

Le applicazioni di XML ai testi contemporanei: problemi di variantistica

Daniele Silvi (silvi@lettere.uniroma2.it)
Università Roma Tor Vergata (Rome, Italy)

Lorenzo Geri
Università La Sapienza Roma (Rome, Italy)

PAPER

La codifica informatica dei testi letterari, almeno in Italia, è stata per lo più realizzata sulle opere "canoniche" della nostra ricca tradizione letteraria (sino al primo Novecento). Ne consegue un modello di edizione elettronica pensato per la costituzione di banche date testuali. In queste codifiche il testo formalizzato è quello delle edizioni correnti, lo scopo della codifica è, da una parte mettere a disposizione ricchi *corpora* di opere della nostra tradizione, dall'altro quello di fornire raffinati strumenti di indagine con i quali indagare tali *corpora* (concordanze e ricerche con operatori logici). Gli esperimenti di codifica per così dire "filologica" si sono concentrati sui problemi, ingenti, che riguardano la formalizzazione dei manoscritti, dato che nella nostra tradizione accademica tutt'oggi è scarsamente presente la filologia dei testi a stampa (*textual bibliography*). La codifica degli avantesti è poi un campo quasi del tutto inesplorato.

Lo scopo del progetto di ricerca *Codifica e analisi informatica di testi letterari* dell'Università di Roma Tre, diretto da Domenico Fiormonte, è quello di codificare e studiare un insieme omogeneo di testi di un autore contemporaneo dei quali disponiamo di diversi stesure dattiloscritte con correzioni autografe sino alle bozze riviste dall'autore. I testi in questione sono i racconti che Vincenzo Cerami nell'estate del 1993 ha pubblicato, con scadenza settimanale, nell'inserto culturale del quotidiano romano *Il Messaggero*, all'interno di un rubrica dal titolo *Indiscreto*, e ha poi raccolto in volume presso l'editore Einaudi, con il titolo di *La gente*. Nel nostro intervento intendiamo discutere i problemi relativi alla codifica di tali avantesti, partendo da un esempio specifico: la codifica del racconto intitolato *La donna serpente*.

Tale racconto è stato in precedenza studiato nell'ambito del progetto *Digital Variants*. In quel contesto è stata realizzata la codifica in HTML del testo definitivo (quello in volume) e di tre stesure precedenti. I testi sono stati visualizzati in *frames* per analizzare le differenze, utilizzando *java-script*. Il risultato è la *Vincenzo Cerami Variants Machine* realizzata dallo studente Mario Macciocca, al quale si è aggiunta la cura linguistica di Cinzia Pusceddu, dell'Università di Edimburgo. La finalità di tale strumento è essenzialmente didattica, dal momento che permette ai docenti di lingua italiana all'estero di mostrare agli studenti in maniera piuttosto semplice ed intuitiva il processo genetico e creativo di un testo letterario, commentando le scelte linguistiche. Il modello di tale visualizzazione è in qualche modo analogo a quello dell'edizione critica, in cui si riporta l'ultima volontà dell'autore, registrando in apparato le varianti. In questo specifico caso, dal testo definitivo, tramite un sistema di *link*, le varianti sono visualizzate nel contesto di ogni singola stesura; come nei casi della *J.J. Machine* e della *Versioning Machine*.

Dopo aver studiato secondo un punto di vista filologico le due edizioni e gli avantesti, abbiamo scelto di codificare le sei versioni in nostro possesso: l'edizione pubblicata all'interno del *Messaggero* del 10/08/1991, le tre redazioni dattiloscritte con correzioni autografe, le bozze riviste dall'autore, e il testo pubblicato in volume.

Analizzando la codifica in HTML già effettuata con i materiali a nostra disposizione abbiamo notato una notevole perdita di informazione, inevitabile in quanto tale codifica non è dichiarativa ma procedurale.

Abbiamo scelto di codificare i testi utilizzando XML, che permette la conservazione di tutti i fenomeni d'autore e, appoggiandosi per la visualizzazione ai fogli di stile, può gestire diversi tipi di *output*. Siamo partiti, per un principio di economicità dalla DTD TEI Lite, dalla quale abbiamo ricavato un *set* dei tag più

adatti a descrivere i fenomeni rilevati. Questa scelta è stata possibile in quanto la tipologia degli avantesti a nostra disposizione non presenta fenomeni eccessivamente stratificati, e non pone, di fatto, il problema delle *overlapping hierarchies*.

Il *tag set* da noi utilizzato è il più possibile semplice, dal momento che abbiamo scelto di concentrare i nostri sforzi nell'utilizzare le non banali possibilità degli attributi. La nostra codifica, infatti, intende essere un'edizione diplomatica, al quale gli attributi affiancano parallelamente una codifica interpretativa. I fenomeni correttivi e variantistici sono stati segnalati da una parte nel loro aspetto sulla pagina: abbiamo registrato il tipo di strumento utilizzato per correggere, penna o matita, il colore di tale strumento; abbiamo distinto le tipologie di segni; abbiamo registrato dei segni non immediatamente interpretabili nel loro significato. Allo stesso tempo tali fenomeni sono stati interpretati: abbiamo distinto la correzione di un refuso da una *variatio*; abbiamo cercato di categorizzare l'eliminazione di parti del testo, distinguendo l'eliminazione di una ripetizione dall'eliminazione di una porzione testuale sentita in fase di revisione dall'autore come scarsamente efficace.

Per raggiungere questo scopo abbiamo utilizzato principalmente due *tags*:

- `<add>`, generalmente usato per marcare parti di testo aggiunte dall'autore è stato arricchito con l'attributo *place* al quale abbiamo assegnato i valori standard che indicano la posizione dell'aggiunta all'interno del testo (per es. soprilineare, a margine, ecc). Queste informazioni si rivelano invece perdute nella codifica in HTML dove le diverse tipologie di aggiunta sono uniformate.
- `<seg>`, un *tag* generico che serve a marcare stringhe di testo e attribuirgli un valore; nel nostro caso è stato utilizzato per quei fenomeni che non riguardano direttamente la lezione del testo, ma che rivelano intenzioni di intervento da parte dell'autore, da noi interpretate basandoci sulle modifiche effettivamente realizzate nelle versioni successive nelle sezioni testuali in questione. Ad esempio abbiamo indicato le sottolineature e le cerchiature in questo modo:
 - `<seg type="ripetizione" rend="sottolineatura a matita">`
 - `<seg type="ripetizione" rend="tratto cerchiato a matita verde">`

In un caso particolare abbiamo individuato la presenza di una serie di tratti a matita verde posizionati a margine sinistro dei paragrafi. L'ipotesi interpretativa, che non può avere una riprova certa, è che tali segni indichino i nuclei narrativi del racconto. Questi fenomeni, perduti nella precedente realizzazione, sono stati invece conservati mediante l'uso del *tag*:

- `<hi rend="tratto a matita verde"></hi>`

Dal momento che a questo scopo la codifica dei fenomeni correttivi si accompagna ad una interpretazione, abbiamo utilizzato il tag `<resp>` per indicare il responsabile della interpretazione, ed eventualmente rimandare ad una nota critica esterna alla codifica.

Per rendere evidenti le potenzialità di questa tipologia di codifica abbiamo scelto di analizzare automaticamente i dati. A questo scopo, per il momento, siamo ricorsi alla soluzione più semplice, eppure, allo stesso tempo, più potente: generare le concordanze su tutti i testi codificati, tramite il programma, che permette di conservare le informazioni contenute nei *tags* e negli attributi. Tale concordanze ci permettono di studiare concretamente le tipologie correttive, ponendoci domande come: quale racconto contiene il maggior numero di aggiunte? E quale il maggior numero di tagli? Quale parola, tra quelle cancellate, ha la maggiore frequenza? Quale rapporto token/types presentano i brani cancellati? Il trattamento informatico dei dati codificati permette di affinare la critica stilistica e/o linguistica.

La nostra codifica, sia pure ancora da affinare e basata su un *tagset* persistente, crediamo mostri quali prospettive si aprono nel caso si decidesse di studiare un *tagset* dedicato alle bozze e ai dattiloscritti. A differenza dei manoscritti, che per la loro multidimensionalità mostrano inesorabilmente i limiti dell'XML, si può immaginare di arrivare ad una "completezza" della codifica. Naturalmente alcuni fenomeni sono in ogni caso destinati ad essere banalizzati. Su testi lunghi tale lavoro rischierebbe di essere inizialmente antieconomico, ma al termine del faticoso processo di formalizzazione fornirebbe un'importante messe di dati, particolarmente stimolante nel caso di *corpora* testuali, e di testi che presentano diverse edizioni, intervallate da numerose stesure.

CLiP 2006
King's College London, 29 June - 1 July, 2006