ANALISI LINGUISTICO-COMPUTAZIONALI DEL *CORPUS* DIALETTALE DELL'*ATLANTE LESSICALE TOSCANO*. PRIMI RISULTATI SUL RAPPORTO TOSCANO-ITALIANO

Simonetta Montemagni

1. Introduzione

«Un discorso sul territorio sfumatissimo in Toscana tra dialetto e lingua e sull'italiano della regione può forse apparire marginale in un lavoro che si pone come scopo primo di chiarire l'organizzazione areale del patrimonio del lessico toscano»: così Teresa Poggi Salani apre il suo contributo dal titolo *Dialetto e lingua a confronto*[1] parlando dell'impresa dell'Atlante Lessicale Toscano, a quei tempi ancora ai suoi albori. La studiosa continua notando che «in Italia in nessun'altra terra come in questa si scopre poi, nell'insieme, così frequentemente che ciò che è dialetto – qui richiesto proprio perché tale – è anche italiano». A impresa ultimata, l'opera avrebbe dovuto accogliere – secondo le aspettative – anche «tanto» italiano di Toscana.

L'Atlante Lessicale Toscano (ALT) è stato pubblicato in versione elettronica su CD-Rom nel 2000[2], e recentemente, nel 2007, è stato riproposto in versione rinnovata e arricchita come risorsa dialettale interrogabile *on-line* la cui denominazione è ALT-Web (ovvero, l'Atlante Lessicale Toscano in rete)[3]. Oggi è dunque possibile provare a quantificare l'italiano di Toscana all'interno dell'ingente *corpus* dei materiali dialettali raccolti con le inchieste sul campo svolte tra

---

[1] T. Poggi Salani, *Dialetto e lingua a confronto,* in G. Giacomelli *et al.*, *Atlante Lessicale Toscano - Note al questionario*, Firenze, Facoltà di Lettere e Filosofia, 1978, pp. 51-65.

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

[3] ALT-Web può essere raggiunto in rete all'indirizzo http://serverdbt.ilc.cnr.it/altweb/. Per una descrizione della risorsa dialettale *on-line* cfr. S. Montemagni - M. Paoli - E. Picchi, *ALT Web: l'Atlante Lessicale Toscano in rete*, in F. Bruni - C. Marcato (a cura di), *Lessicografia dialettale: ricordando Paolo Zolli*, Atti del Convegno di Studi (Venezia, 9-11 dicembre 2004), Roma-Padova, Editrice Antenore, 2006, t. I, pp. 209-241.

il 1974 e il 1986 sulla base di un questionario di 745 domande somministrato a un campione di 2193 informatori differenziati per età, sesso e *status* socio-culturale. In questo contributo, si forniscono i primi risultati di analisi quantitative condotte sull'intero *corpus* ALT affiancate da elaborazioni linguistico-computazionali dei dati dialettali con lo scopo di fornire evidenza utile a una riflessione sul rapporto tra toscano e italiano. L'obiettivo di queste analisi è duplice: da un lato di cominciare a delineare, in materia lessicale, le modalità di sovrapposizione tra dialetto e italiano in Toscana, dall'altro di verificare se dietro al variare dello spessore di questa sovrapposizione vi sia un coerente disegno areale.

Per condurre questo tipo di analisi sono necessari due ingredienti basilari: un *corpus* di testimonianze dialettali – rappresentato dai materiali dell'ALT – e un "metro" italiano da usarsi per misurare le corrispondenze italiano-toscano. Come "metro" è stata utilizzata la lista delle parole italiane di riferimento, originariamente stilata da Matilde Paoli su base lessicografica e successivamente rivista con cura da Annalisa Nesi e Teresa Poggi Salani, in cui per ogni domanda onomasiologica dell'ALT viene riportata la "risposta" italiana. Per quanto tale lista fosse stata originariamente stilata per fare fronte ad aspetti di recupero e accesso ai materiali dialettali da parte dell'utente di ALT-Web che non conoscesse già il progetto e il questionario ALT, si presta a mio avviso a essere assunta a metro di riferimento della norma italiana per quanto riguarda i concetti indagati dall'impresa. Sul versante dialettale, le analisi si sono focalizzate sulle risposte alle domande onomasiologiche del questionario (in totale, 461). Dato che in ALT-Web a ogni attestazione dialettale sono associati diversi livelli di rappresentazione, si è reso necessario selezionare il livello più adeguato in relazione alle finalità della ricerca[4]. Come base di questo studio incentrato sul lessico si è optato per il livello delle rappresentazioni normalizzate, ovvero il livello concepito come primo passo di astrazione rispetto a tratti specifici della realizzazione fonetica del dato come fedelmente registrati dalla trascrizione fonetica e riproposti, ove possibile, dalla sua traslitterazione ortografica. A questo livello di rappresentazione sono state neutralizzate variazioni fonetiche produttive sul territorio toscano: ad esempio, *stiacciàta* e *schiacciàta* sono state ricondotte alla medesima forma normalizzata, lo

[4] Per una illustrazione dettagliata dell'articolato schema di rappresentazione dei materiali dialettali dell'ALT si rinvia a Montemagni *et al.*, *ALT Web*, cit., pp. 223-240.

stesso vale per *vìholo* e *vìcolo*, *schiacciàha*, *schiacciàda* e *schiacciàta*, *fidanzàdo* 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, ad oggi non ancora realizzato: *schiacciàta* e *schiacciàte* rimangono dunque attestazioni distinte così come *schiàccia*, *schiaccétta* e *schiaccìna*. E rimangono distinte anche 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.

Il "metro" italiano è stato proiettato sul *corpus* dei materiali normalizzati dell'ALT per indagare diversi aspetti del rapporto tra i dialetti toscani e l'italiano: nella sezione 2, sono riportati i risultati di una prima indagine sulla sovrapposizione tra toscano e italiano nelle risposte alle domande onomasiologiche dell'ALT; la sezione 3 affronta la questione della copertura geografica delle attestazioni dialettali in relazione alla norma italiana; infine, nella sezione 4 si è cercato di quantificare la "distanza" linguistica che separa le varietà dialettali toscane testimoniate nell'ALT e l'italiano.

2. Toscano e italiano nell'analisi quantitativa delle risposte alle domande dell'ALT

Partiamo in questa breve indagine sul rapporto tra toscano e italiano alla luce dell'evidenza dell'ALT con una domanda alquanto semplice: qual è la proporzione di risposte italiane nell'insieme delle risposte alle domande onomasiologiche dell'ALT? La Tabella 1 fornisce una prima risposta a questo interrogativo:

	TUTTE LE RISPOSTE	RISPOSTE MAGGIORITARIE	RISPOSTE FORNITE DA INFORMATORI		
			ANZIANI	DI MEZZA ETÀ	GIOVANI
RISPOSTE COINCIDENTI CON LA NORMA ITALIANA	19,14%	23,34%	20,30%	21,23%	24,72%
RISPOSTE DIALETTALI	80,86%	76,66%	79,70%	78,77%	75,28%

TOTALE RISPOSTE	238.771	142.610	171.913	183.936	156.185

Tabella 1.

Si noti che i dati contenuti nella Tabella vanno interpretati tenendo in considerazione i criteri che hanno guidato la compilazione del questionario ALT, finalizzato a cogliere lo specifico toscano nella sua complessa articolazione, e che hanno comportato l'esclusione programmatica di «parole che coincidevano dappertutto o quasi con l'italiano»[5]. Nella Tabella le risposte alle domande onomasiologiche dell'ALT sono state ripartite in due classi: quelle coincidenti con la norma italiana, e quelle non coincidenti (qualificate come dialettali). Questa classificazione è stata applicata all'insieme di tutte le risposte fornite dagli informatori – ovvero quelle testimoniate sia in competenza attiva sia in competenza passiva – così come a sottoinsiemi di esse definiti sulla base di criteri di rappresentatività della risposta all'interno della singola località indagata e generazionali: nella seconda colonna (risposte maggioritarie), la classificazione italiano-dialettale è applicata al sottoinsieme delle risposte fornite in competenza attiva da un numero di informatori pari o superiore al 50% del gruppo intervistato in una specifica località; infine, nelle ultime tre colonne ci si è focalizzati sulle risposte fornite – nuovamente in competenza attiva – da tre diverse fasce generazionali, anziani, di mezza età e giovani. Si nota che la sovrapposizione italiano-toscano è dell'ordine del 19% nel caso dell'insieme completo delle risposte, e significativamente più alta nel caso delle risposte che vanno da paritetiche a maggioritarie (23%). Come potevamo aspettarci, la proporzione italiano-toscano varia in modo inversamente proporzionale all'età: se nel gruppo dei giovani le risposte che coincidono con la norma italiana sono quasi il 25%, tale percentuale scende progressivamente con l'aumentare dell'età, fino ad arrivare al 20% di risposte italiane nel caso degli informatori anziani.

Dalla Tabella 1 emerge che la proporzione di risposte italiane nell'insieme delle risposte alle domande onomasiologiche dell'ALT oscilla, a seconda dei criteri di selezione delle risposte, dal 19% al 25%. Un interrogativo legittimo a questo punto è se nella peculiare situazione linguistica toscana questa proporzione sia effettivamen-

[5] G. Giacomelli, *Come e perchè il questionario*, in G. Giacomelli *et al.*, *Atlante lessicale toscano - Note al questionario*, Firenze Facoltà di Lettere e Filosofia, 1978, pp. 19-26.

te veritiera. Sicuramente, i criteri che hanno guidato la definizione del questionario ALT incentrato sui punti chiave in cui si realizza la differenziazione tra le aree toscane, possono fornire una prima risposta a questo interrogativo, ma forse non sono sufficienti a giustificare la proporzione osservata. Al fine di verificare ciò, si è andati a studiare da un punto di vista quantitativo la tipologia di risposte emerse dalle indagini dell'ALT in relazione alle singole domande del questionario. È emerso che il numero di risposte normalizzate diverse per domanda presenta un ampio spettro di variazione sul territorio toscano indagato, oscillando tra 6 e 421. Si va da domande che hanno raccolto complessivamente 6 risposte normalizzate diverse a domande particolarmente produttive, in relazione alle quali sono state raccolte anche più di 400 diverse risposte. La distribuzione delle risposte per domanda è sintetizzata nel Grafico 1:

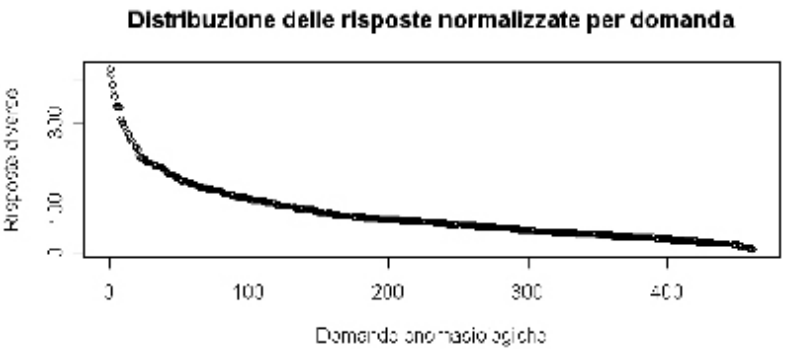

Grafico 1.

All'alta "produttività" di molte domande onomasiologiche dell'ALT sembrano contribuire fattori di varia natura: si va da domande particolarmente "stimolanti" per gli informatori, i cui risultati includono anche creazioni estemporanee molto probabilmente non lessicalizzate, a domande in cui il livello di rappresentazione considerato non consente ad oggi generalizzazioni affidabili. Un esempio del primo caso è costituito dalla domanda 434bis che indagava le denominazioni di 'stupido': delle 372 diverse risposte raccolte, 122 sono hapax che includono usi metaforici del tipo *cetriolo* e *carciofo* così come diverse forme di derivazione come *scemaccio*, *scemalone*, *scemarotto*, *scemarano*, *scemarlotto*, *scemarlaccio*, *scemarello* e *scemerello*, oppure unità lessicali polirematiche più o meno cristallizzate del tipo *mezzo scemo*, *mezzo spostato*, *puro locco*, ecc. Il se-

condo caso è rappresentato da domande il cui focus è costituito dal significato di un verbo: a questa classe appartiene la domanda in assoluto più produttiva dell'ALT (dom. 512) che indagava le diverse denominazioni del 'picchiare forte' e che ha ricevuto ben 421 risposte diverse, tra le quali si riscontrano sia forme flessionali diverse della stessa base verbale (ad es. *maltire*, *maltisce*, *maltito* oppure *ha bordato*, *bordato*, *bordare*) sia varianti formali (del tipo *maltupiàre*, *maltupire*, *matupire*) sia creazioni non si sa fino a che punto lessicalizzate del tipo *bussare forte*, *cardare sodo*, *cardare fino fino*, *caricare di botte*, ecc. In entrambi i casi, al livello di rappresentazione considerato esse si presentano tutte come risposte diverse.

Avendo constatato la peculiare distribuzione delle risposte per domanda nel *corpus* ALT, si è dunque deciso di ricalcolare le percentuali di risposte italiane in relazione a una selezione di domande onomasiologiche che presentavano un ambito di variabilità controllato: in questo modo il "rumore" introdotto da domande particolarmente produttive avrebbe dovuto essere ridimensionato in modo significativo. Sono state selezionate per questo esperimento 165 domande onomasiologiche il cui spettro di variabilità è compreso tra 6 e 50 risposte normalizzate diverse. I risultati ottenuti con questa selezione di domande sono riportati nella Tabella 2:

	TUTTE LE RISPOSTE	RISPOSTE MAGGIORITARIE	RISPOSTE FORNITE DA INFORMATORI		
			ANZIANI	DI MEZZA ETÀ	GIOVANI
RISPOSTE COINCIDENTI CON LA NORMA ITALIANA	36,10%	41,04%	37,22%	38,94%	44,41%
RISPOSTE DIALETTALI	63,90%	58,96%	62,78%	61,06%	55,59%
TOTALE RISPOSTE	63.306	43.957	52.391	55.637	48.539

Tabella 2.

dove si può notare che la proporzione di risposte italiane rispetto a quelle dialettali è quasi raddoppiata, passando dal 19% al 36% nel caso dell'insieme completo delle risposte, al 41% nel caso delle risposte paritetiche e maggioritarie, fino al 37%, 39% e 44% rispettivamente nelle tre fasce generazionali di informatori considerate. L'andamento delle proporzioni toscano-italiano nelle diverse selezioni di risposte appare invece del tutto analogo a quanto osservato in precedenza.

3. Toscano e italiano in rapporto alla copertura geografica delle attestazioni dialettali dell'ALT

Questo secondo esperimento prende le mosse da un articolo che Teresa Poggi Salani ha scritto insieme a Gabriella Giacomelli, ideatrice e direttrice dell'ALT, dal titolo *Parole Toscane*[6]. In questo articolo, le autrici dichiarano di non volersi occupare di «nuclei di risposte che implicassero una eccessiva frammentazione lessicale tra le province toscane» e di essere piuttosto interessate a «parole rappresentative di una larga fascia della regione, incluso il capoluogo». Nel loro studio, hanno condotto un'indagine sincronica dal dialetto alla lingua in terra toscana sulla base di una selezione di otto parole di ambito familiare due delle quali (*acquaio* e *sciocco*) coprono tutta la regione e le rimanenti sei presentano aree di espansione sempre diverse ma che in ogni caso coprono una vasta area della regione.

Come primo passo, si sono classificate le risposte alle domande onomasiologiche dell'ALT in relazione alla loro copertura geografica: il risultato di questa classificazione è sintetizzato nel Grafico 2. Ai fini di questa analisi abbiamo ritenuto opportuno focalizzarci sulle testimonianze fornite in competenza attiva e per le quali l'attestazione all'interno del punto avesse un certo spessore. Sono state dunque considerate solo le risposte da paritetiche a maggioritarie, per un totale di 20.218 risposte:

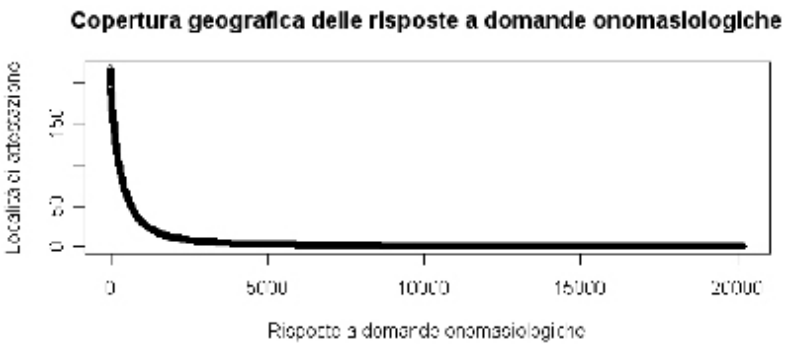

Grafico 2.

[6] G. Giacomelli - T. Poggi Salani, *Parole toscane,* «Quaderni dell'Atlante Lessicale Toscano», 2/3 (1984/85), pp. 123-229.

di queste, solo 2.261 (ovvero poco più del 10%) sono attestate in più di 9 località diverse, e 1.188 sono quelle che hanno una copertura pari o superiore a 25 località. Concentrandoci sulle 1.188 risposte attestate in più di 24 località, abbiamo cercato di vedere quante di esse fossero rappresentate da parole italiane: il risultato di questa analisi è riportato nella Tabella 3, dove per diversi gradi di copertura dell'attestazione dialettale è riportato il numero di risposte paritetiche o maggioritarie diverse caratterizzate dalla copertura specificata nell'intestazione della colonna. Di queste viene fornita la percentuale di risposte che coincidono con la norma italiana e di quelle che invece si differenziano da essa.

	COPERTURA GEOGRAFICA							
	≥ 200	≥ 175	≥ 150	≥ 125	≥ 100	≥ 75	≥ 50	≥ 25
NUMERO TOTALE DI CASI	12	48	109	182	283	418	674	1188
RISPOSTE DIALETTALI	25,00%	20,83%	33,94%	39,01%	45,58%	50,96%	62,76%	73,91%
RISPOSTE COINCIDENTI CON LA NORMA ITALIANA	75,00%	79,17%	66,06%	60,99%	54,42%	49,04%	37,24%	26,09%

Tabella 3.

Si nota che nel caso di attestazioni con una copertura pari o maggiore a 200 località, il 75% coincide con la norma italiana; tale proporzione decresce gradualmente al decrescere della copertuta geografica, fino ad arrivare a una proporzione del 26% nel caso dell'insieme delle risposte testimoniate in più di 24 località diverse. Da questa tabella appare esistere una correlazione positiva tra la copertura geografica delle attestazioni dialettali e la loro italianità: maggiore è la copertura di un'attestazione, maggiori sono le probabilità che questa coincida con la norma italiana. I casi di mancata coincidenza con la norma italiana tra le attestazioni dialettali con una copertura pari o superiore a 175 località sono listati di seguito: *fifa* (dom. 472, 203 loc., ital. *paura*), *sciocco* (dom. 486a, 202 loc., ital. *insipido*), *levare* (dom. 519, 200 loc., ital. *togliere*), *pelato* (dom. 392bis, 196 loc., ital. *calvo*), *pina* (dom. 111, 191 loc., ital. *pigna*), *popone* (dom. 104, 191 loc., ital. *melone*), *pigiare* (dom. 530bis, 187 loc., ital. *comprimere*), *frana* (dom. 34, 187 loc., ital. *smottamento*), *ciccione*, dom. 432a, 185 loc., ital. *grassone*), *becco* (dom. 172, 176 loc., ital. *caprone*).

4. La distanza lessicale tra le varietà dialettali toscane e l'italiano

Nelle precedenti sezioni si è cercato di quantificare la presenza di risposte italiane all'interno dell'insieme delle attestazioni dialettali raccolte sul territorio toscano, sempre facendo astrazione dalla loro distribuzione geografica. In questa sezione ci poniamo un obiettivo più ambizioso, ovvero quello di calcolare la distanza linguistica tra le varietà dialettali toscane emergenti da elaborazioni dialettometriche dei materiali dell'ALT e l'italiano. Gli interrogativi a cui intendiamo fornire una risposta, seppur preliminare, sono essenzialmente due: quali tra le varietà dialettali toscane attestate nell'ALT risultano maggiormente vicine all'italiano e se esista un disegno areale coerente nella risposta a questa domanda.

Diversi sono i metodi e le tecniche per il calcolo della distanza tra varietà linguistiche proposti nella letteratura dialettologico-computazionale. Si parte dall'idea pionieristica avanzata da Seguy[7] e successivamente da Goebl[8] che la similarità tra due varietà dialettali può essere rilevata attraverso la misurazione di concordanze e divergenze linguistiche tra i diversi punti di inchiesta: in particolare, la «prossimità lessicale» tra due località riflette la proporzione di risposte condivise, mentre la «distanza lessicale» deriva dalla proporzione di risposte che differiscono nelle due località. In questo tipo di misure la nozione di similarità è fondata su distinzioni di tipo categoriale: nel confronto di due varietà dialettali solo due tipi di distanze sono riconosciute, una distanza 0 nel caso di uguaglianza di risposta, una distanza 1 nel caso di divergenza. A partire dalla fine degli anni '80, metodi alternativi sono stati proposti per il confronto computazionale di varietà dialettali dal punto di vista fonetico, operanti direttamente sul *corpus* dei materiali dialettali in trascrizione fonetica raccolti sul campo. Tra questi si annovera la distanza di Levenshtein[9] per misurare la similarità tra due stringhe di caratteri,

---

[7] J. Séguy, *La relation entre la distance spatiale et la distance lexicale*, «Revue de Linguistique Romane», 35 (1971), pp. 335-357.

[8] H. Goebl, *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, Tübingen, Max Niemeyer, 1984.

[9] J.B. Kruskal, *An overview of sequence comparison*, in D. Sanko - J. Kruskal (a cura di), *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, Stanford, CSLI, 2nd edition, 1999, pp. 1-44. Prima edizione apparsa nel 1983.

utilizzata in ambito dialettologico a partire da Kessler[10] e ulteriormente raffinata nell'ambito di studi successivi[11].

L'intuizione sottostante alla misura della distanza di Levenshtein consiste nell'immaginare di riscrivere una stringa di caratteri in un'altra. Il processo di riscrittura è condotto attraverso un insieme di operazioni di base a ciascuna delle quali è associato un costo. Le operazioni possibili sono: sostituzione di un carattere con un altro; cancellazione di un carattere; inserimento di un carattere. Si definisce «distanza di Levenshtein» tra due stringhe il numero minimo di operazioni di riscrittura che occorre effettuare per trasformare una stringa di partenza in una stringa finale.

Per quanto il ricorso alla misura di Levenshtein sia più che motivato per quanto riguarda le distanze fonetiche, non appare così scontato per la misurazione di distanze lessicali dove forse potrebbe bastare una nozione binaria di distanza lessicale, definita sulla base della proporzione delle risposte condivise (siano esse in forma tipizzata o meno) in relazione a un insieme di domande in due località. Tuttavia, si dà spesso il caso che le diverse risposte attestate dagli informatori in relazione alla stessa domanda siano forme diverse della stessa base lessicale e includano varianti sia flessionali sia derivazionali. Con l'adozione di una misura binaria della distanza lessicale, attestazioni dialettali diverse riconducibili alla stessa base lessicale appaiono del tutto irrelate. Una possibile soluzione a questo problema, perseguita da Nerbonne e Kleiweg[12] nel loro lavoro sul «Linguistic Atlas of the Middle and South Atlantic States» (LAMSAS) sulla base dei risultati promettenti ottenuti con questa misura per quanto riguarda le distanze fonetiche, consiste nel ricorrere alla distanza di Levenshtein anche per misurare le distanze lessicali delle denominazioni associate a uno stesso concetto. In questo modo varianti morfologiche riconducibili alla stessa base non vengono più trattate come attestazioni diverse e del tutto indipendenti ma

---

[10] B. Kessler, *Computational Dialectology in Irish Gaelic*, in *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), Dublin, s.e., 1995, pp. 60-67.

[11] J. Nerbonne - W. Heeringa - P. Kleiweg, *Edit Distance and Dialect Proximity*, in D. Sanko - J. Kruskal (a cura di), *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, Stanford, CSLI Press, pp.v-xv; W. Heeringa, *Computational Comparison and Classification of Dialects*, Ph.D. thesis, University of Gröningen, 2004.

[12] J. Nerbonne - P. Kleiweg, *Lexical Distance in LAMSAS*, in J. Nerbonne - W. Kretzschmar (a cura di), «*Computers and the Humanities*» *(Special issue on Computational Methods in Dialectometry)*, 37(3), 2003, pp. 339-357.

come risposte correlate, nel senso che presentano una qualche similarità derivante dal fatto di condividere la stessa base lessicale. Si è ritenuta opportuna una scelta analoga anche nel caso dell'ALT: per quanto si operi su materiali dialettali previamente normalizzati, abbiamo visto che il livello di normalizzazione considerato non fa astrazione né da variazioni flessionali e derivazionali, né da variazioni fonetiche non più vitali sul territorio toscano[13] .

Il ricorso alla misura di Levenshtein non è circoscritto a studi della variazione dialettale. Tale misura stata è stata infatti anche utilizzata per calcolare la distanza linguistica tra varietà dialettali quali emergono dai materiali di un atlante linguistico e una norma di riferimento (lingua standard): studi condotti seguendo questo approccio includono analisi del contatto linguistico tra tedesco e olandese[14], e tra il bulgaro e le lingue dei paesi confinanti (macedone, serbo, rumeno e turco)[15]. Nell'esperimento illustrato in questa sezione, la misura di Levenshtein è stata usata per calcolare la distanza linguistica tra le varietà dialettali toscane e l'italiano: sul versante dialettale è stato considerato sia l'insieme di tutte le risposte alle domande onomasiologiche dell'ALT sia le diverse selezioni definite sulla base di criteri di rappresentatività della risposta e generazionali. Per ciascuna località indagata dall'ALT, è stata calcolata la distanza linguistica dall'italiano: tale distanza è costituita dalla media delle distanze ottenute per il campione di attestazioni dialettali selezionato. Le distanze ottenute rispetto all'italiano sono state proiettate su carta[16]: la Figura 1 riporta il risultato ottenuto con l'intero *corpus* di

---

[13] Per uno studio della variazione lessicale in Toscana condotto con tecniche dialettometriche basate sulla distanza di Levenshtein si rinvia a S. Montemagni, *Variazione fonetica e lessicale in Toscana: prime elaborazioni computazionali dei dati dell'Atlante Lessicale Toscano*, in corso di stampa negli Atti del XL Congresso di Studi Linguistica e Modelli Tecnologici di Ricerca - Facoltà di Lettere e Filosofia Università degli Studi del Piemonte Orientale - Vercelli 21-23 settembre 2006.

[14] W. Heeringa - J. Nerbonne - H. Niebaum - R. Nieuweboer - P. Kleiweg, *Dutch-German Contact in and around Bentheim*, in D. Gilbers - J. Nerbonne - J. Schaeken (a cura di), *Languages in Contact*. Volume 28 of Studies in Slavic and General Linguistics, Rodopi, Amsterdam and Atlanta GA, 2000, pp. 145-156.

[15] W. Heeringa - J. Nerbonne - P. Osenova, *Detecting Contact Effects in Pronunciation*, sottoposto per pubblicazione a M. Norde, B. de Jonge, C. Hasselblatt (a cura di), *Language Contact in Times of Globalization*, Amsterdam, Benjamins, disponibile all'indirizzo http://www.let.rug.nl/~nerbonne/papers/detecting-contact-pronunciation.pdf.

[16] Per il calcolo delle distanze linguistiche e per la proiezione dei risultati su carta si è usato il *software* Rug/L04 per analisi dialettometriche e cartografazio-

risposte a domande onomasiologiche, dove i poligoni caratterizzati da toni più scuri di grigio contrassegnano le varietà dialettali più vicine all'italiano, e quelli più chiari quelle più distanti.

Da questa rappresentazione tridimensionale della sovrapposizione italiano-toscano, si nota che l'area di maggiore sovrapposizione coincide con l'area fiorentina, con propaggini in diverse direzioni, verso l'area pratese-pistoiese, verso l'area pisana e quella aretina. Se ne può dedurre che la varietà dialettale toscana più vicina all'italiano è costituita da quella fiorentina. La stessa mappa ottenuta in relazione a diverse selezioni del *corpus* dialettale fornisce risultati simili ma con differenze significative, tutte riguardanti l'area di espansione rispetto al nucleo originario incentrato sulla provincia di Firenze: la più importante, a mio avviso, riguarda la mappa ottenuta considerando solo le attestazioni da paritetiche a maggioritarie dove l'area di maggiore sovrapposizione si estende verso il mare fino a coprire il litorale pisano e gran parte della provincia di Livorno. Per quanto riguarda le province di Siena e Grosseto, l'andamento delle distanze toscano-italiano non sembra presentare una distribuzione areale ben definita.

Ciò che emerge dall'analisi di queste carte è in linea con quanto Teresa Poggi Salani e Gabriella Giacomelli ipotizzano nel loro studio *Parole toscane* sul rapporto tra la varietà fiorentina e le altre varietà dialettali toscane che viene tratteggiato come segue: Firenze rappresenta «il reale centro propulsore di innovazioni di fronte a cui le altre aree si mostrano diversamente recettive»[17]. La diversa recettività delle aree toscane rispetto a termini irradiati da Firenze viene descritta nei seguenti termini: «la provincia di Massa, dialettalmente non toscana, è naturalmente quella che si oppone di più alla 'sopraffazione' del fiorentino»; «l'altro centro di opposizione ci risulta costituito da Arezzo»; «molto recettiva appare la zona pistoiese, almeno quella meridionale, così come quelle di Pisa e Livorno»[18]. In relazione alle province di Siena e Grosseto non sono state rilevate tendenze ben definite. Queste considerazioni, formulate ormai più di due decenni fa, avrebbero potuto essere scritte in commento alla carta di Figura 1, dove le aree di maggiore e minore sovrapposizione tra dialetto e italiano sembrano coincidere in larga parte con quanto

---

ne dei risultati messo a disposizione della comunità scientifica da Peter Kleiweg e liberamente scaricabile dall'indirizzo http://www.let.rug.nl/~kleiweg/L04/. Si ringrazia Peter Kleiweg per l'assistenza prestata per la produzione della carta in Figura 1.

[17] Giacomelli - Poggi Salani, *Parole toscane*, cit., p. 123.

[18] *Ivi*, p. 124.

osservato in relazione all'irradiarsi del fiorentino (che abbiamo visto coincidere in modo significativo con l'italiano) nelle diverse aree toscane. Le studiose affermano che «quello che diciamo è limitato, nel senso che non può essere assunto come una generalizzazione di tendenze toscane: crediamo tuttavia che nella sua solidità possa essere considerato un piccolissimo, ma valido punto di riferimento in questo genere di ricerche»: più di venti anni dopo, i risultati del loro studio, da loro prudentemente definiti come parziali, trovano conferma in analisi condotte sull'intero *corpus* dei materiali ALT con tecniche linguistico-computazionali. Sulla base dell'evidenza acquisita, possiamo concludere che hanno realmente identificato «una generalizzazione di tendenze toscane».

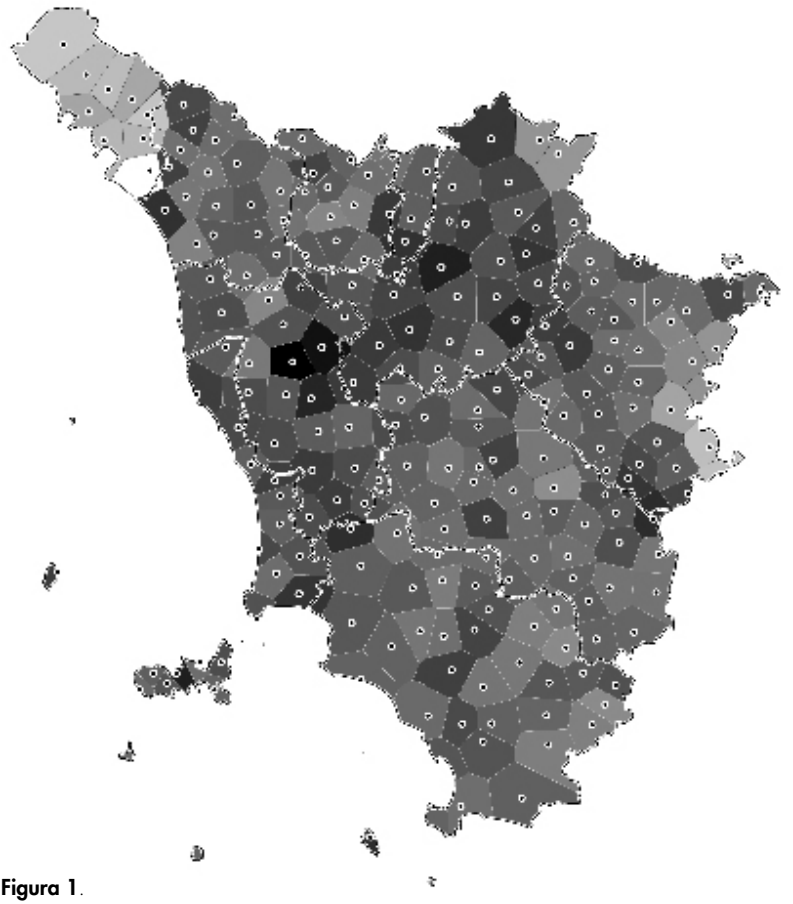

Figura 1.