Un atlante dialettale in rete: ALT-Web

Nella Cucurullo, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi, Eva Sassolini
Istituto di Linguistica Computazionale – CNR
Via Moruzzi 1 – PISA 56124 – Italy

Abstract
The paper presents an on-line dialectal resource, ALT-Web, which gives access to the linguistic data of the Lexical Atlas of Tuscany or *Atlante Lessicale Toscano*, a specially designed linguistic atlas in which lexical data have both a diatopic and diastratic characterisation. The paper illustrates ALT-Web with particular emphasis on: 1) the dialectal data representation model; 2) the access modalities to the ALT dialectal corpus designed to produce an output tailored to the specific needs of the different classes of users (both professionals and common citizens); 3) ontology-based search. These represent three main features which differentiate ALT-Web both from the previous digitalised ALT version and, most interestingly, from other on-line dialectal resources. At the time of writing, this is the first resource of this kind in Italy, and one of the few at the international level.

1 Introduzione

La quantità e varietà di risorse linguistiche accessibili attraverso il Web va via via crescendo; ciò vale anche per le risorse dialettali, siano esse dizionari o atlanti. Tuttavia, da un'attenta analisi di più di 50 siti internet di atlanti linguistici nel mondo[1] emerge che quelli interrogabili on-line rappresentano ancora oggi una rarità. Nel panorama dell'atlantistica dialettale on-line, ALT-Web rappresenta il primo atlante linguistico interrogabile in rete a livello nazionale e uno dei pochi sul piano internazionale. ALT-Web[2] è nato per rendere il patrimonio linguistico-culturale testimoniato dall'*Atlante Lessicale Toscano* (ALT)[3] una risorsa a disposizione della comunità scientifica nazionale e internazionale per lo studio di dinamiche linguistiche – in particolare lessicali – sia a livello areale sia a livello socio-culturale. In que-

[1] Per una ricca lista di siti internet di atlanti linguistici, cfr. http://serverdbt.ilc.cnr.it/altweb/RT_ALT-WEB_links.htm

[2] 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". ALT-Web è raggiungibile al seguente indirizzo: http://serverdbt.ilc.cnr.it/altweb/.

[3] *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: Gabriella Giacomelli. Redazione: Luciano Agostiniani, Patrizia Bellucci, Gabriella Giacomelli, Luciano Giannelli, Simonetta Montemagni, Annalisa Nesi, Matilde Paoli, Teresa Poggi Salani. Pubblicata come Giacomelli et al., 2000.

sto articolo, ci focalizzeremo su quelli che riteniamo essere tre aspetti qualificanti ALT-Web all'interno del panorama degli atlanti linguistici *on-line*. In particolare:

– l'accesso al corpus dei materiali dialettali, che viene fornito secondo modalità differenziate tese a facilitare e a guidare la fruizione dei materiali dialettali ed etnografici in esso contenuti da parte un'utenza variegata (sezione 2);

– l'accesso ai materiali dialettali su base concettuale, per rendere possibile una ricerca dei dati basata sul concetto che esprimono o, più genericamente, a cui si correlano (sezione 3);

– il complesso modello di rappresentazione dei materiali dialettali articolato su diversi livelli, che rende possibili l'interrogazione e il recupero dei materiali a prescindere da dettagli della realizzazione fonetica e secondo le convenzioni dell'ortografia italiana (sezione 4).

2 Modalità di accesso ai materiali ALT

ALT-Web offre modalità di accesso differenziate in relazione alle diverse classi di utenza; nello specifico, una modalità guidata di accesso ai materiali dialettali, mirante a soddisfare richieste "standard", e una modalità avanzata che permette interrogazioni personalizzate. Questi due accessi differenziati possono essere intesi sia come procedure distinte sia come fasi successive di uno stesso processo di esplorazione dei materiali dove all'identificazione immediata degli elementi fondamentali della banca dati segue la possibilità di effettuare interrogazioni personalizzate sull'intero corpus dei materiali dialettali.

La modalità di accesso guidato si rivolge in prima istanza al pubblico meno specializzato, ma anche a coloro che intendono familiarizzare col materiale ALT prima di passare a ricerche più complesse o semplicemente acquisire materiali rappresentativi in modo sintetico 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 richieste canoniche riguardanti 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 di ricerca personalizzati da proiettare sull'intero corpus dei materiali. In questa modalità 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 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, sulla pertinenza rispetto a una varietà e/o registro linguistico, ecc.

2.1 Accesso guidato ai materiali ALT

A questo livello viene offerta la possibilità di scegliere tra due chiavi primarie di accesso ai materiali dell'ALT: per stimolo (ovvero sulla base della domanda di cui l'attestazione dialettale costituisce risposta) e 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: visionando l'intero questionario (modalità tradizionale destinata a chi conosca già il progetto), 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 (cfr. sezione 3). Selezionato l'insieme di domande alla base della propria ricerca, l'utente ha la possibilità di scegliere se prendere visione dei risultati ottenuti in una località di indagine o in un'area specifica (ad esempio, una provincia) oppure se generare la sintesi generale di tutte le risposte reperite in relazione alle domande selezionate sul territorio toscano.[4] Nel primo caso, il risultato è costituito dall'insieme delle schede che registrano le risposte ottenute nelle località selezionate. Nel secondo caso, il risultato della ricerca 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 secondo 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. Si noti che per la visualizzazione dei materiali dialettali l'utente potrà selezionare il tipo di rappresentazione più appropriato ai fini della propria ricerca: in trascrizione fonetica, nella sua traslitterazione in ortografia italiana, in forma normalizzata (cfr. sezione 4).

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. Una volta selezionata l'area geografica di interesse, si potranno scegliere una o più domande di cui si vogliano vedere i risultati raccolti: il risultato sarà una sintesi dei materiali raccolti nell'area prescelta organizzati secondo le domande del questionario selezionate con le modalità illustrate in precedenza.

Le diverse opzioni della modalità di accesso guidato convergono su di un risultato simile, ovvero l'insieme delle attestazioni dialettali ottenute in una data area geografica sulla base di domande specifiche. Nel caso in cui lo stimolo selezionato sia costituito da una singola domanda, è anche possibile ottenere una sintesi dei risultati.

2.2 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 precendenza,[5] con ovvie modifiche derivanti dall'integrazione di nuove informazioni nel corpus dei materiali dialettali (in particolare, i livelli di rappresentazioni normalizzate dei dati dialettali raccolti), oppure legate alla necessità di un'interfaccia amichevole e di facile interpretazione e uso.

[4] La sintesi delle risposte può essere generata solo nel caso la ricerca abbia riguardato una singola domanda.

[5] Per maggiori dettagli sugli aspetti informatici della realizzazione dell'opera pubblicata su CD-Rom, cfr. Montemagni, Paoli, Picchi (2000); Picchi, Montemagni, Biagini (2001).

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 ricorrenza di più elementi, fino ad interrogazioni i cui risultati sono filtrati sulla base di parametri extra-linguistici e/o linguistici. In partenza, all'utente vengono proposte quattro opzioni corrispondenti ai seguenti domini di ricerca:

1. domanda del questionario a cui i materiali si correlano, direttamente o indirettamente;
2. località di inchiesta (o area) in cui sono stati raccolti;
3. forma, in trascrizione fonetica, in traslitterazione ortografica o in forma 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 fedelmente registrati in trascrizione fonetica;
4. 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 raccoglitore o del pre-editore dei materiali.

Anche a questo livello, l'utente può selezionare la o le modalità di rappresentazione preferite dell'attestazione dialettale: fonetica, ortografica o normalizzata.[6]

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 attestazioni all'interno di un insieme definito sulla base del questionario o su base areale. È possibile impostare la selezione, nell'ambito delle risposte a una domanda specifica, dei soli contesti in cui compaia come risposta una voce particolare: ad esempio, si può richiedere la visualizzazione delle risposte alla dom. 233 "Formica rizzaddome" limitatamente alle attestazioni della voce dialettale *pùzzola*. Oppure si può procedere alla ricerca della terminologia alimentare della castagna in area pistoiese: ciò si ottiene combinando due interrogazioni disgiunte, una relativa alle sette domande del settore alimentare riguardanti le castagne, e una relativa alla zona di attestazione, la provincia di Pistoia (corrispondente a quattordici località).

A seconda della chiave primaria di accesso selezionata, nuovi parametri di selezione sono proposti all'utente: una delle novità di ALT-Web consiste nel fatto che i parametri di selezione dei materiali sono dinamicamente generati in considerazione della sequenza delle richieste precedenti effettuate dall'utente. Ad esempio, l'opzione di visualizzazione su mappa dei risultati viene attivata solo nel caso delle chiavi primarie di accesso per domanda (1) e per forma (3). 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 o se possano includere anche materiali integrativi emersi a latere dell'inchiesta. Ul-

[6] Per default, il sistema seleziona la rappresentazione ortografica, aderente al dato dialettale raccolto sul campo e al contempo comprensibile dal più vasto pubblico.

teriori possibili restrizioni nel caso la ricerca sia stata circoscritta alle risposte canoniche riguardano, ad esempio, il recupero delle sole risposte "piene" oppure dei casi di assenza di risposta. Ciò vale per tutte le chiavi di accesso: ciascun parametro di accesso principale ha associate una serie di scelte che vengono proposte all'utente in tempo debito. In pratica, in ALT-Web la formulazione dell'interrogazione è stata organizzata in modo tale da generare una sequenza "a cascata" di richieste che, fase per fase, metta 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.

Ulteriori parametri di selezione delle attestazioni dialettali, che possono essere variamente combinati con i tipi di interrogazione delineati sopra, riguardano: l'uso effettivo da parte dei parlanti; lo status socio-linguistico dell'attestazione dialettale, come il registro, la connotazione, la vitalità d'uso, ecc.; il tipo di fraseologia che interessa in relazione all'oggetto della propria ricerca, ad esempio la ricorrenza all'interno di etnotesti, proverbi e detti; lo status socio-economico e culturale e/o fascia generazionale dell'informatore.

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 per 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). 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 si raggiungono le 745 domande.[7]

Al fine di facilitare ulteriormente il consultatore di ALT-Web nell'identificazione della domanda o delle domande oggetto del proprio interesse, abbiamo costruito un'ontologia[8] che organizza i concetti indagati dall'ALT in gerarchie e reti semantiche articolate su diversi livelli. In particolare, la tipologia dei concetti indagati è stata organizzata in 13 macro-classi derivate dall'originaria strutturazione del questionario in settori, più una classe miscellanea che raccoglie domande non immediatamente riconducibili alle macro-classi identificate. Al livello alto dell'ontologia dell'ALT abbiamo le macro-classi listate nella prima colonna della tabella che segue:

---

[7] Nella precedente versione dell'ALT, si era cercato di ovviare a questa difficoltà attraverso un repertorio di 338 parole chiave utili a individuare raggruppamenti tematici di domande all'interno del questionario. Anche se ciò rappresentava un ausilio per orientarsi all'interno del questionario, richiedeva in ogni caso la consultazione dell'intera lista.

[8] 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 (Gruber, 1995).

Macro-classe	N. raggruppamenti semantici associati	N. 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
vario	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), 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 polirematiche, oppure b) parole dialettali. I nodi terminali dell'ontologia sono costituiti dalle domande del questionario. 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.

I criteri di associazione tra le domande del questionario ALT e le classi concettuali dell'ontologia variano a seconda del tipo di domanda. Le domande onomasiologiche sono ricollegate ai concetti elementari da esse indagati, a loro volta associati a raggruppamenti concettuali progressivamente più ampi fino a raggiungere i nodi alti dell'ontologia. Nel caso delle domande semasiologiche, ovvero le domande volte a indagare i significati associati a uno o più termini dialettali, la tipologia di associazioni al livello dei nodi terminali e preterminali dell'ontologia è diversa: queste domande sono collegate alla classe concettuale più ampia attraverso la parola dialettale indagata. Tali associazioni sono state stabilite sulla base dei risultati attestati per quella domanda sul territorio toscano.

Nel caso delle domande onomasiologiche una domanda viene classificata di pertinenza di una sola macro-classe ed eventuali polisemie sono catturate attraverso l'associazione della domanda a diversi raggruppamenti semantici intermedi: ad esempio, la domanda 2a "Al crepuscolo della mattina", ricondotta alla macro-classe di 'tempo cronologico', è al contempo associata a diverse chiavi semantiche intermedie come 'luce', 'porzione_del_giorno', 'sole' che ne evidenziano diverse sfaccettature di significato. Una stessa domanda semasiologica può essere invece stata classificata come di pertinenza di diverse macro-classi. Ad esempio, la domanda 11 volta a verificare i significati assunti dal termine *dolco* è stata ricondotta, in virtù dell'accentuata polisemia del termine in ambito toscano, alle macro-classi 'alimentazione', 'casa e attività domestiche', 'forme del terreno', 'tempo meteorologico', 'uomo-comportamento e sentimenti', 'uomo-corpo e abbigliamento'.

Una tale strutturazione del questionario risulta utile in quanto permette il recupero dei materiali dialettali su base concettuale: questo tipo di accesso può essere attivato in entrambe le modalità di interrogazione, quella guidata e quella avanzata. In prima istanza all'utente viene sottoposta la lista di macro-classi 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: quello delle domande onomasiologiche, volte a indagare istanze più specifiche del concetto selezionato; quello delle domande semasiologiche, le cui risposte possono includere 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.

Il valore aggiunto derivante dall'uso dell'ontologia illustrata sopra nell'interrogazione della base di dati dell'ALT (accesso per domanda) pare evidente: 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 sostrato concettuale che lega le parole tra loro, mostrandone le relazioni e facendone quindi emergere il significato.

4 Rappresentazione dei materiali dialettali dell'ALT

Ad ogni attestazione dialettale contenuta nella base di dati, ALT-Web associa diversi livelli di rappresentazione[9] articolati come segue:

1. rappresentazione in trascrizione fonetica;

2. rappresentazioni normalizzate in ortografia italiana: ad ogni attestazione dialettale registrata in grafia fonetica sono associati due livelli di trascrizioni normalizzate:

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 pernio 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 il successivo livello di normalizzazione;

2b. normalizzazione di primo livello, che neutralizza tratti specifici della realizzazione fonetica del dato come riportati dalla trascrizione ortografica (ad es. variazioni fonetiche produttive sul territorio toscano) senza però fare astrazione da variazioni morfologiche.

[9] Per maggiori dettagli sull'articolato schema di codifica dei materiali in trascrizione fonetica dell'ALT si rinvia a Agostiniani, Marinai, Montemagni, Paoli (1998) e Montemagni, Paoli, Picchi (in corso di stampa).

In quanto segue, illustreremo brevemente le procedure adottate per la traslitterazione in ortografia italiana e la normalizzazione, i criteri sottostanti e i risultati raggiunti (sezioni 4.1 e 4.2).

4.1 Traslitterazione della forma trascritta foneticamente in ortografia italiana

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. Nella traslitterazione in ortografia italiana delle forme dialettali registrate in trascrizione fonetica si è cercato, ove consentito dalle convenzioni ortografiche italiane, di rendere conto della variabilità effettivamente rilevata con le inchieste sul campo. Tuttavia, in questa operazione di transcodifica non si è raggiunto il rapporto auspicabile di 1:1 pena la riproposizione delle difficoltà di decodifica dell'attestazione dialettale che la traslitterazione 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 traslitterazione non si riesce a rendere conto della spirantizzazione di grado lieve-medio delle occlusive: alle trascrizioni /a'beto/ e /a'βeθo/ è associata la medesima trascrizione ortografica, *abéto*.

La tipologia di neutralizzazioni operate nel passaggio dalla trascrizione fonetica a quella ortografica trova la sua principale motivazione in quanto si riesce a restituire con i grafemi dell'ortografia italiana. Non è stato possibile rendere conto di fenomeni ampiamente diffusi nella regione come ad esempio 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 traslitterazione. Inoltre non si sono potuti rilevare il carattere velare di /l/ e /n/, la consonantizzazione di vocale, la palatalizzazione di /s/ preconsonantica e la palatalizzazione parziale di /l/ davanti a consonante, fenomeni tutti riscontrabili in aree circoscritte.

È stato invece possibile mantenere a questo livello di rappresentazione la distinzione tra *s* e *z* sorde e sonore (/'dzɔlla/ vs /'tsɔlla/, /e'sɔso/ vs /e'zɔzo/), che registra in Toscana un progressivo incremento degli esiti con realizzazione sonora, e l'affricazione di /s/ post-consonantica (/'borsa/ vs /'bortsa/), fenomeno in espansione anche in area fiorentina. L'ortografia italiana ha permesso di rendere conto di un fenomeno molto diffuso come il rotacismo (/'parko/ vs /'palko/) e di uno al contrario territorialmente assai limitato come la realizzazione cacuminale /ɖ/ (/'paɖa/ vs /palla/), traslitterata in modo purtroppo parziale in *d*. Inoltre si è mantenuta l'indicazione del raddoppiamento fonosintattico 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; scelta, questa, utile a rendere immediatamente percepibile il peso della non toscanità 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; ciò è motivato da esigenze di trasparenza in quanto una resa priva dell'indicazione di nasalità poteva dar luogo a incomprensioni.

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 /tʃɛlo/, /ʃɛntsa/ sono state traslitterate come *cèlo*, *scènza* così come /kw/ 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, ànno*) le voci del verbo *avere*.

Concludendo, 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 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 per attestazione 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 trascrizione 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 (ripartite in 200 regole dipendenti da contesto e 89 regole libere da contesto).[10] 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. Il risultato ottenuto in modo automatico 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.

4.2 Normalizzazione di primo livello dei materiali dialettali

La normalizzazione di primo livello è stata intesa come un primo passo di astrazione rispetto a tratti specifici della realizzazione fonetica del dato come riportati dalla trascrizione ortografica. A questo stadio sono state neutralizzate variazioni fonetiche produttive sul territorio toscano: ad esempio, *stiacciàta* e *schiacciàta* sono state ricondotte alla medesima forma normalizzata, lo stesso vale per *vìholo* e *vìcolo*, *schiacciàha*, *schiacciàda* e *schiacciàta*, *fi-*

[10] I criteri, le modalità e i risultati del processo di traslitterazione in ortografia italiana dei materiali dialettali in trascrizione fonetica dell'ALT sono illustrati in Montemagni, Paoli, Picchi (in corso di stampa), sezione 6.1.2.

danzàdo e *fidanzàto*, *diacciàia* e *ghiacciàia*, *cìgghio* e *cìglio*, *mérma* e *mélma*, e così via. Non si è fatto invece astrazione da variazioni morfologiche demandate a un successivo livello di rappresentazione lemmatizzata dove la forma dialettale attestata è ricondotta al relativo esponente lessicale o lemma ma ad oggi non ancora realizzato:[11] *schiacciàta* e *schiacciàte* sono dunque rimaste attestazioni distinte così come *schiàccia*, *schiaccétta* e *schiaccìna*. E sono rimaste distinte forme come *gaglio* e *caglio* che hanno la loro motivazione originaria in una variazione fonetica che oggi non è però più operante in quel particolare territorio della Toscana in cui sono attestate.

Le normalizzazioni operate a questo livello, sono 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 ci siamo avvalsi di conoscenza specifica, in particolare lessicale. Le regole generali di normalizzazione includono tra l'altro la ricostruzione di vocali in corpo di parola e della velare sottoposta a spirantizzazione (*ahàcia* > *acàcia*); la riconduzione della cacuminale a *ll* (*badòtti* > *ballòtti*); l'eliminazione del rafforzamento sintattico (*a vvànvera* > *a vànvera*); la ricostruzione di *n* in luogo di *m* derivante da assimilazione in fonosintassi (*im pròda* > *in pròda*); la neutralizzazione del tratto di sonorità per *s/S* e *z/Z*. Tra le regole lessicali di normalizzazione vale la pena menzionare la riconduzione delle vocali turbate e dell'indistinta alla vocale etimologica; la ricostruzione delle vocali finali e delle vocali iniziali, della sibilante in luogo dell'affricata dopo *l/r/n* (*pigliàrzi* > *pigliàrsi*) e di *t* in contesto voc-*d*-voc in fine di parola (*andàdo* > *andàto*); o ancora la ricostruzione di *l* preconsonantica passata a *i* e scempiamento della consonante (*vóippe* > *vólpe*); la riconduzione della postpalatale sonora a *gli* (*cìgghio* > *ciglio*) e di *nni*+voc a *gn* (*granniòla* > *gragnòla*) a seconda dei casi; l'adeguamento alla norma italiana per *cièlo*, *cièco* e *cuòre*.

Analogamente al caso precedente, questa normalizzazione di secondo livello è stata condotta con l'ausilio di procedure automatiche,[12] 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);

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 DBT (Picchi, 2003);

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 attestazioni dialettali vicine alla forma in corso di normalizzazione e ricorrenti all'interno del corpus dei materiali ALT: tali suggerimenti venivano rintracciati tra le

[11] Tale livello di normalizzazione non è ancora stato realizzato ad oggi per la problematicità dell'identificazione del lemma rilevante, spesso non facilmente identificabile, e che richiede per questo un impegno che andava ben al di là delle risorse allocate per questo compito nell'ambito del progetto. Al momento, i risultati di questo livello di rappresentazione del dato possono essere approssimati attraverso ricerche "sottospecificate" incentrate sulla radice.

[12] Per maggiori dettagli si rinvia a Montemagni, Paoli, Picchi (in corso di stampa), sezione 6.1.3.

forme identificate come foneticamente più vicine sulla base dell'algoritmo per il calcolo della cosiddetta "Levenshtein Distance" o "Edit Distance", una tecnica sempre più largamente usata in letteratura per la manipolazione e il confronto di materiali dialettali.[13]

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 specifica conoscenza lessicale così come di variazioni inter-sistemiche (riguardanti la dimensione diatopica e quella diastratica) e intra-sistemiche.

5 Conclusioni

Abbiamo presentato ALT-Web, l'Atlante Lessicale Toscano in rete. In particolare, ci siamo soffermati su quelli che riteniamo essere tre aspetti cardine dell'opera, che la caratterizzano rispetto alla precedente versione informatizzata dell'ALT e che soprattutto la rendono unica nel panorama degli atlanti linguistici *on-line*. Innanzitutto, ALT-Web offre all'utente modalità di interrogazione flessibili e dinamiche che permettono di definire la propria chiave di accesso al corpus dei materiali dialettali: i dati ALT possono essere recuperati sulla base di un ampio spettro di parametri che vanno ben al di là delle chiavi canoniche di accesso a un atlante linguistico costituite dal questionario e dalla località di inchiesta e includono, tra l'altro, 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. In questo modo, vengono alla luce informazioni che in un atlante linguistico standard rimangono "nascoste" in quanto non immediatamente rilevabili dalla rigida griglia della carta linguistica. Nell'atlantistica dialettale *on-line* le chiavi di accesso sono invece circoscritte al questionario e, talora, alla località: questo è il caso, ad esempio, dell'ALD (Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects), di LAMSAS (Linguistic Atlas of the Middle and South Atlantic States) e di ALPI (Linguistic Atlas of the Iberian Peninsula). Un altro aspetto peculiare è rappresentato dall'accesso ai materiali dialettali su base concettuale: ALT-Web si presenta ad oggi come l'unica risorsa dialettale che supporta un accesso di questo tipo permettendo così interrogazioni formulate su base semantica con risultati più esatti e puntuali; ciò vale in modo particolare per quanto riguarda la classificazione semantica delle risposte raccolte sul campo per le domande semasiologiche, non altrimenti recuperabili senza conoscere a priori l'esito delle corrispondenti domande. Va infine menzionato il complesso e articolato schema di codifica del dato dialettale che rende possibili interrogazioni di tipo diverso, sia incentrate su una pronuncia specifica sia che astraggono da dettagli della realizzazione fonetica. Questa flessibilità non è offerta da nessun altro atlante *on-line*: in primo luogo, gli atlanti considerati non permettono una interrogazione per forme dialettali. Inoltre, nella visualizzazione dei materiali dialettali viene tipicamente riportata la forma attestata in trascrizione fonetica, accompagnata

---

[13] Sull'uso di questa tecnica in campo dialettologico cfr. Kruskal (1999), Nerbonne et al. (1999), Nerbonne (2003).

– talora – dalla sua riproduzione nel caso dei cosiddetti atlanti “sonori”. L’unica eccezione è rappresentata dal LAMSAS che offre all’utente una vista semplificata del dato fonetico raccolto sul campo, ottenuta attraverso la rimozione dei diacritici (Kretzschmar).

Bibliografia

Agostiniani, L., Marinai, E., Montemagni, S., Paoli, M. (1998), *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 disponibile al seguente indirizzo http://serverdbt.ilc.cnr.it/altweb/silfialt.pdf

Giacomelli, G., Agostiniani, L., Bellocci, P., Giannelli, L., Montemagni, S., Nesi, A., Paoli, M., Picchi, E., Poggi Salani, T. (a cura di) (2000), *Atlante Lessicale Toscano*, Roma, Lexis Progetti Editoriali.

Gruber, T.R. (1995), ‘Toward principles for the design of ontologies used for knowledge sharing’, «International Journal of Human and Computer Studies», XLIII, pp. 907-928.

Kretzschmar, W.A., *Linguistic Databases of the American Linguistic Atlas Project* (ALAP), available at citeseer.ist.psu.edu/478606.html.

Kruskal, J.B. (1999), ‘An overview of sequence comparison’, in Sankoff, D., Kruskal, J. (a cura di), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, CSLI Publications.

Montemagni, S., Paoli, M., Picchi, E. (2000), *DBT-ALT. Manuale di Riferimento*, Roma, Lexis Progetti Editoriali.

Montemagni, S., Paoli, M., Picchi, E. (in corso di stampa), ‘ALT-WEB: l’Atlante Lessicale Toscano in rete’, in Atti del Convegno “Lessicografia Dialettale. Ricordando Paolo Zolli” tenutosi a Venezia in data 10-11 dicembre 2004.

Picchi, E., Montemagni, S., Bigini, L. (2001), ‘DBT-ALT: A System for Storing and Querying the Data of the Atlante Lessicale Toscano (ALT)’, in *Dialectologia et Geolinguistica (DiG)*, vol. 9, pp. 85-103.

Nerbonne, J., Heeringa, W., Kleiweg, P., ‘Edit Distance and Dialect Proximity’, in Sankoff, D., Kruskal, J. (a cura di), *op. cit.*

Nerbonne, J. (2003), *Linguistic Variation and Computation*, in «*Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*», pp. 3-10.

Picchi, E. (2003), ‘PiSystem: sistemi integrati per l’analisi testuale’, in Zampolli, A. *et al.* (eds.), *Computational Linguistics in Pisa. Linguistica Computazionale*, Special Issue, XVIII-XIX, Pisa-Roma, IEPI, 2003, pp. 597-627.