INTRODUZIONE

I

L'applicazione dell'informatica all'analisi testuale, in particolare ad attività lessicografiche o alla creazione automatica di *thesauri*, frequenze, concordanze, *indices locorum*, rimari, incipitari, può considerarsi, dai pionieristici lavori di Roberto Busa[1] in poi, una scelta obbligata più che un'alternativa felice alla schedatura e allo spoglio manuale di testi, sia per economia di costi, che per celerità ed esattezza di risultati[2]. Il crescente sviluppo di tecnologie *hardware e software* e il conseguente affermarsi dell'informatica distribuita permette oggi di affiancare il *computer* agli strumenti più tradizionali del lavoro quotidiano dello studioso. La possibilità di consultare anche su piccoli calcolatori corpi cospicui di testi con sistemi di interrogazione semplici e veloci apre così alla ricerca un terreno fertile e insondato perché offre alle esigenze della moderna analisi testuale risultati esatti e totali, non più frutto di parziali lavori di scavo affidati alla fatica e all'approssimazione dello spoglio manuale.

Ma va sottolineato che i vantaggi dell'automatismo, dalla velocità al rigore 'oggettivo' della macchina, si misurano soprattutto sulla qualità dei materiali elaborati, per il cui trattamento è necessario operare nel rispetto assoluto della materia linguistica originale. Nel caso di lingue antiche la cura nel trattamento dei dati testuali deve inoltre confrontarsi con le particolari difficoltà poste da testi che sfuggono per loro natura ad ogni tipo di normalizzazione, ad esempio per ricchezza di varianti grafiche, e presentano particolari fenomeni linguistici non più attestati nelle lingue moderne.

Esemplare per la ricchezza di variazioni grafemiche, la lingua delle lettere di Michelangelo ha costituito in questo senso un terreno ideale di sperimentazione metodologica, costringendoci ad affinare strumenti di analisi e procedure di lavoro.

Dopo una preliminare indagine sulle edizioni disponibili e la scelta del testo di riferimento nell'edizione critica curata da Paola Barocchi e Renzo Ristori per l'Istituto Nazionale di Studi sul Rinascimento di

---

[1] R. Busa, *Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milano 1951. Idem, *Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae...*, Stuttgart 1979-80.

[2] Su questi problemi si veda: R. Busa, *Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum... cit.*; Idem, *Concordances* in *Encyclopedia of Library and Information Science*, New York 1971, vol. 5, 592-604; A. Zampolli, "Trattamento automatico di dati linguistici e linguistica quantitativa" in AA.VV., *Dieci anni di linguistica italiana (1965-1975)*, a cura della Società di Linguistica Italiana, SLI, Roma 1977, 349-370; S. Hockey, *A Guide to Computer Applications in the Humanities*, London 1980; R. Busa, *Fondamenti di informatica linguistica*, Milano 1987.

Firenze tra il 1965 e il 1983, l'analisi dei problemi linguistici del testo di Michelangelo ci ha indotto a rilevare l'importanza degli interventi critici del filologo e la particolare consistenza di fenomeni tipici della lingua del tempo - come il raddoppiamento fonosintattico - oltre alla straordinaria varietà delle scelte grafiche dell'artista anche riguardo allo stesso fenomeno fonetico (*accordo/achordo; giugno/g[i]ugno; guerra/ghuerra*). Un approccio di tipo conservativo, rispettoso della materia linguistica originale, ci ha spinti a costituire un archivio testuale strutturato in modo da permettere all'utente di rilevare gli interventi filologici sul testo, i fenomeni linguistici, e tutte le varianti grafiche, evitando ogni forma di normalizzazione. Queste scelte sono state rese possibili dall'adozione di un *software* - DBT (Data Base Testuale) - sviluppato da Eugenio Picchi presso l'Istituto del C.N.R. di Linguistica Computazionale di Pisa - capace di offrire, oltre alle funzionalità proprie di ogni sistema di analisi testuale, una serie di opzioni volte alla rilevazione delle varianti filologiche (integrazioni, espunzioni, ecc.) e morfologiche, e alla gestione diacronica del materiale linguistico.

All'analisi dei requisti e alla scelta del *software* ha fatto seguito il trasferimento del testo michelangiolesco in una forma leggibile dall'elaboratore, in *machine readable form*. Alcuni programmi di lettura automatica - chiamati OCR (*optical character recognition*) - hanno permesso di ridurre i tempi di immissione, acquisendo i testi a stampa direttamente da apparecchi per lettura ottica. L'efficienza e la funzionalità di questi sistemi, influenzate da fattori come la qualità della stampa, il tipo dei caratteri tipografici, lo stato e il colore della carta, sono state sperimentate con esito felice per l'edizione critica di Paola Barocchi e Renzo Ristori. La qualità della stampa dei volumi infatti ha pienamente consentito l'automatizzazione della fase di immissione e ha reso meno gravoso il controllo dei dati, sempre necessario ad eliminare eventuali errori di lettura.

Il testo così acquisito è stato poi sottoposto alle codifiche necessarie al programma di elaborazione per evidenziare gli elementi linguistici da trattare e per stabilire la collocazione delle singole occorrenze. Per questa sono stati utilizzati due diversi tipi di indicazione: il riferimento logico e il riferimento topografico. Le singole lettere sono state prescelte come unità logiche di riferimento; il riferimento topografico di ogni occorrenza è invece costituito dal volume, dalla pagina e dalla riga dell'edizione prescelta.

A sostegno del lavoro manuale di codifica procedimenti di tipo automatico hanno permesso di segnalare in breve tempo con codici opportuni i fenomeni che intendevamo rilevare: integrazioni ed espunzioni, accenti, *trait d'union*, apici, ecc. Abbiamo trattato con simboli particolari le parole interessate dal fenomeno del raddoppiamento fonosintattico per facilitarne la ricerca e per ottenere l'elenco completo di tutte le occorrenze sintagmatiche. Accanto alle forme abbreviate, un codice apposito permette al programma di distinguere il punto di abbreviazione dal punto fermo di interpunzione. I numeri

e le stringhe da non indicizzare sono stati racchiusi da appositi delimitatori.

Confermando il nostro assoluto rispetto della grafia dell'edizione critica, siamo intervenuti nei lemmi delle concordanze e dell'indice di frequenza per togliere alle parole non più contestualizzate le maiuscole combinatorie, cioè apposte loro nel testo in conseguenza di un punto fermo precedente.

Le parole latine sono state codificate in modo da apparire visualizzate in corsivo.

Preparato con queste codifiche, il testo è stato quindi indicizzato e sono state estratte le prime liste di frequenza per verificare gli errori ed eseguire controlli incrociati sui testi. Dopo la fase di correzione, il testo è stato infine predisposto alla consultazione interattiva con l'elaboratore. La flessibilità del programma di gestione dell'archivio testuale memorizzato, disponibile su disco magnetico, offre agli studiosi la possibilità di fare al sistema richieste personali e di ottenere risultati corrispondenti a esigenze di studio specifiche. Oltre a tale possibilità la stampa dell'indice di frequenza e delle concordanze per forma, eseguita in modo totalmente automatico (cioè con rinuncia all'intervento manuale nella lemmatizzazione e anche alla lemmatizzazione automatica[3], per la quale manca un dizionario di riferimento della lingua antica), consegna all'utente strumenti di lavoro più tradizionali ma più utili in proporzione della rapidità di approntamento[4].

Quanto alle concordanze, la piccola quantità dei dati testuali non ha posto, nella edizione a stampa, il problema dell'abbattimento delle parole di più alta frequenza, e si è mantenuta la loro completa attestazione sia nella stampa che nell'archivio memorizzato su supporto magnetico.

SONIA MAFFEI

II

Nato dalle decennali esperienze maturate nell'Istituto di Linguistica Computazionale (ILC) del Consiglio Nazionale delle Ricerche di Pisa, DBT (Data Base Testuale) si propone di offrire strumenti moderni e specifici nel campo dell'analisi testuale utilizzando le potenzialità offerte dalle nuove tecnologie informatiche sia *hardware* che *software*.

---

[3] Si veda, da ultimo, per l'uso della lemmatizzazione automatica T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*, Milano 1993.

[4] Cfr. G. Nencioni, "Concordanze per forma o lemmatizzate?", Bollettino d'Informazione del Centro di Ricerche Informatiche per i Beni Culturali della Scuola Normale Superiore di Pisa, III, 1 (1993), 53-56.

Nel rispetto delle procedure già seguite presso l'ILC e delle ricerche in corso, le linee guida, la progettazione e lo sviluppo del sistema sono state le seguenti:

* totale rispetto delle qualità lessicografiche del materiale testuale da elaborare;
* capacità di ottenere, in tempo reale ed in maniera interattiva, tutte le funzioni tipiche di un sistema automatizzato di analisi testuale (ricerca all'interno del testo, calcolo delle frequenze, concordanze, *index locorum*, ecc.);
* ottimizzazione, in termini di occupazione, delle memorie di massa per la gestione degli archivi indispensabili al funzionamento del sistema;
* ottimizzazione dei tempi di risposta del sistema;
* possibilità di accesso diretto e di applicazione delle funzioni non ai singoli testi ma su *corpora* di grandi dimensioni;
* facilità di integrazione del sistema testuale con altre procedure più complesse ed articolate per svolgere ricerche ed elaborazioni specializzate.

DBT è stato concepito come un sistema di analisi testuale attento ai problemi filologici e adatto alla ricerca di fenomeni linguistici particolari. L'insieme delle sue funzioni offre allo studioso uno strumento agevole e duttile, capace di adattarsi alle esigenze più specifiche. La natura modulare del sistema consente l'integrazione con procedure di analisi ulteriori, ad esempio funzioni per la redazione di lessici, dizionari o glossari (di autore, genere o periodo) basati su archivi testuali di riferimento, lemmatizzazione semiautomatica ecc. Oltre alle liste a stampa delle frequenze e delle concordanze del testo vasariano, la consultazione diretta con DBT consentirà quindi all'utente di utilizzare la gamma più ampia di strumenti per l'analisi del testo, rendendo disponibile anche la parte del materiale non inserita nella stampa.

Una presentazione completa del programma di interrogazione, delle singole funzioni disponibili e delle modalità operative di utilizzo sarà inserita in un manuale d'uso distribuito insieme all'archivio testuale informatizzato. Per dare una dimostrazione delle potenzialità dello strumento e dei tipi di ricerca possibili, riteniamo comunque utile presentare una breve descrizione delle principali funzioni di interrogazione disponibili:

** ricerca di parole;
** ricerca di tutte le parole del testo contenenti una o più stringhe di caratteri (tale funzione può ad esempio permettere la ricerca di tutte le parole relative ad un determinato paradigma, ovviando alla mancanza di una lemmatizzazione del testo);
** calcolo delle frequenze di ogni parola ritrovata, sia in un singolo testo che nell'intero *corpus*;
** visualizzazione dei contesti di una parola con diverse modalità di presentazione (possibilità di modificare l'ampiezza del contesto; visualizzazione con modalità KWIC (Keyword in Context), cioè unica riga e parola-chiave al centro, oppure con la parola-chiave situata al centro e un ugual numero di parole che precedono e seguono);

** definizione di funzioni di ricerca nel testo con più parole e con operatori logici (l'operatore AND identifica la funzione per le co-occorrenze, mentre OR determina e riunisce in una unica unità gruppi di parole omogenee).

Seguono delle figure che illustrano alcuni esempi di interrogazione sull'intero *corpus* vasariano: la ricerca della forma "maniera" (fig.1), la visualizzazione dei contesti ristretti della stessa parola (fig.2), e del contesto ampliato di una delle occorrenze (fig.3).

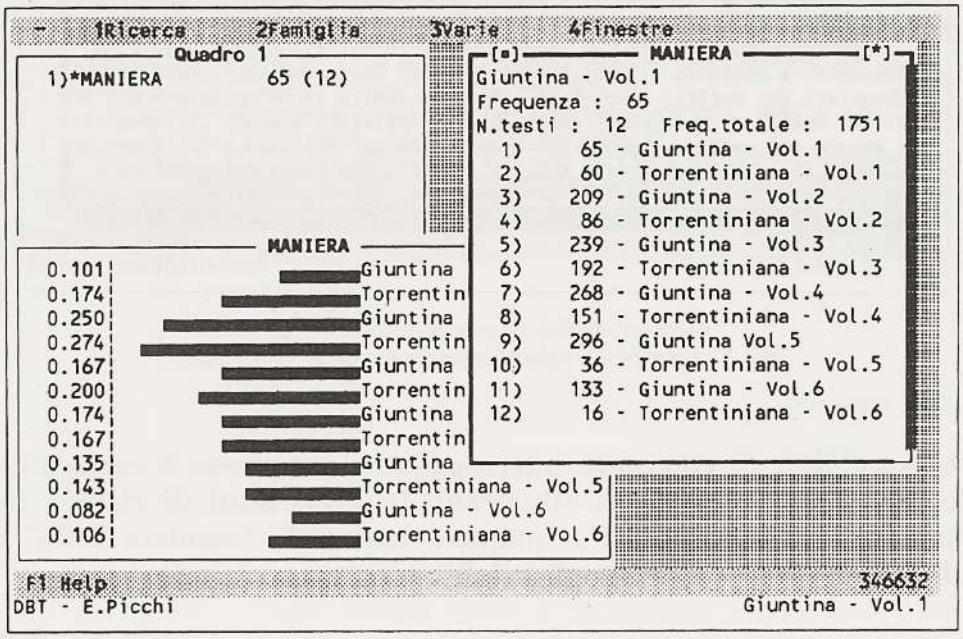

fig. 1 - Esempio di ricerca della forma "maniera"

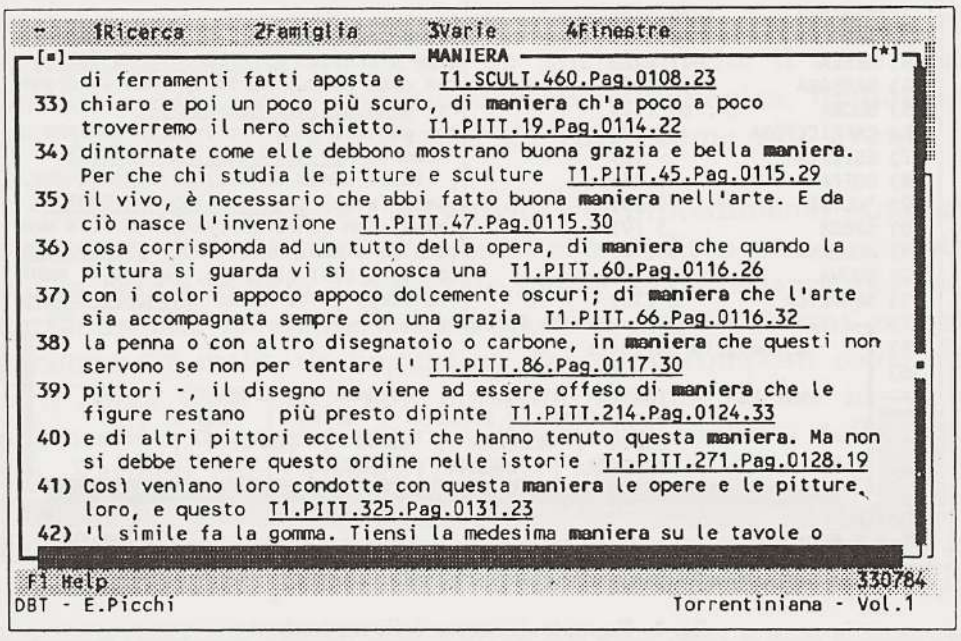

fig. 2 - Esempio di visualizzazione di contesti ristretti

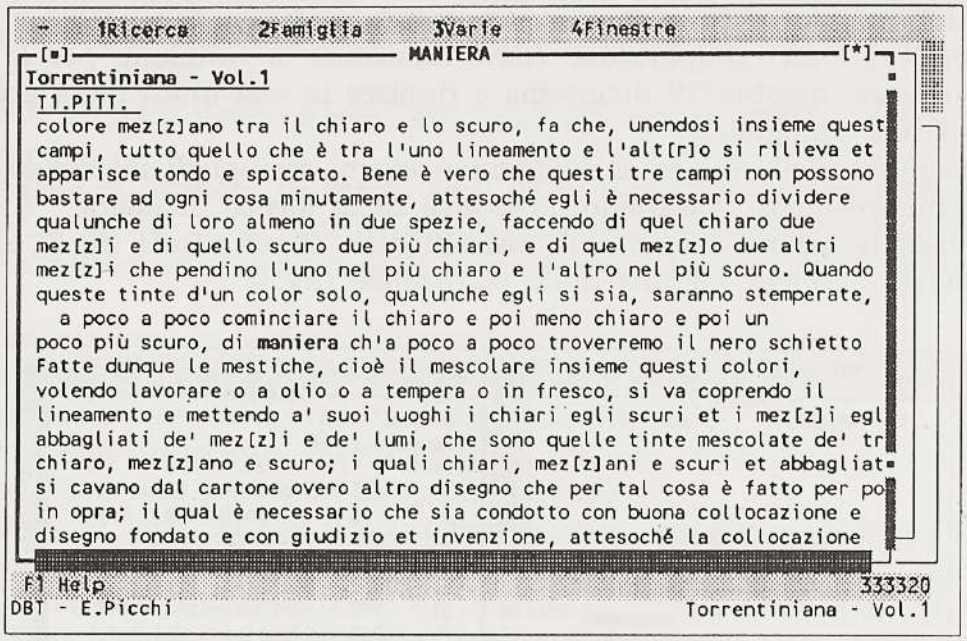

fig. 3 - Esempio di visualizzazione di contesti ampliati

La possibilità di compiere interrogazioni complesse è esemplificata dalle figure successive che illustrano le procedure di ricerca delle espressioni "maniera antica", "maniera barbara", "maniera bella" ecc. (fig.4) e i risultati dello spoglio dell'intera opera (fig.5).

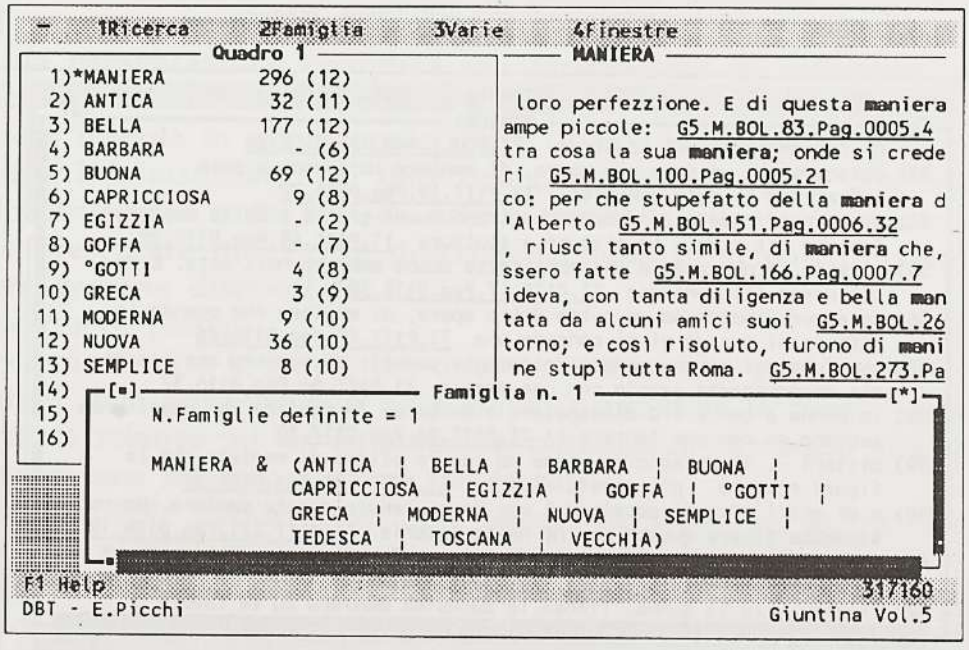

fig. 4 - Esempio di ricerca delle espressioni
"maniera antica", "maniera barbara", "maniera bella"

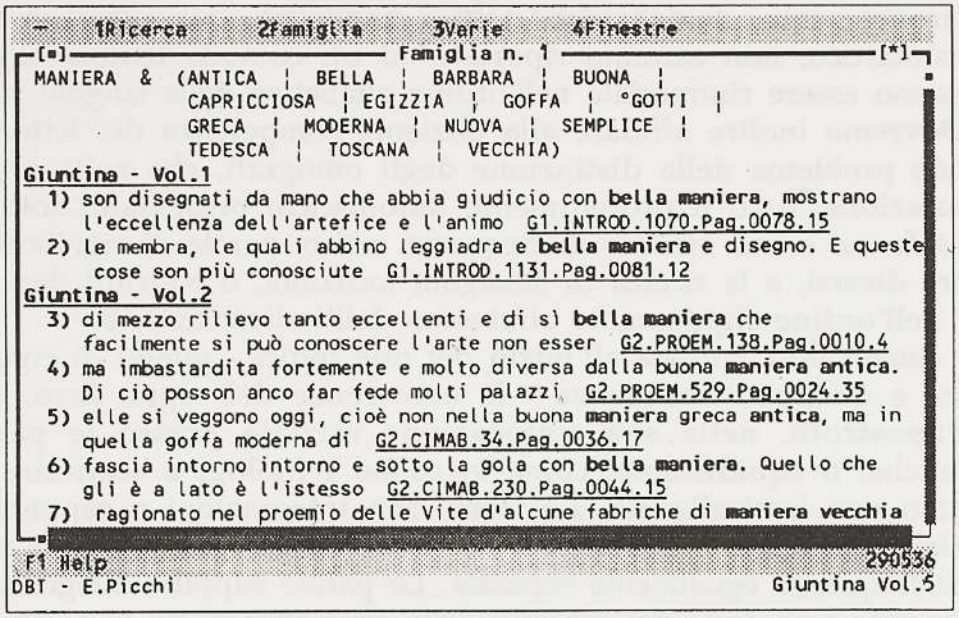

fig.5 - Esempio di spoglio dell'intera opera

EUGENIO PICCHI

III

Gl'indici e le concordanze che presentiamo rispettano fedelmente il testo dell'edizione curata da Paola Barocchi e Renzo Ristori per l'Istituto Nazionale di Studi sul Rinascimento di Firenze tra il 1965 e il 1983[5]. Di quel carteggio sono state sottoposte ad elaborazione informatica soltanto le lettere scritte da Michelangelo. Il testo è stato riprodotto con assoluto rispetto dell'opera filologica, limitandosi i nostri interventi a fatti estranei ad essa, come più avanti sarà debitamente specificato.

Il criterio di conservare l'oggettività del procedimento automatico e la speranza di fornire in tempi brevi allo studioso uno strumento di lavoro ci hanno indotto a realizzare indice e concordanze per forma, rinunciando alla lemmatizzazione anche a scapito di una maggiore dispersione del materiale linguistico. Toponimi composti come *Monte Lupo* o *Borgo Ogni Santi* non potranno dunque essere identificati direttamente e dovranno essere verificati nel contesto della concordanza. Varianti come *Genova/Gienova; Michaelangelo/ Michelagniol/ Michelagniolo/ Michelagnolo/Michelagnuolo/Michelangelo/Michelangiolo; accettato/accectato; accordo/achordo/a cordo; pazienza/patienza; araldo/araudo; avuto/auto*, per non parlare delle varianti sintagmatiche, quali quelle

[5] *Il Carteggio di Michelangelo*, edizione postuma di Giovanni Poggi a cura di Paola Barocchi e Renzo Ristori, Firenze 1965-1983.

del troncamento, dell'elisione, della *i* prostetica, del rafforzamento fonosintattico, non saranno riportate ad un comune lemma base e dovranno essere rintracciate nell'ordine alfabetico delle singole parole. Dovremo inoltre affidare alla paziente competenza del lettore il grande problema della distinzione degli omografi, che nella nostra elaborazione, dato il procedimento automatico, presentano sotto la stessa forma e con numero di frequenza unico, parole di significato e natura diversi, e la ricerca di sintagmi, locuzioni, o costrutti disseminati nell'ordine rigidamente alfabetico dell'indicizzazione.

Il consultatore troverà all'inizio dei due indici - quello di concordanza e quello di frequenza - le occorrenze del segno zero, cioè dell'apostrofo, nella sua collocazione iniziale presso le parole aferetiche, o equidistante, come lo usano i filologi a sostituire un articolo non lessicalizzato. Le forme con integrazioni o espunzioni parziali compaiono separatamente dalle forme integre, con indicazione di frequenza egualmente separata. Le parole supplite integralmente vengono riprodotte nel contesto delle concordanze ma non compaiono a lemma né nelle concordanze né nell'indice di frequenza, non costituendo realtà linguistica. Le pochissime parole espunte integralmente, indicate con parentesi uncinate, compaiono nelle concordanze e nell'indice di frequenza con propria indicazione di frequenza. I frammenti autografi di parole non ricostruibili in luoghi lacunosi del testo compaiono a lemma nelle concordanze e nell'indice di frequenza e sono contrassegnati da parentesi quadre rivolte nel senso della lacuna. Il punto in alto, utilizzato nell'edizione per segnalare il raddoppiamento fonosintattico, è stato conservato e compare accanto a ciascuno dei due termini interessati dal fenomeno[6], che nell'indice di frequenza il lettore potrà analizzare nella sua casistica e incidenza, essendogli presentato, oltre che negli elementi isolati, nelle sue diverse combinazioni sintagmatiche. Le parole latine compaiono, distinte dal corsivo, nello stesso indice di quelle volgari.

La forma adottata per le concordanze è un indice di tipo *KWIC* (*Key word in context*) con la parola-chiave al centro e il contesto non più ampio di una riga. Il taglio del contesto è stato attuato per via automatica, ma tiene conto della suddivisione del testo in frasi; non compaiono infatti nella riga di contesto le parole che appartengono a frasi diverse da quella in cui si trova la parola riportata a lemma. Il numero di frequenza compare tra parentesi tonde accanto al lemma in grassetto. All'interno di ogni lemma, presentato in grassetto, l'ordine dei contesti procede secondo la successione dei volumi dell'edizione di riferimento. Sul margine destro della pagina sono i riferimenti logici e topografici al testo: il volume dell'edizione di riferimento; l'indicazione in numeri romani della lettera a cui si riferisce il contesto; l'anno di datazione; la pagina e la riga del volume (non contando la prima riga quando contiene il solo numero della lettera).

UMBERTO PARRINI

[6] Ad esempio le due parole del sintagma *co·lui* compariranno come *co·* e *·llui*.