

Stampa: Bertoncello Artigrafiche - Maggio 2006

EDITRICE ANTENORE S.r.l.

00193 ROMA - VIA VALADIER 52 - TEL. 06-3260.0370
FAX 06-3223.132 - E-MAIL ANTENORE@EDITRICEANTENORE.IT

BIBLIOTECA VENETA · 23 - 24

LESSICOGRAFIA DIALETTALE RICORDANDO PAOLO ZOLLI

*Atti del Convegno di Studi
Venezia, 9-11 dicembre 2004*

A CURA DI
FRANCESCO BRUNI E CARLA MARCATO



EDITRICE ANTENORE
ROMA-PADOVA · MMVI

CARLA MARCATO, <i>Presentazione</i>	VII
TOMO I	
STEFANO PATRON, <i>Paolo Zolli bibliofilo nel ricordo di un bibliotecario</i>	1
MANLIO CORTELAZZO, <i>L'avventura lessicografica con Paolo Zolli</i>	5
FABIO MARRI, <i>Paolo Zolli italianista "revisionista"</i>	9
TULLIO TELMON, <i>La recente lessicografia amatoriale in Piemonte</i>	25
REMO BRACCHI, <i>Nomi della paura nelle valli dell'Adda e della Mera</i>	45
MARIO PIOTTI, <i>Il primo vocabolario del dialetto bresciano (1759)</i>	71
CORRADO GRASSI, <i>Implicazioni teoriche e di metodo di un rapporto simbiotico tra museo etnografico e lessicografia dialettale: l'esempio trentino</i>	83
GIOVANNI KEZICH-ANTONELLA MOTTI, <i>Il Trentino dei contadini. Piccolo Atlante sonoro della cultura materiale. Note di Presentazione</i>	95
PATRIZIA CORDIN-TIZIANA GATTI, <i>Dai dizionari dialettali su carta ai dizionari in rete. Aspetti metodologici e questioni aperte</i>	109
CHIARA SCHIAVON, <i>Dal pavano nei vocabolari al vocabolario del pavano</i>	135
FRANCO CREVATIN, <i>Caratteri generali della 'Raccolta' di F.Z. Muzazzo in dialetto veneziano</i>	151
ANGELA CARACCILO ARICO', <i>Per la storia dell'edizione del 'Dizionario del dialetto veneziano' di Giuseppe Boerio</i>	167
GIANNA MARCATO, <i>Le locuzioni in G. Boerio: veneziano e italiano a confronto</i>	173
FEDERICO VICARIO, <i>Fonti documentarie tardomedievali e studi lessicografici sul friulano</i>	189
FLAVIA URSINI, <i>Un dialetto al tramonto e la sua rappresentazione lessicografica: il Vocabolario del dialetto di Rovigno d'Istria'</i>	201
SIMONETTA MONTEMAGNI-MATILDE PAOLI-EUGENIO PICCHI, <i>ALT-Web: l'Atlante Lessicale Toscano' in rete</i>	209
NERI BINAZZI, <i>Per una lessicografia dalla parte del parlante: il Vocabolario del fiorentino contemporaneo'</i>	243
FABRIZIO FRANCESCHINI, <i>"Parole d'Acciaio": neologismi, forestierismi e riflessi dialettali nel lessico delle acciaierie di Piombino (LUSID)</i>	265
ANTONIO BATINTI-FERDINANDO GRANDE-GIOVANNA SAMBUCINI, <i>Il lessico nella produzione poetica (1980-2002) in dialetto perugino di C. Spinelli a confronto con i vocabolari dialettali di area</i>	285
ENZO MATTESINI, <i>Forestierismi nei dialetti dell'Umbria: i francesismi</i>	297
NICOLA DI NINO, <i>Uno sguardo alla lessicografia romanesca</i>	319
FRANCESCO AVOLIO, <i>Comuna Finamore e la lessicografia dialettale abruzzese</i>	

SIMONETTA MONTEMAGNI-MATILDE PAOLI-EUGENIO PICCHI
 ALT-WEB: L'ATLANTE LESSICALE TOSCANO IN RETE

1. INTRODUZIONE

Scopo di questo articolo è la presentazione di *ALT-Web*, ovvero l'*Atlante Lessicale Toscano* in rete.¹ In una società a tecnologia avanzata, la produzione di cultura passa inevitabilmente attraverso una maggiore e più capillare diffusione grazie alle tecnologie informatiche che mettono a disposizione di un sempre più vasto pubblico le potenzialità di risorse formative di ogni disciplina. *ALT-Web* è stato ideato per rendere il patrimonio linguistico-culturale testimoniato dall'*Atlante Lessicale Toscano* una risorsa educativa realmente disponibile in modo che possa fornire un contributo alla conservazione della memoria dell'identità culturale toscana e al contempo costituisca un prezioso punto di riferimento per lo studio di dinamiche linguistiche sia a livello areale sia a livello socio-culturale. La sua collocazione in rete porta inevitabilmente *ALT-Web* a rivolgersi a una vasta gamma di utenti non più circoscritta agli addetti ai lavori (ovvero dialettologi, linguisti, etno-linguisti), ma che include anche insegnanti, operatori culturali (ad esempio, personale di musei e di istituzioni culturali pubbliche e private) fino al cittadino navigatore di Internet che voglia capire di più della propria identità linguistica e culturale. Il vasto e variegato bacino di utenza a cui intende rivolgersi *ALT-Web* ha portato alla trasformazione della versione informatizzata dell'*Atlante Lessicale Toscano* (conosciuta come *DBT-ALT*) in una rete ipertestuale con modalità e funzionalità di accesso differenziate in relazione alle diverse classi di utenza; a questo aspetto, è legata l'altra interpretazione dell'acronimo *ALT-Web*, ovvero quella di «*ALT* come rete».

Avendo schematicamente delineato le motivazioni sottostanti alla creazione di *ALT-Web*, in quanto segue ci concentreremo su aspetti di progettazione e realizzazione che rivestono un qualche interesse

1. La realizzazione di *ALT-Web* è stata finanziata dalla Regione Toscana (U.O.C. «Musei, Paesaggio e Attività Culturali») nell'ambito del progetto «Strumenti per l'integrazione e la valorizzazione dei sistemi museali e per la ricerca sul patrimonio culturale».

per il linguista e il dialettologo. In particolare, dopo un breve excursus che riepiloga le caratteristiche principali delle risorse di partenza (sezione 2), ci soffermeremo sulla progettazione e sulla realizzazione di *ALT-Web*, partendo dall'analisi dei requisiti (sezione 3) e la definizione delle caratteristiche generali (sezione 4) per arrivare ad aspetti più specifici che riguardano le modalità di accesso ai materiali (sezione 5) e la normalizzazione dei materiali dialettali in trascrizione fonetica (sezione 6).

2. LE RISORSE DI PARTENZA

2.1. *L'Atlante Lessicale Toscano*

*L'Atlante Lessicale Toscano (ALT)*² è stato progettato come un atlante linguistico, specificamente lessicale, di ambito regionale, con scopo di rilevare e definire le condizioni di variabilità diatopica e diastratica che, relativamente al lessico, sussistono all'interno del repertorio dei parlanti della regione, tenendo anche conto del rapporto del tutto peculiare esistente tra le diverse parlate locali e la lingua nazionale. Il suo impianto, giunto a una forma definita nel questionario del 1973 (cui si aggiunge l'ulteriore stesura del 1975 che amplia la griglia ma non ne modifica la sostanza), trova le sue basi in ricerche di seminario e studi precedenti condotti e coordinati da Gabriella Giacomelli. Le oltre settecento domande, differenziate per ambiti semantici ma anche per obiettivi in relazione alle diverse istanze della ricerca, sono state sottoposte nei 224 centri indagati in Toscana, in un arco di tempo che va dal 1974 al 1986 (con la maggior concentrazione delle inchieste però negli anni 80 e con la ripetizione delle prime più distanti nel tempo), a ben 2193 parlanti locali.³

2. *Atlante Lessicale Toscano*, opera svolta con il sostegno della Regione Toscana, in collaborazione con l'Accademia Toscana di Scienze e Lettere «La Colombaria». Direzione dell'impresa: G. GIACOMELLI. Redazione: L. AGOSTINIANI, P. BELLUCCI, G. GIACOMELLI, L. GIANNELLI, S. MONTEMAGNI, A. NESI, M. PAOLI, T. POGGI SALANI.

3. Per maggiori notizie sull'*ALT* e la sua storia cfr. G. GIACOMELLI, *L'Atlante Lessicale Toscano. Presentazione*, in «Quaderni dell'Atlante Lessicale Toscano», n. 0 1982, p. 275; ID., *Storia, criteri, metodi, prospettive dell'Atlante Lessicale Toscano*, ivi, n. 5/6 1987/1988, pp. 7-25.

2.2. *L'ALT* in versione elettronica: *DBT-ALT*

Così come le inchieste di seminario e gli studi sulle varietà locali produssero il progetto dell'Atlante, la conclusione del lavoro di ricerca sul campo, i cui risultati erano stati discussi e vagliati durante la progressione dell'opera e di conseguenza avevano già mostrato il loro potenziale di informazione, ha portato come conseguenza l'evoluzione del progetto originario verso la creazione di una banca dati che contenesse l'intero *corpus* dei materiali dialettali raccolti. Nel 1985 inizia così il rapporto di collaborazione con l'Istituto di Linguistica Computazionale (*ILC*) del CNR di Pisa che ha portato alla creazione dei programmi di archiviazione e del sistema di interrogazione attraverso cui i dati sono fruibili dal 2000, anno della pubblicazione dell'*Atlante Lessicale Toscano* su CD-rom.⁴

La Banca Dati dell'*ALT* contiene l'intero *corpus* dei materiali lessicali reperiti con le inchieste sul campo, codificati in più di 350.000 schede che oltre alle forme ottenute in risposta alle domande del questionario includono materiali accessori che vanno da contesti linguistici tipici e annotazioni di varia natura, dalla rilevazione di differenziazioni di tipo semantico alla messa a fuoco di variazioni di registro, stile, e così via. A questi materiali «canonici» si affiancano materiali lessicali integrativi emersi in associazione alle risposte quantificabili in circa 30.000 schede: in tutto, la banca dati dell'*ALT* contiene dunque circa 380.000 schede.

Il programma di gestione ed interrogazione della Banca Dati, *DBT-ALT*, rappresenta una specializzazione del software di interrogazione *DBT*.⁵ Il nucleo di partenza, costituito dalle funzionalità di

4. G. GIACOMELLI ET ALII, *Atlante Lessicale Toscano*, Roma, Lexis Progetti Editoriali, 2000.

5. *DBT (Data Base Testuale)*. Autore: E. PICCHI. Copyright: Consiglio Nazionale delle Ricerche. Il *DBT* è un software di analisi testuale e di interrogazione «full-text»; esso costituisce ormai un punto di riferimento nel panorama letterario e linguistico italiano per le ricerche di tipo testuale. Per maggiori dettagli cfr. E. PICCHI, *Esperienze nel settore dell'Analisi di corpora testuali: software e strumenti linguistici*, in *Informatica e Scienze Umane. Mezzo Secolo di Studi e Ricerche*, a cura di M. VENEZIANI, Firenze, Leo S. Olschki Editore, 2003, pp. 129-55; E. PICCHI, *PiSystem: sistemi integrati per l'analisi testuale*, in *Computational Linguistics in Pisa-Linguistica Computazionale a Pisa. Linguistica Computazionale*, a cura di A. ZAMPOLLI, N. CALZOLARI e L. CIGNONI, Special Issue XVIII-XIX, Pisa-Roma, IEPI, 2003, to. II pp. 597-627. Per una breve descrizione del *DBT* inclusiva di demo si rinvia anche a all'indirizzo web <http://www.ilc.cnr.it/pisystem/>.

base per la gestione di una base di dati testuali, è stato specializzato per il trattamento dei materiali dell'*Atlante Lessicale Toscano*, un corpus « bilingue » che include dati sia in trascrizione fonetica sia in ortografia italiana. In *DBT-ALT*, tutte le funzioni di recupero di informazioni possono essere effettuate su materiali sia in trascrizione fonetica sia in ortografia italiana. Inoltre, i materiali possono essere recuperati sulla base di un ampio spettro di parametri che vanno dalla domanda del questionario a cui si correlano (direttamente o indirettamente), la località di inchiesta in cui sono stati raccolti alla forma registrata (in trascrizione fonetica) sul campo e alle notazioni e commenti forniti a integrazione delle risposte. L'insieme delle funzionalità di base è stato integrato con altre funzioni di ricerca più complesse che permettono la selezione dei materiali sulla base di parametri extra-linguistici: ad esempio, qualsiasi ricerca può essere associata a restrizioni sulla tipologia del punto di inchiesta e/o degli informatori. Infine, *DBT-ALT* include una funzionalità specifica basata per la gestione dei materiali di un atlante geolinguistico, ovvero la possibilità di proiettare su carta i risultati ottenuti attraverso le diverse funzioni di ricerca.

3. ALT-WEB: ANALISI DEI REQUISITI

3.1. Le caratteristiche dell'ALT

Un atlante linguistico rimane in genere circoscritto ai materiali strettamente pertinenti alle domande del questionario di raccolta. Questo non vale per l'*Atlante Lessicale Toscano*, atlante nel nome ma molto di più nei fatti, ovvero nei dati che esorbitano di molto rispetto al progetto. Lo stesso progetto originario, che già si discostava da atlanti linguistici tradizionali per il considerare in modo sistematico anche le dimensioni diacronica, diastratica e diafasica del dato, pur nella sua complessità di intenti, non è sufficiente di per sé stesso, a giustificare l'abbondanza e varietà di materiali raccolti che includono, oltre alle canoniche risposte al questionario, notazioni integrative degli informatori, fraseologia più o meno cristallizzata, testimonianze paremiologiche o di letteratura popolare, nonché brevi etnotesti.

Un esempio di questa ricchezza e varietà di informazioni all'interno dell'*ALT* è rappresentato dai materiali che a qualche titolo rivestono interesse nel campo della ricerca etnografica: l'opera è nata con

impostazione di tipo geolinguistico senza alcuna esplicita finalità in ambito demoantropologico,⁶ ma ovviamente, trattandosi di ricerca sul lessico, l'elemento di cultura materiale risulta implicito nel dato. Già a livello del questionario l'implicazione etnografica è presente, anche se in modo discontinuo: alcune domande la sottendono oggettivamente, altre, di per sé meno significative, convogliano informazioni che in aggiunta ad altre concorrono a fornire frammenti utili se non alla ricostruzione completa di una attività tradizionale (per esempio un antico processo di lavorazione) almeno a tracciarne un sommario profilo. Questa dimensione implicita nelle domande si amplifica nelle risposte degli informatori e nel corredo di informazioni accessorie che vengono fornite specie laddove quella specifica attività ha o ha avuto rilevanza particolare (per ragioni storiche, geografiche, ecc.).

La progettazione e costruzione della banca dati dell'*ALT*, avvenuta a completo reperimento dei materiali, è stata condotta nel rispetto per ogni dato attestato, anche se non immediatamente rilevante ai fini del prodotto Atlante. A fianco dei materiali canonici (ovvero le risposte al questionario), sono stati codificati e dunque resi accessibili anche tutti i dati non facilmente riproducibili su carta, quali contesti linguistici ed annotazioni relative alla referenza delle risposte al questionario; non solo, sono state pure incluse testimonianze etnografiche e materiali lessicali emersi a margine dell'inchiesta i quali, a loro volta, possono essere corredati di specificazioni contestuali. In *DBT-ALT* tutti questi elementi sono stati resi reperibili attraverso una serie di dispositivi quali parole chiave, codici, rinvii interni e altro ancora.

Nella progettazione di *ALT-Web* si è voluto continuare a rendere conto della poliedricità del dato *ALT*, in ogni sua sfaccettatura e valenza. A questo proposito, abbiamo fatto tesoro dell'esperienza di anni di consultazione di *DBT-ALT* in contesti diversi e per finalità differenziate: fra le molte strade percorribili attraverso i dati sono state individuate le « vie maestre » dell'interrogazione, sia dal punto di vista della loro produttività in relazione ai risultati ottenuti sia dal punto di vista della loro immediatezza e facilità di accesso. In con-

6. Cfr. A. NESI, *Le potenzialità d'uso dei materiali ALT sotto il profilo demo-antropologico*, in AA.VV., *Atlante Lessicale Toscano Presentazione*, Firenze, Olschki, s.d. ma 1985, par. 3.

creto, in *ALT-Web* la complessa rete di rapporti che lega tra loro i materiali *ALT* è stata resa più evidente e di conseguenza più fruibile anche da un'utenza specificamente interessata al settore.

3.2. I bisogni dell'utenza finale

L'*ALT* in versione elettronica pubblicato nel 2000 si rivolgeva a una ristretta cerchia di addetti ai lavori, primariamente linguisti e dialettologi. Tra le caratteristiche che rendono difficilmente fruibile quest'opera da parte di un'utenza non specialistica va innanzitutto annoverato il fatto che i materiali dialettali sono accessibili soltanto attraverso la trascrizione fonetica: ciò costituisce senza dubbio una difficoltà non trascurabile per i non addetti ai lavori, ma non soltanto. Infatti, il recupero di materiali dialettali in trascrizione fonetica può non essere banale anche da parte dello specialista in quanto il dettaglio di rappresentazione imposto dalla trascrizione fonetica frammenta una stessa attestazione lessicale in tante varianti che l'utente deve sapersi prefigurare in partenza per poter formulare la propria interrogazione in modo adeguato.⁷ Ad esempio, se l'utente è interessato al tipo lessicale *proda* in ambito toscano, deve predisporre diverse interrogazioni, dove ad esempio la vocale /o/ presenta diversi gradi di apertura, l'occlusiva dentale sonora può essere spirantizzata, così come /p/ in posizione iniziale se la forma è stata prodotta come parte di un contesto frasale. Nel caso di *proda* l'utente dovrebbe così prefigurarsi una serie di possibili esiti quali /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/, /prɔ̃da/. Quindi, anche per l'addetto ai lavori il dettaglio imposto dalla trascrizione fonetica può costituire una difficoltà ai fini del recupero automatico dei materiali dialettali trascritti.

Infine, non va sottovalutato il fatto che il sistema di interrogazione di una base di dati dialettali come quella dell'*ALT*, per quanto versatile e ricco, richiede un utente già familiare sia con la tipologia dei materiali dialettali di un atlante linguistico sia con software di interrogazione di banche di dati linguistici.

7. Cf. L. AGOSTINIANI-E. MARINAI-S. MONTEMAGNI-M. PAOLI, *Una procedura informatica di accesso intelligente a materiali in trascrizione fonetica: l'esperienza dell'Atlante Lessicale Toscano*, intervento tenuto al V Congresso SILFI, (Catania, 15-17 ottobre 1998), manoscritto reperibile al seguente indirizzo web <http://serverdbt/altweb/silfialt.pdf>.

La situazione tratteggiata sopra mal si concilia con l'ampio e variegato bacino di utenza prospettato per *ALT-Web*. In *ALT-Web* abbiamo cercato di ricomporre questa frattura attraverso la definizione di procedure differenziate di codifica, accesso e interrogazione dei materiali dialettali, illustrate nelle sezioni che seguono.

4. ALT-WEB: DEFINIZIONE DELLE CARATTERISTICHE GENERALI

La progettazione di *ALT-Web* è stata guidata dalla necessità di conciliare due esigenze apparentemente in contrasto:

- rendere conto della ricchezza, della complessità, e della diversa qualità e varietà delle informazioni raccolte con le inchieste dell'*Atlante Lessicale Toscano*;

- fornire un prodotto di facile ed immediata fruizione per la varia tipologia dei futuri consultatori.

A tal fine, le risorse di partenza sono state integrate con informazioni e funzionalità aggiuntive volte a potenziare e al contempo facilitare la consultazione del prodotto finale. In particolare:

- l'accesso ai materiali dell'*Atlante Lessicale Toscano* viene fornito secondo modalità differenziate intese a facilitare e a guidare la fruizione dei materiali dialettali ed etnografici in esso contenuti. In particolare, le funzioni di accesso ai materiali in rete sono state sviluppate secondo modalità diverse tra loro complementari: all'identificazione immediata degli elementi fondamentali della banca dati (es. le attestazioni dialettali raccolte in una località o in relazione ad uno specifico stimolo), è affiancata la possibilità di effettuare interrogazioni personalizzate sull'intero *corpus* dei materiali dialettali presenti nella base di dati;

- è stato predisposto un accesso ai materiali dialettali anche su base concettuale, che permette una ricerca dei dati basata sul concetto che esprimono o, nel caso di materiali integrativi, a cui si correlano;

- i materiali dialettali registrati in grafia fonetica nella risorsa originaria sono stati affiancati da diversi livelli di trascrizioni normalizzate in ortografia italiana per rendere possibile l'interrogazione e il recupero dei materiali a prescindere da dettagli della realizzazione fonetica.

In quanto segue, viene fornita una breve descrizione delle funzionalità di accesso ai materiali (sezione 5) e della normalizzazione

in ortografia italiana dei materiali registrati in grafia fonetica (sezione 6).

5. MODALITÀ DI ACCESSO AI MATERIALI *ALT*

Per offrire modalità di accesso differenziate in relazione alle diverse classi di utenza sono stati realizzati due diversi livelli di interrogazione, ovvero:

1. accesso guidato ai materiali;
2. interrogazione personalizzata della banca dati.

Questi due accessi differenziati possono essere intesi come procedure distinte ma anche come fasi successive di uno stesso processo di esplorazione dei materiali.

La modalità di accesso guidato si rivolge al pubblico meno specializzato, ma anche a coloro che intendono familiarizzare col materiale *ALT* prima di passare a ricerche più complesse o semplicemente acquisire in modo sintetico materiali rappresentativi della situazione della Toscana o di sue sub-aree; in questa modalità, al navigatore della rete dell'*ALT* vengono proposti pochi « sentieri » obbligati incentrati sulle interrogazioni canoniche di una banca dati contenente i materiali di un atlante linguistico: le attestazioni dialettali raccolte in una località e le risposte raccolte in relazione a una specifica domanda.

La modalità di accesso avanzato permette all'utente di formulare liberamente le proprie richieste creando percorsi personalizzati di ricerca da proiettare sull'intero *corpus* dei materiali *ALT*. In questa modalità di accesso viene proposta al navigatore una vasta gamma di parametri per la definizione di percorsi complessi nel *corpus* dei materiali dialettali raccolti così come nelle glosse e nei commenti a corredo delle attestazioni dialettali registrate, inclusa una batteria di filtri basati sulle caratteristiche generazionali e socio-culturali degli informatori, sull'uso e la competenza dichiarati in relazione alla voce dialettale, sulla pertinenza rispetto a una varietà e/o registro linguistico, ecc.

5.1. Accesso guidato ai materiali *ALT*

A questo livello viene offerta la possibilità di scegliere tra due chiavi primarie di accesso ai materiali dialettali dell'*ALT*:

1. per stimolo, ovvero sulla base della domanda di cui l'attestazione dialettale costituisce risposta;
2. per localizzazione areale.

La prima chiave di accesso ai materiali *ALT* è costituita dalle domande del questionario sulla base del quale è stata effettuata la rilevazione. L'utente può arrivare a rintracciare le domande di interesse secondo due modalità diverse: visionando l'intero questionario *ALT* (modalità tradizionale destinata a chi conosca già il questionario *ALT*), oppure percorrendo una gerarchia concettuale che porta all'identificazione delle domande contenute nel questionario che presentano una qualche attinenza con i propri interessi di ricerca (maggiori dettagli sulla modalità di accesso su base concettuale sono fornite nella sezione 5.3). Selezionato l'insieme di domande alla base della propria ricerca, l'utente ha la possibilità a questo punto di scegliere se visionare a) i risultati ottenuti in una località di indagine o in un'area specifica (ad esempio, una provincia), o b) la sintesi generale di tutte le risposte ottenute in relazione alle domande selezionate sull'intero territorio toscano (ciò vale solo nel caso la ricerca abbia riguardato una singola domanda). Nel caso a) il risultato è costituito dall'insieme delle schede che registrano le risposte ottenute nelle località selezionate. Nel caso b) è rappresentato dalla lista di tutte le attestazioni dialettali raccolte in relazione alla domanda selezionata, corredate di informazioni accessorie (il numero di località in cui sono state raccolte e il numero di schede che ne registrano l'attestazione), che può essere ordinata in due diverse modalità: sulla base della frequenza di occorrenza sul territorio toscano, oppure alfabeticamente. Per ciascuna voce di questa lista è possibile scegliere se proiettare la singola attestazione dialettale sulla carta della regione in modo da poter vedere l'area di diffusione di quel particolare termine, oppure se prendere visione delle schede corrispondenti. Nella consultazione delle schede, l'utente potrà decidere la modalità di visualizzazione delle attestazioni dialettali, in trascrizione fonetica, ortografica o nella forma normalizzata (cfr. sezione 6.1).

L'accesso guidato ai materiali può anche avvenire a partire dalla localizzazione geografica. Scegliendo questa opzione si avrà modo di esaminare l'insieme delle attestazioni dialettali raccolte in una data località in relazione a uno specifico insieme di domande. L'utente potrà effettuare la propria selezione visionando la carta della regione

con i riferimenti delle singole località indagate oppure l'elenco delle stesse località suddivise per provincia. Una volta selezionata l'area geografica di interesse, si potrà scegliere se prendere visione del materiale raccolto attraverso la griglia del questionario, oppure se ottenere una sintesi generale delle attestazioni dialettali testimoniate dagli informatori intervistati in quel particolare centro. Nel primo caso, l'utente otterrà una sintesi dei materiali raccolti nell'area organizzati per le domande del questionario selezionate. Nel secondo caso, si otterrà invece la lista alfabetica di tutte le attestazioni raccolte corredate di informazione accessoria (ad esempio le domande di cui costituiscono risposta). Di ogni singola attestazione raccolta nella località sarà possibile visualizzare la diffusione sul territorio della regione, attraverso la generazione automatica di mappe dialettali.

Come si può notare, le diverse opzioni della modalità guidata convergono su di un risultato simile, ovvero l'insieme delle attestazioni dialettali ottenute in una data area geografica sulla base di domande specifiche. A seconda della chiave primaria di accesso selezionata è anche possibile ottenere una sintesi dei risultati ottenuti in relazione a una domanda specifica oppure in una località determinata.

5.2. Interrogazione avanzata dei materiali ALT

In *ALT-Web*, la funzionalità di interrogazione avanzata della banca dati dell'*ALT* è stata messa a punto sulla falsariga di quanto sviluppato in *DBT-ALT*⁸, con ovvie modifiche derivanti – ad esempio – dall'integrazione di nuove informazioni nel *corpus* dei materiali dialettali oppure legate alla necessità di un'interfaccia amichevole e di facile interpretazione ed uso.

ALT-Web fornisce procedure di ricerca dinamiche che permettono all'utente di definire interattivamente la chiave di accesso al *corpus* dei materiali *ALT*. Si va da interrogazioni incentrate su singoli elementi, a interrogazioni più complesse volte alla verifica della co-occorrenza

di più elementi, fino ad interrogazioni i cui risultati sono filtrati sulla base di parametri extra-linguistici e/o linguistici.

I materiali dell'*Atlante Lessicale Toscano* possono essere recuperati sulla base di un ampio spettro di parametri:

- domanda del questionario a cui si correlano, direttamente o indirettamente;

- località di inchiesta in cui sono stati raccolti;

- forma, in trascrizione fonetica, ortografica o in versione normalizzata, sia che essa costituisca risposta a domande del questionario, sia che faccia parte del *corpus* di materiali integrativi, sia che ricorra all'interno di descrizioni, fraseologia o commenti degli informatori che sono stati fedelmente registrati in trascrizione fonetica;

- contenuti dell'apparato descrittivo e di commento, che include notazioni degli informatori relative al termine dialettale testimoniato, testimonianze di pratiche tradizionali, o osservazioni di vario genere così come commenti del raccogliitore o del pre-editore dei materiali.

Le funzionalità di accesso di base elencate sopra possono essere variamente combinate per la formulazione di interrogazioni più complesse alla ricerca della co-occorrenza di diversi tipi di informazione in relazione alla stessa attestazione dialettale, oppure per la ricerca dell'occorrenza di una attestazione tra un insieme di varianti.

I risultati dei tipi di interrogazione delineati sopra (semplici e complesse) possono essere filtrati sulla base di:

- status socio-economico e culturale e/o fascia generazionale dell'informatore che li ha attestati;

- area geografica in cui l'attestazione è stata raccolta;

- status socio-linguistico dell'attestazione dialettale, ad esempio registro, connotazione, uso, ecc.;

- uso effettivo da parte dei parlanti.

Vale la pena fare notare che i parametri di selezione sono qui dinamicamente generati tenendo in considerazione la sequenza delle richieste precedenti effettuate dall'utente. Ad esempio, l'opzione di visualizzazione su mappa dei risultati viene attivata solo quando la ricerca includa una interrogazione del *corpus* dei materiali dialettali; oppure, nel caso l'interrogazione abbia riguardato i risultati di una domanda, viene richiesto se i materiali recuperati debbano essere circoscritti alle risposte canoniche alla domanda oppure se possano

8. Cfr. S. MONTEMAGNI-M. PAOLI-E. PICCHI, *DBT-ALT. Manuale di Riferimento*, Lexis Progetti Editoriali, Roma, 2000; E. PICCHI-S. MONTEMAGNI-L. BIAGINI, *DBT-ALT: A System for Storing and Querying the Data of the 'Atlante Lessicale Toscano' (ALT)*, in «Dialectologia et geolinguistica (DiG): Journal of the International Society for Dialectology and Geolinguistics», vol. 9 2001, pp. 85-103.

includere anche materiali integrativi emersi a latere dell'inchiesta. In pratica la formulazione dell'interrogazione è stata organizzata in modo tale da generare una sequenza «a cascata» di richieste che, fase per fase, mette in luce i possibili percorsi da imboccare in modo produttivo. In questo modo si è ritenuto di aiutare il consultatore della banca dati dell'ALT a eliminare il «rumore» che la scelta multipla, particolarmente ricca e differenziata, può causare qualora i parametri su cui può essere basata siano offerti in modo indifferenziato e simultaneamente.

5.3. Accesso ai materiali dialettali su base concettuale

Una delle chiavi canoniche di accesso ai materiali di un atlante linguistico, quando la rilevazione sia stata condotta sulla base di un questionario, è la domanda: all'interno del *corpus* dei materiali raccolti si ricercano tutte le attestazioni che sono state reperite in risposta alla domanda prescelta, oppure che sono emerse in relazione ad essa (nel caso di materiali integrativi ad essa correlati). Una ricerca di questo tipo presuppone però la conoscenza del questionario usato per le inchieste, spesso articolato in centinaia di domande: nel caso specifico dell'ALT, il questionario di raccolta conteneva 745 domande.

Al fine di facilitare il consultatore di *ALT-Web* nell'identificazione della domanda o delle domande oggetto del proprio interesse, abbiamo costruito un'ontologia che organizza i concetti indagati dall'ALT in gerarchie e reti semantiche articolate su diversi livelli. Il termine «ontologia» è qui usato nell'accezione corrente nell'ambito delle tecnologie dell'informazione per denotare un repertorio strutturato di concetti rilevanti per la descrizione e organizzazione di un certo dominio di conoscenza.⁹ La tipologia dei concetti indagati dall'ALT è stata organizzata in 13 macro-classi derivate dall'originaria strutturazione del questionario in settori, più una classe miscelanea che raccoglie domande non immediatamente riconducibili alle macro-classi identificate. Al livello alto dell'ontologia dell'ALT abbiamo le macro-classi (designate dal termine «settori» in *ALT-Web*) listate nella prima colonna della tabella che segue:

9. Cfr. T.R. GRUBER, *Toward principles for the design of ontologies used for knowledge sharing*, in «International Journal of Human and Computer Studies», XLIII 1995, pp. 907-28.

Macro-classe	Numero raggruppamenti semantici associati	Numero domande associate
agricoltura	40	378
alimentazione	31	331
allevamento	16	182
animali selvatici	12	106
bosco e raccolta della legna	24	233
casa e attività domestiche	38	290
forme del terreno	27	143
piante e frutti	23	185
tempo cronologico	8	40
tempo meteorologico	17	107
uomo: attività e relazioni sociali	32	178
uomo: comportamento e sentimenti	73	464
uomo: corpo e abbigliamento	55	407
varia	9	26


Per ciascuna macro-classe o settore sono stati identificati un insieme di classi intermedie o raggruppamenti concettuali più specifici (la cui entità numerica relativamente a ciascun settore è riportata nella seconda colonna della tabella), per un totale di 405 associazioni: ciascuna macro-classe è articolata, in media, in 29 classi concettuali di livello intermedio. Scendendo nell'ontologia, a ciascuna classe concettuale di livello medio sono associati a) concetti elementari espressi attraverso parole italiane singole o espressioni multilessicali, oppure b) parole dialettali. I nodi terminali dell'ontologia sono costituiti dalle domande del questionario: le domande onomasiologiche sono ricollegate ai concetti elementari da esse indagati; le domande semasiologiche sono ricollegate alla classe concettuale più ampia attraverso la parola dialettale indagata. Nella terza colonna della tabella sono riportate il numero di domande ricondotte a ciascun settore, per un totale di 3070 associazioni macro-classe > raggruppamento concettuale intermedio > parola italiana o dialettale > domanda.

Esemplifichiamo l'utilità di quanto illustrato sopra con un esempio concreto. Supponiamo di essere interessati alla terminologia alimentare della castagna. Dopo aver selezionato la macro-classe «alimentazione» e successivamente il raggruppamento concettuale di «castagna», si arriva al livello pre-terminale della gerarchia (rappresentato nella figura che segue) articolato in:

– sei concetti elementari (« castagnaccio », « farina di castagne », « farinata di castagne », « caldarrosta », « castagna bollita » e « polenta di castagne ») indagati attraverso otto domande onomasiologiche distinte (il concetto di castagna bollita è indagato da tre domande distinte, incentrate su diverse modalità di cottura e conservazione delle castagne);

– quattro termini dialettali indagati attraverso tre domande semasiologiche distinte.

Domande relative al raggruppamento concettuale: **castagna**

continua... 

La parola usata per... (domande onomasiologiche)		Cosa significa la parola... (domande semasiologiche)	
castagnaccio	<input checked="" type="checkbox"/> Torta di farina di castagne. (dom. 3048)	pattona	<input checked="" type="checkbox"/> "pattona". (dom. 304a)
farina_di_castagne	<input checked="" type="checkbox"/> Farina di castagne. (dom. 303)	pula	<input checked="" type="checkbox"/> "pula". (dom. 138)
farinata_di_castagne	<input checked="" type="checkbox"/> Farinata di castagne. (dom. 305)	neccio	<input checked="" type="checkbox"/> "neccio", "niccio". (dom. 306)
caldarrosta	<input checked="" type="checkbox"/> Castagne arrostiti. (dom. 307)	niccio	<input checked="" type="checkbox"/> "neccio", "niccio". (dom. 306)
castagna_bollita	<input checked="" type="checkbox"/> Castagne lessate con la buccia. (dom. 308)		
	<input checked="" type="checkbox"/> Castagne lessate senza la buccia. (dom. 309)		
	<input checked="" type="checkbox"/> Castagne secche lessate. (dom. 310)		
polenta_di_castagne	<input checked="" type="checkbox"/> Polenta di farina di castagne. (dom. 304)		

L'utilità di una tale strutturazione del questionario dovrebbe essere evidente, in quanto permette il recupero dei materiali dialettali su base concettuale.¹⁰ In prima istanza all'utente, che si presuppone non conosca la lista delle domande, viene sottoposta la lista di macroclassi concettuali o settori. Una volta selezionata la macro-classe rilevante ai fini della propria ricerca, l'utente otterrà una lista di parole chiave corrispondenti a raggruppamenti concettuali più granulari, che esprimono concetti più specifici. Con la selezione di un singolo concetto si arriva all'identificazione dell'insieme delle domande riconducibili al concetto selezionato, articolato in due sottoinsiemi:

1. quello delle domande onomasiologiche, volte a indagare istanze più specifiche del concetto selezionato;

10. L'accesso su base concettuale può essere attivato in entrambe le modalità di accesso ai materiali dell'ALT, quella guidata e quella avanzata.

2. quello delle domande semasiologiche, le cui risposte includono attestazioni rilevanti per la propria ricerca. Nel caso delle domande semasiologiche, le associazioni attivate tengono conto della tipologia di significati raccolti sul campo per il termine indagato.

Dovrebbe essere chiaro al lettore a questo punto quale sia il valore aggiunto dall'uso dell'ontologia illustrata sopra nell'interrogazione della base di dati dell'ALT per domanda. Si può pensare alla differenza esistente tra la consultazione di un semplice elenco di parole italiane e un dizionario concettuale. Il primo ci fornisce la lista delle parole chiave per l'accesso alla base di dati, il secondo fornisce un substrato concettuale che lega le parole tra loro, mostrandone le relazioni e facendone quindi emergere il significato.

6. RAPPRESENTAZIONE DEI MATERIALI DIALETTALI DELL'ALT

Come già ricordato nella sezione 3.2, in *DBT-ALT* i materiali lessicali sono registrati in trascrizione fonetica; ciò costituisce senza dubbio una barriera non trascurabile per i non addetti ai lavori. Per rendere fruibile la testimonianza linguistica e culturale dell'ALT anche da parte di un'utenza non specialistica, in *ALT-Web* a ogni forma dialettale registrata in grafia fonetica sono stati associati diversi livelli di trascrizioni in ortografia italiana al fine di rendere possibili ricerche che astraggano dalla realizzazione fonetica e/o morfologica del dato lessicale.

Questa opera di trascrizione secondo le convenzioni dell'ortografia italiana dei materiali di *ALT-Web* non si è limitata a garantire la fruibilità dell'opera da parte dei non addetti ai lavori, ma costituisce, a nostro avviso, anche un importante arricchimento dei materiali di base per il linguista e il dialettologo. Infatti, essa è stata condotta nei termini di una « tipizzazione » vera e propria dei materiali dialettali dell'ALT. Di fatto, i criteri per la tipizzazione di materiali dialettali variano profondamente a seconda dei parametri a cui ci si richiama. Se, come spesso si verifica, la tipizzazione tende a riunire tutte le forme etimologicamente collegate, astraendo da variazioni morfologiche quali la suffissazione, il tipo ottenuto sarà diverso da quello cui si arriva a partire da interessi di tipo morfologico (in questa prospettiva, per esempio, diverse suffissazioni danno luogo a tipi diversi); e ancora diverso da quello che si ottiene quando si pertinentizzi-

no meccanismi di incrocio, di paretimologia o altro ancora. D'altro canto, sui criteri per la tipizzazione incide anche la tipologia dei materiali da tipizzare: la stessa forma può ricevere tipizzazioni diverse se si richiamano le forme raccolte come risposta alla stessa domanda in altre località, oppure attestazioni raccolte in risposta ad altre domande nell'ambito dello stesso punto di inchiesta. Da quanto brevemente accennato sopra emerge l'auspicabilità di un modello stratificato di tipizzazione, in cui ogni dimensione trovi la sua rappresentazione adeguata. È a questo tipo di modello che ci siamo ispirati nella definizione dello schema di codifica dei materiali dialettali dell'*ALT*.

In quanto segue illustreremo i diversi livelli di rappresentazione dei materiali dialettali in *ALT-Web*, ne forniremo le motivazioni sottostanti per arrivare a mostrare la loro utilità nelle procedure di interrogazione e recupero dei materiali dialettali dalla banca dati dell'*ALT*.

6.1. Livelli di rappresentazione del dato dialettale in *ALT-Web*

Ad ogni attestazione dialettale contenuta nella base di dati, *ALT-Web* associa diversi livelli di rappresentazione articolati come segue:

1. rappresentazione in trascrizione fonetica;
2. rappresentazioni normalizzate in ortografia italiana: ad ogni attestazione dialettale registrata in grafia fonetica sono associati diversi livelli di trascrizioni normalizzate procedendo per astrazioni successive. I livelli di normalizzazione previsti sono:

2a. traslitterazione in ortografia italiana dell'attestazione dialettale: questa rappresentazione è stata concepita come guida alla lettura e alla decodifica della forma in trascrizione fonetica per l'utente non addetto ai lavori; va comunque rilevato che essa gioca un ruolo centrale in questo schema di codifica in quanto costituisce il perno che mette in comunicazione i due macro-livelli di rappresentazione, in trascrizione fonetica da un lato e in ortografia italiana dall'altro. In virtù di questo ruolo di «cerniera», questo rappresenta il livello base da cui siamo partiti i successivi livelli di normalizzazione;

2b. normalizzazione di primo livello, che neutralizza tratti specifici della realizzazione fonetica del dato come riportati dalla tra-

scrizione ortografica (ad esempio, variazioni fonetiche produttive sul territorio toscano) senza però fare astrazione da variazioni morfologiche, demandate al livello di rappresentazione successivo;

2c. lemmatizzazione, ovvero riconduzione della forma attestata al relativo esponente lessicale o lemma: a questo livello si astrae da variazioni di tipo morfologico, riguardanti sia la morfologia flessiva sia quella derivazionale. In questo modo vengono creati i prerequisiti per interrogazioni che, a partire da un lemma dato, permettono il recupero di tutte le forme flesse e derivate a esso riconducibili e attestate nel *corpus* dei materiali dialettali.

Nella versione corrente di *ALT-Web* i materiali dialettali sono rappresentati ai livelli 1, 2a e 2b; per quanto riguarda il livello 2c, esso rappresenterà una naturale estensione dell'esistente.¹¹

6.1.1. Rappresentazione della trascrizione fonetica

Il sistema di trascrizione fonetica adottato dall'*ALT* si rifà sostanzialmente – con le modifiche e le integrazioni richieste per la codifica dei materiali toscani – al sistema cosiddetto «Ascoli-Merlo» nella configurazione a suo tempo stabilita per le inchieste della Carta dei Dialetti Italiani.¹² In quanto segue, illustreremo il tipo di codifica digitale adottato per i simboli dell'alfabeto fonetico dell'*ALT*, che è stato definito non solo sulla base delle esigenze del calcolatore ma anche e soprattutto delle elaborazioni e analisi che si prospettavano per le attestazioni dialettali dell'*ALT*.

Nel definire criteri e modalità di questa codifica sono infatti emerse esigenze differenziate e contrastanti, ciascuna delle quali legata ad aspetti particolari del processo elaborativo cui i materiali trascritti dell'*ALT* dovevano essere sottoposti (dalla digitalizzazione del

11. Al momento i risultati di questo livello di rappresentazione del dato possono essere approssimati attraverso ricerche «sottospecificate» incentrate sulla radice.

12. I motivi della scelta, rispetto al sistema di trascrizione fonetica standard costituito dal sistema *IPA*, furono a suo tempo dettati da opportunità di congruenza con la tradizione italiana degli studi dialettologici. Al di là di questo, di fatto la scelta di un sistema che, come quello «Ascoli-Merlo», affidasse buona parte delle distinzioni foniche all'impiego di diacritici (per esempio, segni di apertura/chiusura per le vocali: /ɛ/ ~ /ɛ̃/, /ɔ/ ~ /ɔ̃/), rispetto a un sistema che, come quello *IPA*, affidasse le medesime distinzioni all'impiego di segni diversi (cioè, /e/ ~ /ẽ/, /o/ ~ /õ/), è risultata di fatto operativamente più produttiva.

dato dialettale originario al suo ordinamento, recupero e stampa). Ne è derivato uno schema di codifica complesso e articolato che affianca rappresentazioni analitiche (o composizionali) e rappresentazioni sintetiche del dato fonetico originario. Concretamente, se da un lato ogni unità della grafia fonetica è stata scomposta in una base fonica (corrispondente, con una certa approssimazione, al fonema sottostante la realizzazione fonica registrata) modificata, opzionalmente, da segni diacritici (caratterizzanti la variante allofonica),¹³ dall'altro a ogni simbolo dell'alfabeto fonetico, considerato nella sua globalità, è stato associato un codice unico. A seconda del tipo di elaborazione, viene invocato l'uno o l'altro tipo di rappresentazione; la conversione da un tipo all'altro di rappresentazione è generata automaticamente.¹⁴

In questa sede, ci concentreremo sulla rappresentazione composizionale del dato fonetico dell'*ALT*, in quanto essa costituisce il livello di rappresentazione a cui si fa riferimento nelle procedure di interrogazione e recupero dei dati. Attraverso questo tipo di codifica vengono creati i presupposti per un'esplorazione del *corpus* dei materiali *ALT* con finalità ed interessi diversi da quelli fonetici. Infatti, una trascrizione fonetica troppo dettagliata può talora rendere difficoltoso il recupero dei materiali linguistici: maggiore è il livello di dettaglio nella trascrizione, maggiori saranno i fattori di cui tenere conto nella formulazione di ricerche che vogliano astrarre dalle particolari realizzazioni fonetiche. Come visto nella sezione 3.2, la ricerca di un determinato tipo lessicale diventa alquanto difficoltosa in quanto si dovrebbero già conoscere in partenza tutte le sue realizzazioni, cioè avremmo bisogno in partenza del dato di arrivo. Con una rappresentazione composizionale del dato fonetico come quella dell'*ALT*, è invece possibile circoscrivere la ricerca solo alle basi sottostanti alla realizzazione fonica del dato dialettale, astraendo dai tratti codificati attraverso i diacritici (ad esempio, riguardanti il grado di

13. Vale la pena notare che questo sistema di codifica è una derivazione immediata del sistema di trascrizione fonetica adottato nell'*ALT* - quello «Ascoli-Merlo» - che affida buona parte delle distinzioni foniche all'impiego di diacritici.

14. Per una descrizione dettagliata di questo schema di codifica ibrido si rinvia a S. MONTMAGNI-M. PAOLI, *Dalla parola al bit (e ritorno): percorsi dall'inchiesta sul campo alla banca dati dell'Atlante Lessicale Toscano*, in «Quaderni dell'Atlante Lessicale Toscano», n. 7/8 1989-1990, Olschki Editore, Firenze, pp. 7-52 (partic. cfr. pp. 36-43).

apertura della vocale oppure l'eventuale spirantizzazione di occlusiva in contesto intervocalico). Una rappresentazione di questo tipo presenta il duplice vantaggio di mantenere integra la rappresentazione del dato fonetico originario e allo stesso tempo di poter astrarre da alcune delle sue specificazioni nella formulazione di ricerche.

La nozione chiave di questo tipo di codifica è quella di «classe di equivalenza». L'insieme dei segni della grafia fonetica dell'*ALT* è stato ripartito in 20 classi di equivalenza dove i foni all'interno della stessa classe sono ricondotti alla stessa base fonica. La tabella che segue riporta le diverse classi di equivalenza su cui si fonda questo tipo di rappresentazione: nella colonna «Classe di equivalenza» è riportata la composizione di ciascuna classe di equivalenza identificata, mentre nella colonna contrassegnata dall'intestazione «Base» è specificata la base fonica astratta corrispondente, così come codificata negli archivi dell'*ALT*.

Classe di equivalenza	Base	Classe di equivalenza	Base
a, á ã, â, ä, ă	a	l, ł	l
b, ɸ	b	n, ñ, ɲ	n
ç, ʧ	C	o, ó, ô, õ, ö ō, ȝ, ȝ̄, ȝ̅, ȝ̆, ȝ̇	o
d, ɖ	d	p, ɸ	p
e, é, ê, ẽ, Ț ē, ē̄, ē̅, ē̆, ē̇, ē̈, ē̉	e	r, ʀ	r
ə, ə́	@	š, ś	x
ǰ, ǰ̄	G	ʃ, ʃ̄	X
g, ɣ, ǰ̄	g	t, ʈ	t
i, í, î, ï i̇	i	u, ú, û ũ, ũ̄, ũ̅, ũ̆, ũ̇	u
k, k̄	k	ʉ z, ź	Z

Tornando all'esempio di *proda*, una ricerca formulata come sequenza delle basi p+r+o+d+a porta al recupero simultaneo di tutte le varianti menzionate nella sezione 3.2, ovvero tutte le forme che mostrino differenze per quanto riguarda il grado di apertura della vocale e/o l'eventuale spirantizzazione delle occlusive. Da un'attenta analisi della tabella si evince che le classi di equivalenza di foni definite per que-

sto livello di rappresentazione permettono di astrarre, nell'interrogazione e nel recupero del dato dialettale trascritto, dai seguenti tratti fonici: grado di apertura della vocale (/prɔ̄da/ = /prɔ̄da/ = /prɔ̄da/); spirantizzazione di occlusiva (/abɛ̄to/ = /abɛ̄to/); perdita di occlusione nelle affricate (/a bačio/ = /a bačio/); nasalizzazione di vocale (/bɔ̄mba/ = /bɔ̄mba/); turbamento di vocale (/lúna/ = /lúna/); consonantizzazione di vocale (/viɔ̄ttolo/ = /viɔ̄ttolo/; /čɛ̄duo/ = /čɛ̄duo/); carattere velare di /l/ e /n/ (/altalɛ̄na/ = /altalɛ̄na/); posizione dell'accento (/krɔ̄nɔ̄lo/ = /krɔ̄nɔ̄lo/). Gli esempi riportati riguardano tutti alternanze che si verificano in corpo di parola; vale la pena segnalare, comunque, che queste neutralizzazioni operano anche al livello della fonosintassi. Ad esempio, si astrae dalla spirantizzazione di occlusiva in contesto frasale, neutralizzando la distinzione tra /prɔ̄da/ e /la) pɔ̄da/.

6.1.2. Traslitterazione della forma in trascrizione fonetica in ortografia italiana

Come già detto, la forma in trascrizione ortografica è stata concepita come guida alla lettura della forma originaria in trascrizione fonetica che in effetti non sostituisce ma affianca.¹⁵ Nella translitterazione in ortografia italiana delle forme dialettali registrate in trascrizione fonetica si è dunque cercato, ove possibile (ovvero quando consentito dalle convenzioni ortografiche italiane), di rendere conto della variabilità effettivamente rilevata con le inchieste sul campo. Tuttavia, in questa operazione di transcodifica dalla trascrizione fonetica all'ortografia italiana, non si è raggiunto il rapporto auspicabile di 1:1 pena la riproposizione delle difficoltà di decodifica dell'attestazione dialettale che la translitterazione stessa si proponeva di eliminare. Per quanto si sia cercato di riprodurre tutti i tratti di pronuncia registrati, a questo livello intervengono una serie di inevitabili neutralizzazioni dovute all'indisponibilità dei corrispondenti grafemi nell'ortografia italiana. Ad esempio, nella translitterazione non si riesce più a rendere conto della spirantizzazione di grado lieve-medio delle occlusive: al-

¹⁵ L'associare la translitterazione ortografica alla trascrizione fonetica gioca anche un ruolo importante dal punto di vista didattico: permette a chi si avvicina alla disciplina linguistica di familiarizzare con la grafia fonetica, così come aiuta a rendere evidente la convenzionalità dell'ortografia.

le trascrizioni /abɛ̄to/ e /abɛ̄to/ viene associata la medesima trascrizione ortografica, *abeto*.

La tipologia di neutralizzazioni operate nel passaggio dalla trascrizione fonetica a quella ortografica trova quindi la sua principale motivazione in quanto si riesce a restituire con i grafemi dell'ortografia italiana. A questo livello non è stato possibile rendere conto di fenomeni ampiamente diffusi nella regione come la spirantizzazione di occlusiva (con l'eccezione del grado *h* dell'occlusiva velare che viene mantenuto distinto), o la perdita di occlusione nelle affricate palatali; fenomeni, comunque, che proprio per la forte caratterizzazione come marca di toscanismo possono considerarsi un'informazione quasi universalmente acquisita e per così dire essere « dati per scontati » come sottostanti alla translitterazione. Un altro caso di neutralizzazione riguarda le articolazioni pre-palatali dei nessi /kɪ/, /gɪ/, /tɪ/ e /dɪ/ – rappresentate rispettivamente da /č/, /ğ/, /t/ e /d/ – alle quali viene associata la stessa rappresentazione ortografica, ovvero *chi*, *ghi*, *ti* e *di*. In alcuni casi si perde l'informazione di gradi attenuati di particolari fenomeni: nel caso di realizzazioni intermedie o oscillanti (usate ad esempio per indicare la pronuncia leggermente lenita delle occlusive sorde in certe zone della Toscana)¹⁶ viene translitterata la dominante; nel caso di articolazioni attenuate o ridotte (a cui si ricorre ad esempio per rendere il grado vocalico ridotto)¹⁷ viene ripristinata l'articolazione « piena ». Altri tratti che non è stato possibile translitterare fedelmente hanno invece una ridotta diffusione all'interno della regione, come ad esempio il carattere velare di /l/ e /n/, la palatalizzazione di /s/ preconsonantica oppure la palatalizzazione parziale di /l/ (rappresentata come /l'/) davanti a consonante (fenomeno, quest'ultimo, ormai residuale in ridottissime aree). Infine, la resa in ortografia italiana non dà conto della consonantizzazione di vocale.

Riepiloghiamo di seguito la tipologia di neutralizzazioni operate nel passaggio dalla resa in trascrizione fonetica all'ortografia italiana:

- spirantizzazione di occlusiva (escluso il grado *h* che si mantiene

¹⁶ Le articolazioni intermedie o « oscillanti » sono indicate nella grafia *ALT* mediante la sovrapposizione, a cavallo della riga di due segni, con prevalenza del valore espresso in alto.

¹⁷ Le articolazioni attenuate o ridotte, realizzate cioè con intensità minore, sono indicate nella grafia *ALT* in carattere minore e sollevato rispetto alla riga.

distinto) (/akáča/ = /ak'áča/); perdita di occlusione nelle affricate (/a bačío/ = /a bačío/);

- carattere velare di /l/ e /n/ (/altaléna/ = /altałéna/);

- consonantizzazione di vocale (/viǝttolo/ = /viǝttolo/; /čęđuo/ = /čęđuo/);

- palatalizzazione di /s/ preconsonantica (/štradélllo/ = /stradéllo/);

- realizzazioni intermedie: viene selezionata la dominante (/abéǝo/ = /abéto/);

- realizzazioni di grado ridotto: viene ripristinata l'articolazione «piena» (/abét°/ = /abéto/);

- le articolazioni pre-palatali /č/, /ǝ/, /t/ e /d/ sono traslitterate rispettivamente come i nessi /ki/, /gi/, /ti/ e /di/ (/dǝvolo/ = /d'ávolo/; /ǝiáččo/ = /ǝáččo/; /kiáve/ = /čáve/; /stǝáččata/ = /st'áččata/);

- palatalizzazione parziale di /l/ davanti a consonante (/pál'ko/ = /pálko/).

Vale la pena notare che i primi tre tipi di neutralizzazioni della lista coincidono con classi di equivalenza definite per il recupero dei materiali in trascrizione fonetica (cfr. sezione 6.2).

Passiamo adesso in rassegna le distinzioni che sono state invece mantenute a questo livello di rappresentazione. Si è mantenuta la distinzione tra *s* e *z* sorde e sonore (/zzǝlla/ vs /ǝǝǝlla/, /ešǝfo/ vs /ešǝso/), che registra in Toscana un progressivo incremento degli esiti con realizzazione sonora, e l'affricazione di /s/ post-consonantica (/bǝrsa/ vs /bǝrza/), fenomeno in espansione anche in area fiorentina. L'ortografia italiana ha permesso di rendere conto di un fenomeno molto diffuso come il rotacismo (/párko/ vs /pálko/) e di uno al contrario territorialmente assai limitato come la realizzazione cacuminale /d/ (/páda/ vs /pálla/), traslitterata in modo purtroppo parziale in *d*. Inoltre, a questo livello di rappresentazione si è mantenuta l'indicazione del raddoppiamento fonosintattico (/a ppoventá/ vs /a ppoventá/) e dell'accento (anche secondario). Per ciò che riguarda il sistema vocalico, vengono mantenute le sette vocali di base, oltre alle turbate *ö*, *ü* ed *ë* che indica l'indistinta;¹⁸ scelta, questa, utile a

18. L'utilizzo delle vocali turbate ci è parso consentito dalla familiarità ormai invalsa con grafie di forme di origine anglosassone entrate nell'uso.

rendere immediatamente visibile il peso della non toscaneità linguistica di aree come la Lunigiana.

A questo livello avviene anche la ricostruzione delle consonanti nasali in posizione finale, rappresentate al livello della trascrizione fonetica nei termini di vocali nasalizzate. Questa soluzione si motiva con esigenze di trasparenza in quanto una resa priva dell'indicazione di nasalità poteva dar luogo a incomprensioni: ci è parso infatti che *pàn* fosse preferibile rispetto a *pà* per 'pane' e inoltre svolgesse la non trascurabile funzione di (parziale) decodifica del diacritico usato per indicare la realizzazione nasale della vocale.

La volontà di adesione alla trascrizione fonetica ha implicato anche alcune deroghe alla norma ortografica italiana, per cui si è rimandato al livello successivo di normalizzazione l'aggiustamento all'ortografia italiana di *zzi+voc* e di *zz* in posizione iniziale; inoltre, forme del tipo /čęlo/, /ššęnza/ sono state traslitterate come *čelo*, *šęnza* così come /ku/ è stato sempre reso come *qu* (da cui la legittimità a questo livello di forme come *quòre*). Infine, si sono sempre rese con accento (*à*, *ò*, *ài*, *anno*) le voci del verbo *avere*.

Abbiamo visto che nel passaggio dalla trascrizione fonetica alla sua codifica in ortografia italiana si sono rese necessarie alcune neutralizzazioni; al contempo, è stato possibile rappresentare adeguatamente un'ampia gamma di fenomeni quali il rotacismo, l'affricazione di /s/ post-consonantica, il raddoppiamento fonosintattico, ecc. Alcuni dati numerici possono aiutarci a questo punto a capire l'impatto delle inevitabili neutralizzazioni sulla resa in ortografia italiana della trascrizione fonetica. Il *corpus* delle attestazioni dialettali in trascrizione fonetica nell'*ALT* è costituito da 380.348 occorrenze (che includono anche fraseologia di vario tipo), corrispondenti a 84.075 attestazioni diverse (con una frequenza media di 4,5). Nel passaggio all'ortografia italiana, le attestazioni diverse si sono ridotte a 74.105, con un fattore di normalizzazione di 1,13 (calcolato come rapporto tra il numero di attestazioni diverse in trascrizione fonetica e in ortografia italiana). Tale fattore mostra che, per quanto in questo passaggio si sia verificata una forma alquanto ridotta di normalizzazione, la resa in ortografia italiana dei materiali dialettali dell'*ALT* ha permesso di riprodurre in modo abbastanza fedele le caratteristiche della realizzazione fonetica da parte dei parlanti.

Per la traslitterazione in ortografia italiana delle attestazioni in tra-

scrizione fonetica ci siamo avvalsi di procedure automatiche che, a partire dalla versione in trascrizione fonetica, hanno generato automaticamente le corrispondenti forme in ortografia italiana sulla base di 289 regole di traslitterazione. I dati numerici riportati sopra possono fornire un'idea dell'utilità di una scelta del genere, sia in termini di tempo sia di resa qualitativa in relazione al considerevole abbattimento del margine di errore che ne consegue. Lo strumento ideale per realizzare questo tipo di operazione è stato identificato nelle *espressioni regolari* (*regular expressions* o più brevemente *regex*),¹⁹ una notazione algebrica che permette di definire in maniera formale e rigorosa schemi (tecnicamente, *pattern*) di stringhe di caratteri. Le 289 regole di traslitterazione (ripartite in ~200 regole dipendenti da contesto e ~90 regole libere da contesto) sono state dunque codificate nei termini di un insieme ordinato di espressioni regolari, che sono state applicate al *corpus* dei materiali dialettali. Il risultato ottenuto è stato verificato manualmente. In questa fase di revisione finale, particolare attenzione è stata dedicata alle attestazioni alle quali in fase di traslitterazione automatica era stata assegnata una marca di «problematicità» in quanto non si disponeva di informazione sufficiente per una affidabile e univoca traslitterazione automatica (ad esempio, i casi di forme ricostruite). Vale la pena fare notare che questi casi problematici hanno riguardato solo il 9% delle traslitterazioni generate automaticamente.

6.1.3. Rappresentazioni normalizzate in ortografia italiana

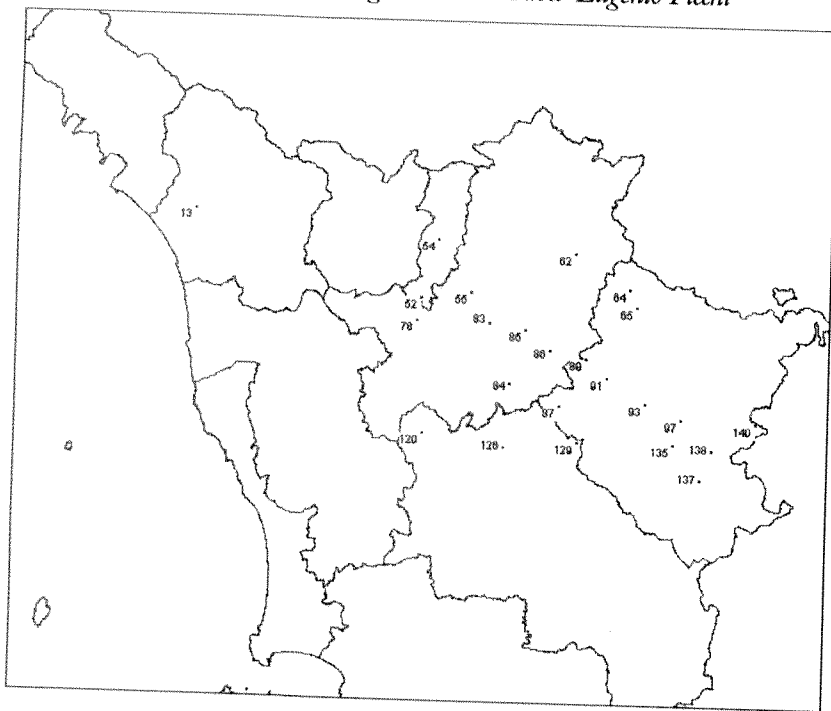
Il recupero di materiali lessicali in trascrizione ortografica si scontra con la stessa difficoltà osservata per quelli in trascrizione fonetica: proprio per il fatto di essere trascritte secondo un sistema di rappre-

19. Le espressioni regolari (*ER*) sono state originariamente sviluppate dal logico matematico S.C. Kleene (1909-1994) nel 1956 come notazione algebrica per rappresentare insiemi di stringhe di simboli. Molti programmi (ad es. Emacs, Word, ecc.) e linguaggi di programmazione (ad es. Perl) supportano le espressioni regolari, ovvero permettono di specificare *pattern* di stringhe usando la sintassi delle *ER* e di verificare su un testo se esistono stringhe che soddisfano tali *pattern*. Per una introduzione alle espressioni regolari si rinvia a: L. KARTTUNEN-J.P. CHANOD-G. GREFFENSTETTE-A. SHILLER, *Regular Expressions for Language Engineering*, in «Journal of Natural Language Engineering», 2, 4, 1996, pp. 305-28; A. LENCI-S. MONTEMAGNI-V. PIRRELLI, *Testo e computer*, Carocci, Roma, 2005 (Cap. 4).

sentazione grafica che tende a riprodurre differenze fonetiche anche minute, le forme assunte localmente da una stessa parola possono essere (anche molto) distanti tra di loro. Si considerino, ad esempio, alcune delle forme reperite sul territorio toscano per indicare il caglio, così come vengono rese nella trascrizione ortografica: *caio*, *càgio*, *càglio*, *càddio*, *càgghio* e *càlio*. Nonostante la resa in ortografia italiana, ci troviamo nuovamente di fronte alla situazione – per certi versi paradossale – in cui l'utente deve prefigurarsi in partenza il risultato della propria ricerca. Questo livello di rappresentazione normalizzata dei materiali *ALT* costituisce un primo passo che crea i presupposti per ricerche che astraggano da tratti della realizzazione foneticomorfologica ed anche lessicale del dato.

In genere, nella normalizzazione di materiali dialettali si sceglie un livello piuttosto alto di astrazione la cui utilità è direttamente proporzionale alla variabilità lessicale riscontrata, a sua volta solitamente più alta più vasta è l'area indagata. La questione in sostanza è stabilire il livello di astrazione più utile a rappresentare adeguatamente l'area indagata. Per tornare all'esempio citato, accanto a *caglio*, che assume le forme elencate sopra, esiste anche nel *corpus ALT* *gaglio*, come rappresentante di analoga classe di forme, che non può considerarsi, in una visione attuale delle parlate toscane, variante fonetica allo stesso livello delle altre rappresentate dalla classe. La sonorizzazione della velare in posizione iniziale non è più operante in Toscana e le forme che la presentano sono ormai lessicalizzate; basta dare uno sguardo alla diffusione della forma *gaglio* sulla mappa della regione (vedi figura nella pagina seguente) per rendersi conto che questa assume una significativa definizione areale.

Un caso diverso è costituito dai termini designanti la *nepitella* sul territorio toscano: le forme attestate al livello 2a sono *empitella*, *epitel-la*, *lempitella*, *lepitella*, *limpidella*, *nempitella*, *nepitella*, *riempitella*, *vempitella* e altre ancora. Il peso numerico delle attestazioni reali ad esse riconducibili e le aree definite che individuano le qualificano non come realizzazioni legate alle condizioni contingenti, ma come tutte ugualmente interessanti da essere studiate singolarmente, ancorché proficuamente riconducibili a *nepitella* in un più astratto cambio di prospettiva. Casi come questo, che rappresentano a nostro parere la grande ricchezza di un atlante proiettato su un territorio relativamente ridotto come quello di una regione, ci hanno indotto a sce-



gliere per il livello di rappresentazione 2b un livello di astrazione relativamente basso.

In concreto, la normalizzazione di livello 2b è intesa come un primo livello di astrazione rispetto a tratti specifici della realizzazione fonetica del dato come riportati dalla trascrizione ortografica. A questo stadio vengono neutralizzate variazioni fonetiche produttive sul territorio toscano: ad esempio, *stiaciàta* e *schiaciàta* vengono ricondotte alla medesima forma normalizzata, lo stesso vale per *viholo* e *vi-colo*, *schiaciàha*, *schiaciàda* e *schiaciàta*, *fidanzàdo* e *fidanzàto*, *diacciàia* e *ghiacciàia*, *cìgghio* e *cìglio*, *mérma* e *mélma*, e così via. Non si astrae invece da variazioni morfologiche: *schiaciàta* e *schiaciàte* rimangono attestazioni distinte così come *schiaciàcia*, *schiaciàcetta* e *schiaciàcina*. E rimangono distinte forme come i citati *gaglio* e *caglio* che hanno la loro motivazione in una variazione fonetica che però oggi in quel particolare territorio della Toscana non è più operante. Anche la classe sopra elencata delle forme riconducibili a *nepitella* è distintamente rappresentata a questo livello.

Ricapitoliamo di seguito la tipologia di normalizzazioni operate a questo secondo livello, organizzate in due insiemi disgiunti: quelle basate su regole generali, che sono state applicate « a tappeto » sul corpus dei materiali ALT, e quelle per le quali è richiesta conoscenza specifica, in particolare lessicale.

a) Regole generali:

- ricostruzione di vocali in corpo di parola: /bìgli/ - *bìg-li* ricondotto a *bìgoli*;
- riconduzione a *sch-* di /sç/ - *s-c(i)*: *s-ciàfòn* ricondotto a *schiaffòne*;
- ricostruzione della velare sottoposta a spirantizzazione (grafia *h*): *ahàcia*, *abbahàre*, *albihòcca* ricondotti a *acàcia*, *abbacàre*, *albicòcca*;
- ricostruzione della dentale sorda intervocalica realizzata come fricativa velare (grafia *h*): *abbandonàho*, *aggranchiàho*, *battùho* ricondotti a *abbandonàto*, *aggranchitò*, *battùto*;
- la cacuminale (trascritta *d*) viene ricondotta a *ll*: *agnèdo*, *badòtti*, *pipistrèdo* ricondotti a *agnèllo*, *ballòtti*, *pipistrèllo*;
- eliminazione del rafforzamento sintattico: *a ciancanèlla*, *a ppaggi-no*, *tu ccapíssi* ricondotti a *a ciancanèlla*, *a paggíno*, *tu capíssi*;
- ricostruzione di *n* in luogo di *m* derivante da assimilazione in fonosintassi davanti a *p/b* o *m*: *im bìlico*, *nom prèsta*, *pam mòlle* ricondotti a *in bìlico*, *non prèsta*, *pan mòlle*;
- neutralizzazione del tratto di sonorità per *s/S* e *z/Z*: *abbòzza* e *abbòZZa* convergono su *abbòzza*;
- *zz* in posizione iniziale passa a *z*: *zzòlla*, *zzàzzera* e *ZZitèlla* resi come *zòlla*, *zàzzera* e *zitèlla*;

b) Regole lessicali:

- le vocali turbate vengono ricondotte alla vocale etimologica: *lúna*, *cunòta* ricondotti a *lúna*, *cunètta*;
- l'indistinta viene ricondotta alla vocale etimologica: *quàrtè d lúna*, *cuntravòntè* ricondotti a *quàrto di lúna*, *contravènto*;
- ricostruzione delle vocali finali: *làmp*, *balén*, *lúm* ricondotti a *làmpo*, *baléno*, *lúme*;
- ricostruzione delle vocali iniziali: *ntepàtiho*, *ncòtta*, *mbròdola* ricondotti a *antepàtico*, *incòtta*, *imbròdola*;
- ricostruzione di *l* preconsonantica passata ad *i* e scempiamento della consonante: *aibbatrèllo*, *gòippe*, *càiddu* ricondotti a *albatrèllo*, *gòlpe*, *càldo*;

- (t)ti viene mantenuto tale oppure ricondotto a (c)chi a seconda dei casi: *béstie* mantenuto tale, *gragnolístio* ricondotto a *gragnolíschio*;
- (d)di viene mantenuto tale oppure ricondotto a (g)ghi o a gl(i) a seconda dei casi: *àddio* ricondotto a *àglio*, *cíndia* a *cínghia*;
- la postpalatale sonora viene ricondotta a gli a seconda dei casi: *cíggio*, *bargèggi* ricondotti a *ciglio* e *bargègli*;
- nni+voc viene ricondotta a gn a seconda dei casi: *granniòla* passa a *gragnòla*;
- ricostruzione della sibilante in luogo dell'affricata dopo l/r/n: *gèlzo*, *addormírzi*, *ànzia* ricondotti a *gèlso*, *addormírsi*, *ànsia*;
- ricostruzione di t in contesto voc-d-voc in fine di parola (caso tipico costituito dalla desinenza del participio passato): *venúdo*, *nformigólido*, *acchittàdo* ricondotti a *venúto*, *informicolíto*, *acchittàto*;
- riconduzione a l di r dovuta a rotacismo: *vórpe*, *àrba*, *càrdo* ricondotti a *vólpe*, *àlba*, *càldo*;
- zzi seguito da vocale passa a zi+vocale: *agitazióne*, *barbuzziènte* ricondotti a *agitazióne*, *balbuziènte*;
- adeguamento alla norma italiana per *cièlo*, *cièco* e *cuòre* e simili;
- inserimento degli apostrofi: *c è* e *l àrba* normalizzati rispettivamente in *c'è* e *l'alba*.

Analogamente al caso precedente, questa normalizzazione di secondo livello è stata condotta con l'ausilio di procedure automatiche, come illustrato di seguito:

1. per ogni attestazione dialettale traslitterata in ortografia italiana, è stata automaticamente generata un'ipotetica forma normalizzata sulla base di un ampio insieme di regole di normalizzazione (precisamente 414) espresse in termini di espressioni regolari;
2. il risultato di questa procedura di generazione di potenziali forme normalizzate ha costituito il punto di partenza della fase di normalizzazione vera e propria, condotta manualmente sull'intero corpus dei materiali dialettali raccolti con l'ausilio di una versione specializzata della procedura interattiva di lemmatizzazione integrata all'interno del *PI-SYSTEM*;²⁰
3. al fine di garantire la coerenza della normalizzazione di varianti inter- o intra-sistemiche di uno stesso tipo lessicale di base, la procedura interattiva di normalizzazione suggeriva al normalizzatore atte-

20. Cfr. PICCHI, op. cit.

stazioni dialettali vicine alla forma in corso di normalizzazione e ricorrenti all'interno del corpus dei materiali ALT: tali suggerimenti venivano rintracciati tra le forme identificate come foneticamente più vicine sulla base dell'algoritmo per il calcolo della cosiddetta « *Levenshtein Distance* » o « *Edit Distance* ». ²¹

La complessità del percorso di normalizzazione tratteggiato sopra si motiva con un tipo di normalizzazione ben più complesso rispetto a quello operato al livello precedente, caratterizzato da corrispondenze 1 : n e m : 1 tra le rappresentazioni di partenza e quelle di arrivo e per la computazione delle quali si doveva tener conto di conoscenza lessicale così come di variazioni inter-sistemiche (riguardanti la dimensione diatopica e quella diastratica) e intra-sistemiche.

6.2. Recupero dei materiali dialettali in ALT-Web

I diversi livelli di rappresentazione associati ai materiali dialettali contenuti nella base di dati di *ALT-Web* creano i presupposti per ricerche che astraggano progressivamente da dettagli della realizzazione fonetica del dato da parte del parlante. In questa sezione, cercheremo di mostrare attraverso l'ausilio di esempi l'utilità di questo complesso e articolato schema di codifica ai fini dell'interrogazione e del recupero dei materiali dialettali dalla banca dati dell'ALT. Come vedremo, la produttività della ricerca varia in modo considerevole a seconda del livello di rappresentazione sul quale viene proiettata l'interrogazione.

Supponiamo di voler recuperare le attestazioni diverse ottenute in risposta alla domanda 331 del questionario ALT, finalizzata alla raccolta dei termini designanti il « caglio ». Proiettando questa richiesta sul livello della rappresentazione fonetica (o livello 1), si ottengono 90 risposte diverse, con una frequenza di attestazione nelle località indagate che oscilla tra 182 (nel caso della risposta maggioritaria /ká-l'ò/) e 1 (i casi di risposte con attestazione unica sono ben 56, corri-

21. Cfr. J.B. KRUSKAL, *An overview of sequence comparison*, in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, a cura di D. SANKOFF e J. KRUSKAL, Stanford, Center for the Study of language and information, 1999; J. NERBONNE-W. HEERINGA-P. KLEIWEG, *Edit Distance and Dialect Proximity*, ivi, pp. v-xv; J. NERBONNE, *Linguistic Variation and Computation*, in *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*, April 2003, pp. 3-10.

spondenti al 62% dell'insieme delle risposte). Vale la pena notare che solo le prime 6 risposte della lista ordinata per frequenza decrescente presentano più di 4 attestazioni: rispettivamente, /kál'p'o/, /presáme/, /presúra/, /gál'p'o/, /kájjo/ e /akkuétta/.

Situazione analoga si osserva proiettando la stessa interrogazione sul livello di rappresentazione 2a, ovvero quello della traslitterazione in ortografia italiana della forma raccolta sul campo. In questo caso, le forme diverse appaiono essere 85 (contrapposte alle precedenti 90) con una distribuzione di frequenza alquanto simile.

Passando all'interrogazione del livello di rappresentazione 2b, ovvero quello dei materiali normalizzati, si constata invece uno scarto notevole nella produttività della ricerca. Infatti, a partire dalla stessa richiesta, le risposte diverse ottenute in relazione alla domanda 331, si riducono a poco più della metà di quelle registrate originariamente in trascrizione fonetica, ovvero 51. Questo deriva dall'accorpamento di risposte diverse che sono state ricondotte alla stessa forma normalizzata secondo i criteri illustrati nelle sezioni precedenti.

Vediamo ora i tipi di normalizzazioni operate nel passaggio da un livello di rappresentazione all'altro. Nella traslitterazione in ortografia italiana delle forme in trascrizione fonetica sono state neutralizzate distinzioni alquanto sottili come quella tra /kál'p'o/, /kál'o/ e /ká^{ll}g'o/ dove la differenza riguarda la realizzazione forte o debole della laterale palatale oppure un'oscillazione tra la realizzazione con /l'p'/ e quella con /gǵ/, tutti casi che non era possibile rendere linearmente con l'ortografia italiana. Il numero di neutralizzazioni operate a questo livello è comunque molto ridotto, riguardando soltanto 5 forme, la maggior parte delle quali contenenti realizzazioni intermedie come quella esemplificata sopra.

Procedendo al passaggio dal livello di rappresentazione 2a a quello 2b, ovvero il livello che astrae da tratti specifici della realizzazione fonetica corrispondenti a variazioni fonetiche produttive sul territorio toscano, abbiamo visto che la tipologia delle risposte si è ridotta quasi alla metà. Vediamo alcuni esempi: alla forma normalizzata *cà-glio* sono state ricondotte forme alquanto diverse come *càio*, *càiè*, *càgio*, *càddiè*, *càio*, *càgghio*, *càiu*, *cài* e *càgliu*; in relazione alla forma normalizzata *gàglio* si osservano simili neutralizzazioni in quanto essa sussume forme come *gài*, *gàddio*, *gàgghio*. Rimangono tuttavia distinte a questo livello attestazioni come *presuràia*, *presàme*, *presina*, *presúra*, *presúri*, *pre-*

súria, *presúro*, *presóia*, *presóio*, *presóre*, *presóro*, catterizzate da diverse suffissazioni e tenute distinte da *présa* così come da *parsúro* e *persúro*.

Abbiamo visto l'impatto dei diversi livelli di rappresentazione del dato su una ricerca incentrata sulla domanda, dove abbiamo osservato la tipologia dei risultati ridursi progressivamente nel passaggio da un livello all'altro. Passiamo ora a una ricerca di tipo diverso, incentrata sulla forma dialettale raccolta sul campo, e vediamo la progressione della produttività della ricerca attraverso i diversi livelli.

Supponiamo di voler recuperare al livello di rappresentazione in trascrizione fonetica le attestazioni della forma /skjaččáta/. Formulando l'interrogazione come sequenza delle sottostanti basi foniche, ovvero s+k+i+a+C+C+a+t+a, le forme diverse recuperate sono due: /skjaččáta/ e /skjaččata/ con la fricativa dentale in sillaba finale, per un totale di 57 occorrenze. Se invece l'utente era interessato, ad esempio, solo alla variante con fricativa dentale in sillaba finale poteva circoscrivere il recupero alle attestazioni che corrispondevano esattamente alla richiesta, con un risultato di 25 occorrenze.

Passando a interrogare il livello di rappresentazione 2a, si constata uno scarto nella produttività della ricerca. Infatti, a partire dalla richiesta *schacciata* le forme diverse in trascrizione fonetica recuperate sono numerose, tra cui: /skjaččáta/, /skjaččáta/, /sčaččáta/, /sčaččata/, ecc. Ovvero, le forme recuperate a questo livello presentano diversi gradi di palatalizzazione del nesso /kj/ e della /s/ preconsonantica, combinati tra di loro e con la realizzazione fricativa dell'occlusiva dentale. Nel caso specifico, si raggiungono 173 attestazioni, ampliando di ben tre volte il risultato ottenuto tramite l'interrogazione operante al livello di rappresentazione precedente.

Procedendo infine all'ultimo livello di rappresentazione, quello che astrae da variazioni di tipo fonetico produttive sul territorio della regione, la mole dei risultati della ricerca effettuata a partire dalla forma normalizzata *schacciata* aumenta ulteriormente, ovvero quasi raddoppia passando da 173 a 310 attestazioni recuperate. A questo livello, vengono ricondotte allo stesso tipo astratto forme alquanto diverse quali *schacciàta*, *schiacciàda*, *stiacciàta*, *schiacciàha* così come *s-ciasséda*. Rimangono invece distinte a questo livello attestazioni quali *schìaaccia*, *schiaccina*, *schiacçetta*, ecc.

Per quanto ulteriori livelli di normalizzazione siano auspicabili, crediamo che la codifica dei materiali dialettali in *ALT-Web* sia già al-

lo stato attuale utile sia al non addetto ai lavori, che interrogherà la base di dati rendendo attraverso le convenzioni dell'ortografia italiana la pronuncia locale, sia al dialettologo e al linguista che troveranno in questa codifica stratificata in livelli caratterizzati da astrazioni incrementali un primo, sebbene elementare, livello di tipizzazione dei materiali dell'*ALT*.

7. PER CONCLUDERE

In questo articolo abbiamo presentato *ALT-Web*, un progetto il cui acronimo è suscettibile di due interpretazioni:

1. *l'Atlante Lessicale Toscano* in rete: l'idea nasce dalla convinzione che il patrimonio linguistico e culturale toscano testimoniato nell'*ALT* non potesse e non dovesse rimanere chiuso nel « cofanetto » della pubblicazione su CD-rom, ma che dovesse essere messo a disposizione della comunità scientifica degli studiosi così come della comunità più vasta di cui rappresenta testimonianza linguistica;

2. *l'Atlante Lessicale Toscano* come rete: il vasto e variegato bacino di utenza a cui ha inteso rivolgersi l'opera ha portato alla trasformazione della versione informatizzata dell'*Atlante Lessicale Toscano* (*DBT-ALT*) pubblicata nel 2000 in una rete ipertestuale con modalità e funzionalità di accesso differenziate in relazione alle diverse classi di utenza. A tal fine, le risorse di partenza sono state potenziate con informazioni e funzionalità aggiuntive volte a migliorare la fruibilità dei dati per ricerche linguistiche e dialettologiche e al contempo facilitare la consultazione del prodotto finale.

Particolare attenzione è stata dedicata al problema del recupero dei materiali dialettali originariamente registrati in trascrizione fonetica. Per quanto in linea di principio i calcolatori rappresentino strumenti che facilitano e rendono più efficiente l'accesso a grandi archivi di dati, nella pratica ciò non è più totalmente vero nel caso di una base di dati lessicali dialettali in cui i materiali siano trascritti foneticamente. In questo caso, il dettaglio imposto dalla trascrizione fonetica nella rappresentazione del dato lessicale può costituire una difficoltà per il suo recupero tramite procedure automatiche in quanto l'utente dovrebbe essere in grado di prefigurarsi in anticipo la varietà di possibili esiti di un dato tipo lessicale nell'area dialettale coperta dalle testimonianze raccolte nella base di dati. Nella sezione 6 del-

l'articolo abbiamo inquadrato il problema della rappresentazione dei materiali lessicali in trascrizione fonetica in una base di dati lessicali dialettali in vista del loro recupero attraverso procedure automatiche e abbiamo illustrato la strategia adottata in *ALT-Web* per la normalizzazione dei materiali dialettali secondo uno schema di codifica articolato su più livelli. Con la tipologia di normalizzazioni messe in atto, la ricerca dei materiali dialettali in *ALT-Web* può astrarre da tratti specifici della realizzazione fonetica del dato attestato; inoltre, la possibilità che l'interrogazione avvenga secondo livelli differenziati di astrazione incrementale, da selezionarsi da parte dell'utente a seconda dei fini della propria ricerca, permette di volta in volta di ampliare o restringere il raggio d'azione del recupero, con un ovvio incremento del « rendimento scientifico » dai dati dialettali raccolti.