

Il soccorso informatico per lo studio di uno scrittore difficile, C.E. Gadda.

Ceccotti Maria Luigia, Sassi Manuela, Pardelli Gabriella
Istituto di Linguistica Computazionale del CNR
Via Alfieri 1
56100 PISA
luigia@ilc.pi.cnr.it, manuela@ilc.pi.cnr.it, gabriella@ilc.pi.cnr.it

ABSTRACT

Nella prima parte di questo contributo si illustreranno le caratteristiche dell'Archivio Gadda in DBT, frutto di un lungo lavoro redazionale, di transcodifica e di codifica. Nella seconda sarà brevemente descritta e motivata la realizzabilità di un sito web su Gadda, un modello di 'laboratorio culturale' che, costituito inizialmente da alcuni brani gaddiani, da nostre recenti pubblicazioni, da dati bibliografici, potrebbe essere arricchito dall'apporto del lettore di Gadda, studioso, studente, curioso...

Parte prima

Il sistema DBT è un sistema interattivo testuale, che tramite funzioni accessibili mediante appositi menu, permette la lettura e la ricerca testuale in:

- un **archivio**, che può corrispondere ad un unico testo o a più testi,
- un **corpus**, un insieme di archivi.

Tale sistema, che permette la lettura dinamica in quanto ogni parola del testo può essere utilizzata per eseguire ricerche nell'archivio o nel corpus, propone vari livelli di analisi, che sono attivati solamente dall'inserimento corretto di codici. Per semplificare distinguiamo un livello base ed un livello codificato.

Livello base

Un file, contenente un testo in machine-readable form, diventa un file-input-DBT con l'inserimento di un unico codice, il codice di riferimento logico che, inserito nel primo record del file, è l'identificatore di un'unità logica del testo, unità logica che può coincidere con tutto il testo.

Il sistema riconosce i segni di interpunzione e le stringhe alfabetiche delle quali fornisce le concordanze, l'index locorum, l'incipitario, l'explicitario, l'indice alfabetico, l'indice inverso, l'indice decrescente, l'indice di leggibilità; produce l'analisi di sequenze di parole, di sequenze di caratteri e di preposizioni; inoltre mette a disposizione la funzione che calcola le co-occorrenze statistiche (mutual information).

Livello codificato

Un testo non è costituito di semplici stringhe alfabetiche e di segni di interpunzione: ci sono i nomi propri, i numeri, i segni speciali quali il trait d'union, sono spesso presenti parole di altre lingue, disegni, formule, note che devono essere distinte dal testo e così via. E' l'inserimento quindi di codici, segnalatori (tags) di fenomeni linguistici, che consente di trasformare un testo piatto (raw text) in un testo strutturato, dati in informazioni, da selezionare ed estrarre tramite funzioni di ricerca, che vengono attivate dal sistema solo in presenza nel testo input dei codici richiesti.

Conviene precisare che tutte le funzioni di ricerca del livello base sono attive anche a questo livello, permettendo, ad esempio, l'index locorum delle parole francesi presenti nell'archivio, concordanze

distinte per le forme (Guisa e guisa, Oliva e oliva) che il sistema a livello base riconoscerebbe come unica forma (guisa, oliva).

L'Archivio elettronico delle Opere di C.E.Gadda è stato realizzato partendo da file per macchine tipografiche¹ che, 'ripuliti' dai codici predisposti per la fotocomposizione, hanno avuto una vita brevissima perché hanno subito delle modificazioni riguardo all'impaginazione per vincoli imposti dalla versione del DBT disponibile nel 1994, quando il progetto di un Archivio elettronico delle Opere di Gadda ha avuto inizio.

Ogni data, numero o sigla, ad esempio, in un file-input-DBT, deve essere compresa in un unico record per permettere l'applicazione dei relativi codici; le note a piè di pagina sono da collocare fisicamente nel file-input subito dopo la parola che le 'chiama'; i disegni e le formule sono state inserite in appositi file. Successivamente è stata affrontata la fase di codifica² dei testi.

Riferiamo qui brevemente i codici applicati e ciò che il loro corretto inserimento permette in fase di consultazione.

- Codice di **riferimento topografico**: consente la localizzazione fisica dell'occorrenza di una parola nel testo ed è visualizzato, dopo il contesto di ogni occorrenza della parola, insieme con il riferimento logico.
- Codice **corsivo**: informa il DBT della presenza nel testo cartaceo di una sottostringa, di una parola o di un insieme di parole in corsivo.
- Codice **segno speciale**: precede quei segni speciali, quali '-', '!', che non sono da valutare come segni di punteggiatura ma trait d'union, segno d'abbreviazione.
- Codice **maiuscola**: permette di poter distinguere la parola iniziante con maiuscola dopo il punto dalla parola iniziante con maiuscola per norma o per volontà gaddiana, non nella visualizzazione della pagina elettronica, dove la maiuscola compare sempre tale, ma nella ricerca della parola.
- Codice **legame**: permette di trattare come unità lessicale più stringhe di caratteri (sic_et_simpliciter).
- Codice **personaggio**: permette di attribuire a un personaggio la sua parte di testo, fino a quando un codice analogo venga utilizzato per un altro personaggio o fino a quando si ritorna al testo narrativo.
- Codice **poesia**: è una coppia di codici, di cui il primo segnala l'inizio di un testo poetico, il secondo il ritorno alla prosa, se necessario.
- Codice **linguaggio**: questo codice può essere utilizzato per costruire non solo sottoinsiemi di lingue naturali ma anche di linguaggi speciali o di parole che si vogliono raggruppare per una particolare caratteristica. E' stato utilizzato:
 - per le lingue naturali
 - per sottoinsiemi quali il linguaggio editoriale, perigrafico, dei titoli o delle note.Il testo così codificato, in fase di consultazione, può essere interrogato per un linguaggio alla volta o possono essere attivati tutti i linguaggi insieme, quindi l'intero testo; in quest'ultimo caso il DBT segnala comunque per ogni parola il sottoinsieme di cui fa parte.
- Codice **data, numero, abbreviazione**: permette di indicizzare numeri, date, abbreviazioni. E' obbligatorio che ogni codice sia aperto e chiuso sullo stesso record, record che può contenere anche più di un codice di questo tipo. Il DBT permette anche una versione di questi codici cosiddetta

¹ La Garzanti Editrice s.p.a. li ha concessi all'ILC insieme all'autorizzazione ad utilizzarli per motivi di studio e di didattica.

² Utilizziamo qui il termine codifica e rimandiamo a Orlandi [1990], p.119, per la distinzione tra codifica e transcodifica di testi.

'normalizzata' in modo da costruire elenchi di date ordinate cronologicamente, di sigle o di numeri che nel testo cartaceo presentano differenze tipografiche (ad esempio la presenza o la mancanza di punti in una sigla, lo stesso numero in lettere o cifre), rilevanti per il computer.

- Codice **nota**: permette, durante la lettura di un'unità logica o di un contesto allargato, la visualizzazione delle note segnalate da icone (links). Il testo di tali note può essere indicizzato o solamente visualizzato.

Nei testi gaddiani:

- le note esplicative/storiche sono contrassegnate da un numero, lo stesso che compare nel testo, e sono in alcune opere a piè di pagina, in altre a fine testo,
- le varianti alternative/evolutive sono contrassegnate da una lettera in corsivo, la stessa che compare nel testo, e sono a piè di pagina.

Negli archivi elettronici gaddiani:

- le note di un archivio ricevono dal sistema una numerazione progressiva,
- tutte le note possono essere visualizzate durante la lettura di un'unità logica o di un contesto allargato.
- tutti i numeri e le lettere che nel testo richiamano una nota sono stati indicizzati utilizzando il codice **numero**.
- il contesto allargato di una parola presente in una nota indicizzata visualizza contemporaneamente il testo che la richiama.

Inoltre le note esplicative/storiche che si trovano a fine testo (che in Gadda spesso sono degli excursus storico-letterari) fanno parte del sottoinsieme del linguaggio **nota**, mentre per poterle visualizzare lì dove il testo le richiama, sono state duplicate e corredate del codice nota non indicizzata.

Le note a piè di pagina se esplicative/storiche sono corredate del codice nota indicizzata, se varianti alternative o evolutive del codice nota non indicizzata.

Si veda la Fig.1 nella pagina seguente dove appare la riproduzione di una schermata con la visualizzazione di un brano, tratto da *Meditazione milanese*, prima stesura, parte prima, paragrafo primo, intitolato "Il dato e l'inizio dell'attività relatrice", pp. 628-29, e dei due tipi di note appena descritte.

- Codice **immagine**: permette di visualizzare delle immagini inserite nel testo cartaceo tramite la precedente creazione di un file-album per ogni archivio contenente immagini. Un'icona segnala la presenza di un'immagine quando viene attivata la lettura di un'unità logica o di un contesto allargato. Cliccando sull'icona, appare sullo schermo l'immagine che può essere manipolata usando le funzioni elementari dei programmi di gestione grafica (ingrandimento, riduzione, adeguamento della finestra, contrasto e luminosità). Una volta attivato un file-album si può passare dall'immagine corrente a quella precedente o a quella successiva, con la contemporanea visualizzazione della finestra contenente il testo in cui vi è l'immagine selezionata. Segnaliamo inoltre che l'immagine può essere identificata con una breve descrizione.

Si veda l'esempio proposto in Fig.2 alla pagina seguente, dove si riproduce la schermata del brano tratto da *Meditazione milanese*, seconda stesura, paragrafo terzo, intitolato "La materia e la molteplicità" con la relativa illustrazione di p. 879.

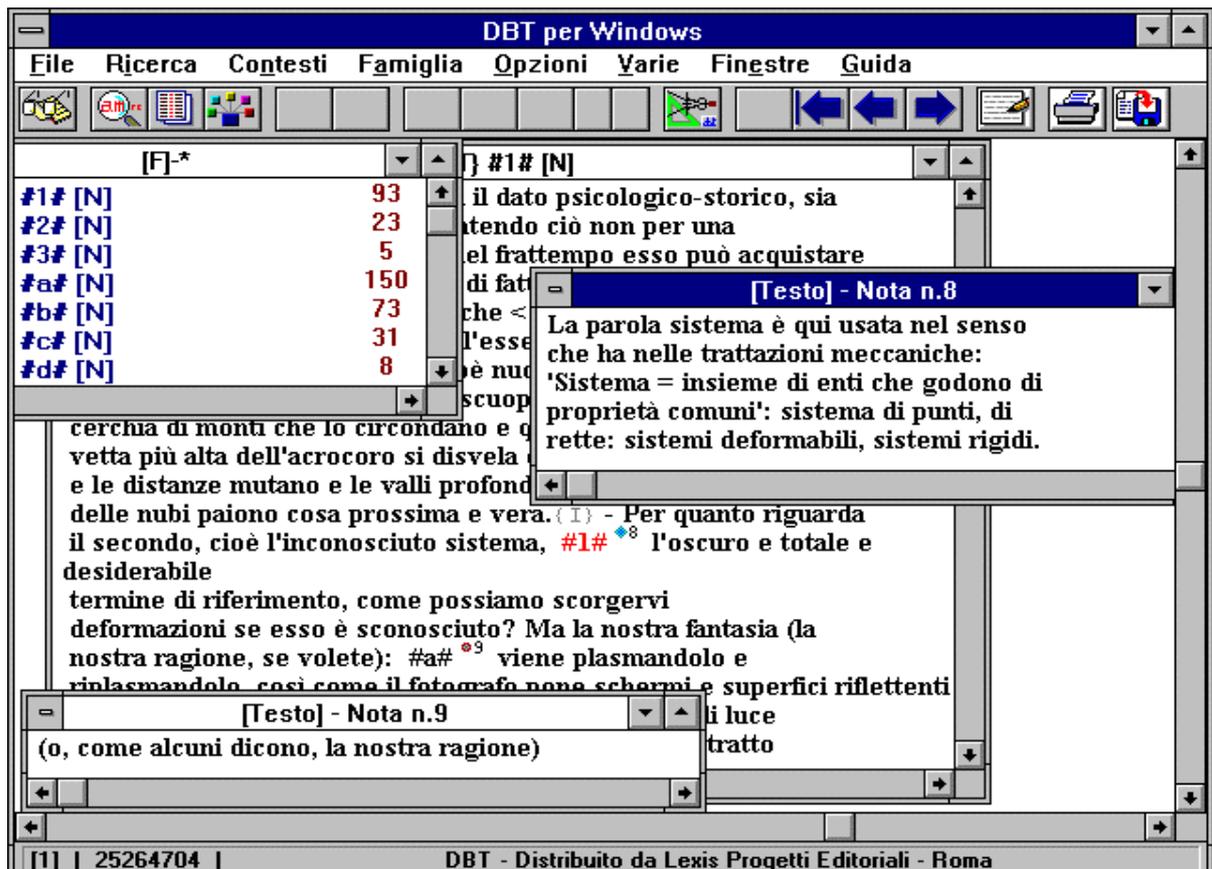


Fig. 1 - Brano con due tipi di note.

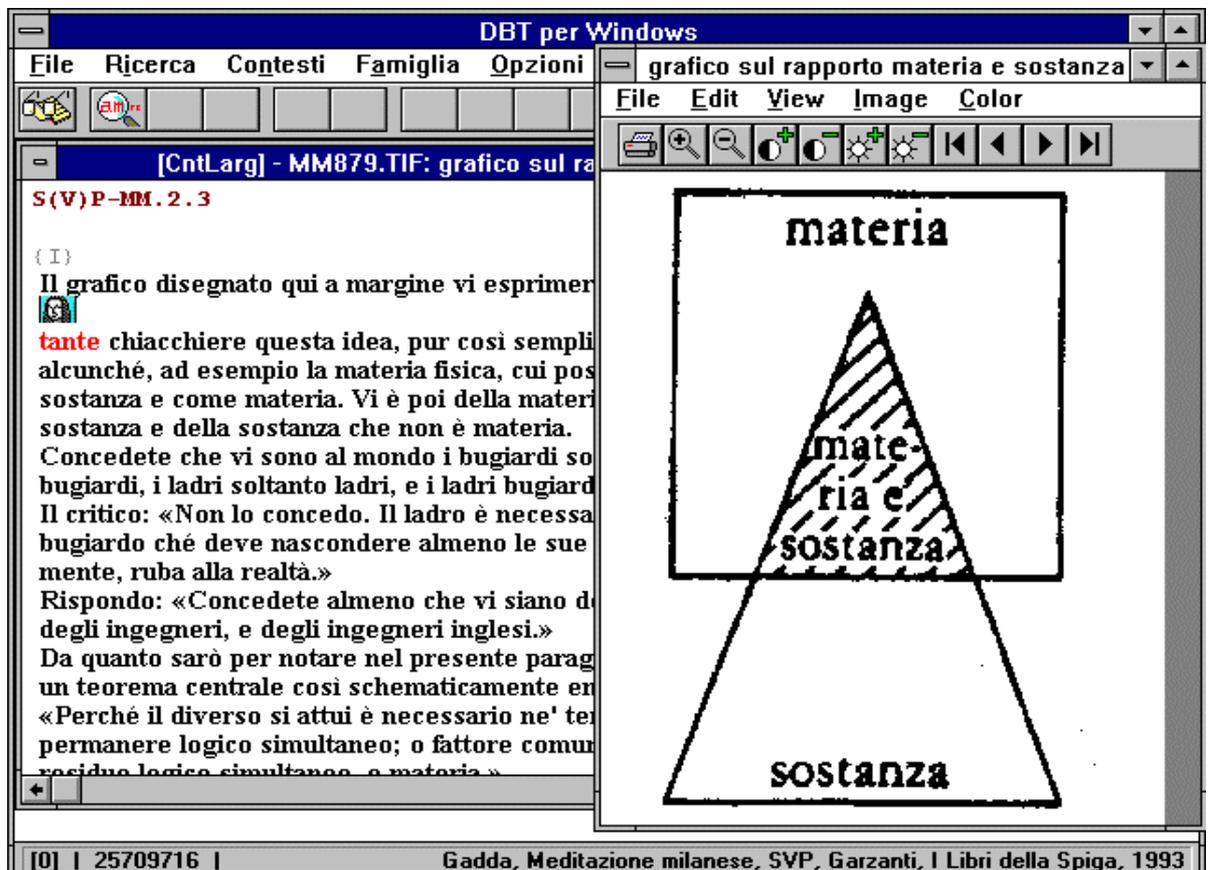


Fig. 2 - Esempio di visualizzazione di immagine collegata al testo

In questa sede non si ritiene opportuno una descrizione dettagliata di come si consulta l'Archivio Gadda³, le cui possibilità, si ribadisce, rispetto ad un archivio per il quale sia stato codificato solo il livello base, sono dovute all'attivazione delle risorse del sistema DBT mediante un faticoso lavoro di inserimento di codici.

Attualmente presso l'ILC di Pisa si sta procedendo con due intenti: costruire strumenti lessicografici che siano ragionevolmente proponibili e su carta e su fogli di lavoro elettronici, lavorare in vista della diffusione dell'Archivio Gadda, qualora la Casa Editrice Garzanti possa acconsentire alla richiesta che da più parti le e ci perviene, tramite un CD e/o un sito Internet.

Parte seconda

Questa sezione del lavoro è dedicata alla presentazione di un progetto che, nell'ambito di uno studio di fattibilità sulla realizzazione di un sito Internet dedicato alle opere di su Gadda, prevede l'implementazione, in formato XML, di un modello di 'laboratorio culturale', i cui elementi iniziali siano alcuni brani gaddiani, nostre pubblicazioni, dati bibliografici, immagini, laboratorio che dovrebbe essere arricchito in modo interattivo dall'apporto del lettore di Gadda, studioso, studente, curioso ...

Carlo Emilio Gadda non è uno scrittore molto famoso e popolare, come lo è ad esempio, nel villaggio globale in cui viviamo, Ernest Hemingway. Nonostante ciò la query su Internet, lanciata (autunno del 1999) su siti italiani con un cercatore diffuso (Excite) e contenente la sola parola "Gadda" ha dato come risultato numerico 346 documenti, mentre quella con "Hemingway" ha dato 359.

Nel 1998 il Web è diventato, pare, il più ricco archivio di documenti testuali superando persino la Library of Congress, ma il Web in cui noi ora navighiamo non è la Biblioteca digitale che auspichiamo per il futuro. Superfluo, ci pare, soffermarci sull'obiettivo, proclamato e ricercato in modo esplicito sin dagli anni '60 e a partire dagli Stati Uniti, di realizzare grandi progetti di biblioteche virtuali costituite da metadati (i cataloghi) e da dati (il contenuto di libri, riviste). I progetti, se portati a termine, nel volgere di pochi anni si sono dimostrati obsoleti per la nascita di strumenti tecnologici più sofisticati e non sempre compatibili con la strumentazione della precedente generazione. Una fatica degna di Sisifo, in breve. E si diffondeva intanto il World Wide Web in modo spontaneo, incontrollato. Il futuro è forse stato concretamente avviato dagli addetti ai lavori nel corso della più importante manifestazione a livello internazionale di Internet, la Internet World, dove, nel dicembre del 1996, è stata proposta una rivoluzionaria tecnologia: la 'information push' che privilegia l'utente-spettatore rispetto all'utente-navigatore in quanto un sito è un canale che trasmette dati automaticamente all'utente interessato a riceverli, utente che, in questa rivoluzione, diventa soggetto, spettatore attivo in quanto riceve dati selezionati ab origine a seconda delle sue richieste prestabilite.

Web è stato ed è ancora un fenomeno spontaneo, in cui solo da pochi anni si sta utilizzando una conquista dell'information technology che è il testo strutturato logicamente. Infatti nel 1986 con la norma ISO 8879 è nato ufficialmente lo Standard Generalized Mark-up Language (SGML), un formato che prevede una codifica logica e non procedurale mediante marcature (tags) e un documento che contiene le istruzioni per interpretare la codifica del documento.

Dal 1997 la Microsoft, la Netscape Corporation ed altri contendenti sono impegnati in questo settore innovativo.

La Microsoft attraverso Internet Explorer 4 offre le sottoscrizioni e i canali. Nel primo caso vengono trasferite periodicamente informazioni, ovviamente quelle aggiornate, da uno o più siti Web a chi ha sottoscritto un abbonamento, nel secondo sono disponibili in ogni momento le informazioni dei canali ai quali si è abbonati.

La Netscape propone sia un servizio di posta elettronica in HTML (In-box direct), e quindi anche messaggi con contenuto multimediale, sia l'attivazione di canali attraverso Java.

Permettete a questo punto una digressione riproponendo anche Gadda attraverso le parole di Giulio Ferroni⁴ che conclude la sua presentazione dell'opera dell'ingegnere-scrittore affermando che “è probabile che le prospettive *ecologiche* permetteranno di riconoscere ancora tutta la grandezza e la

³ In Ceccotti [1997].

⁴ In Ferroni [1991], p. 335.

difficile attualità di Gadda”, e che osserva in seguito⁵: “il contributo della letteratura all'ecologia ... può rivolgersi non a un impossibile recupero di valori tradizionali, ma a una più rigorosa definizione del rapporto tra la società umana e l'ambiente naturale. In molte espressioni della letteratura moderna e contemporanea si può in effetti trovare una acutissima percezione, superiore a quella di molte teorie filosofiche o scientifiche, delle modificazioni subite dall'ambiente e dalle stesse condizioni di vita ...: basta pensare a Saba, preoccupato di recuperare, al di là della degradazione del quotidiano, il valore delle cose, dell'esperienza che si deposita in esse; a Gadda, ossessionato dalla minaccia degli oggetti che si sovrappongono l'uno all'altro, che invadono ogni spazio vitale; Oltre a questa capacità di conoscenza critica di un mondo saturo di oggetti, di un ambiente naturale e storico radicalmente trasformato, la letteratura può avere un'altra funzione *ecologica* essenziale: quella di ricondurre la comunicazione alle sue condizioni naturali, di allontanarci dalla sovrabbondanza di discorsi, di parole, di immagini che percorrono il pianeta, dall'accumulo di comunicazione indifferente che assedia ogni momento della nostra vita. Il rapporto con la letteratura può agire come una sorta di depurazione dall'inquinamento del linguaggio; può aiutare a ritrovare gerarchie e distinzioni tra le diverse sfere dell'esperienza, facendosi strumento di quella che lo psichiatra e antropologo americano Gregory Bateson (1904-1980) ha definito ecologia della mente. Perché ciò sia possibile è necessario ritrovare il piacere e la capacità della lettura: ma per farlo occorre forse porsi anche il problema dell'*ambiente* letterario e librario, della sempre crescente quantità di opere e testi, della sterminata produzione di scritture, non più controllabili da nessun critico e da nessun lettore. Questa eccessiva espansione quantitativa della letteratura pone nuovi, molteplici problemi, che probabilmente dovranno essere affrontati in futuro da una vera e propria ecologia della letteratura che aiuti a liberarsi dai messaggi culturali inutili e indifferenti.”

Proviamo ad estendere queste preoccupazioni alla società dell'informazione nel suo complesso, alle labirintiche pagine Web in cui si può navigare a gratis o a pagamento, e concorderete che ecologia della letteratura fa parte dell'iperonomo ecologia dell'informazione.

Bibliografia

- Bosak J., Bray T., *XML and the second-generation Web*, «Scientific American», 5 (1999).
<http://www.sciam.com/1999/0599issue/0599bosak.html>
- Ceccotti M.L., Sassi M., *L'Archivio elettronico delle Opere di C.E.Gadda: come è stato costruito, come si consulta*, ILC-CNR, S.T.A.R., Pisa, 1997.
- Ceccotti M.L., Sassi M., *Alla ricerca dei termini gaddiani. Una pre-concordanza*, ILC-CNR, S.T.A.R., Pisa, 1999.
- Ferroni G., *Storia della Letteratura italiana - Il Novecento*, Einaudi, Torino, 1991.
- Gadda C. E., *Opere*, Edizione diretta da Dante Isella, collana “I Libri della Spiga”, vv. I-V, Garzanti Editore, Milano, 1988-93.
- Orlandi T., *Informatica umanistica*, La Nuova Italia Scientifica, Roma, 1990.
- Pardelli G., Lanza C., *Sviluppo delle raccolte e procedure di gestione nelle biblioteche dell'ILC e dell'ICAS*, Convegno “La Biblioteca un servizio di rete”, Follonica, 1-2 ottobre 1998.
- Pardelli G., Rini S., *Progetto BIBLOS: realizzazione del prototipo per la messa in linea con INTERNET della banca dati bibliografica dell'Istituto di Linguistica Computazionale*, CNR/ILC/RT 1/98.
<http://soi.cnr.it/ilc/prog.html>
- Picchi E., *Informatica e scienze umane: Procedure di analisi testuale*, in Parola e Immagine, a cura di Maria Assunta Zanetti, (Pubblicazione della Facoltà di Lettere e Filosofia dell'Università di Pavia; 88), Firenze, La Nuova Italia Editrice, 1999, p.181-190.
- Seta E., *Digitalizzazione e linguaggi di marcatura*, Bollettino AIB 39 (1999), n.1/2, p. 63-77.

⁵ Ib. nella scheda dedicata a 'Ecologia e letteratura', p.616-17.