DENTRO LA "LIZ", OVVERO L'EDIZIONE DI MILLE TESTI

PASQUALE STOPPELLI

La storia che sto per raccontare ha inizio in un pomeriggio della primavera del 1985 in un'aula del Dipartimento di Italianistica dell'Università di Roma "La Sapienza", dove ricoprivo da professore associato l'insegnamento di Filologia italiana. Dopo una prima parte di corso dedicata al metodo, per addestrare gli studenti sulle tecniche di edizione avevo scelto quell'anno il testo della *Mandragola*. Qualche anno prima era stata pubblicata nella "Biblioteca Universale Rizzoli" una buona edizione della commedia, a cui uno dei due suoi curatori aveva fatto seguire nel 1983 un saggio di critica testuale. C'erano dunque materiali freschi su cui esercitarsi.

Il lavoro in aula doveva ripercorrere, per come era possibile in poco più di venti ore, tutto il processo che dalla collazione dei due testimoni significativi della commedia portava alla costituzione del testo. Quei documenti presentavano delle varianti che non era agevole ricomporre in un unico testo, ma neppure erano a tal punto divergenti da imporre una scelta netta di uno sull'altro. Capitava spesso dinanzi all'incertezza della soluzione di invocare l'*usus scribendi* dell'autore: ma come accertarlo rispetto a un corpus testuale vicino alle tremila pagine? Cominciavano in quegli anni a diffondersi i personal computer; si avevano invece notizie vaghe sugli scanner, macchine che dicevano essere delle fotocopiatrici in grado di "leggere" il testo e restituirlo digitato. Dinanzi all'impossibilità di valutare la maggiore pertinenza delle lezioni concorrenti rispetto alle abitudini di scrittura di Machiavelli a un certo punto mi arresi, dichiarando agli studenti che probabilmente della questione testuale della

Pasquale Stoppelli (Università di Roma "La Sapienza") ha lavorato su testi letterari di area quattro-cinquecentesca e su filologia attributiva, filologia dei testi a stampa, lessicografia e applicazioni informatiche ai testi letterari. Sua recente pubblicazione, *La 'Mandragola': storia e filologia* (2005).

Mandragola si sarebbe potuti venire a capo solo quando si avesse avuto a disposizione l'intera opera dell'autore in formato elettronico.

Alla fine della lezione uno di loro mi si avvicinò, dicendomi che suo padre, un ingegnere con la passione per le nuove tecnologie, disponeva di un computer e di uno scanner e che se avessi voluto avremmo potuto fare delle prove di acquisizione del testo. Cominciammo e la cosa con mia sorpresa riuscì più facile di quanto immaginassi. Naturalmente il corso era finito ancora prima che potessimo disporre del primo file di testi machiavelliani, ma l'esperimento fu comunque positivo ed effettivamente qualcuno dei dubbi sul testo della *Mandragola* per quella via già allora poté essere sciolto. Quello studente si chiamava Lorenzo Bartoli e insegna oggi lingua e letteratura italiana all'Università Autonoma di Madrid. In seguito avrei saputo che il suo nonno materno era stato Umberto Bosco, l'illustre studioso che con Contini aveva mostrato più sensibilità di chiunque altro in ambito italianistico nel riconoscere l'importanza delle concordanze nel lavoro critico, e aveva lui stesso promosso la redazione, allora con metodi manuali, delle concordanze dei *Canti* di Leopardi e addirittura del *Decameron* di Boccaccio. Dato che la realizzazione dei file machiavelliani può essere considerata la posa della prima pietra della LIZ, quella significativa ascendenza familiare, sebbene allora a me sconosciuta, sarebbe stata di ottimo auspicio.

Se il caso non ci avesse messo la mano, non posso tuttavia oggi dire se l'iniziativa sarebbe finita lì o avrebbe comunque avuto un seguito. La circostanza fortunata fu un convegno di aggiornamento del CIDI per insegnanti di lingue straniere, organizzato a Cagliari nel maggio del 1988. Tra gli argomenti dell'incontro era l'uso in classe del dizionario. La mia presenza in quel convegno era stata richiesta dalla Garzanti, casa editrice per la quale avevo curato il rifacimento del *Dizionario della lingua italiana*, pubblicato nel 1987. All'incontro di Cagliari era anche presente Lorenzo Enriques, amministratore delegato della Zanichelli, che era lì per lo Zingarelli e gli altri dizionari della casa editrice bolognese. Enriques e io rappresentavamo in quell'occasione interessi concorrenti, ma nacque tra noi una simpatia personale che sarebbe stata determinante per la messa in cantiere e poi la realizzazione della LIZ. Ma il fatto determinante, seppure fortuito, fu che tra i relatori del convegno fosse anche Eugenio Picchi, allora ricercatore dell'Istituto di Linguistica computazionale del CNR di Pisa, che mostrò in una delle sedute il prototipo del DBT (*Data Base Testuale*, come lui l'aveva battezzato), un software ancora sperimentale da lui sviluppato in linguaggio Pascal, in grado però già di produrre concordanze con una velocità di esecuzione che per i tempi faceva impres-

sione. Ricordo che lo standard dei processori di quelli che allora si chiamavano IBM compatibili era il 286, che il sistema operativo era il DOS, che Windows ancora non si conosceva, che i lettori di CD-ROM erano rari e che i masterizzatori erano apparecchi del costo di molti milioni di lire, di cui si sapeva l'esistenza in poche software house.

Il primo progetto di quella che si sarebbe chiamata LIZ (acronimo di *Letteratura Italiana Zanichelli*, ma anche gradevole ipocoristico femminile di Elizabeth) fu pronto entro il 1989 e con esso la relativa contrattualizzazione. L'obiettivo minimo era quello di concordare elettronicamente, attraverso appunto DBT, almeno cento testi letterari italiani. La riuscita dell'impresa non era sicura, anche se l'esperienza della trasposizione digitale dei testi di Machiavelli costituiva comunque un precedente incoraggiante. Il budget messo a disposizione dalla Zanichelli era consistente per quello che era allora l'ammontare dei finanziamenti pubblici della ricerca in ambito umanistico, ma era comunque ristretto rispetto al lavoro necessario alla preparazione di una mole così imponente di testi. A questo proposito è il caso di anticipare che non solo la prima release, ma tutte le successive sono state realizzate con finanziamenti della casa editrice; strumenti o laboratori universitari non sono stati mai impegnati. Quello che è intervenuto col mondo accademico e istituzionale è stato fin dagli inizi uno scambio alla pari di testi. Mi piace ricordare a questo proposito la contemporanea attività pionieristica di Amedeo Quondam come direttore dell'Istituto di Studi Rinascimentali di Ferrara nel promuovere la digitalizzazione di testi della tradizione petrarchista. Molti testi preparati per la LIZ furono messi a disposizione dell'Istituto ferrarese, e così reciprocamente testi realizzati a Ferrara finirono nella LIZ.

Oggi le biblioteche digitali sono argomento di attualità. Nell'ambito della cosiddetta informatica umanistica esiste una corrente d'opinione che sostiene che il testo elettronico debba riprodurre la fonte a stampa secondo modalità che ne ripetono finanche i particolari più insignificanti. E siccome a questa impostazione la LIZ sia per impossibilità oggettiva sia per scelta non ottempera, spunta di tanto in tanto qualcuno ad atteggiare la bocca a una smorfia di disgusto. Ma a parte l'assurdo di pretendere che ciò che è stato fatto ieri corrisponda a caratteristiche che si vogliono imprescindibili oggi, la LIZ non ha mai aspirato a essere una biblioteca digitale. E non si capisce perché mai dovrebbe esserlo, visto che quasi tutti i testi in essa registrati sono in tutte le biblioteche pubbliche, anche comunali e di quartiere. La LIZ altro non è che un sistema di con-

cordanze elettroniche intratestuali e intertestuali. Ma siccome l'equivoco continua, a evitare che in futuro si travisi ancora la realtà, spero che aprire la porta del suo retrobottega possa servire non solo a chiarire molte cose che dall'esterno appaiono forse inspiegabili, ma anche a partecipare un patrimonio non proprio trascurabile di esperienza, tanto più utilmente oggi che l'acquisizione digitale di testi letterari interessa, oltre che le università, le istituzioni culturali nazionali.

Cominciamo dal protocollo seguito nella preparazione dei testi, rimasto sostanzialmente identico dalla prima release del 1993 fino alla quarta del 2001. Il primo atto è costituito dal passaggio del testo allo scanner. Si era alla fine degli anni Ottanta. Gli scanner e i software di riconoscimento dei caratteri lavoravano con standard di precisione lontani da quelli attuali. Questo comportava che il testo acquisito presentasse una percentuale piuttosto alta di errori. Per la qualità del risultato era allora decisivo (oggi molto meno) il tipo di carta dell'edizione scansionata: quella patinata aveva una resa migliore di quella ruvida o spugnosa; analogamente, la composizione digitale era di gran lunga preferibile a quella in linotype o addirittura in caratteri mobili. I libri stampati prima del 1960 di conseguenza non consentivano in genere un'acquisizione sufficientemente corretta. Questo ha fatto sì che nella scelta del testo da scansionare non sia stato sempre possibile utilizzare l'edizione di riferimento negli studi, ma si è dovuti talora ripiegare su un'altra edizione che si atteneva a quel testo avendo però una migliore resa alla scansione. Perciò dunque nella bibliografia della LIZ si incontra qualche volta il rinvio a volumi di collane divulgative di classici come i "Grandi Libri Garzanti", la "Biblioteca Universale Rizzoli" o gli "Oscar Mondadori" e non direttamente al testo critico originale. In particolare i volumi più vecchi della collana degli "Scrittori d'Italia" di Laterza, depositari delle edizioni critiche di molte opere, avevano una resa pessima allo scanner. Naturalmente il testo nel passaggio dalle pagine dell'edizione critica a quelle della successiva economica poteva avere subìto rimaneggiamenti circa gli accidentali del testo (tipi di accento, di virgolette, paragrafatura ecc.), che rimanevano pertanto tali anche nel testo LIZ. Per non dire degli errori, pressoché ineliminabili in ogni passaggio di copia, presenti nelle edizioni divulgative e che era difficile intercettare a meno che non fossero evidenti. Né, come sa chi conosce l'abbiccì del lavoro editoriale, era possibile dare in correzione un testo scansionato su una fonte su quello di un'altra fonte. Ma c'era un ulteriore problema. Un libro passato allo scanner ne esce fisicamente malconcio. Non si poteva certo chiedere in prestito a una biblioteca un'edizione di pregio o di valore stori-

co e poi restituirla squinternata. Opere di questo genere neanche vanno in prestito. In questo caso o il libro si acquistava se era in commercio o, se non era in commercio, si ripiegava su un'altra soluzione. Questa è la ragione per cui, per citare un caso fra gli altri, le poesie di Carducci si leggono nella LIZ secondo il testo di una apparentemente poco autorevole edizione dei "Classici popolari Bietti", l'unica allora disponibile che stampasse l'intero corpus poetico secondo il testo dell'edizione canonica zanichelliana, della quale invece non si riusciva a disporre.

Già da queste prime note emerge un criterio di pragmaticità adottato nella costituzione del corpus testuale della LIZ. Avere la possibilità di concordare tutte le più importanti opere degli autori più significativi della letteratura italiana sarebbe stato un vantaggio talmente grande per gli studi che, di fronte al fare solo rispettando con assoluto rigore i crismi dell'operare filologico o al non fare, si è preferito comunque fare come meglio si poteva. L'utilità dimostrata negli studi dalla LIZ in questi anni credo abbia confortato la scelta fatta.

Ritornando al protocollo di preparazione dei testi, il file che risultava dall'acquisizione era quanto mai disordinato: spazi superflui, lettere improprie in corrispondenza di macchie della carta, numerazione dei versi o dei paragrafi da eliminare, numeretti di richiamo alle note a piè di pagina da cancellare, e poi soprattutto errori testuali, più o meno abbondanti a seconda della tipologia della fonte di acquisizione. Siamo agli inizi degli anni Novanta: i processori 286 erano intanto stati sostituiti dai 386, che avrebbero a loro volta lasciato il posto ai 486, ma il software di riconoscimento dei caratteri non aveva ancora fatto sostanziali passi avanti. Il testo in questa forma non poteva andare in correzione: era necessario un preliminare trattamento che conferisse a esso per lo meno l'aspetto di una bozza. Il testo, uscito dallo scanner in formato ASCII, veniva aperto con un editor che avevo conosciuto alla Garzanti, allora molto usato nelle tipografie e nelle redazioni dei giornali. Il suo nome era XYWrite e aveva un ingombro di appena 300 kb: non aveva alcuna amichevolezza, ma garantiva prestazioni straordinarie. Soprattutto consentiva a chi non aveva alcuna competenza di programmazione di creare con facilità delle macro, che avevano il vantaggio di poter essere raggruppate in grappoli e lanciate con un unico comando. In poco tempo mettemmo insieme un sistema di correzione consistente in un insieme di parecchie centinaia di macro, ciascuna delle quali eseguiva in automatico delle operazioni di ricerca e sostituzione. Per effetto di questo trattamento il testo da una condizione informe acquisiva un assetto che lo rendeva trattabile manualmente. Soprattutto venivano corretti per

questa via rapidamente gli errori ricorrenti dell'acquisizione ottica. L'abbandono necessario anni più tardi di XYWrite per WinWord non sarebbe avvenuto senza un po' di rammarico.

A questo punto il testo, dopo un rapido scorrimento a video per controllarne l'integrità (facilmente durante la scansione potevano saltare delle pagine o delle pagine potevano essere acquisite due volte), l'eliminazione con altre macro degli accapo di stampa se si trattava di prosa, ma la loro conservazione in caso di poesia o dei capoversi dei paragrafi in prosa, era pronto per essere stampato e dato in correzione manuale come una normale bozza.

La preparazione dei testi di LIZ3 e LIZ4 sarebbe avvenuta in una struttura adeguata, la Lexis, una società di editoria elettronica che avevo intanto contribuito a fondare con alcuni miei collaboratori proprio per realizzare questo lavoro. La prima e la seconda LIZ erano state invece letteralmente fatte in casa, con una rumorosa stampante ad aghi e carta a modulo continuo. La correzione delle bozze per la LIZ3 e 4 fu fatta da correttori di professione. I testi confluiti nella prima e seconda LIZ erano stati invece affidati per la correzione per lo più a giovani laureati, desiderosi di cominciare a guadagnare qualcosa dopo la conclusione degli studi. Alcuni di loro sono oggi accademici brillanti o comunque studiosi noti, ma certo (buon per loro!) non potevano garantire quel tipo di precisione che è propria del correttore di bozze professionale. Non che i testi alla fine del procedimento fossero ancora particolarmente scorretti, ma certo qualche errore restava. Anche perché la quantità dei materiali da processare era comunque elevata. Molti errori sfuggiti alla correzione manuale venivano intercettati successivamente attraverso un controllo alfabetico di tutte le forme del testo fatto attraverso lo stesso DBT. Se la forma scorretta dava luogo a una parola inesistente, era molto probabile che venisse intercettata; se invece corrispondeva a una parola esistente, l'errore sfuggiva a questo controllo. Con l'editore avevamo concordato un limite di tolleranza di un errore sostanziale ogni 10.000 battute. In questa media siamo abbondantemente restati anche se alcuni testi sono del tutto privi di errori e altri possono eccedere seppure di poco la percentuale orientativamente stabilita. C'è da aggiungere che lavorando con finanziamenti privati il raggiungimento dell'obbiettivo non poteva essere procrastinato *sine die*, come avviene non raramente quando la fonte di finanziamento è pubblica. I due anni preventivati diventarono quattro, ma la prima LIZ nel 1993 sarebbe uscita.

La fase più delicata della preparazione del testo era tuttavia la successiva, quella della sua codifica. Oggi sulla codifica del testo letterario si

possono leggere addirittura delle monografie. Esiste una associazione internazionale (la TEI, *Text Encoding Iniziative*) che ha fissato, credo, oltre 400 tipi di codifica, che dovrebbero coprire qualsiasi aspetto della testualità letteraria. L'esigenza che è dietro l'iniziativa è quella di creare uno standard del testo elettronico, e dunque delle biblioteche digitali, indipendente sia dalla lingua del testo sia dai formati sia dal tipo di data base che dovrà acquisirlo. Sulla materia sono nate intanto competenze specifiche e con esse anche speranze di avviare per questa via carriere accademiche. Nei primi anni Novanta si era alla preistoria di questi problemi. Oggi capita ancora di trovare chi col senno di poi svaluta i testi della LIZ perché sarebbero rispondenti a un formato proprietario (quello appunto del DBT), dunque non standard. Un nuovo richiamo alle date è superfluo, ma è soprattutto discutibile l'assunto che è dietro questa riserva e che fa della codifica la questione cruciale del testo elettronico, quasi che l'elemento decisivo di un archivio digitale non fosse, oltre ovviamente alla correttezza testuale, il software che lo interroga. È il motore di ricerca l'anima di una base di dati testuale: la sua efficienza si valuta nella capacità di estrarre dai testi la maggior quantità e la miglior qualità di informazione partendo da una codifica la più leggera possibile. Nel recupero di informazione da archivi testuali la ricerca in questo settore si indirizza oggi verso l'intelligenza artificiale, non verso motori di ricerca a basso rendimento, dunque in grado di essere efficaci solo su dati strutturatissimi. Il tipo di codifica è questione meramente strumentale, addirittura secondaria. Codificare un testo in profondità è dispendioso e dunque insostenibile in relazione a una grande quantità di testi. Ma un archivio testuale che non consista di una grande quantità di testi non ha valore.

Con questo ritorno al DBT. Il prototipo mostrato a Cagliari era intanto cresciuto. Alle funzioni di base si erano aggiunte numerose altre modalità di ricerca, alcune (come la ricerca sequenziale) molto sofisticate. Si restava in ambiente DOS anche se le prime versioni di Windows cominciavano a diffondersi. Nella prima versione ufficiale del software si emulava tuttavia l'ambiente Windows. I dati venivano restituiti in finestre che potevano restare attive sullo schermo in gran numero. I testi per essere macinati in DBT necessitavano di informazioni sussidiarie semplicissime: il titolo dell'unità testuale, la marca distintiva di prosa o verso, l'indicazione di inizio paragrafo per la prosa e un'annotazione aggiuntiva per la poesia in ottava rima (dove i versi andavano conteggiati appunto per ottave e non singolarmente). Il sistema avrebbe consentito tante altre possibilità di classificazione: per esempio, il tipo di lingua (ita-

liano, latino ecc.), i nomi propri ecc. Scegliemmo soltanto di distinguere il tondo dal corsivo. Le parole latine o in altre lingue sarebbero state interrogate insieme a quelle italiane, senza alcuna distinzione. Questa scelta essenziale rendeva possibili le operazioni di codifica con procedure semplificate, dunque sufficientemente rapide. DBT restituiva automaticamente nelle concordanze il numero del verso o dell'ottava e il numero di paragrafo, computato a ogni capoverso, per la prosa. La semplificazione della codifica non era soltanto opportuna per velocizzare la preparazione dei testi: avrebbe reso più facile la consultazione, aspetto non trascurabile considerata la destinazione della banca dati a un'utenza che non poteva certo dirsi naturalmente votata alla tecnologia. Il numero del verso, dell'ottava o del paragrafo era comunque un'indicazione di secondo livello; il primo livello, quello superiore, era dato dal titolo dell'unità testuale, cioè una sequenza da apporre manualmente all'inizio di ognuna di esse. Lo schema strutturale a due livelli doveva essere applicato a tutti i testi. Era l'elemento di omogeneità formale che connotava imprescindibilmente l'intera banca dati. Questo ha comportato talvolta, soprattutto nei testi in poesia della prima LIZ, il sacrificio di materiali paratestuali, come dediche, date ecc., che non appartenevano alla "poesia" e dunque non pertinenti a finire tra le parole concordate. Naturalmente per alcuni testi questa classificazione a due livelli era di per sé pertinente (per esempio, il *Canzoniere* di Petrarca: numero del componimento e numero del verso; per i poemi in ottave: numero del canto e numero dell'ottava); per altri era necessario forzare con degli accorgimenti le gerarchie originarie (il *Decameron*, strutturato su tre livelli – giornata, novella, paragrafo – veniva ricondotto a due livelli ripetendo a ogni riferimento il numero della giornata prima di quello della novella; i quattro livelli della *Scienza nuova* di Vico venivano anch'essi ridotti a due); per altri ancora la scansione dell'opera in unità testuali e in paragrafi poteva comportare la manomissione dell'architettura originaria del testo, e con questo dar luogo a piccoli arbìtri filologici. Questa manomissione riguardava soprattutto quelle opere di secondo Ottocento che esorbitavano dai generi canonici o non rispettavano le tradizionali partizioni in capitoli.

In qualche caso è stato necessario forzare l'edizione di riferimento anche in relazione alla paragrafatura. I capitoli della *Nuova cronica* di Giovanni Villani nell'edizione Porta presentano, per esempio, un testo continuo, senza alcuna scansione in paragrafi, che si estende in molti capitoli per dieci-quindici pagine. In queste condizioni il reperimento del luogo sarebbe stato oltremodo complicato. Per facilitare il riscontro con

l'edizione di riferimento ci siamo permessi noi di scandire in paragrafi i capitoli più lunghi. L'operazione non si può definire ortodossa, ma le esigenze di praticità erano in questo caso prevalenti. Certo sarebbe stato molto più semplice se avessimo potuto indicare sempre la pagina dell'edizione, ma la cosa avrebbe finito per interessare il copyright. Lo abbiamo fatto per lo *Zibaldone* di Leopardi, ma solo perché le pagine citate non sono quelle dell'edizione critica ma del manoscritto originale. L'impossibilità per ragioni legali di rimandare nei contesti alla pagina del libro da cui il testo era assunto o ad altri elementi dell'edizione costituisce sicuramente un impaccio nell'eventuale riscontro del dato elettronico sull'edizione cartacea. Soprattutto per quei testi in prosa che nelle edizioni di riferimento hanno una paragrafatura ormai unanimemente accettata negli studi. Nella LIZ infatti, come si è già detto, i paragrafi vengono computati automaticamente a ogni capoverso; nelle edizioni cartacee i paragrafi si conteggiano in genere per unità sintattiche o di contenuto, indipendentemente dal capoverso. Qualche studioso ha giustamente lamentato, soprattutto in relazione al *Decameron*, la mancata rispondenza del numero dei paragrafi della LIZ a quello dell'edizione originaria. Una paragrafatura manuale che ripetesse la numerazione delle edizioni sarebbe stata impossibile da realizzare, ma per un testo dell'importanza del *Decameron* un'eccezione si sarebbe potuta fare se questo non avesse comportato un problema di diritti.

Di fatto la costituzione delle biblioteche digitali è costretta a fare i conti pesantemente con la legislazione sul diritto d'autore, non solo nel senso ovvio dell'impossibilità di riprodurre i testi di autori non di pubblico dominio, ma anche per quel che riguarda l'assetto della pagina di stampa, anch'essa tutelata, e oggi in maniera più chiara che nel passato pure per il testo dell'edizione critica. Negli anni Novanta la tutela del testo critico viveva in un limbo d'incertezza. In ogni caso, a evitare possibili contestazioni, abbiamo quasi sempre modificato gli accidentali del testo, così da creare la maggiore omogeneità possibile tra tutti i testi che confluivano nell'archivio. Di fatto il corpus della LIZ è esso stesso un megatesto nato dall'unione di tanti testi, la cui omogeneità è funzionale al reperimento dell'informazione indipendentemente dalla varietà delle forme materiali in cui elementi identici fossero in origine rappresentati. Abbiamo perciò cercato, fin dove è stato umanamente possibile, di rendere in maniera unitaria parole che erano scritte in testi affini e talora anche nello stesso testo in forma diversa, per esempio ora con accento grave, ora con acuto o circonflesso; abbiamo eliminato la dieresi nei versi, che costituiva un'informazione metrica ma che era di forte disturbo nella

generazione delle concordanze, e altre cose di questo genere. In questa maniera i testi della LIZ sono diventati riconoscibili oltre che, ovviamente, per i pochi refusi residui (comunque non più numerosi di quelli di qualsiasi pubblicazione cartacea), per questa sorta di riverniciatura degli accidentali. Proprio per aver modificato taluni aspetti accidentali dei testi rispetto alle edizioni di riferimento, quattro giovani leoni dell'allora nascente informatica umanistica mi accusarono sul web, in un comunicato a quattro mani che aveva il tenore di una *fatwa*, di esser venuto meno all'etica filologica. Non avevano alcuna idea dei problemi che comportava, proprio sul piano filologico, l'allestimento di un corpus testuale digitale omogeneo di enormi dimensioni.

Nella primavera del 1993, come già detto, finalmente la LIZ usciva. La prima reazione sia degli addetti ai lavori sia dei patiti dell'informatica fu molto favorevole. Poter disporre di un corpus di 362 opere di 109 autori in un unico dischetto era allora un fatto straordinario. I mezzi di comunicazione furono attenti a registrare la novità. La presentazione ufficiale dell'opera avvenne presso l'Accademia dei Lincei. Quell'evento fu preceduto da un episodio che vale la pena di raccontare. La bravissima responsabile dell'ufficio stampa della Zanichelli aveva diffuso, come avviene di solito in questi casi, un comunicato stampa che conteneva una descrizione dell'opera e alcune pagine che ne illustravano con esempi le principali funzioni di ricerca. Tra le stampate dimostrative era una concordanza della sequenza *pargoletta mano*, sintagma che risultava essere stato usato da Tasso prima che nella celeberrima *Pianto antico* di Carducci. In un lancio d'agenzia il dato fu stravolto nel senso che Carducci avrebbe rubato *Pianto antico* al Tasso e fu attribuita a me curatore della LIZ la sensazionale scoperta. I tre principali quotidiani italiani ("Corriere della Sera", "la Repubblica" e "La Stampa") abboccarono e il giorno successivo diedero con rilievo la notizia nella pagina della cultura. Il "Corriere" l'accompagnò addirittura con un'intervista a Cesare Segre, il quale naturalmente ricondusse la cosa a quello che effettivamente era: un normale rapporto intertestuale, peraltro segnalato finanche nelle antologie per la scuola media. Fui costretto a smentire per non mettere il mio nome su una tale sciocchezza.

L'esistenza di 362 testi letterari in formato digitale elettrizzava in ogni modo quanti di formazione umanistica provavano una forte attrazione per le nuove tecnologie e soprattutto per il mondo di Internet, che allora si avviava a esplodere. Dalle liste di discussione, che cominciavano allora a essere frequentate, veniva insistente la richiesta alla Zanichelli di

mettere liberamente in rete i file dei testi, che erano comunque fuori copyright, e sui quali perciò a opinione dei primi naviganti la casa editrice non aveva il diritto di rivendicare alcun privilegio. La Zanichelli e noi curatori eravamo contrari non perché quelle opere non fossero di pubblico dominio, ma solo per proteggere il nostro lavoro dalla concorrenza sleale. Questo diniego diede origine a una forte corrente di antipatia nel web nei confronti della LIZ, che avrebbe avuto il suo culmine in un episodio non si sa se più comico o grottesco, che più avanti racconterò. I detrattori più attivi erano un gruppo di docenti di lingua e letteratura italiana disseminati in università europee e nordamericane, spalleggiati in Italia più o meno copertamente da altri studiosi che, abbacinati dalle nuove tecnologie, erano persuasi che per effetto dell'informatica presto sarebbe tutto cambiato negli studi letterari e forse non erano entusiasti che una novità che comunque riguardava l'informatica non uscisse dai loro circoli. In un primo momento le critiche riguardarono il prezzo di vendita del disco giudicato troppo alto, poi si concentrarono sulla messa in evidenza di qualche bug immancabilmente presente nel programma, quindi sul fatto che fosse un sistema chiuso, cioè non incrementabile dall'utente. In realtà costoro non capivano o fingevano di non capire che la LIZ era solo uno strumento per generare concordanze, non una biblioteca digitale né un giocattolo elettronico, dunque da questo punto di vista uno strumento che, malgrado la novità tecnologica, era assolutamente tradizionale, e dunque di scarsa utilità per chi non aveva alcun interesse agli studi linguistici e filologici, ed era invece attratto irresistibilmente dalla novità allora ritenuta rivoluzionaria degli ipertesti, o discettava sulle specificità semiologiche del testo digitale, oppure faceva previsioni sulle sorti magnifiche e radicalmente innovative della filologia informatica.

La pretesa di disporre liberamente dei testi elettronici si concretizzò comunque in un atto di forza. In un sito non registrato della Università dello Utah di Salt Lake City fu messo in rete senza alcuna autorizzazione l'intero corpus testuale della LIZ, estratto violando le protezioni del disco. L'autore dell'impresa chiese a una sua amica in Italia di divulgare la notizia nelle liste di discussione. Costei, con fare da Biancaneve, scrisse a tutti che navigando in Internet si era imbattuta per caso in un sito interessantissimo, di cui dava l'indirizzo, che conteneva i testi di tutti i maggiori classici italiani. Il corpus era perfettamente riconoscibile e bastò una e-mail di protesta dell'amministratore delegato della Zanichelli al webmaster dell'università americana per ottenerne la chiusura. Ma intanto i buoi erano usciti dalla stalla. Quei testi sarebbero rispuntati qua

e là nel tempo in innumerevoli altri siti, anche se non nella forma ingenua con cui era stato fatto la prima volta.

Ma l'episodio più singolare, come ho anticipato tra il comico e il grottesco, di quest'assalto alla diligenza si sarebbe verificato qualche tempo più tardi. Avrebbe avuto come protagonisti un docente di un'università austriaca, che chiameremo A, e un suo collega di un'università statunitense, che chiameremo B. Inutile dirlo: tutt'e due italiani. A e B avrebbero partecipato entrambi di lì a poco allo stesso convegno. A avrebbe presentato una relazione con dei dati statistici di provenienza LIZ; B glieli avrebbe contestati, mettendone in evidenza l'inesattezza; A, sconfessato pubblicamente, si riprometteva di citare in giudizio la Zanichelli chiedendo di essere risarcito per il danno subito dalla sua immagine di studioso. C'era però un particolare non trascurabile: A era in possesso di una copia illegale della LIZ e per aver diritto al risarcimento doveva per lo meno esibire una copia legale. E poiché B possedeva invece una copia regolarmente acquistata, A chiedeva di passargliela. L'intrigo era più adatto a una Spectre pasticciona che non a quella che una volta si sarebbe detta un po' pomposamente la repubblica delle lettere. Ma come si scoprì? La mail con i dettagli del piano spedita da A a B raggiunse per un'operazione maldestra tutti gli iscritti di una lista di discussione, con effetto immaginabile. Mai come in questo caso il diavolo aveva insegnato a fare le pentole ma non i coperchi. A e B da allora in poi non si sarebbero più occupati della LIZ, e forse per loro non è stato un male.

Quando nella primavera del 1995 fu pubblicata la seconda release, i testi da 362 salirono a 500. Tranne qualche aggiunta di scarso rilievo, la novità sostanziale era data dalla presenza della quasi totalità delle opere di Pirandello e D'Annunzio. Pirandello era morto nel 1936, D'Annunzio nel 1938. La legge sul diritto d'autore allora vigente rendeva di pubblico dominio le opere degli autori a 50 anni della loro morte. Nel caso italiano agli anni canonici erano da aggiungerne sei per la seconda guerra mondiale. Il copyright sulle opere di Pirandello era già scaduto due anni prima. Il 31 dicembre 1994 sarebbe scaduto quello di D'Annunzio. Ma il primo luglio del 1995 sarebbe andata in vigore la legislazione europea sul diritto d'autore che portava a 70 anni dalla morte degli autori la tutela delle loro opere. Questo significava che a quella data Pirandello e D'Annunzio sarebbero entrati nuovamente in copyright. Esisteva insomma un corridoio di soli sei mesi per pubblicare le opere di D'Annunzio, fino ad allora di proprietà della sola Mondadori. In questo corridoio si sarebbero infilate molte case editrici italiane e con esse la LIZ, che aggiun-

geva a D'Annunzio Pirandello. Le opere pubblicate quando l'autore era di pubblico dominio potevano essere continuate a pubblicare quando lo stesso fosse ritornato sotto tutela. Con l'aggiunta di Pirandello e D'Annunzio, il corpus testuale della LIZ acquisiva due autori fondamentali per la prosa e la poesia novecentesche, ma l'estensione a 70 anni del copyright avrebbe pregiudicato pesantemente la possibilità di allargare in seguito ad altri autori novecenteschi la banca dati testuale nelle uscite successive. Il DBT, per quanto perfezionato e arricchito di altre funzioni, funzionava ancora in DOS, anche se intanto andava sempre più affermandosi Windows come sistema operativo dei personal computer.

In LIZ3, pubblicata sul finire del 1997, i testi passarono da 500 a 770. Gli incrementi riguardarono tutti i secoli. La preparazione dei testi e l'allestimento del disco erano stati curati dalla Lexis, dunque con standard di lavorazione meno fai-da-te di quelli precedenti. A questo proposito devo ricordare l'apporto di alcuni giovani collaboratori senza i quali sarebbe stato difficile ottenere risultati importanti in così poco tempo: Francesca Ferrario e Angelo Pagliardini per la parte filologica, Florestano Pastore per quella informatica. Picchi aveva intanto lavorato per trasferire in Windows il DBT. Nuove modalità di codifica consentivano di curare più di quanto non fosse avvenuto nelle prime due edizioni l'aspetto del testo, dunque la funzione biblioteca: conservazione di elementi paratestuali, maggiore eleganza di presentazione, possibilità di trasferimento automatico dei materiali della ricerca in WinWord. La nuova LIZ assumeva insomma un volto nuovo, anche se continuava a conservare l'impostazione di fondo delle precedenti, derogando proprio in ragione delle sue origini in DOS a taluni standard che intanto con l'uso di Windows andavano affermandosi. Quanto ai testi era praticamente impossibile rimettere mano, per ragioni di costo e di tempo, ai file più vecchi, in modo da renderli omogenei alla modalità dei nuovi. Tutt'al più si potevano correggere i refusi via via individuati. Per questo aspetto il caso più rappresentativo è quello di Goldoni, di cui la prima LIZ registrava solo 14 commedie. Il testo era stato tratto dai volumi degli "Struzzi" Einaudi, curati da Marzia Pieri, le cui pagine erano risultate allora le uniche leggibili allo scanner. Quando nella LIZ3 decidemmo di inserire l'intero corpus delle commedie e delle tragicommedie dell'autore veneziano (ben 133 testi) fu necessario far riferimento all'edizione mondadoriana di *Tutte le opere di Goldoni* a cura di Giuseppe Ortolani, che per il miglioramento della tecnologia del software di riconoscimento dei caratteri era divenuta intanto leggibile allo scanner. Ci si pose il problema se acquisire nuovamente le 14 commedie già inserite dalla fonte prece-

dente o mescolare le due fonti. In considerazione dell'onere economico decidemmo di conservare comunque i vecchi file, costruendo così a Goldoni un vestito a due colori. Si verificavano dunque condizioni analoghe a quelle che nelle tipografie rinascimentali impedivano, dato il costo della carta, di distruggere i fogli già tirati, malgrado la presenza in essi di errori poi corretti. Qualcosa di simile avveniva per i nostri file. È vero che in pochissimi casi avremmo sostituito il testo già inserito con un altro divenuto in seguito di riferimento negli studi, ma questa è stata l'eccezione non la regola.

Quando nel 1998 il settimanale "L'Espresso" avrebbe ottenuto dalla Zanichelli la licenza di pubblicare a puntate LIZ3 insieme alla rivista, smembrandola in sei cd che ne distribuivano per secoli i materiali testuali, raggiungendo una media di ottantamila copie vendute a disco, quello sarebbe stato il primo grande successo in Italia di distribuzione di prodotti informatici in accoppiata a un periodico. La LIZ fu portata a conoscenza del grande pubblico, ma fu fruita soprattutto come biblioteca digitale, quindi con uno stravolgimento parziale delle sue funzioni originarie che erano quelle di uno strumento di studio e di ricerca. Cambiato l'editore, nei dischi de "L'Espresso" non poterono più essere inclusi Pirandello e D'Annunzio. Alcuni acquirenti, considerato l'orientamento del settimanale, attribuirono l'assenza di D'Annunzio a censura ideologica.

La quarta release della LIZ, venuta alla luce nel 2001, è di fatto una prosecuzione di LIZ3, con l'aggiunta di 330 testi che ne portano al numero canonico di mille l'ammontare complessivo. La novità funzionale più rilevante è la lemmatizzazione del corpus. Con questo diventò possibile interrogare i testi per lemmi oltre che per forme. Ovviamente erano state lemmatizzate le forme, non i singoli contesti, operazione che sarebbe stata impossibile. Un lemmatizzatore dell'italiano contemporaneo sviluppato dallo stesso Picchi per l'Istituto di Linguistica computazionale aveva attribuito a ogni forma del corpus uno o più lemmi (in caso di omografia). Per le forme antiche o particolari i cui lemmi non erano riconosciuti in automatico, l'assegnazione fu fatta manualmente. I risultati della ricerca per lemma contengono necessariamente del rumore e come in ogni operazione che comporta un intervento manuale possono essere stati commessi degli errori, ma l'opportunità offerta da questa funzione su un corpus di circa 300 milioni di caratteri è davvero straordinaria. E devo rammaricarmi che non tutti gli utenti della LIZ, un po' per la pigrizia di esplorare modalità nuove non presenti nelle preceden-

ti release, un po' per l'insufficiente amichevolezza del sistema e la non sempre sua trasparenza terminologica, si siano accorti di questa potenzialità.

Mi avvio alla conclusione, ma non prima di qualche osservazione aggiuntiva su due aspetti non secondari finora soltanto sfiorati: la costituzione del corpus e la presenza di refusi. Comincio dal primo, ricordando che negli studi di teoria della letteratura la questione del canone è stata oggetto in questi ultimi anni di un interesse precipuo. Nel nostro caso non vale tuttavia la pena di scomodare la bibliografia sull'argomento. Il criterio con cui autori e opere sono entrati nella LIZ ha risposto fondamentalmente a una logica empirica: starei per dire di buon senso se questa espressione non avesse appunto poco senso. Anzitutto gli autori maggiori (Dante, Petrarca, Boccaccio ecc.) con il corpus il più possibile largo delle loro opere; quindi gli autori minori con le loro opere più significative; infine i minimi con l'opera più importante. Comunque, più facile a dire che a fare. Questo criterio ha dovuto infatti confrontarsi con le situazioni più diverse: come già anticipato, l'esistenza o meno di edizioni moderne affidabili, la loro disponibilità pratica, la loro leggibilità allo scanner ecc. Per non dire che di testi letterariamente e storicamente importanti mancavano addirittura edizioni moderne. Così, per esempio, per il *Cannocchiale aristotelico* di Emanuele Tesauro, o per i sonetti del Burchiello o i *Canti di Ossian* di Melchiorre Cesarotti. Di questi due ultimi testi si è allestita un'edizione provvisoria *ad hoc* sulla base delle edizioni sette e ottocentesche. La mole del *Cannocchiale* non ha invece reso possibile qualcosa di analogo. Ritornando ai criteri di costituzione del corpus, fin dalla prima LIZ ci si era preoccupati di non restare ancorati a un concetto rigido di letteratura. Costruendo una banca dati testuale dell'italiano dal Duecento ai primi decenni del Novecento, era necessario aprire, seppure limitatamente, anche a scritture non catalogabili come strettamente letterarie (la poesia, la novella, il romanzo ecc.). Testi religiosi, storiografici, pratici, trattatistica, relazioni di viaggio ecc. hanno trovato tutti rappresentanza; per non dire della serie completa delle riviste "Il Caffè" e "Il Conciliatore". È evidente che quanto più ci si allontanava dal cuore di quelle due-trecento opere la cui significatività nel quadro della nostra storiografia letteraria è da tutti riconosciuta, si navigava sempre più in acque incerte, nelle quali ogni scelta diventava in qualche modo arbitraria, e di conseguenza la presenza di un testo invece di un altro poteva dipendere da ragioni esterne, qualche volta persino dalla disponibilità o meno del volume nella propria biblioteca.

Chiunque potrebbe lamentare delle assenze o ritenere al contrario superflue delle presenze: oggi che mi avrebbe fatto comodo disporne, mi rammarico io stesso di non aver inserito nella LIZ le *Osservazioni sulla morale cattolica* del Manzoni, che avrebbero forse meritato più di altri testi di essere registrate.

Infine i refusi. L'errore di stampa è stato l'incubo di tutte le fasi di lavorazione, la nostra vera e propria ossessione quotidiana. Chi ha seguito fin qui questa nota spero si sia reso conto della difficoltà di costruire un sistema di queste dimensioni garantendo un assetto filologicamente accettabile. Esercitare una vigilanza sulla correttezza del testo di un corpus così ampio, quando bastava sfiorare un tasto sbagliato per vanificare questo assunto, ha comportato una fatica enorme. Ovvio che la perfezione non sarebbe stata in alcun modo raggiungibile. Tutt'al più si poteva aspirare a restare entro i confini di una sufficiente correttezza filologica. In questo, dalla prima all'ultima release, è stata per noi di grande aiuto la collaborazione di quegli utenti, peraltro abbastanza numerosi, che annotavano nel corso del loro lavoro gli errori testuali incontrati e li comunicavano via via a noi privatamente. Così, nel passaggio da un'uscita all'altra, abbiamo avuto la possibilità di eliminare molti errori fastidiosi. Ricordo ancora *il fiero astigiano* (l'Alfieri) vissuto per un certo tempo nella LIZ come *fiero artigiano*. Su un *Melisenda* divenuto *Melisanda* in una poesia di Carducci viene invece scrivendo da un po' di tempo in qua con invidiabile tenacia Tito Orlandi[1], coptologo convertito all'informatica, il quale però come tutti i convertiti professa ahimè con estremismo la nuova religione. Ma un buon contributo nella stessa direzione lo aveva già dato un giovane linguista[2], pubblicando un elenco di errori riscontrati nei testi della LIZ. Una sua riflessione conclusiva esprime con evidenza icastica il presupposto teorico a fondamento del suo studio: «L'infallibilità della LIZ [...] è tale nell'individuazione delle forme a condizione che sia corretto il testo immesso». Mi rammarico soltanto che

[1] Tito Orlandi, «Jaufré Rudel, ovvero Le disgrazie di un navigatore», *La Cultura*, XLII (2004), pp. 495-505. Dello stesso autore e con argomenti ripetuti, «I testi della letteratura italiana e la loro digitalizzazione: un problema aperto», in corso di stampa in *La cultura italiana. Ricerca, Didattica, Comunicazione. Percorsi formativi per l'insegnamento dell'italiano* (Atti del Convegno Paris, 16-18 ottobre 2003), a cura di L. Begioni, C. Cazalé Bérard, G. Gerlini, CIRRMI - Université la Sorbonne Nouvelle - Paris 3, 2004; scritto quest'ultimo a cui l'autore ha dato ampia diffusione già in *preprint*. Ma non posso garantire la completezza dell'informazione bibliografica: altri articoli di Orlandi su *Melisanda* potrebbero essere in corso di stampa in altre sedi.

[2] Stefano Telve, «Alcune correzioni alla LIZ4», *Studi linguistici italiani*, XXVIII (2002), pp. 97-110.

agli errori da lui segnalati non abbia avuto la possibilità di aggiungere un mio elenco, e così rendere un servizio più completo a quegli studiosi incapaci di distinguere tra un refuso e una lezione corretta.

Al tirare delle somme credo di poter dire che quella che appariva un'impresa temeraria ha dato luogo a uno strumento che, malgrado tutto e al pari di dizionari storici, grammatiche storiche, repertori bibliografici ecc., è entrato ormai stabilmente nell'attrezzatura d'obbligo degli studi di lingua e letteratura italiana. La LIZ, come risulta dalla bibliografia degli studi, è oggi usata proficuamente in tutte le università del mondo in cui si studia l'italiano. Si può dire che non ci sia edizione o commento di testo, saggio filologico che non ne dichiari o comunque dimostri l'impiego. Ho scritto in altra occasione che per ottenere risultati significativi l'interrogazione deve muovere da una precisa strategia di ricerca, prendere l'avvio da ipotesi già presenti nella mente di chi interroga. Il mezzo elettronico non servirà soltanto a confermarle o a smentirle, potrà indurre nel corso del procedimento interattivo a esplorare altre vie, a mettersi su percorsi ai quali non si era pensato. Questo il valore euristico del mezzo, che per dare il meglio di sé richiede un interlocutore che unisca sapere e creatività a capacità di formalizzazione. Con qualche avvertenza soprattutto per i meno esperti, che potrebbero essere propensi ad attribuire allo strumento tecnologico più valore di quanto in effetti non ne abbia. I dati reperiti infatti non sono i risultati della ricerca: per diventare significativi essi devono essere inseriti in un quadro argomentativo che tenga conto di ragioni sempre complesse di ordine storico, linguistico, filologico. Inoltre la possibilità di accedere in pochi istanti a un'informazione altrimenti difficilmente attingibile può far nascere la sensazione del dominio del testo. Ma il testo si domina solo attraverso la lettura, non ci sono scorciatoie possibili. Il testo frammentato, parcellizato, così come appare nelle concordanze, può in certi casi dare addirittura un'impressione fuorviante. I risultati della ricerca vanno dunque sempre riportati al testo nella sua integrità. Si potrebbe ancora fare un invito alla prudenza nello stabilire relazioni intertestuali o ricordare che i risultati delle ricerche che riguardano la poesia hanno maggiore significatività e rappresentatività di quelli della prosa. Ma mi fermo qui.

Chiudo con una domanda: potrà in futuro la LIZ crescere nel numero dei testi e migliorare la sua amichevolezza d'uso? Ci sarà insomma una LIZ5? Dipende da vari fattori, tra i quali la ripresa del settore produttivo dell'editoria elettronica dopo il crollo seguito all'ubriacatura iniziale. Il matrimonio tra filologia e informatica si consuma fertilmente solo nel-

la costituzione di basi di dati testuali, ma i pochi investimenti pubblici in questo campo continuano ad andare in altre direzioni. Ritornando alla LIZ non posso fare altro che notare con un po' di preoccupazione che, se non si fa qualcosa, prima o poi il suo cd diventerà obsoleto e dunque non più utilizzabile. È nel trasferimento in rete l'unica garanzia di sopravvivenza. Ma chi vorrà o potrà assumersene l'onere?