lessema che le compone. Le polirematiche, a parte le grammaticali (accanto a, nei pressi di, dal momento che ecc.), come è noto caratterizzano gli usi o colloquiali o tecnici della lingua. Nel corpus le occorrenze complessive sono 214.774, riconducibili a 12.916 tipi di polirematica, con un'incidenza di circa tre polirematiche ogni cento parole. Se si tiene conto del fatto che in gran parte si tratta di polirematiche

grammaticali, il ricorso a queste strutture appare modesto ed è sintomo di una certa distanza, oltre che dai linguaggi tecnici, dalla colloquialità." [2]
Questi dati diventano ancor più interessanti se riprendiamo due altre pietre miliari dell'indagine linguistica: il "Lessico di frequenza dell'italiano contemporaneo"[3] e il "Lessico di frequenza dell'italiano parlato"[4]. Entrambi si basano sullo spoglio di mezzo milione di forme; il primo è basato su un corpus di italiano scritto che copre l'arco di tempo compreso tra il 1947 e il 1968, i testi provengono da 5 settori distinti: teatro, romanzi, cinema, periodici, sussidiari[5]. Il sottoinsieme "romanzi" permette dunque tutta una serie di raffronti diretti con i dati del "Primo Tesoro della Lingua Letteraria Italiana del Novecento"; l'unica limitazione è data dalle poliremathiche[6] che all'epoca del primo Lessico di frequenza non erano ancora presenti come categoria nella riflessione scientifica.
Nella prossima puntata presenterò da un lato tutta una serie di progetti e corpora, sia dell'italiano scritto sia del parlato, e dall'altro alcuni confronti e riflessioni sugli elenchi delle 5.000 parole più frequenti dei dizionari menzionati.

Grazie per l'attenzione
dal vostro Giuliano Merz

e-mail: giuliano.merz@uibk.ac.at

---

[1] "12.285 sono antroponimi (first o family names), nomi propri di personaggi storici, reali, o, più spesso, di personaggi fittizi: tra di essi 1.656 sono ipocoristici o vezzeggiativi, soprannomi, cinonimi e altri zoonimi propri, 6.490 sono nomi geografici (di località, regioni, paesi) e toponimi (odonimi, plateonimi ecc.), 3.440 sono nomi di istituzioni, partiti politici, associazioni ecc., 2.503 sono titoli di libri, film, opere, 1.824 sono crononimi (date, orari)"
[2] T. De Mauro: Il corpus, in: T. De Mauro, Primo Tesoro della Lingua Letteraria Italiana del Novecento. Torino/Roma, UTET/Fondazione M. e G. Bellonci onlus 2007; pp. 14-17
[3] U. Bortolini, C. Tagliavini, A. Zampolli: Lessico di frequenza della lingua italiana contemporanea. Milano, IBM, 1971; XXX, 532 pp. (ediz. fuori commercio, cm. 32) e Milano, Garzanti, 1972; 852 pp. (22 cm.). Per l'italiano scritto esiste un lessico parallelo, anch'esso basato sul canonico mezzo milione di parole, quello di Alphonse Juilland / Vincenzo Traversa: Frequency dictionary of Italian words. The Hague, Mouton/de Gruyter 1973
[4] T. De Mauro et al.: Lessico di frequenza dell'italiano parlato. Milano, ETASLIBRI 1993 (con 2 dischetti contenenti tutte le trascrizioni); il corpus integrale è consultabile su un sito dell'Università di Graz (Austria), v. http://languageserver.uni-graz.at/badip/
[5] Il lessico dello Juilland si basa su 5 sottogruppi, ciascuno di 100.000 forme, e precisamente: drammi (6 testi), novelle (39 testi), saggistica (23 testi), stampa (286 articoli), monografie scientifiche (38 opere), tutti compresi nel periodo 1920-1939
[6] Ripropongo per chiarezza la definizione come la fornisce Tullio De Mauro: "poliremathica: gruppo di parole che ha un significato unitario, non desumibile da quello delle parole che lo compongono, sia nell'uso corrente sia in linguaggi tecnico-specialistici, come in italiano vedere rosso "adirarsi" o scala mobile "crescita dei salari al crescere dell'inflazione", ecc.". Il termine è stato introdotto nell'indagine linguistica dallo stesso De Mauro (1995)