Una procedura informatica di accesso intelligente a materiali in trascrizione fonetica: l'esperienza dell'Atlante Lessicale Toscano

Luciano Agostiniani, Elisabetta Marinai,
Simonetta Montemagni, Matilde Paoli
Centro di Documentazione del Lessico Toscano Moderno - Firenze
Istituto di Linguistica, Università degli Studi di Perugia
Istituto di Linguistica Computazionale, CNR - Pisa

1 Introduzione

In linea di principio, i calcolatori rappresentano strumenti che facilitano e rendono più efficiente l'accesso a grandi archivi di dati. In pratica, ciò non è più totalmente vero quando i dati oggetto della ricerca siano trascritti foneticamente. In questo caso, il dettaglio imposto dalla trascrizione fonetica nella rappresentazione del dato lessicale può costituire una ulteriore difficoltà per il suo recupero tramite procedure automatiche. Questo articolo illustra una procedura informatica di accesso intelligente a materiali in trascrizione fonetica che è stata concepita e sviluppata per l'interrogazione dei materiali dell'Atlante Lessicale Toscano (ALT).

Ci limiteremo a tratteggiare in modo molto sommario che cosa sia l'ALT, a quali intenti risponda e quali siano i numeri che definiscono le dimensioni del corpus raccolto, ed illustremo solo brevemente la configurazione della banca dati elettronica che raccoglie attualmenti i dati del corpus nonché del sistema di interrogazione con cui a tali dati si accede, dal momento che esistono trattazioni esaurienti di entrambi gli argomenti. Successivamente ci soffermeremo sulla procedura della tipizzazione dei materiali lessicali che viene di solito adottata in ambito linguistico per risolvere il problema del recupero dei materiali in trascrizione fonetica e per facilitare la loro successiva elaborazione. In questa sezione illustreremo con esempi tratti dal corpus dell'ALT, le problematiche che tale procedura pone e la possibile soluzione cui ci ha condotti l'esperienza del nostro lavoro. Nei paragrafi successivi passeremo ad affrontare quello che costituisce l'argomento sostanziale della trattazione: il recupero dei materiali trascritti foneticamente a prescindere da una preventiva tipizzazione. In primo luogo si cercherà di chiarire, a partire da dati concreti, quali siano i termini del problema, che difficoltà si incontrino nell'affrontarlo e quali tecniche siano state elaborate per risolverlo. In seguito illustremo come le tecniche esistenti siano state adattate alla realtà dell'ALT e ne sia stata ricavata una procedura originale. Nelle sezioni che seguono si chiarirà quale sia il modello di funzionamento della procedura, su quali presupposti essa sia basata; quali siano in dettaglio i suoi meccanismi di base ed i criteri che la guidano; infine se ne discute in modo critico l'efficacia proponendo esempi reali per misurarne la produttività.

2 L'Atlante Lessicale Toscano

L'Atlante Lessicale Toscano (ALT)[1] è un atlante linguistico regionale che si è proposto di rilevare e definire le condizioni di variabilità diatopica e diastratica che, relativamente al lessico, sussistono all'interno del repertorio dei parlanti della Regione. L'esistenza di condizioni del genere è legata a due fattori: da una parte, la ben nota variegazione dialettale che, come ormai è ampiamente noto, segna le parlate tradizionali della Toscana; dall'altra, il rapporto del tutto peculiare che sussiste tra queste parlate e la lingua nazionale.

Ciò ha avuto ovviamente delle ricadute sulle modalità con cui le ricerche per l'Atlante sono state impostate e condotte. In particolare, il questionario impiegato risulta finalizzato non solo a riconoscere e qualificare lo "specifico toscano" nel lessico del repertorio tradizionale della Regione (e naturalmente le sue differenziazioni interne), ma anche, e soprattutto, a segnalare quei particolari nodi del lessico in cui ciò che è costitutivo della tradizione toscana, o di un suo particolare filone, entra in conflitto – all'interno del repertorio dei parlanti – con la lingua italiana. E' per questo che, all'interno della lista di 745 domande che costituiscono il questionario impiegato, l'angolatura adottata non è uniforme, ma varia a seconda del livello su cui si situa la frizione tra varietà diverse o tra esse e la lingua. La difformità più evidente – ma certo non l'unica – è che mentre alcune domande sono sono finalizzate a rilevare quali forme sono impiegate in riferimento a un dato concetto, altre tendono invece ad appurare quali significati, localmente, corrispondano a una data forma.

La ricerca si è svolta in 224 punti di inchiesta, costituiti da centri abitati che il progetto di ricerca non prevedeva necessariamente omogenei per dimensioni e caratteristiche socio-economiche, ma piuttosto rappresentativi sotto vari aspetti: non ultimo, appunto, quello di presentare diversificate caratteristiche socio-economiche (anche, ma non soltanto, collegate alla diversa dimensione), e differenti condizioni per quanto riguarda il centro egemone di riferimento. In questa rete piuttosto fitta di punti le inchieste sono state svolte, via il questionario, con un campione di informatori che per ogni centro varia, in rapporto soprattutto alla dimensione del centro stesso, mediamente tra 4 e 10, rappresentativo per quanto possibile delle variabili età, sesso e grado di istruzione, e che assomma per l'intero territorio a 2082 unità. Tenendo conto del fatto che ogni informatore intervistato ha fornito, nella maggior parte dei casi, più di una risposta, il materiale raccolto è costituito da un insieme di entrate lessicali di notevole entità: una volta sottoposto a codifica, l'insieme delle risposte risulta strutturato in un corpus lessicale di circa 380.000 schede, delle quali 350.000 compattano le risposte alle domande del questionario, mentre in circa 30.000 altre schede sono raccolti materiali forniti come integrazione alle risposte, o comunque frutto di sollecitazioni scaturite nel corso dell'inchiesta[2].

[1] *Atlante Lessicale Toscano* – Regione Toscana, Accademia Toscana di Scienze e Lettere "La Colombaria". Redazione: Gabriella Giacomelli (Direttore), Luciano Agostiniani, Patrizia Bellucci, Luciano Giannelli, Simonetta Montemagni, Annalisa Nesi, Matilde Paoli, Teresa Poggi Salani.

[2] Per una bibliografia particolareggiata aggiornata all'ottobre 1989, si rimanda alla Bibliografia pertinenente all'Atlante Lessicale Toscano curata da G.Giacomelli e pubblicata in coda a Giacomelli (1987/1988).

3 L'ALT in versione elettronica: DBT-ALT

DBT-ALT è una versione specializzata del DBT, un sistema di database testuale per la memorizzazione, gestione ed interrogazione di grandi archivi di testi[3]; il sistema gestisce una varia tipologia di dati linguistici strutturati, che contengono rappresentazioni sia in trascrizione fonetica - forme ottenute in risposta alle domande o a margine di esse, fraseologia o etnotesti - sia in ortografia italiana - descrizioni o note discorsive.

Nella progettazione e sviluppo di DBT-ALT, si è inoltre dovuto far fronte alle specifiche esigenze della ricerca geolinguistica e sociolinguistica poste dall'ALT, della gestione di un sistema integrato di archivi sussidiari contenenti informazioni riguardo alle località indagate, agli informatori intervistati ed al questionario di raccolta. Inoltre DBT-ALT include la possibilità di proiettare su carta i risultati di una ricerca.

L'Atlante Lessicale Toscano (ALT) è così oggi interrogabile secondo le canoniche chiavi di accesso ad un corpus di materiali dialettali raccolti sul campo tramite questionario: la domanda di cui i materiali costituiscono risposta e la località in cui sono stati raccolti. Inoltre, gli stessi dati possono essere filtrati sulla base delle caratteristiche socio-culturali dell'informatore che li ha attestati.[4]

A queste modalità di accesso se ne affiancano altre che permettono esplorazioni personalizzate, a partire ad esempio da un concetto espresso tramite una chiave semantica, o semplicemente da una attestazione dialettale. L'intervento che qui proponiamo si incentrerà sull'accesso alla Banca Dati dell'ALT (BD-ALT) a partire da attestazioni dialettali riportate in 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*. 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: /e̩/ ~ /e̩/, /o̩/ ~ /o̩/), rispetto a un sistema che, come quello IPA, affidasse le medesime distinzioni all'impiego di segni diversi (cioè, /e/ ~ /ɛ/, /o/ ~ /ɔ/), è risultato di fatto – per motivi che si chiariranno più avanti – operativamente più produttivo (si veda sezione 6.2).

[3] DBT ovvero Data Base Testuale è una elaborazione di Eugenio Picchi, dell'Istituto di Linguistica Computazionale del CNR (Pisa); una descrizione del sistema nelle sue funzionalità si trova in Picchi (1991; in corso di stampa). La versione specializzata del programma DBT-ALT è descritta in Picchi, Montemagni, Biagini (in corso di stampa). Alla realizzazione di DBT-ALT hanno contribuito Lisa Biagini, Elisabetta Marinai e Davide Merlitti.

[4] Per l'architettura del sistema DBT-ALT ed i processi di codifica dei materiali dialettali contenuti nella BD-ALT, si rimanda alle seguenti trattazioni: Agostiniani (1985: § 2); Montemagni (1986); Agostiniani, Montemagni, Poggi Salani (1989); Montemagni, Paoli (1989/1990); Agostiniani *et al.*, (1992).

4 Recupero di materiali lessicali in trascrizione fonetica: generalità

Il recupero dei materiali lessicali in trascrizione fonetica si scontra con una ovvia difficoltà di fondo: proprio per il fatto di essere trascritte secondo un sistema di rappresentazione 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 tratta di una difficiltà che concerne sia il recupero automatico che quello manuale, e che in genere si cerca di superare mettendo in opera procedure di "tipizzazione" lessicale: ad ogni forma trascritta viene associato un "tipo" lessicale astratto, assunto poi ad oggetto della ricerca, e ricavato da esame comparativo delle forme assunte dalla parola. In genere, l'esame comparativo è più "spinto" di quello necessario a superare le divergenze foniche tra le forme, in modo da permettere anche il recupero di forme divergenti (anche) quanto a morfologia (flessiva e derivazionale).

Di fatto, i criteri per la tipizzazione possono variare 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 pertinentizzino 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.

Alcuni esempi tratti dall'ALT possono essere chiarificatori rispetto a quanto appena detto: esaminiamo la forma /orlę̀ttu/ data a Piancastagnaio (P.198) in risposta alla domanda 288 'cantuccio del pane'; se si ha intenzione di contrapporla alle altre forme concorrenti nel resto della Toscana, per esempio /kantúc̆c̆o/ e simili, /krostę̀llo/ e simili ed altre ancora, verrà ricondotta assieme alle altre attestazioni di /orlę̀ttu/ ed a quelle di /orlę́tto/-/orlę̀tto/ ad un unico tipo rappresentato da *orletto*.Ma se l'intento è caratterizzare dal punto di vista linguistico il punto in questione ed in particolare la testimonianza della tradizionalità, occorrerà rilevare la sistematicità con cui il singolare maschile della prima classe di sostantivi e aggettivi esce in *–u*; per cui la rappresentazione astratta dovrà essere *orlettu*.

Per mostrare invece come incidano le diverse prospettive che vanno adottate a seconda dell'obbiettivo della ricerca, si prendano ad esempio le risposte alla domanda 67 'nepitella': nel caso in cui l'interesse sia etimologico, legato alla fortuna dei continuatori del latino, le forme attestate dall'ALT per la Toscana – prescindendo da apax e casi sospetti di confusione tra piante diverse (sempre isolati, comunque) – sono riconducibili a due soli grandi raggruppamenti: quello dei derivati di *menta* e quello dei derivati di *népeta*. Se invece l'interesse è di tipo morfologico, ed in particolare legato alla suffissazione, la serie dei tipi lessicali da ricostruire è più complessa e articolata. Mentre *menta* ha continuatori derivati attraverso suffissazioni diversificate (*mentina*, *mentone*, *méntola*, *mentolina*, *mentolone*, *mentastro*, *mentuccia/mentuzza*), la derivazione da *nepeta* si esprime solo tramite il suffisso *–ella*, in un raggruppamento rappresentato a questo livello dal tipo *nepitella*.

Ma la rappresentazione delle risposte alla domanda 67 cambia ancora quando si pertinentizzino i meccanismi di variazione delle forme. In questo caso la rappresentazione dei derivati di *menta* resta identica a quella ipotizzata nel caso

precedente, mentre il raggruppamento rappresentato da *nepitella* si frammenta nei seguenti tipi:

empitella
epitella
lempitella
lepitella
limpidella
mepitella
nempitella
nepitella
nicotella
nipotella
pipinella
pipitella
repitella
riempitella
vempitella

in cui una sorta di instabilità della sostanza fonica – tipica di certo lessico riguardante soprattutto piccoli animali selvatici, ma comuni, specie insetti, o piante spontanee che fanno comunque parte del conosciuto, o ancora lessico connotato in qualche modo affettivamente in quanto legato all'infanzia – viene riequilibrata dalla costanza della struttura sillabica e dalla invariabilità del suffisso.

Non bisogna poi dimenticare che nel territorio toscano esistono condizioni di variegazione dialettale, circostanza che (in particolare ad un grado di astrazione medio bassa) pone ulteriori problemi. Infatti realizzazioni che in un luogo rappresentano esiti di fenomeni fonetici tuttora operanti, in altro luogo possono essere residui lessicalizzati di situazioni simili ormai pregresse o dovute a prestiti da altra area.

Si prenda il caso, per esempio, dei suffissi *-uccio, -uzzo*. Se consideriamo le risposte alla domanda 359 dell'ALT '(portare i bambini) sulle spalle', si possono individuare fra i derivati di *a cavallo* due serie parallele di attestazioni, una nella Romagna Toscana, l'altra in area sud-occidentale, entrambe costituite da forme in cui il suffisso presenta l'affricata dentale. Ma mentre le attestazioni tosco-romagnole sono con tutta evidenza la realizzazione fonetica locale di *a cavalluccio*, e quindi da questo tipo adeguatamente rappresentate, per il secondo raggruppamento le condizioni dialettali locali, che non contemplano la presenza dell'affricata dentale nel suffisso in questione, indirizzano piuttosto verso la ricostruzione di un tipo autonomo *a cavalluzzo*. E si considerino poi i rapporti diversi che le varietà di aree diverse all'interno della Toscana hanno con la lingua: in questo caso particolare, per esempio, vi sono buoni motivi per assegnare all'influenza della lingua – o se vogliamo del tipo di area fiorentina *a cavalluccio* assunto dalla lessicografia come di lingua – l'attestazione tosco romagnola; mentre quella grossetana si configura come locale e, per di più, coscientemente contrapposta dai parlanti "maremmani" al tipo toscano-italiano.

A completare, infine, il quadro della molteplicità delle tipizzazioni possibili, possiamo utilizzare un esempio limite, anche se non certo unico. La domanda 18 del questionario ALT richiedeva il nome della 'neve', che – a parte aree marginali in cui si ha il tipo *neva* – è ovunque, secondo le aspettative, rappresentato da *neve*; ciò che

contrappone diverse aree all'interno della regione è il grado di apertura della vocale tonica, nonché in alcuni casi il dittongamento: è evidente che una rappresentazione adeguata dell'insieme sincronico di queste risposte deve tenere distinte le varianti /nɛ̨́ve/, /nɛ̨́ve/ e /niɛ̨́ve/.

Dalle schematiche esemplificazioni che abbiamo dato sopra emerge chiaramente, crediamo, quanto complessa sia, in partenza ed in generale, quella procedura di astrazione che è la tipizzazione lessicale. Nel caso specifico dell'Atlante Lessicale Toscano, poi, l'operazione è ulteriormente complicata da una serie di circostanze. Costituisce un problema, prima di tutto, la quantità dei materiali: come abbiamo visto, si tratta di manipolare ± 400.000 schede, per giunta qualitativamente disomogenee. Ci si deve poi confrontare con il fatto che le attestazioni provengono da un sistema composito di varietà dialettali, quale quello compreso entro i confini del territorio toscano. Infine, incide considerevolmente l'impostazione stessa del questionario dell'Atlante, che come si è visto sposta di volta in volta il fulcro della domanda a seconda dell'obbiettivo che si vuole centrare: la definizione delle differenziazioni dialettali delle parlate toscane e le discordanze con la lingua.

E' fuori discussione l'opportunità di applicare all'ALT – in prospettiva - un modello stratificato di tipizzazione, in cui ogni dimensione trovi la sua rappresentazione adeguata, da estendere a tutti i materiali, anche a quelli per i quali la pluralità dei livelli di rappresentazione non appaia significativa nella dimensione della carta. Ma crediamo sia emerso abbastanza, da quanto precede, che tipizzare l'intero corpus dei materiali ALT prima della sua prima pubblicazione era un'operazione che richiedeva tempi talmente lunghi da risultare antieconomica e inopportuna. Si è dunque cercato di risolvere il problema del recupero automatico di dati trascritti foneticamente aggirando l'ostacolo di una preventiva tipizzazione: seguendo cioè una serie di procedure, che verranno illustrate e discusse nei paragrafi che seguono.

5 Recupero automatico di materiali lessicali in trascrizione fonetica senza tipizzazione preventiva

Le procedure canoniche di accesso automatico ai dati di un archivio testuale presuppongono in partenza una "corrispondenza forte" (cioè "esatta") tra l'oggetto della ricerca e quanto recuperato. Nel caso di materiali in trascrizione fonetica, ciò implica che si dovrebbe conoscere in anticipo la realizzazione fonetica esatta delle parole oggetto della ricerca: quando l'ambito della ricerca è costituito da forme codificate in trascrizione fonetica, anche la richiesta deve essere formulata secondo uno schema di codifica compatibile. Ciò vale a dire che se oggetto della ricerca sono le varie attestazioni della parola *neve* l'utente deve formulare la sua interrogazione secondo la sua trascrizione fonetica, e dunque specificando la posizione dell'accento, il grado di apertura della vocale tonica e così via, a seconda del dettaglio imposto dal sistema di trascrizione fonetica adottato.

Nel caso specifico di un atlante linguistico, la formulazione dell'interrogazione si fa ancora più difficile: l'utente dovrebbe potersi prefigurare in anticipo la varietà di possibili esiti di un dato tipo lessicale nell'area dialettale coperta dall'atlante. Quindi, se l'utente è interessato al tipo lessicale *proda* in ambito toscano, deve prefigurarsi diverse interrogazioni, dove 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 /prɔ̨́da/, /prɔ̨́da/, /prɔ̨́da/,

/prǫ̨ɖa/, /prǫ̇ɖa/, /prǫ̈ɖa/, /ꝑrǫ̨da/, /ꝑrǫ̇da/, /ꝑrǫ̈da/, /ꝑrǫ̨ɖa/, /ꝑrǫ̇ɖa/, /ꝑrǫ̈ɖa/. Per potere recuperare questa varietà di attestazioni, l'utente deve formulare una interrogazione per ciascuna di queste, oppure può formulare una interrogazione complessa che contiene tutte le forme in una relazione di disgiunzione. Comunque sia, con sistemi di interrogazione basati su corrispondenze forti, ci troviamo di fronte alla situazione – per certi versi paradossale – in cui l'utente deve prefigurarsi in partenza il risultato della ricerca.

Una possibile alternativa è costituita dalle cosiddette "fuzzy matching techniques", cioè tecniche che non impongono la restrizione di una corrispondenza forte tra l'oggetto della ricerca e l'informazione recuperata, ma che procedono piuttosto mediante l'identificazione di "corrispondenze deboli", o "approssimate".[5] Tecniche di questo tipo sono usate per il recupero in una base di dati testuali di una stessa parola attestata secondo diverse varianti ortografiche, o per l'identificazione e l'interpretazione di errori di digitazione o generati durante il riconoscimento automatico di testi tramite lettore ottico. Queste tecniche, se da un lato accrescono la quantità di informazioni recuperate, dall'altro introducono molto rumore nei risultati della ricerca. L'utente interessato alle attestazioni del tipo lessicale *proda* dovrebbe in questo caso, sulla base della conoscenza della variabilità tra dialetti sul territorio toscano, formulare la sua interrogazione come /?r??a/, cioè sostituendo un "?" ai foni ipotizzati come soggetti a diverse realizzazioni fonetiche. A partire da questa interrogazione, il risultato della ricerca includerà senza dubbio l'insieme delle forme ricercate, ma non soltanto. Insieme a queste ve ne saranno molte altre che non hanno alcuna relazione con l'oggetto della ricerca come /práȼa/, /práta/, /prę́da/, /prę́ta/, /prę́ʃa/, /príma/, /prǫ́na/, /prǫ̨va/, /prúla/, /prúńa/, /brúna/, /krúna/, etc. La similarità fonetica delle forme recuperate è circoscritta ai foni pienamente specificati nell'interrogazione, cioè /r/ e /a/. Riguardo ai foni lasciati sottospecificati, ovvero quelli codificati mediante "?", il recupero non è circoscritto a foni in relazione di alternanza, ma include anche coppie i cui elementi non hanno alcuna relazione di similarità fonetica, ad esempio, /p/ ~ /k/, /d/ ~ /ʃ/.

Per superare il problema di risultati che includono molto rumore, o almeno per ridurlo in modo drastico, sono state messe a punto tecniche di recupero di informazioni ancora basate su corrispondenze deboli, ma dove la corrispondenza tra la forma ricercata e quanto conosciuto dal sistema è calcolata sulla base della similarità fonetica. Questo è il caso dei sistemi di correzione ortografica o dei sistemi per l'accesso a grosse banche dati di nomi e indirizzi in cui, nel caso di mancata corrispondenza forte, vengono ricercate forme vicine a quella desiderata facendo uso di tecniche di corrispondenza debole basata su similarità fonetica.[6] Questi sistemi vedono insiemi di nomi come {*Smith*, *Smit*, *Smythe*, *Smeeth*} per l'inglese o {*Balsini*, *Balzini*}, {*Tognolo*, *Toniolo*} per l'italiano in relazione di corrispondenza debole supportata da similarità fonetica. Supponiamo che un utente ricerchi il cognome *Balsini* in un archivio di dati ma l'archivio a sua volta contenga soltanto attestazioni di *Balzini*. Il calcolatore può restituire *Balzini* come risultato della ricerca sulla base del fatto che /s/ e /z/, quando immediatamente precedute da consonante liquida o

[5] Per una descrizione delle tecniche di "fuzzy matching" si rinvia alla pagina Web all'indirizzo http://law.house.gov/data-src/help/htmlsrc/fuzadv.html.

[6] Si veda ad esempio il "Soundex Coding System", un sistema per la codifica di cognomi basato sulla pronuncia adottato presso i National Archives and Record Administration, Washington DC, USA; dettagli su questo sistema di codifica possono essere trovati presso la pagina Web al seguente indirizzo http://www.nara.gov/genealogy/coding.html.

nasale, hanno realizzazioni fonetiche “simili”, o più precisamente realizzazioni che si alternano liberamente nello stesso contesto. Questa inferenza è possibile per il calcolatore in quanto /s/ e /z/ sono definite come appartenenti alla stessa classe di equivalenza; ed è per questa ragione che al livello della rappresentazione interna /s/ e /z/, quando precedute da liquida o nasale, non sono più viste dal calcolatore come entità diverse.

I sistemi incentrati su corrispondenze deboli basate su similarità fonetica effettuano un riallineamento della rappresentazione ortografica e di quella fonologica tramite la creazione di classi di equivalenza di grafemi o sequenze di grafemi, eventualmente corredati di specificazioni contestuali, alle quali sono associate rappresentazioni interne unificate. Le corrispondenze deboli, quando calcolate sulla base di similarità fonetica, permettono una interrogazione più efficiente in quanto il rumore introdotto dalle “fuzzy matching techniques” è ridotto in modo significativo migliorando qualità e precisione del risultato della ricerca.

La domanda che viene spontaneo porsi a questo punto è se e in che misura tecniche di corrispondenza debole guidata da similarità fonetica possano essere adottate per il recupero di materiali in trascrizione fonetica nell’ambito di atlanti dialettali. Vi sono due aspetti fondamentali che rendono problematica, o almeno non immediata, l’adozione di queste tecniche nel caso specifico. In primo luogo, quanto visto finora opera su rappresentazioni ortografiche che non si rapportano necessariamente in modo isomorfico alle corrispondenti rappresentazioni fonetiche: la stessa rappresentazione fonetica può essere associata a diversi grafemi o sequenze di grafemi. Ciò non si verifica al livello della trascrizione fonetica dove ad ogni segno dell’alfabeto fonetico corrisponde una ed una sola rappresentazione fonetica. Inoltre, le tecniche sopra descritte operano all’interno di un sistema linguistico, mentre nel caso di materiali dialettali di un atlante i sistemi linguistici in ballo sono inevitabilmente più di uno. Questi due fattori impongono una profonda revisione e riadattamento delle tecniche di “fuzzy matching” guidato da similarità fonetica quando si vogliano adottare per il recupero di materiali in trascrizione fonetica di un atlante dialettale.

6 Una possibile soluzione per il recupero di materiali lessicali in trascrizione fonetica non tipizzati

Non disponendo (ancora) di materiali tipizzati, nello sviluppo di DBT-ALT abbiamo dovuto affrontare lo specifico problema del recupero di dati in trascrizione fonetica tramite procedure informatiche. Nel caso specifico dell’ALT tali difficoltà sono ulteriormente accresciute, come abbiamo visto, dal fatto che il corpus dei materiali raccoglie attestazioni della plurarità dei dialetti toscani. Se da un lato ci dobbiamo aspettare che la BD-ALT contenga materiali che presentano una certa variabilità sia da un punto di vista fonetico che morfologico, dall’altro non possiamo dare per scontato che l’utente abbia le conoscenze necessarie per potersi prefigurare la varietà dei possibili esiti di un dato tipo lessicale sul territorio toscano, in sostanza per formulare un’ipotesi attendibile e fondata circa i risultati della ricerca condotta su una vasta e complessa area dialettale quale l’intera Toscana.

Come far fronte alla situazione paradossale secondo la quale l’utente dovrebbe conoscere in anticipo la realizzazione fonetica esatta delle parole oggetto della propria ricerca (situazione che non si verifica necessariamente)? Dati i problemi descritti nella sezione precedente, si è data la necessità di superare i limiti intrinseci delle tecniche

di “fuzzy matching” tout cour. Si è dunque deciso di sviluppare una procedura ad hoc di accesso intelligente ai dati in trascrizione fonetica dove la corrispondenza debole tra la forma oggetto della ricerca e quelle recuperate è calcolata sulla base della conoscenza dei dialetti toscani, conoscenza supportata da anni di analisi dei materiali dell’Atlante Lessicale Toscano.

Questa procedura presenta analogie ma anche profonde differenze rispetto ai sistemi di recupero di informazioni basati su corrispondenze deboli guidate da similarità fonetica: i livelli di rappresentazione in gioco sono diversi, così come la tipologia di corrispondenze stabilite tra livelli.

Come più volte ripetuto, nel caso dei materiali dell’ALT il livello di base su cui operare è quello della trascrizione fonetica, e non della rappresentazione ortografica. A questo livello, per definizione non si verifica che diversi grafemi o sequenze di grafemi siano associati allo stesso fono. Il livello di rappresentazione interna, al quale avviene la neutralizzazione di differenze che si osservano al livello di base, è più “astratto” che nel caso precedente: nel caso di un atlante linguistico stiamo operando all’interno di una complessa architettura di sistemi dialettali, e dunque la rappresentazione interna deve astrarre non solo da varianti intra-sistemiche (come nel caso precedente) ma anche inter-sistemiche.

Per quanto riguarda la rappresentazione interna, si è inoltre pensato di articolarla su più livelli, ed in particolare livelli incrementali di rappresentazioni astratte: al livello più basso l’astrazione rispetto al dato attestato è minore rispetto ai livelli successivi. Ciò si riflette sull’estensione delle classi di equivalenza stabilite per i vari livelli, più ristrette ai livelli più bassi, più estese quanto maggiore è l’astrazione rispetto alla realtà del dato.

In questo modo, all’utente è offerta la possibilità di selezionare livelli differenziati per l’accesso ai dati:

- un primo livello (0) che richiede una corrispondenza forte tra l’oggetto della ricerca e il dato recuperato;
- una serie di livelli che si basano su corrispondenze deboli, permettendo così all’utente di astrarre da tratti specifici della realizzazione fonetica e/o morfologica del dato dialettale:
  - livello 1, che astrae da segni diacritici quali apertura di vocale, spirantizzazione di occlusiva etc.;
  - livello 2, che si basa su relazioni complesse di equivalenza tra fonie (semplici e composte);
  - livello 3, che astrae da variazioni di tipo morfologico, riguardanti sia la morfologia flessiva sia quella derivazionale.

Quindi, ad un primo livello 0 che impone la corrispondenza esatta tra l’oggetto della ricerca e il dato linguistico recuperato, se ne affiancano altri (1-3) che permettono di astrarre nell’interrogazione da tratti fonetici e/o morfologici della realizzazione attestata. Il ricorso ai livelli 1-3 non è da considerarsi come una strategia di ripiego nel caso di risposta mancata o esigua ad una interrogazione di livello 0; anzi spesso sono proprio le modalità di accesso di livello 1-3 da privilegiarsi rispetto alla modalità 0, per la loro maggiore produttività. Non a caso, il livello di accesso 1 è quello predefinito nel caso del programma di interrogazione DBT-ALT.

Ad oggi, il programma DBT-ALT permette interrogazioni che coprono i livelli 1-3: ai livelli 1 e 2 si astrae da aspetti particolari della realizzazione fonetica, al livello 3 la neutralizzazione riguarda varianti morfologiche, ed in particolare della morfologia flessiva di nomi e aggettivi.

6.1 Dietro alla procedura di accesso intelligente a dati in trascrizione fonetica dell'ALT

Una nozione chiave della procedura dell'ALT per un accesso intelligente ai dati lessicali è quella di "classe di equivalenza". Le corrispondenze deboli calcolate per l'accesso di livello 1 e 2 sono basate sulla definizione di classi di equivalenza di fonie, foni singoli e/o sequenze di foni. Al livello di accesso 3, le classi di equivalenza raccolgono morfi alternanti.

Le classi di equivalenza definite raccolgono alternanze di fonie e/o morfi ricorrenti nella varietà dei dialetti toscani (ad esclusione delle aree marginali[7]), sia all'interno dello stesso sistema dialettale sia tra sistemi diversi. La composizione interna di una classe di equivalenza può essere complessa, vale a dire può includere varianti intra-sistemiche così come inter-sistemiche. Le varianti inter-sistemiche riguardano sia la dimensione diatopica sia quella diastratica, essendo il campione di informatori intervistati differenziato rispetto all'età ed allo status socio-culturale. A livello intra-sistemico viene considerata simultaneamente la produzione di una parola sia in isolamento sia in contesto di frase. Tutte queste varianti, ad oggi, sono rappresentate sullo stesso piano all'interno della stessa classe, anche se – come verrà argomentato in seguito – una strutturazione interna delle classi di equivalenza potrebbe migliorare il rendimento della procedura di accesso (si veda la sezione 6.3).

In quanto segue, le classi di equivalenza sottostanti ai diversi livelli di accesso della procedura dell'ALT sono descritte mediante esempi (sezione 6.2) insieme ai criteri adottati per la loro formazione (sezione 6.3). Segue, nella sezione 7, una esemplificazione della loro utilità nell'interrogazione della BD-ALT.

6.2 Classi di equivalenza

6.2.1 Classi di equivalenza di foni di Livello 1

Le classi di equivalenza di Livello 1 operano a livello del singolo fono dove l'appartenenza di un dato fono ad una classe specifica non è condizionata dal contesto in cui esso occorre. Da notare il fatto che le classi definite a questo livello sono disgiunte, ovvero ogni segno dell'alfabeto fonetico appartiene ad una sola classe.

Le classi di equivalenza di Livello 1 sono definite attraverso un sistema di codifica della trascrizione fonetica che si basa su rappresentazioni composizionali, cioè dove ogni simbolo dell'alfabeto fonetico è codificato mediante una base che può

[7] Le aree marginali con il loro riferirsi a sistemi completamente diversi da quelli toscani, cumulano all'interno di una stessa forma una serie di fenomeni fonetici che impongono regole di corrispondenza e conseguenti classi di equivalenza proprie; una volta proiettate sull'intero territorio toscano le classi di equivalenza specifiche delle aree marginali interferivano con le corrispondenze delle varianti toscane: il rumore introdotto era tale da rendere il risultato della ricerca di difficile lettura. Una possibile soluzione attualmente in esame è la possibilità di specializzare le classi di equivalenza, o sottoinsiemi di esse, in rapporto a specifiche aree di riferimento (si veda sezione 6.3).

essere ulteriormente specificata per mezzo di diacritici.[8] 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 (si veda sezione 3). Ad esempio, per quanto riguarda le vocali, la stessa base si riferisce a vocali accentate e non, chiuse o aperte, e/o palatalizzate e/o nasalizzate dove quest'ultima tipologia di specificazioni è codificata mediante diacritici associati alla base.

La tabella che segue riporta le classi di equivalenza definite per l'accesso di Livello 1 ai dati dell'ALT: nella colonna "Classe di equivalenza" sono riportate le diverse classi di equivalenza definite per questo livello di accesso, mentre nella colonna - contrassegnata dall'intestazione "Base" - è specificata la base astratta corrispondente.

Classe di equivalenza	Base
a, á ã, ắ, ä̋, a̋̃	a
b, ƀ	b
č, ȼ	C
d, đ	d
e, é, ẹ́, e̋, ę́ ẽ, ệ, e̋̃, ę̃, ḛ́, e̋́, ę́, ë, ë̃, e̋̃	e
ə, ə́	@
ǧ, ǵ	G
g, ǥ, ġ	g
i, í, í̯, ǐ i̯	i
k, ķ	k

Classe di equivalenza	Base
l, ł	l
n, ṅ, ṉ	n
o, õ̜, ọ́, ó̱, ó̜, ő̜ õ, ọ̃, õ̱, ö, ő, ö̃, ő̃	o
p, ᵽ	p
r, ṛ	r
š, ś	x
ʃ, ʃ́	X
t, ŧ	t
u, ú, ú̜ ũ, ũ̜, ü, ü̃, ű̃ u̯	u
ʒ, ʒ́	Z

La tabella riporta le distinzioni attestate nella grafia ALT che sono neutralizzate al Livello 1 dove si astrae, ad esempio, dal grado di apertura delle vocali o dalla spirantizzazione di occlusive. Quindi, a questo livello le attestazioni del tipo lessicale *proda* ipotizzate come plausibili in ambito toscano nella sezione 5, cioè /pró̜da/, /pró̜da/, /pró̱da/, /pró̜ɖa/, /pró̜ɖa/, /pró̱ɖa/, sono tutte percepite dal sistema come equivalenti in quanto i foni alternanti appartengono alla stessa classe di

[8] Per una descrizione della codifica della trascrizione fonetica nell'ambito di DBT-ALT, si rimanda a Montemagni, Paoli (1989/1990:36-43).

equivalenza. Ciò implica che non devono essere ricercate singolarmente le possibili varianti dialettali in cui un dato tipo lessicale può apparire, ma basta formulare una unica interrogazione incentrata sulle basi sottostanti.

Le classi di equivalenza di foni di Livello 1 permettono dunque di astrarre, nell'interrogazione e nel recupero del dato dialettale trascritto, dai seguenti tratti fonici:

- grado di apertura della vocale (/prǫ̨da/ ≡ /prǫ́da/ ≡ /prǫ̨́da/)
- spirantizzazione di occlusiva (/abę́to/ ≡ /ab̨ę́to/)
- perdita di occlusione nelle affricate (/a bačío/ ≡ /a bă̦ćío/)
- nasalizzazione di vocale (/bǫ́mba/ ≡ /bǫ̨́mba/)
- turbamento di vocale (/lúna/ ≡ /lu̇na/)
- consonantizzazione di vocale (/vi̯ǫ̨ttolo/ ≡ /viǫ̨ttolo/; /čę̨́du̯o/ ≡ /čę̨́duo/)
- carattere velare di /l/ e /n/ (/altalę́na/ ≡ /altaɬę́na/)

Gli esempi riportati riguardano tutti alternanze che si verificano in corpo di parola. E' importante segnalare, comunque, che questo livello opera anche in fonosintassi: ad esempio, si astrae dalla spirantizzazione di occlusiva in contesto frasale, neutralizzando la distinzione tra /prǫ́da/ e /(la) p̶rǫ́da/.

Infine, al Livello 1 si astrae anche dall'accento: /krǫ̨ńńolo/ e /krońńǫ̨́lo/ sono percepite come equivalenti, nonostante la posizione dell'accento sia diversa nei due casi.

6.2.2 Classi di equivalenza di fonie di Livello 2

Al Livello 2, le classi di equivalenza operano sia a livello del singolo fono sia di sequenze di foni, e per questo motivo sono genericamente designate come classi di equivalenza di fonie. A differenza delle classi di equivalenza di Livello 1, al Livello 2 l'appartenenza ad una classe specifica può essere contestualmente condizionata, ovvero l'equivalenza tra due foni singoli o sequenze di foni può essere ristretta a contesti specifici. Inoltre, le classi definite si basano su classificazioni multidimensionali, per cui lo stesso fono o sequenza di foni può appartenere a più classi, a seconda della dimensione classificatoria considerata. Questo implica che, a differenza delle classi di Livello 1 che sono disgiunte, le classi di equivalenza di Livello 2 possono presentare intersezioni.

Queste caratteristiche di base rendono le classi di equivalenza di Livello 2 molto più complesse e dunque anche più difficili a descriversi. Ad esempio, mentre al livello 1 ogni classe fa riferimento ad una sottostante base fonica, in questo caso l'identificazione di una base che sussuma l'insieme di fonie raccolte nella stessa classe diventa impresa più difficoltosa ed anche arbitraria. Quindi, a differenza del caso precedente, ogni classe è illustrata nella tabella che segue tramite insiemi di attestazioni lessicali che sono in relazione di equivalenza al Livello 2.

CLASSE di EQUIVALENZA	ESEMPI
č, čč, ć, š (sse #_V // V_V)	/a bačío/ ≡ /a bačč́io/ ≡ /a baćío/ ≡ /a bašío/
ǧ, ǵ, ǧǧ,	/ǧákka/ ≡ /ǧǧákka/ ≡ /ʤákka/

CLASSE di EQUIVALENZA	ESEMPI
ʃ (sse #_V // V_V)	/čeráǵa/ ≡ /čeráğa/
k, ǩ, h	/bákera/ ≡ /báǩera/ ≡ /báhera/
kⱼ, ǩ ⱼ, kkⱼ, tⱼ, ŧⱼ, ttⱼ, č, čč, ť, ťť	/mákkⱼa/ ≡ /máča/ ≡ /máčča/ ≡ /máťťa/ /grańńolískⱼo/ ≡ /grańńolístⱼo/ ≡ /grańńolíščo/ /pettⱼére/ ≡ /peččére/
gⱼ, ǧⱼ, ggⱼ, dⱼ, đⱼ, ddⱼ, ğ, ğğ, d', d'd'	/gⱼaččáⱼa/ ≡ /dⱼaččáⱼa/ ≡ /ğaččáⱼa/ ≡ /d'aččáⱼa/ /ad'd'áⱼa/ ≡ /ağğáⱼa/ /aggindáto/ ≡ /ağğindáto/
l, r (sse _C[≠ r]), ⱼ (sse _C_i C_i), ľ (sse _C)	/dǫ́lko/ ≡ /dǫ́rko/ ≡ /dǫ́ľko/ ≡ /dǫ́ⱼkko/
ll, ḍ, ḍḍ	/ballǫ́tti/ ≡ /baḍǫ́tti/
ľ, ľľ, lⱼ, llⱼ, ğ, ğğ, d', d'd'	/číľľo/ ≡ /číľo/ ≡ /čílⱼo/ ≡ /číğğo/ /kaľľáta/ ≡ /kad'd'áta/
ń, ńń, nⱼ, nnⱼ	/pańńǫ́ne/ ≡ /panⱼǫ́ne/ /pannⱼére/ ≡ /panⱼére/ ≡ /pańére/ ≡ /pańńére/
r, rr (sse V_V)	/karrarę́čča/ ≡ /kararę́čča/
s, ʃ,ʃ, ś, š (sse _C), ʃ (sse _C), z (sse l_ // r_ // n_)	/esǫ́so/ ≡ /eʃǫ́ʃo/ /roʃmaríno/ ≡ /roʃmaríno/ /suʃína/≡/suʃína/ /bⱼastę́ma/ ≡ /bⱼaśtę́ma/ ≡ /bⱼaštę́ma/ /ánsimo/ ≡ /ánzimo/
š, šš	/kášša/ ≡ /kaša/
z, zz, ʒ, ʒʒ	/zzázzera/ ≡ /zázzera/ ≡ /ʒʒázzera/ ≡ /ʒʒáʒʒera/

Come si può notare da un'analisi attenta delle singole classi, questo secondo livello di accesso, basandosi su relazioni di equivalenza tra diversi segmenti fonologici, permette un maggiore grado di astrazione rispetto al Livello 1: ad esempio, a questo livello viene stabilita l'equivalenza tra /kⱼ/, /tⱼ/, /č/ e /ť/, cioè tra foni singoli e sequenze di foni. Talora l'equivalenza è circoscritta a contesti specifici; /z/ è alternante con /s/ limitatamente a certi contesti, in quanto l'affricazione della sibilante ha luogo soltanto quando preceduta da liquida o nasale. Nella colonna degli esempi, ogni riga include forme in relazione di corrispondenza debole a questo livello; l'interrogazione di Livello 2 formulata a partire da una qualsiasi di queste forme porta automaticamente anche al recupero delle altre.

Anche a questo livello, le alternanze sono ricercate sia a livello di parola singola, sia in contesto di frase. In fonosintassi si astrae, nell'interrogazione e nel recupero del dato dialettale trascritto, dai seguenti fenomeni:

- raddoppiamento fonosintattico: /a solatío/ ≡ /a ssolatío/; /a bačío/ ≡ /a bbačío/;
- spirantizzazione di occlusiva: /tornár di ǩása/ ≡ /torná ddi hása/ ≡ /andá n kása/;
- perdita dell'elemento occlusivo delle affricate: /ai ć́ipręssi/ ≡ /ai čipręssi/;
- affricazione delle sibilanti: /al zolatío/ ≡ /al solatío/;

– caduta della vocale finale: /ákku̯ e nnẹ́ve/ ≡ /ákku̯a e nnẹ́ve/; /fra llúsk e bbrúsko/ ≡ /fra llúsko e bbrúsko/.

6.2.3 Classi di equivalenza di morfi di livello 3

Le classi di equivalenza di Livello 3 raccolgono morfi alternanti. Al momento, le classi definite ed operanti per l'accesso ai materiali ALT per questo livello sono circoscritte alla morfologia flessiva di nomi e aggettivi, vista dalla duplice prospettiva inter- ed intra-sistemica.

A questo livello, tutte le vocali atone occorrenti in fine di parola sono ricondotte ad un esito unificato. Ciò vale a dire che, sul piano intra-sistemico, si comincia ad astrarre da varianti morfologiche quali variazioni di genere e numero laddove esse siano espresse da vocale finale: /miľľàččo/ e /miľľàčči/, /skia̯ččàta/ e /skia̯ččàte/ sono viste a questo livello come istanziazioni di un unico tipo lessicale, e dunque come equivalenti.

Le classi di equivalenza di Livello 3 includono anche varianti morfologiche inter-sistemiche: si considerino ad esempio le forme designanti la cantina, /čiľľẹ́re/ e /čiľľẹ́ri/, entrambe al singolare ma appartenenti a sistemi morfologici diversi. Sempre tra le varianti inter-sistemiche neutralizzate a questo livello sono da menzionare i casi di dileguo della vocale finale in parola singola, ovvero non contestualmente condizionato: in questo modo, le forme /bakkàn/, /balugàn/ e /pàn/ sono assimilate alle corrispondenti /bakkàno/, /balugàno/ e /pàne/ e dunque recuperate a questo livello mediante la stessa interrogazione.

6.2.4 Classi di equivalenza a confronto

La distinzione tra classi di equivalenza di Livello3 da un lato e classi di Livello 1 e 2 dall'altro si basa sulle unità linguistiche che raggruppano, rispettivamente, morfi e foni (o segmenti costituiti da sequenze di foni).

La distinzione tra classi di Livello 1 e 2 è invece più sottile. Ricapitoliamo di seguito le caratteristiche distintive dei due raggruppamenti: le classi di equivalenza di Livello 1 operano a livello del singolo fono, l'appartenenza di un dato fono ad una classe specifica non è condizionata dal contesto in cui occorre e le classi definite sono disgiunte; le classi di equivalenza di Livello 2 operano sia a livello del singolo fono sia di sequenze di foni, l'appartenenza ad una classe specifica può essere contestualmente condizionata e le classi definite possono presentare intersezioni. Quindi, la distinzione tra Livelli 1 e 2 si fonda sulla tipologia dei dati raggruppati così come la loro classificazione, mono-dimensionale nel caso delle classi di Livello 1 e multi-dimensionale nel caso delle classi di Livello 2.

La multi-dimensionalità delle classi di equivalenza di Livello 2 ha una serie di ripercussioni per quanto concerne la loro relazione con le classi del livello precedente. Si danno fondamentalmente due casi:

– la classe di Livello 2 può essere un'ampliamento di una classe di Livello 1: ad esempio, la classe di Livello 1 che raggruppa l'occlusiva velare sorda e la sua variante lievemente spirantizzata { liv1: k, kʻ } al Livello 2 viene estesa tramite l'inclusione di /h/, risultando nella classe che segue { liv2: k, kʻ, h };

— le classi di Livello 2 possono codificare distinzioni ortogonali rispetto a quelle codificate a Livello 1: ad esempio, la sibilante palatale /š/, che al Livello 1 appartiene alla classe { liv1: š, ś }, al Livello 2 rientra in più classi, quella delle affricate palatali { liv2: č, čč, ć, š }, delle sibilanti { liv2: s, ʃ, ś, š, ʃ, z } e delle sibilanti palatali di diverso grado { liv2: š, šš }.

Quanto osservato sulla relazione tra le classi di Livello 1 e 2 dovrebbe dare ragione della diversa produttività di una stessa interrogazione effettuata ai due livelli; nella sezione 7 che segue questo punto verrà illustrato con un esempio concreto.

6.3 Criteri sottostanti la definizione delle classi di equivalenza

I criteri che hanno guidato la definizione delle classi di equivalenza per i diversi livelli non sono soltanto di tipo teorico, come illustrato finora, ma rispondono anche a esegenze pratiche. In particolare, per quanto concerne gli ultimi, si è mirato all'ottimizzazione del rapporto tra la quantità di materiali lessicali recuperati e il rumore prodotto da ricerche basate su corrispondenze deboli.

Una conseguenza immediata dell'applicazione di questo criterio come guida per la formazione delle classi di equivalenza sta nell'esclusione dalle classi di equivalenza definite di variazioni allofoniche di una certa area alle quali corrispondono opposizioni fonologiche di un'altra area. Infatti, l'inclusione all'interno di classi di equivalenza di opposizioni fonologiche, anche se ristrette ad un'area dialettale specifica, avrebbe causato inevitabilmente una elevata quantità di rumore nei risultati di ricerche condotte sulla base di queste classi di equivalenza ibride. Ad esempio, se ad un qualche livello avessimo neutralizzato l'opposizione di sonorità per le occlusive quando intervocaliche, richiamandoci alla fenomenologia della lenizione, avremmo potuto recuperare tramite la stessa interrogazione due varianti quali /četrɨǫ̨lo/ e /čedrɨǫ̨lo/; allo stesso tempo, avremmo però assimilato forme distinte quali /brǫ́to/ 'burrone' e /brǫ́do/. Altro tipo di fenomeni escluso dalle classi di equivalenza definite ad oggi riguarda la variazione vocalica (ad oggi trattata solo al livello 3, per quanto concerne la variazione morfologica): di nuovo, mentre avremmo a ragione assimilato /miľľáččo/ e /miľľéččo/, avremmo anche messo insieme voci quali /álba/ ed /ę́rba/. Casi come quelli appena riportati ci hanno indotto a ridurre il numero di fenomeni neutralizzati tramite le classi di equivalenza. Ciò implica anche che le classi di equivalenza definite per i diversi livelli non possono considerarsi esaustive, ma semplicemente il risultato di un compromesso tra la produttività della ricerca e la presenza di rumore nei dati forniti in risposta.

Una possibile soluzione al problema di variazioni allofoniche che hanno come controparte opposizioni fonologiche starebbe nella specializzazione geografica delle classi di equivalenza, o sottoinsiemi di esse: vale a dire che l'astrazione da un certo fenomeno potrebbe essere ristretta ad una data sub-area toscana, quella cioè in cui tale fenomeno è il risultato di variazione allofonica. Una strategia di questo tipo, che nel caso del toscano rappresenta soltanto una possibile opzione, nel caso di imprese che coprono una più vasta area dialettale, caratterizzata da una maggiore variabilità, questa condizione diventa un requisito fondamentale per la definizione di una procedura di accesso intelligente ai dati in trascrizione fonetica.

7 Un esempio

Cerchiamo di illustrare con un esempio reale la progressione della produttività della ricerca secondo la procedura illustrata in queste pagine. In particolare, vediamone l'effetto nella pratica attraverso il contronto dei risultati di interrogazioni di livello diverso.

7.1 Interrogazione di Livello 0

Il Livello 0 è quello in cui la corrispondenza tra l'oggetto della ricerca ed il suo risultato deve essere esatta. Supponiamo di voler recuperare le attestazioni della forma /skʲaččáta/ all'interno della banca dati ALT; si otterranno solo le attestazioni che corrispondono esattamente alla richiesta, cioè una realizzazione fonetica con sibilante dentale, occlusiva velare sorda e occlusiva dentale sorda. A questo livello la forma recuperata è solo una, quella di partenza, per un totale di trenta attestazioni all'interno dell'intera BD-ALT.

7.2 Interrogazione di Livello 1

Se formuliamo la stessa richiesta di /skʲaččáta/ tramite un'interrogazione di Livello 1, basata cioè su corrispondenze deboli stabilite sulla base delle classi di equivalenza definite per questo livello, le forme diverse recuperate sono due: /skʲaččáta/, come al Livello 0, ma anche /skʲaččaŧa/ con la fricativa dentale in sillaba finale. A questo livello di accesso ai dati, il risultato della ricerca è quasi raddoppiato, ammontando a cinquantacinque occorrenze.

7.3 Interrogazione di Livello 2

Passando all'interrogazione di Livello 2, si constata uno scarto notevole nella produttività della ricerca. Infatti, a partire dalla stessa richiesta - /skʲaččáta/ - le forme diverse recuperate sono molte, tra cui: /skʲaččáta/, /skʲaččáŧa/, /stʲaččáta/, /stʲaččáŧa/, /sčaččáta/, /sčaččáŧa/, /sťaččáta/, /sťaččáŧa/, /skʲačáta/, etc. Ovvero, le forme recuperate a questo livello di accesso presentano diversi gradi di palatalizzazione del nesso /kʲ/ e della /s/ preconsonantica, combinati tra di loro e con la realizzazione fricativa dell'occlusiva dentale, nonché altre variazioni quali la realizzazione forte o debole dell'affricata palatale. Nel caso specifico, si raggiungono trecentonove attestazioni, ampliando di ben sei volte il risultato ottenuto tramite l'interrogazione di Livello 1, e decuplicando quello ottenuto al Livello 0. Inoltre, con la ricostruzione della vocale finale, il risultato include anche locuzioni quali /skʲaččát alla fʲorentína/, con elisione prodotta in fonosintassi.

E' importante notare che lo stesso risultato si sarebbe ottenuto a partire da qualsiasi altra delle forme recuperate.

7.4 Interrogazione di Livello 3

Procedendo infine all'ultimo livello di accesso, quello che astrae da variazioni di natura morfologica, la mole dei risultati della ricerca effettuata a partire da /skʲaččáta/ aumenta ulteriormente, anche se non in modo significativo come nel caso precedente. Ma questo deriva principalmente dalla formulazione della domanda di cui questa

forma costituisce tipicamente risposta, che richiedeva la parola al singolare. A questo livello, infatti, vengono recuperate attestazioni della foma al plurale, nonché il sintagma /páne skɹaчча́to/ dove addirittura la categoria morfosintattica della forma ricercata è diversa (si tratta di un aggettivo). E vi sono perciò casi in cui l'interrogazione di Livello 3 è maggiormente produttiva: ad esempio il risultato della ricerca di /páne/ contiene anche attestazioni al plurale e casi di caduta della vocale finale (cioè /pán/ lunigianese e garfagnino).

In ogni caso una valutazione accurata della produttività di interrogazioni effettuate a questo livello è ancora prematura in quanto, ripetiamo, il Livello 3 deve essere in gran parte ancora elaborato. I primi risultati ottenuti, comunque, si presentano come promettenti.

8 Conclusioni

In questo articolo, viene prefigurata una possibile risposta ad un problema cruciale dell'accesso tramite procedure informatiche a dati in trascrizione fonetica non tipizzati. La procedura illustrata è stata concepita e realizzata per l'accesso ai dati dell'ALT, ma è comunque estensibile e specializzabile – con i dovuti correttivi – per il recupero di materiali dialettali di altre aree.

Con le modalità di accesso messe in atto, la ricerca può astrarre da tratti specifici della realizzazione fonetica del dato attestato; inoltre, la possibilità che l'interrogazione avvenga secondo livelli differenziati di astrazione incrementale, che coprono variazioni fonetiche e morfologiche, 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.

Quanto raggiunto potrà anche essere utilmente sfruttato come punto di partenza del processo di tipizzazione manuale dei materiali contenuti nella BD-ALT; infatti, tramite questa procedura è possibile recuperare, in modo automatico, materiale sparso altrimenti difficilmente reperibile. In questo modo, la tipizzazione del dato dialettale di un atlante linguistico può andare al di là del ristretto orizzonte della singola carta linguistica e spaziare attraverso il vasto panorama dell'intero corpus dei materiali.

Inoltre, la costruzione e l'impiego di classi di equivalenza di fonie e morfi ai fini dell'interrogazione ci ha offerto l'opportunità di una valutazione, anche se indiretta, della copertura e validità dell'analisi dei dialetti toscani supportata dall'esperienza dell'ALT.

RIFERIMENTI BIBLIOGRAFICI

AGOSTINIANI, L. (1985), «Dalle inchieste all'Atlante: per la costituzione di una banca dati». In *Atlante Lessicale Toscano,* AA.VV., Olschki Editore.

AGOSTINIANI, L., MONTEMAGNI, S., POGGI SALANI, T. (1989), «Atlante Lessicale Toscano: il lavoro di preedizione e la costituzione di una banca dati». In *Atti del XV Convegno di Studi Dialettali Italiani. "Gli atlanti regionali: aspetti metodologici, linguistici e etnografici"*, Pisa, Pacini, pp.1-7.

AGOSTINIANI, L., MONTEMAGNI, S., PAOLI, M., PICCHI, E., POGGI SALANI, T. (1992), «La costruzione di un sistema integrato per il trattamento dei dati dell'Atlante

Lessicale Toscano; esperienze, problemi, prospettive». In *Atti del Congresso Internazionale del Centro di Studi Filologici e Linguistici Siciliani. "Atlanti linguistici italiani e romanzi"*, a cura di G. Ruffino, Palermo, pp. 357-393.

GIACOMELLI, G. (1987/1988), «Storia, criteri, metodi, prospettive dell'Atlante Lessicale Toscano». In *Quaderni dell'Atlante Lessicale Toscano,* n. 5/6, pp. 7-25.

MONTEMAGNI, S. (1986), *Un esperimento di Dialettologia Computazionale: elaborazione ed analisi dellle risposte ad un gruppo di domande dell'ALT*. Tesi inedita, Facoltà di Lettere dell'Università di Firenze, a.a. 1985-1986, Relatore G. Giacomelli.

MONTEMAGNI, S., PAOLI, M. (1989/1990), «Dalla parola al bit (e ritorno): percorsi dall'inchiesta sul campo alla banca dati dell'ALT». In *Quaderni dell'Atlante Lessicale Toscano,* n., 7/8, pp. 7-52.

PICCHI, E. (1991), «DBT: a textual Database system». In *Linguistica Computazionale. Computational Lexicology and Lexicography,* VII, 2, Pisa, Giardini Editore, pp. 77-105.

PICCHI, E. (in corso di stampa), *Linguistica Computazionale: Analisi testuale e lessicale*, Bulzoni Editore.

PICCHI, E., MONTEMAGNI, S., BIAGINI, L. (in corso di stampa), «DBT-ALT: a System for Storying and Querying the Data of the Atlante Lessicale Toscano (ALT)». Negli *Atti del 2nd International Congress of Dialectologists and Geolinguists*, Amsterdam, 28/7-1/8-1997.