La Lessicografia della Crusca in Rete*

Massimo Fanfani
(Accademia della Crusca – Università di Firenze)
mfanfani@unifi.it

Marco Biffi
(Accademia della Crusca – Università di Firenze)
biffi@crusca.fi.it

Abstract
The five editions of the *Vocabolario degli Accademici della Crusca* (Vocabulary of the Crusca Academics) are an incomparable treasure of Italian language history and for history of lexicography.
With Lessicografia della Crusca in rete (Crusca's Lexicography on the net) the Crusca Academy publishes on the internet this fundamental heritage.
The first four editions (1612, 1623, 1691, 1739-1738) have been made ready in an electronic edition, appropriately marked to highlight sub-fields of linguistic and lexicographic interest, and supplied by digital reproductions of the originals. Of the fifth edition (1863-1923) has been made a version of images (over 10.000) so as to virtually leaf through the volumes or research specific voices.
The text is in an electronic format, transcribed with philological care, marked with XML/TEI tags and can be inquired through a search engine purposely conceived for a diachronic interrogation.

1 Le «Crusche» nel buratto informatico

Il *Vocabolario della Crusca* ha costituito, fino agli inizi del XX secolo, la roccaforte della lingua letteraria di matrice toscana e una ineliminabile pietra di paragone per gli scrittori e i lessicografi italiani. E nell'Europa del Seicento e del Settecento è stato considerato – per impianto concettuale, metodo di compilazione, soluzioni tecniche innovative – un modello per i grandi vocabolari delle altre lingue nazionali. Frutto della vivace stagione di speculazioni linguistiche e filologiche che aveva caratterizzato la cultura fiorentina del tardo Rinascimento, l'opera apparve subito come un monumento alle glorie della tradizione toscana. Gli accademici della Crusca si erano infatti fondati su un accurato spoglio di diverse centinaia di scrittori, a cominciare dai più celebri: Dante, Boccaccio e Petrarca. Ma ben rappresentati erano anche gli anonimi volgarizzatori dal latino e dal francese, i cronachisti e gli storici, gli autori di prediche e di scritti religiosi, diversi letterati più o meno eterodossi, dal Sacchetti al Lasca. Non mancavano, infine, alcuni non toscani (Bembo e Ariosto) che apparivano non discordanti dalla linea complessiva del lavoro.

* Il lavoro è stato progettato congiuntamente; il paragrafo 1 è stato redatto da Massimo Fanfani, il 2 da Marco Biffi.

Tuttavia non si trattava solo di un vocabolario letterario e arcaizzante, costruito come un "tesoro" storico sui lemmi affioranti dalle opere incluse nel canone. Dentro le caselle di quel lemmario fondamentalmente trecentesco, utilizzando materiali tratti da autori moderni, esempi dell'uso parlato, modi di dire e proverbi, definizioni e commenti di dotti e illetterati, i compilatori avevano voluto riconnettere alla lingua antica quella viva, l'oralità all'uso scritto, ciò che era proprio del fiorentino a ciò che aveva circolazione comune, latinismi e tecnicismi a voci popolari, gergali o d'origine locale. Ne risultava un lessico complesso e stratificato, assai più ricco di quanto a prima vista apparisse dal lemmario, proprio perché la macrostruttura del vocabolario, e anche la disposizione della materia nei singoli lemmi, privilegiava le parole e gli esempi della tradizione antica assunta a riferimento. Tale sottile commistione di vecchio e di nuovo abilmente celata nel *Vocabolario della Crusca* non costituì certo un ostacolo alla sua fortuna e, anzi, fu una risorsa che non pochi sfruttarono per ricavare da quel gran pozzo lessicografico parole, forme, accezioni sommerse, comunque significative e riutilizzabili.

Oggi l'operazione in certo senso analoga che viene realizzata attraverso Internet risponde ad altre necessità, intendendo recuperare e rendere fruibile, e non solo per gli storici e i filologi, uno dei capisaldi della cultura linguistica italiana. Negli ultimi decenni si è infatti compreso che il *Vocabolario della Crusca* oltre a essere stato oggetto di discussioni e polemiche, ha anche avuto una grande influenza come miniera e fucina di lingua. Poterlo consultare e ristudiare in modo sistematico fin nelle sue nervature, classificare ordinatamente i suoi materiali, è quindi un'impresa utile anche per comprendere più a fondo la vicenda dell'italiano e la lingua degli scrittori.

L'idea di una lettura "elettronica", destrutturata e capillare, del *Vocabolario* maturò all'interno dell'Accademia della Crusca alla fine degli anni settanta del secolo scorso, quando, per impulso di Giovanni Nencioni, si decise di riconfrontarsi con la storia della sua tradizione lessicografica. Il progetto del "rovesciamento", la trascrizione e indicizzazione informatica della prima edizione del *Vocabolario* (1612), messo allora in cantiere da Mirella Sessa, con l'assistenza di Umberto Parrini del Centro di calcolo della Scuola Normale, dal 2001 è a disposizione degli studiosi nella pagina Internet dell'Accademia: strumento di grande versatilità, consente di leggere il vocabolario nella sua interezza, distinguendo fra i macrocontesti (lemma, definizione, esempio) e diversi più specifici microcontesti.

Dopo tale traguardo, il lavoro è stato completato approntando una banca dati che, per le cinque impressioni del *Vocabolario della Crusca* (1612, 1623, 1691, 1729-1738, 1863-1923), comprendesse la riproduzione fotografica del testo, oltre alla sua trascrizione informatica. Il progetto, denominato *La lessicografia della Crusca in rete*, che ha preso avvio nel 2001 grazie a un finanziamento della Presidenza del Consiglio dei Ministri e dal 2006 è consultabile in Internet, presenta almeno due significativi vantaggi. Il primo è la possibilità di compiere confronti e indagini parallele su più edizioni del *Vocabolario*, in modo da valutare l'evoluzione dall'una all'altra, anche nella prospettiva di una ricostruzione della diacronia linguistica e metodologica degli accademici. Il secondo è costituito dalla presenza delle immagini di tutte le circa 20.000 pagine delle cinque edizioni del *Vocabolario*, che, acquisite in formato digitale, possono esser sfogliate al computer come si farebbe avendo i volumi a portata di mano; e che, inoltre, sono richiamabili quando si cerchi un lemma o si desideri ricontrollarne la trascrizione elettronica. Occorre aggiungere tuttavia che, mentre sono consultabili per immagini

le pagine di tutte le edizioni del *Vocabolario*, riguardo alla quinta (1863-1923), di impianto diverso dalle precedenti e interrotta alla O, per il momento non si è proceduto alla digitazione informatica, che dunque è limitata alle prime quattro impressioni, le sole interrogabili a tutto campo.

Il motore di ricerca per la *Lessicografia della Crusca in rete* è stato pensato in funzione della particolare natura dei testi informatizzati. Anche i parametri di marcatura dei lemmi vogliono evidenziarne i settori di rilievo strutturale o caratteri linguistici peculiari, specie in relazione alle possibilità di confronto intertestuale fra le diverse edizioni. In sostanza sono stati scorporati i seguenti campi: lemma (distinguendo gli omografi), definizione, esempio, indicazione bibliografica con rimando alle abbreviature, nota o glossa dei compilatori negli esempi, parola straniera (latina; greca o d'altra lingua), sottolemma, sottolemma senza esempi, modo di dire, proverbio, rinvio a lemma e a sottolemma. Oltre a tali marcature volte a sezionare il testo, si son indicizzati a parte tutti gli interventi esterni, cioè le integrazioni e le correzioni dovute ai curatori. Si tratta in genere di ritocchi minimi, effettuati solo se necessari per la ricerca elettronica; ritocchi che tuttavia sono stati distinti dal testo originale, riprodotto esattamente com'è, coi suoi errori e le sue imperfezioni. In questo modo sarà sempre possibile, attraverso la forma corretta indicizzata a parte recuperare dati che altrimenti resterebbero muti all'interrogazione informatica. Ad esempio, sistematiche integrazioni accompagnano gli approssimativi rimandi bibliografici del *Vocabolario*, proprio per far sì che il programma elettronico, che non può intuire ciò che manca, ritrovi tutti gli esempi di uno stesso autore o di uno stesso testo e sappia ricollegarli fra loro.

La banca dati della *Lessicografia della Crusca in rete*, per la vastità dei materiali considerati e per la sua innovativa architettura, consentirà quindi di percorrere liberamente, nelle strutture portanti e nei rifacimenti da un'edizione all'altra, il glorioso *Vocabolario* in cui gli accademici avevano tentato di ordinare l'universo di cose pensieri e affetti racchiuso sotto la crosta delle parole.

2 La banca dati e il motore di ricerca

L'operazione di informatizzare il testo di un vocabolario già stampato o manoscritto è fondamentale per aumentarne le potenzialità di risposta come strumento lessicografico. Di fatto, un vocabolario costituito da voci che ripetono costantemente la loro struttura (lemma esponente, categoria grammaticale, definizione, esempi, ecc.) è una base di dati che, nel formato cartaceo, si presenta forzatamente indicizzata solo in base al lemma. Trasformare il vocabolario in un insieme di schede interrogabili con procedure informatiche, trasformarlo cioè in base di dati effettiva, significa restituirgli la sua vera natura, e renderlo indicizzabile anche in funzione dei diversi campi, più o meno impliciti, contenuti nella struttura delle sue voci. D'altro lato, se in formato elettronico, il vocabolario diventa anche un testo libero, da consultare come una normale banca dati testuale: è possibile cioè ricercare forme all'interno della trattazione delle voci così come in un testo letterario o in un trattato.

2.1 Preparazione della banca dati: strategie e metodo

In relazione al *Vocabolario degli Accademici della Crusca* gli aspetti evidenziati sono particolarmente importanti e consentono di attribuire allo strumento una potenzialità di ricer-

ca sorprendente. Come si sa il *Vocabolario* nasce dall'esigenza di proporre un modello linguistico di riferimento per tutta l'Italia basato sul fiorentino trecentesco. Questa impostazione ha una duplice conseguenza: 1) a lemma (vale a dire la zona indicizzata e quindi rintracciabile alfabeticamente) sono riportate le parole del fiorentino del Trecento, ma all'interno delle definizioni i compilatori impiegano anche forme della lingua contemporanea (che però ovviamente sfuggono all'indicizzazione cartacea); 2) per motivare la correttezza della parola proposta si allegano esempi antichi, e quindi si ha una scissione cronologica tra la lingua dei compilatori e quella degli esempi. È evidente che la possibilità di interrogare informaticamente il testo consente di recuperare tutto il lessico contenuto nelle definizioni e assente a lemma (restituisce cioè circolarità al dizionario); e che la possibilità di individuare come campi distinti le definizioni e gli esempi permette di individuare all'interno del corpus generale dei vocabolari i due sotto-corpora della lingua degli esempi e di quella dei compilatori. Partendo da queste due istanze generali si è pensato di arricchire le possibilità di ricerca puntando su altre zone specifiche del *Vocaboalario*: parole greche e latine, forestierismi, locuzioni, proverbi, parole dell'uso vivo, fonti. Sono stati inoltre marcati i sottolemmi e i rinvii.

Un'analisi della struttura delle voci – che, per quanto rispondenti a certi schemi generali, non erano pensate come rigide griglie razionali – ha suggerito che l'approccio migliore per restituire tutte le informazioni richieste era quello della marcatura del testo libero con tag XML/TEI (*Text Encoding Initiative*);[1] un approccio che avrebbe consentito l'impiego di una marcatura standard, potenzialmente disponibile a un dialogo con strumenti simili. Si sono quindi individuati l'insieme dei marcatori TEI per dizionari utili ai fini del lavoro (32 elementi) e la loro gerarchia, definendo l'opportuna DTD (*Document Type Definition*);[2] e si è predisposta la marcatura del testo, trascritto con cura filologica. I pochi interventi (quasi sempre introdotti, come si è detto, per normalizzazioni funzionali alla ricerca informatica) sono stati segnalati con un apposito marcatore, che in fase di interrogazione consente quindi di discriminare il testo tenendo conto o meno delle correzioni. Infine la segnalazione del confine di pagina è stata razionalizzata al massimo per rendere possibile l'aggancio dell'immagine in facsimile, così da restituire immediatamente il testo nella sua struttura grafica originaria, particolarmente importante per un dizionario.

Particolare cura è stata dedicata alle fonti citate per gli esempi, indicate con abbreviazione. Delle abbreviature esiste sempre un tavola esplicativa in ogni edizione, ma spesso vi è incongruenza fra queste e le abbreviature effettivamente usate. Le indicazioni di fonti sono state suddivise in due parti, una fissa (corrispondente all'indicazione dell'opera) e una mobile (corrispondente all'indicazione del passo): sulla base della lista delle indicazioni fisse è stata poi realizzata una tabella di conversione che consente da un lato di individuare tutti gli esempi di un dato autore, o di una data opera, a partire da una base di dati di tipo bibliografico disponibile in fase di interrogazione, e dall'altro di risalire in modo rapido e sistematico dal-

[1] Cfr. Sperberg-McQueen e Burnard 2002; cfr. anche il sito www.tei-c.org. La TEI prevede, fra l'altro, convenzioni specifiche di codifica per i dizionari, con una ricca serie di marcatori.

[2] Alla progettazione della banca dati e alla definizione della marcatura XML/TEI ha collaborato anche l'Istituto dell'Opera del Vocabolario Italiano del CNR di Firenze.

l'abbreviazione presente nel testo alla sua esplicitazione nella *Tavola delle Abbreviature* della rispettiva edizione (vedi anche il *Paragrafo 2.1*).[3]

2.2 Il motore di ricerca

Il motore di ricerca che gestisce l'immensa banca dati della *Lessicografia* è stato realizzato dal MICC (Centro per la Comunicazione e l'Integrazione dei Media, dell'Università degli Studi di Firenze) sulla base della banca dati realizzata dall'Accademia della Crusca e in collaborazione con il suo Centro Informatico. Non potendo illustrare nel complesso tutte le caratteristiche di interrogazione del *Vocabolario* in forma elettronica, ci soffermeremo su alcuni punti essenziali che ci sembrano di particolare rilevanza sul fronte dell'informatica linguistica e della lessicografia.

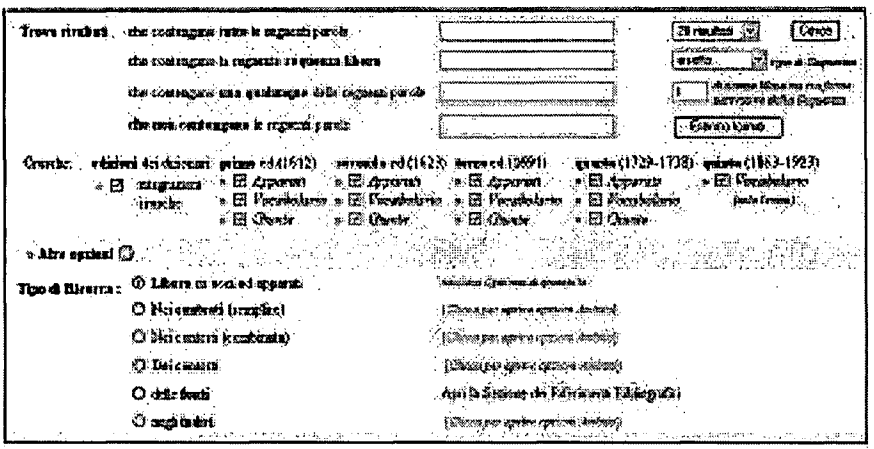

Figura 1. La pagina di "ricerca esperta"

Il motore di ricerca, che nel corso della progettazione e realizzazione è stato battezzato con il nomignolo di *Cruscle*, lavora per forme: l'operazione di lemmatizzazione di un *corpus* diacronicamente connotato come quello dei *Vocabolari*, per giunta distribuito su quattro secoli di storia della nostra lingua, sarebbe stata realizzabile solo con procedure largamente manuali, e avrebbe quindi richiesto risorse e tempi enormi. Del resto un motore per forma, sviluppato con opportune accortezze, mette in grado l'utente esperto di effettuare ricerche proficue e sistematiche anche su corpora che spaziano in ampie fasce cronologiche, come ci ha dimostrato e ci dimostra tuttora uno strumento come la *LIZ Letteratura Italiana Zanichelli*, la cui interrogazione è gestita dal DBT di Eugenio Picchi. *Cruscle* consente la ricerca con

[3] Trascrizione del testo, marcatura e tabelle sono a cura di Claudia Bichi, Silvia Dardi, Fiammetta Fiorelli, Rossella Gasparrini e Cecilia Palatresi; la trascrizione del testo greco è di Elena Bonaccini, Mariella Canzani e Giulio Niccoli.

caratteri *jolly*, e pertanto individua rapidamente anche le eventuali varianti diacroniche di lingua. Inoltre rende possibile ricercare forme indipendentemente o meno dai caratteri accentati: un'opzione, ad esempio, che consente, anche al consultatore meno avvertito sulle convenzioni grafiche dei *Vocabolari* più antichi, di individuare gli avverbi latini terminanti in *-e*, sistematicamente indicati con *-è* (*abditè*, *abiectè*, *absolutè* ecc.).

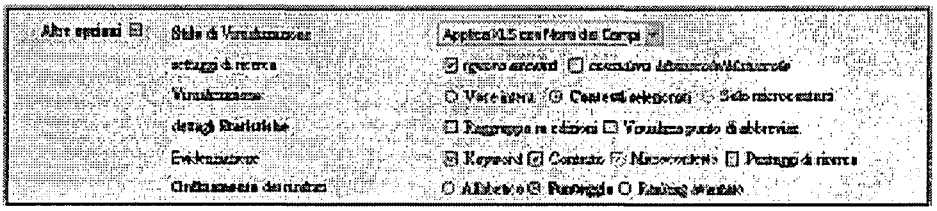

Figura 2. Le opzioni di ricerca.

Anche la *tokenizzazione* del testo – l'indicizzazione per l'individuazione delle forme – è stata gestita con accortezza diacronica in relazione alla gestione degli apostrofi e degli asterischi; e la punteggiatura, opportunamente indicizzata, può essere ricercata al pari delle forme in tutti i tipi di interrogazione, in modo da allargare il campo di indagine linguistica anche ad alcuni aspetti sintattici.

Nella banca dati della *Lessicografia* è possibile ricercare gruppi di due o più forme, precisandone la distanza in numero di parole, impiegando tutti i caratteri *jolly* e integrando eventualmente uno o più segni interpuntivi nel gruppo di oggetti ricercati (ad esempio: "fior di farina", "a modo di", ". Ma anche", ecc.).

Se quelle viste finora sono caratteristiche generali del motore, che offrono un'idea delle sue potenzialità in relazione al *Vocabolario* come banca dati testuale da interrogare *full text*, molte sono le funzionalità di ricerca legate alla sua natura di base di dati. Innanzi tutto è possibile individuare vari sotto-corpora di ricerca: in base all'edizione (o gruppi di edizioni a scelta dell'utente, che può anche decidere se limitare la sua indagine agli apparati, ai lemmari, alle *Giunte* di completamento e correzione), in base ai macrocontesti (lemma, definizione, esempi, commenti degli accademici), in base ai microcontesti (abbreviazione bibliografica, parole d'uso vivo, proverbi, locuzioni, parole latine, parole greche, parole di altre lingue straniere); o, infine, intersecando tutti i parametri tra loro per individuare specificatamente una sezione da indagare.

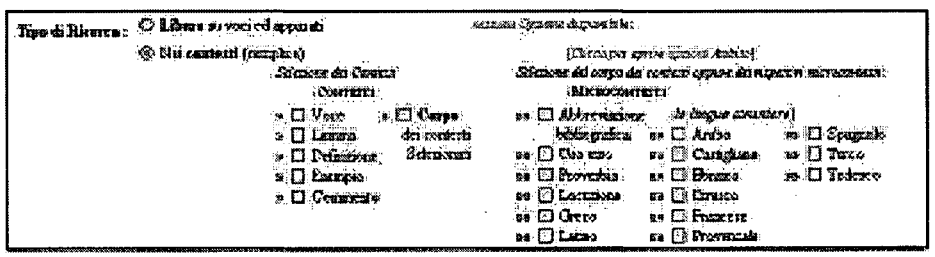

Figura 3. Le ricerche nei macrocontesti e nei microcontesti.

Una volta impostata una ricerca è sempre possibile consultare la lista delle forme corrispondenti o richiamare direttamente il contesto immediato; le liste, particolarmente utili nel caso di ricerche a risposta multipla, possono essere ordinate alfabeticamente o in funzione della frequenza (nelle modalità più avanzate di ricerca è possibile anche stabilire un punteggio di *ranking* che dia maggiore o minore peso alle forme cercate in relazione agli specifici macrocontesti: lemma, definizione, esempio, commento). Dal contesto immediato si risale alla finestra che contiene l'intera voce, in cui è evidenziata sia la forma ricercata sia il macrocontesto che la contiene; con un apposito collegamento è poi possibile visionare la riproduzione in facsimile dell'originale.

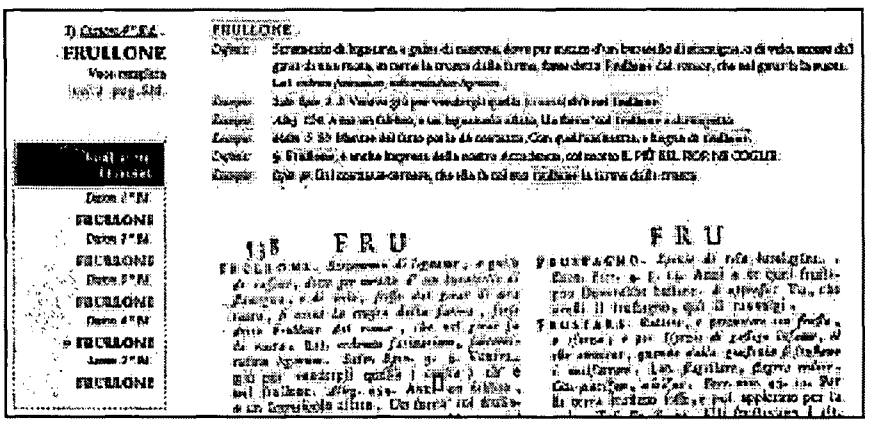

Figura 4. Esempio di "finestra voce" con apertura del facsimile corrispondente.

Nella spalla sinistra della finestra della voce completa è situato uno schema sinottico che riassume la situazione nelle altre edizioni del *Vocabolario*: se il lemma è presente in una delle altre Crusche, è possibile accedervi direttamente con un semplice clic del *mouse*. In questo modo è sempre possibile seguire la storia della trattazione di una voce in modo rapido e sistematico, e la *Lessicografia della Crusca in rete*, oltre che per la storia della lingua italiana, diviene strumento fondamentale anche per lo studio della sua lessicografia.

La navigazione ipertestuale all'interno delle voci prevede collegamenti automatici con i rinvii inseriti dagli accademici e l'aggancio delle abbreviature dei citati alla tabella delle fonti, una completa base di dati che riporta, in modo formalizzato, tutte le indicazioni della *Tavola delle abbreviature* più altre informazioni aggiunte (datazione, elenco delle abbreviature effettivamente usate ecc.). D'altra parte questa importante tabella è consultabile, in modo tale da poter risalire, dato un autore o un'opera, a tutti gli esempi relativi contenuti nelle varie edizioni del *Vocabolario* (o nelle edizioni scelte da chi consulta).

Secondo un approccio generalmente perseguito nella progettazione degli strumenti informatici dell'Accademia della Crusca, a lato alla ricerca mirata e puntuale è lasciata aperta la possibilità di consultare liberamente i materiali: non solo sfogliando i singoli volumi dei *Vocabolari*, ma, con ricerche incluse nella sezione "Ricerche guidate", scorrendo gli indici pre-

confezionati di proverbi, parole dell'uso vivo, modi di dire, sottolemmi e così via, fino a coprire l'intera gamma dei microcontesti marcati.

Bibliografia

A. Dizionari

Lessicografia della Crusca in rete: www.accademiadellacrusca.it/biblioteca_virtuale.shtml.

B. Altri testi

Alisi, T. M., Becchi, G., Becchi, N., Biffi, M., D'Amico, G., Evangelisti, A., Fanfani, M., Maraschio, N. (2006), 'Advanced search facilities for accessing Crusca Academy of Italian Language', in Cappellini V., Hemsley, J. (eds.), *Electronic Imaging & the Visual Arts EVA 2006 Florence Proceedings*, Bologna, Pitagora Editrice, pp. 164-69.

Parodi, S. (1974), *Gli atti del primo Vocabolario*, Firenze, Sansoni.

Parodi, S. (1983), *Quattro secoli di Crusca. 1583-1983*, Firenze, Accademia della Crusca.

Picchi, E., Stoppelli, P. (2001) *Letteratura Italiana Zanichelli 4.0. CD-ROM dei testi della letteratura italiana*, Bologna, Zanichelli. (LIZ).

Sessa, M. (1982), 'Saggio di "rovesciamento" del primo Vocabolario della Crusca', *Studi di lessicografia italiana*, IV, pp. 269-333.

Sessa, M. (1991), *La Crusca e le Crusche. Il Vocabolario e la lessicografia italiana del Sette-Ottocento*, Firenze, Accademia della Crusca.

Sessa, M. (1999), 'Il lessico delle commedie fiorentine nel Vocabolario degli accademici della Crusca (nelle prime tre edizioni)', *Studi di lessicografia italiana*, XVI, pp. 331-377.

Sessa, M. (2001), 'Il "rovesciamento" del primo Vocabolario della Crusca (1612)', *Crusca per voi* 22, pp. 3-18.

Sperberg-McQueen, C. M., Burnard, L. (eds) (2002), *Guidelines for Text Encoding and Interchange*. Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford.