Aree fonetiche e lessicali toscane a confronto

Prime elaborazioni computazionali dei dati dell'Atlante Lessicale Toscano

Simonetta Montemagni
ILC-CNR

Dialettologia e computer

1. il computer è usato per rendere più efficienti le procedure di ordinamento e ricerca, giocando un **ruolo di rafforzamento degli apparati metodologici tradizionali**
   - questo approccio porta alla creazione di risorse elettroniche quali:
     - atlanti linguistici
     - dizionari dialettali
     - corpora dialettali di lingua scritta o trascrizioni di parlato

2. lo strumento informatico orienta e ridefinisce le possibilità di elaborazione dei dati: il suo ruolo non è più circoscritto alla sfera tecnica ma ha un **impatto innovativo sul piano metodologico**
   - questo caso è tipicamente ricondotto a metodi e tecniche per la creazione di visioni “sintetiche” della variazione linguistica:
     - classificazione delle varietà dialettali
     - cartografazione della variazione linguistica

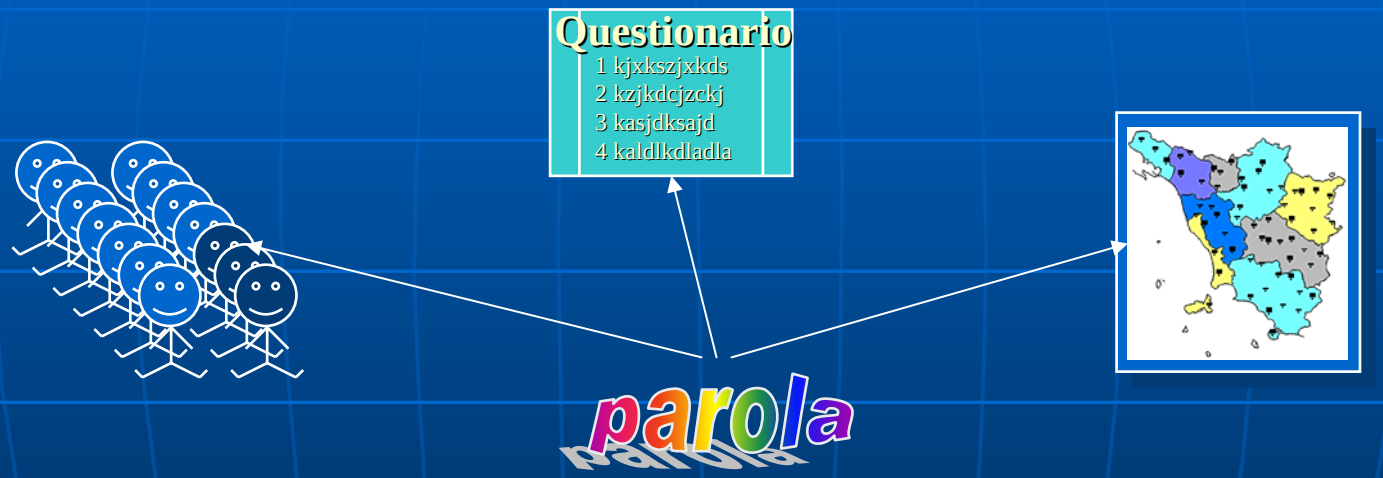
Approcci dialettometrici allo studio della variazione dialettale

- L'uso di tecniche statistiche per il calcolo della distanza linguistica proposto agli inizi degli anni '70 da Seguy (1971) e ulteriormente sviluppato da Goebl (1984)
- Misura della distanza dialettale basata esclusivamente su **distinzioni categoriali**
- Kessler (1995) segna una svolta in questo filone di studi, applicando misure di similarità tra stringhe di caratteri direttamente sul dato dialettale in trascrizione fonetica
  - differenza di pronuncia espressa come **differenza numerica**
  - **additività** delle singole differenze numeriche
- Nerbonne et al. (1999), Heeringa (2004) hanno esteso l'uso di questa tecnica a diversi tipi di rappresentazioni linguistiche
- Ad oggi, l'uso combinato di tecnologie linguistico-computazionali con tecniche di analisi statistica multivariata appare promettente nello studio della variazione linguistica di diverse lingue (olandese, inglese, irlandese, sardo, norvegese, bulgaro)

L’Atlante Lessicale Toscano (ALT)

(redazione: Giacomelli [direttore], Agostiniani, Bellucci, Giannelli, Montemagni, Nesi, Paoli, Picchi, Poggi Salani)

- L’ALT è un atlante linguistico in cui ogni dato lessicale è rapportato alle dimensioni **diatopica** e **diastratica**



- Il corpus dialettale dell’ALT contiene i risultati di interviste svolte in 224 località dellaToscana, con 2193 informatori sulla base di un questionario di 745 domande, per un totale di:
  - circa 350.000 schede che compattano le risposte alle domande del questionario
  - circa 30.000 schede contenenti materiali integrativi

Il modello di rappresentazione dei materiali dialettali dell'ALT

- Modello di rappresentazione multi-livello che rende possibili esplorazioni del corpus incentrate su una pronuncia specifica e che possono fare astrazione da dettagli della realizzazione fonetica
- A ogni attestazione dialettale sono associati diversi livelli di rappresentazione articolati come segue

1. trascrizione fonetica

- basata su una versione del sistema di trascrizione CDI specializzata per la codifica dei materiali toscani

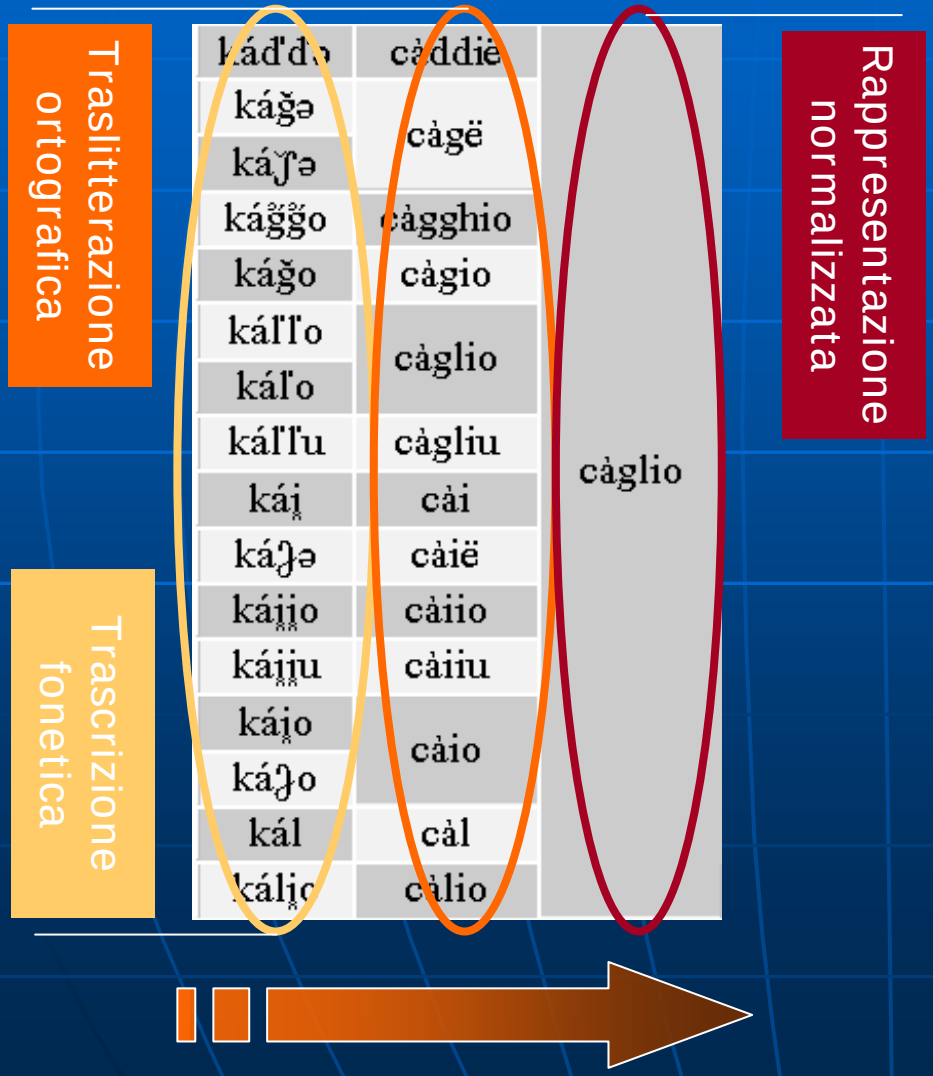
2. traslitterazione in ortografia italiana

- concepita come guida alla lettura e alla decodifica della forma in trascrizione fonetica per l'utente non addetto ai lavori
- si propone rendere conto della variabilità effettivamente rilevata con le inchieste sul campo (compatibilmente con le convenzioni ortografiche italiane)

3. rappresentazione normalizzata

- neutralizza tratti specifici della realizzazione fonetica del dato (tipicamente variazioni fonetiche produttive sul territorio toscano) senza fare astrazione da variazioni morfologiche

Il modello di rappresentazione dei materiali dialettali dell'ALT



- La definizione delle aree fonetiche e lessicali toscane sfrutta questo articolato schema di rappresentazione
  - trascrizione fonetica > aree fonetiche
  - rappresentazioni normalizzate > aree lessicali

Misura della distanza tra varietà dialettali

- basata sulla **Levenshtein distance** o **Minimum edit distance**
- misura la distanza tra due stringhe s_1 e s_2 in termini del "costo" della trasformazione di s_1 in s_2
- operazioni possibili:
  - sostituzione
  - cancellazione
  - inserimento
- diversi costi per diverse operazioni
- diverse sequenze di operazioni sono possibili per trasformare s_1 in s_2
  - **Levenshtein distance** = numero minimo di operazioni per passare da s_1 a s_2
- misura sensibile alla lunghezza delle stringhe coinvolte: normalizzazione della distanza rispetto alla lunghezza dell'allineamento (es. 4/11=0,36)

s	t	-	-	i	a	c	c	i	a	-	-
s	-	c	h	i	a	c	c	i	a	t	a
.	*	*	*	.	.	.	.	.	.	*	*

s	-	t	i	a	c	c	i	-	-	a
s	c	h	i	a	c	c	i	a	t	a
.	*	*	.	.	.	.	.	*	*	.

5 vs **4**

La Levenshtein distance per la misura di similarità fonetiche

- Introdotta da Kessler (1995) per lo studio dei dialetti irlandesi; ulteriormente raffinata da Heeringa (2004)
- Distanze fonetiche calcolate su
  - **Rappresentazioni fonetiche**
  - **Rappresentazioni a tratti**
  - Spettrogrammi acustici
- Problemi e soluzioni
  - Rappresentazioni fonetiche:
    - due unità sono uguali o diverse
    - d(ɛ,e) = d(a,z)
  - Rappresentazioni a tratti:
    - possibile soluzione al problema delle rappresentazioni fonetiche: ogni fonema rappresentato mediante un vettore contenente i valori associati ai singoli tratti
    - stretta dipendenza dal sistema di tratti selezionati

	i	e	u	i-e	i-u
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
lip-rounding	0(none)	0(none)	1(rounded)	0	1

d(i,e) = 1
d(i,u) = 5

La Levenshtein distance per la misura di similarità lessicali

- Nozione binaria di distanza lessicale
  - proporzione di risposte condivise in relazione a un insieme di domande in due località (cfr Seguy, Goebl)
  - problema: varianti lessicali morfologicamente correlate trattate come risposte diverse

- Possibile soluzione (cfr Nerbonne e Kleiweg sui dati del LAMSAS)
  - Ricorso alla distanza di Levenshtein anche per misurare le distanze lessicali

- Scelta sensata anche nel caso dell'ALT in quanto il livello di normalizzazione considerato non fa astrazione da
  - variazioni flessionali e derivazionali:
    - *schiacciàta* vs *schiacciàte*
    - *schiàccia* vs *schiaccétta* vs *schiaccìna*
  - variazioni fonetiche non più vitali sul territorio toscano:
    - *gaglio* vs *caglio*

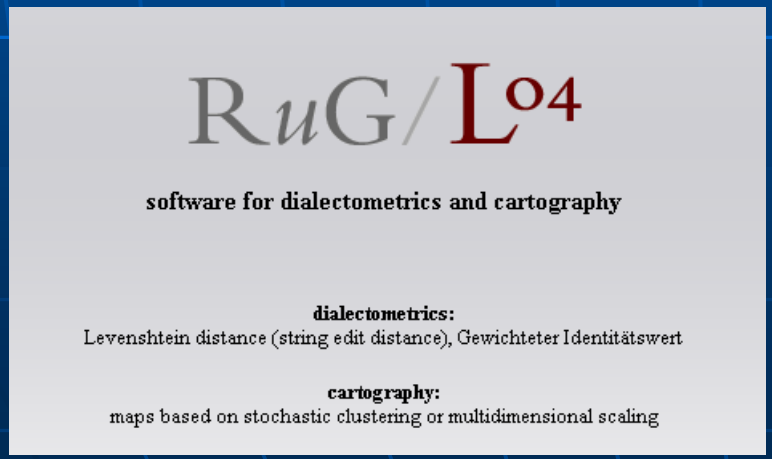
La matrice delle distanze

	Pontremoli	Aulla	Filetto	Licciana Nardi	Fivizzano
Pontremoli	0	0,286706	0,270004	0,2739	0,268461
Aulla	0,286706	0	0,227969	0,202063	0,216318
Filetto	0,270004	0,227969	0	0,275059	0,284294
Licciana Nardi	0,2739	0,202063	0,275059	0	0,198557
Fivizzano	0,268461	0,216318	0,284294	0,198557	0

- Il confronto tra varietà dialettali è condotto sulla base di un insieme di parole che sono oggetto di misurazioni di similarità linguistica per ciascuna coppia di varietà dialettali
- La distanza tra due varietà è calcolata come la media delle distanze ottenute per il campione di parole selezionate
- Le distanze tra le località indagate sono rappresentate in una matrice bidimensionale N x N dove ogni cella specifica la distanza linguistica tra due località

Alla scoperta dei confini e continua dialettali toscani

- Per l'analisi delle matrici delle distanze fonetiche e lessicali abbiamo seguito Nerbonne et al. usando due tecniche complementari di analisi statistica multivariata:
  - clustering
  - scaling multidimensionale



http://www.let.rug.nl/~kleiweg/L04/

Aree lessicali toscane

- Selezione dei dati
  - Risposte a domande onomasiologiche
  - Livello di rappresentazione: rappresentazione normalizzata
  - Numero di risposte diverse per domanda nell'ALT compreso tra 421 e 1
    - Ambito di variabilità selezionato: 5-50 (corrispondente a una selezione di 165 domande onomasiologiche)

DOM. 91 Riccio della castagna

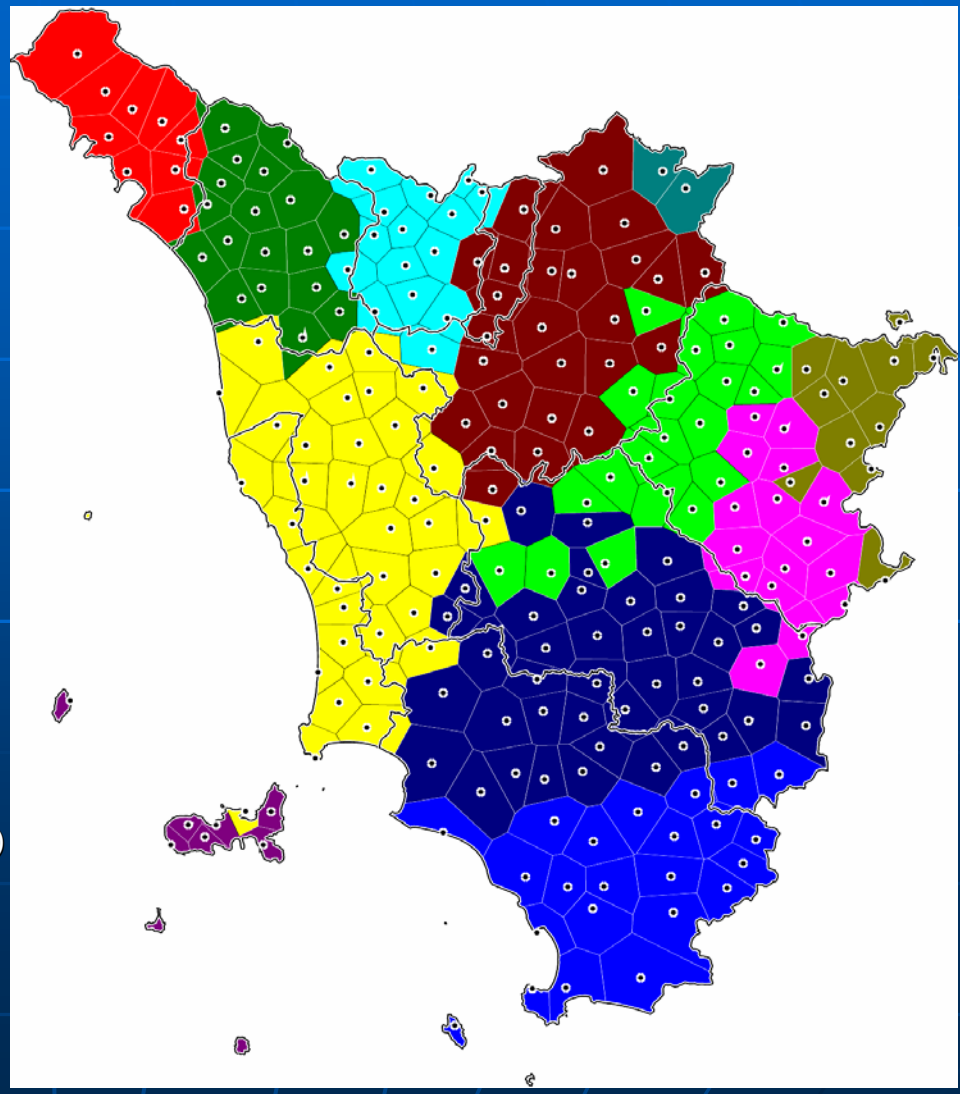
: 1 Pontremoli
- rìccia
- rìccio
: 10 Antona
- el càrdo
: 100 Caprese M
- péglia
: 102 Anghiari
- péglia
: 103 Sansepolc
- péglia
: 104 Badia Ted
- péglia
: 105 Sestino
- péglia
- rìccio
: 106 Antignanc
- rìccio
: 108 Cecina
- càrdo
- rìccio
: 11 Arni
- càrdo
...

DOM. 519b Raccattare

: 113 Querceto
- còglie
- raccattàre
: 114 Laiatico
- raccattàre
: 115 Montecatini Val di Cecina
- còglie
- raccòglie
- raccògliere
- raccàtta
- raccattàre
: 116 Volterra
- còglie
- raccòglie
- raccògliere
- raccàtta
- raccattàre
: 117 Pomarance
- cògliere
- raccògliere
...

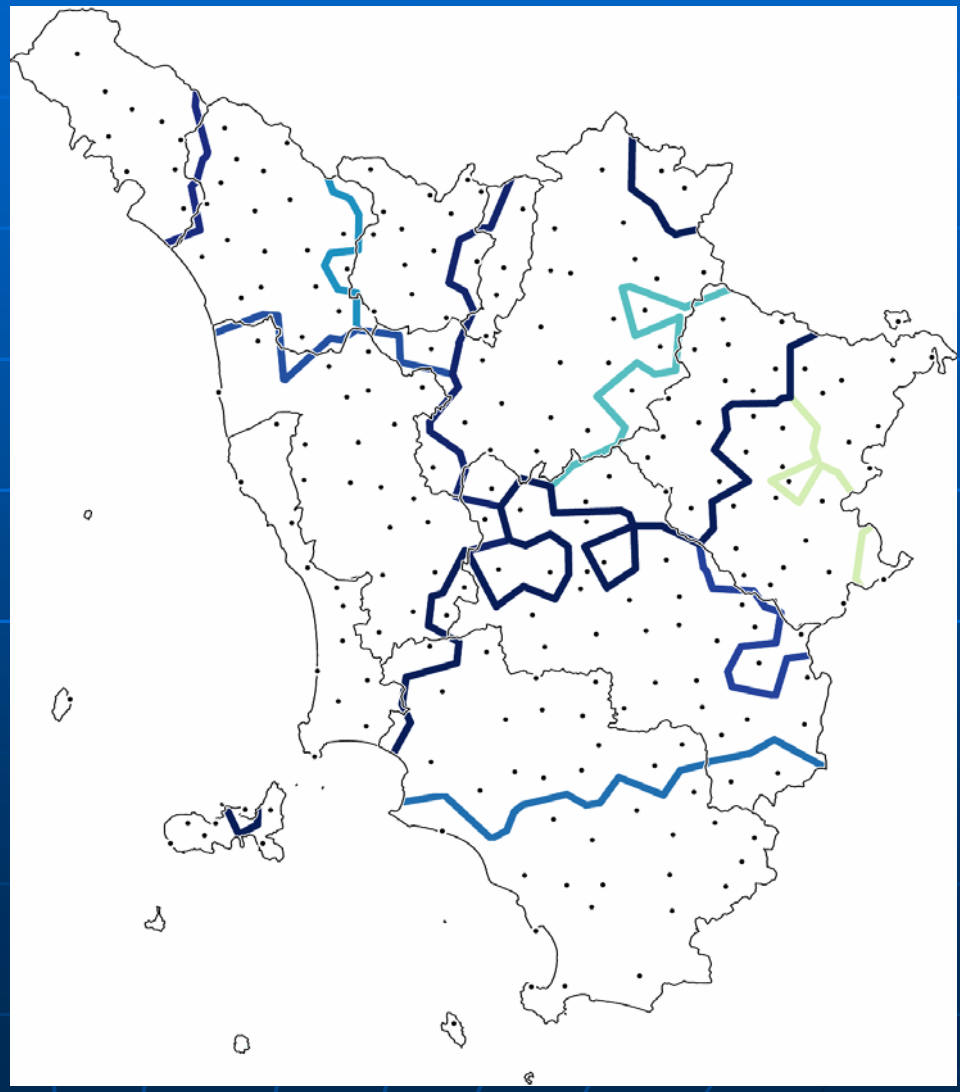
Aree lessicali toscane

- Tecnica usata:
  - clustering gerarchico agglomerativo
- Attestazioni lessicali considerate:
  - 61.714
- Cronbach Alpha:
  - 0,941676
- 12 aree lessicali identificate:
  - dialetti non toscani
    - romagnolo
    - lunigianese
  - dialetti toscani
    - fiorentino
    - pistoiese
    - lucchese
    - pisano-livornese
    - elbano
    - aretino suddiviso in
      - Valdarno superiore, Casentino
      - Val di Chiana
      - Val Tiberina toscana
    - senese (con propaggini nel grossetano)
    - Maremmano e Amiatino
- Ma cosa sta dietro ai confini dialettali identificati? Rappresentano tutti confini linguistici ugualmente significativi?



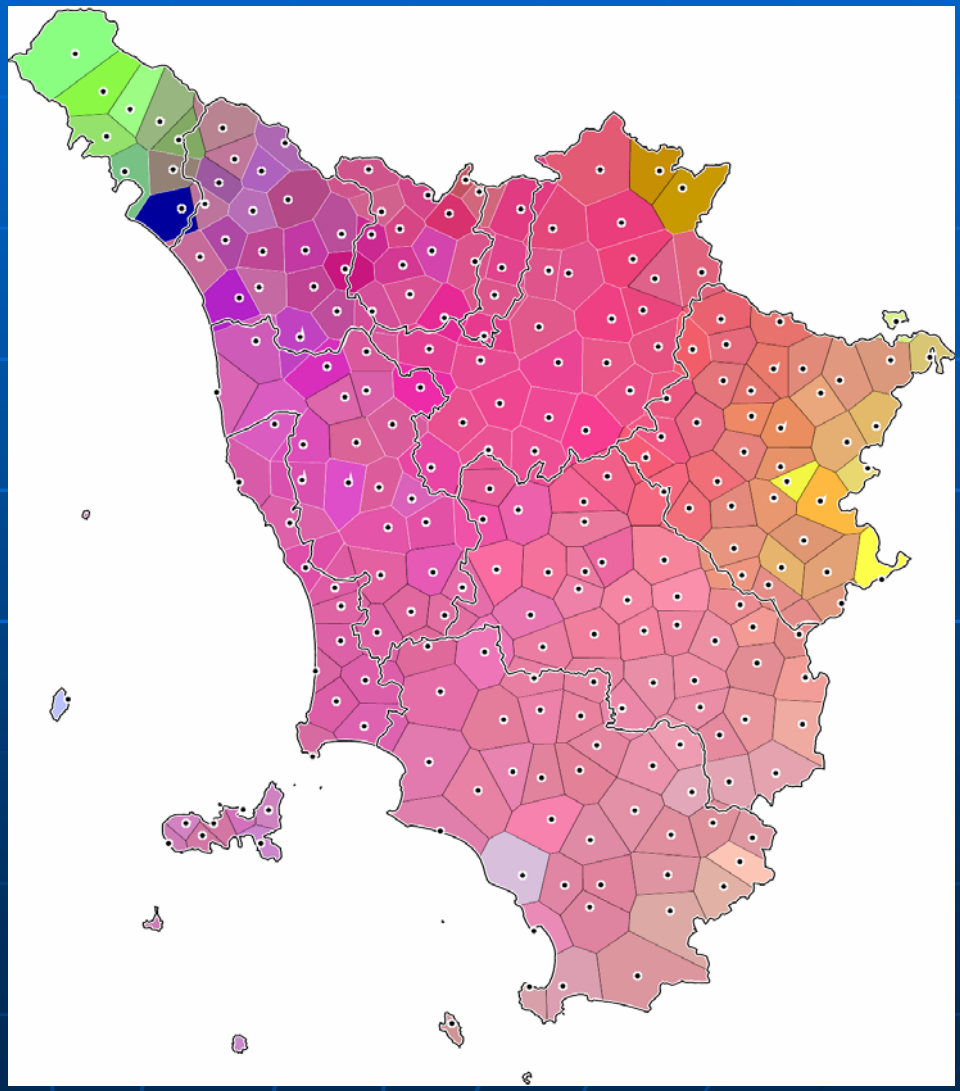
Aree lessicali toscane

- “Composite Cluster Map” (Kleiweg), ottenuta attraverso operazioni di clustering successive
- L’intensità del colore delle linee di confine riflette la significatività del confine identificato
  - Il confine più significativo taglia la Toscana in due separando l’area grossetana, senese e aretina (ad esclusione del Casentino e del Valdarno) e dei dialetti romagnoli dal resto della regione
  - Seguono, in ordine di significatività, i confini che identificano:
    - i dialetti lunigianesi
    - il fiorentino
    - il senese-grossetano
    - il pisano-livornese
    - il pistoiese e lucchese
    - il Maremmano e Amiatino
    - …



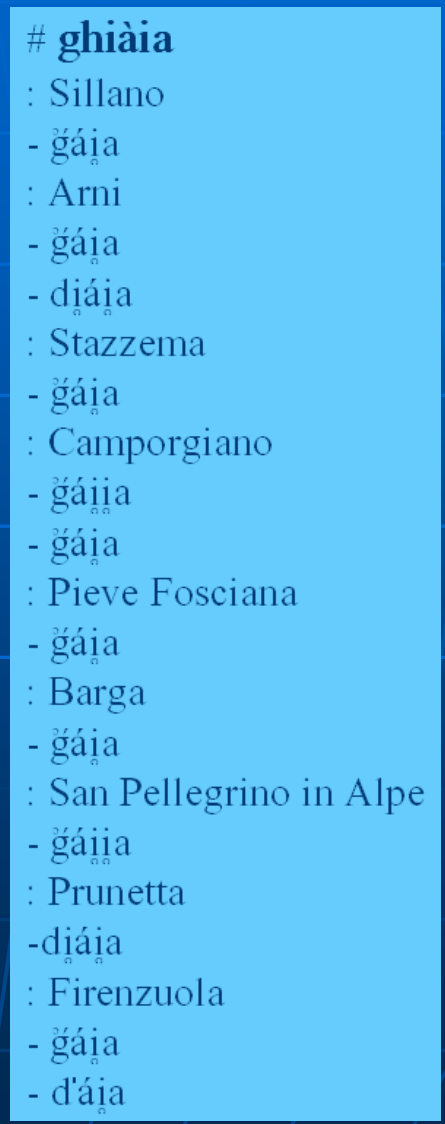
Aree lessicali toscane

- Tecnica usata:
  - scaling multidimensionale (MDS)
  - dalla matrice delle distanze può essere ricavato un sistema di coordinate che riflette la vicinanza/distanza linguistica tra varietà dialettali
    - da 224 a 3 dimensioni
- Oltre i confini:
  - “continuum map”
  - il colore che contrassegna ciascun poligono riflette le tre coordinate assegnate attraverso MDS
    - forti contrasti di colore indicano differenze linguistiche marcate
    - colori simili contrassegnano varietà linguistiche vicine



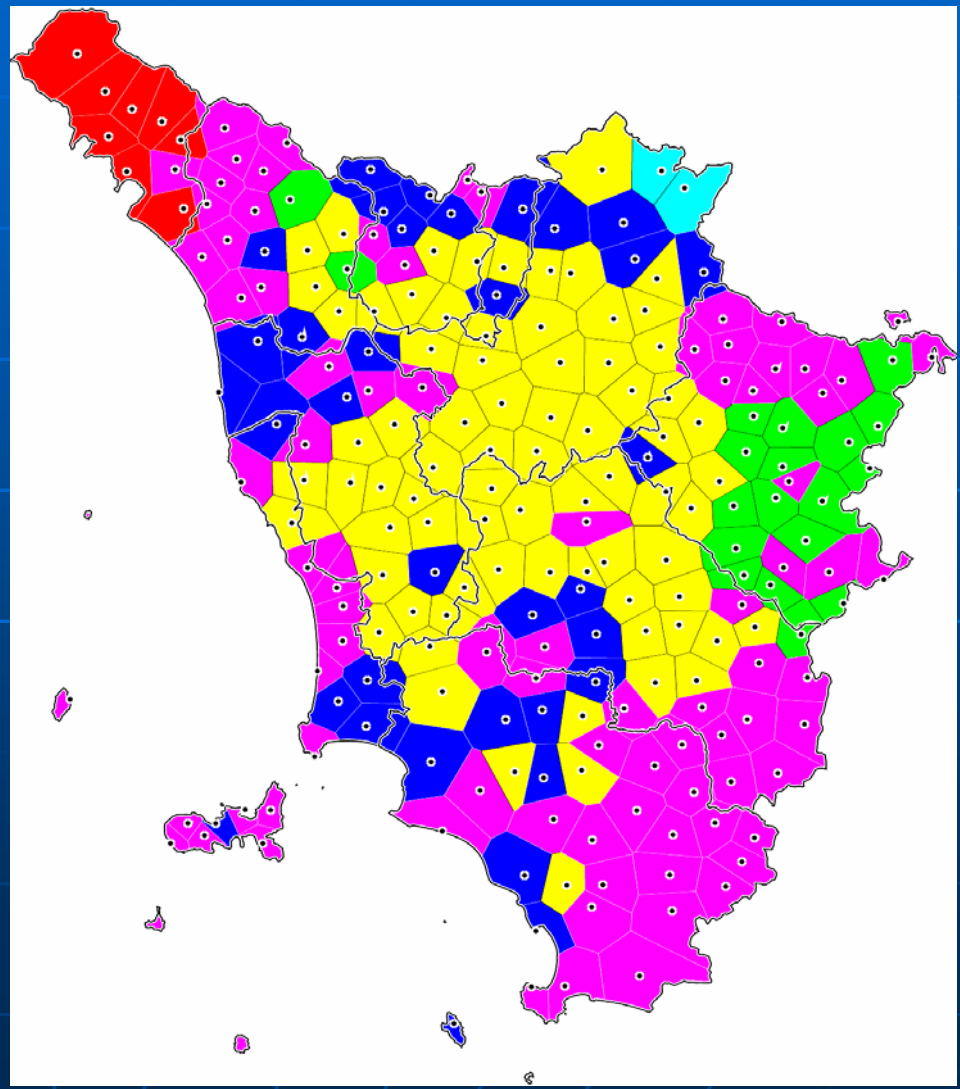
Aree fonetiche toscane

- Selezione dei dati
  - 592 tipi lessicali selezionati tra le rappresentazioni normalizzate
  - Livello di rappresentazione: trascrizione fonetica
  - Copertura geografica minima: 100 località
  - Varietà di realizzazione fonetica compresa tra 34 e 5



Aree fonetiche toscane

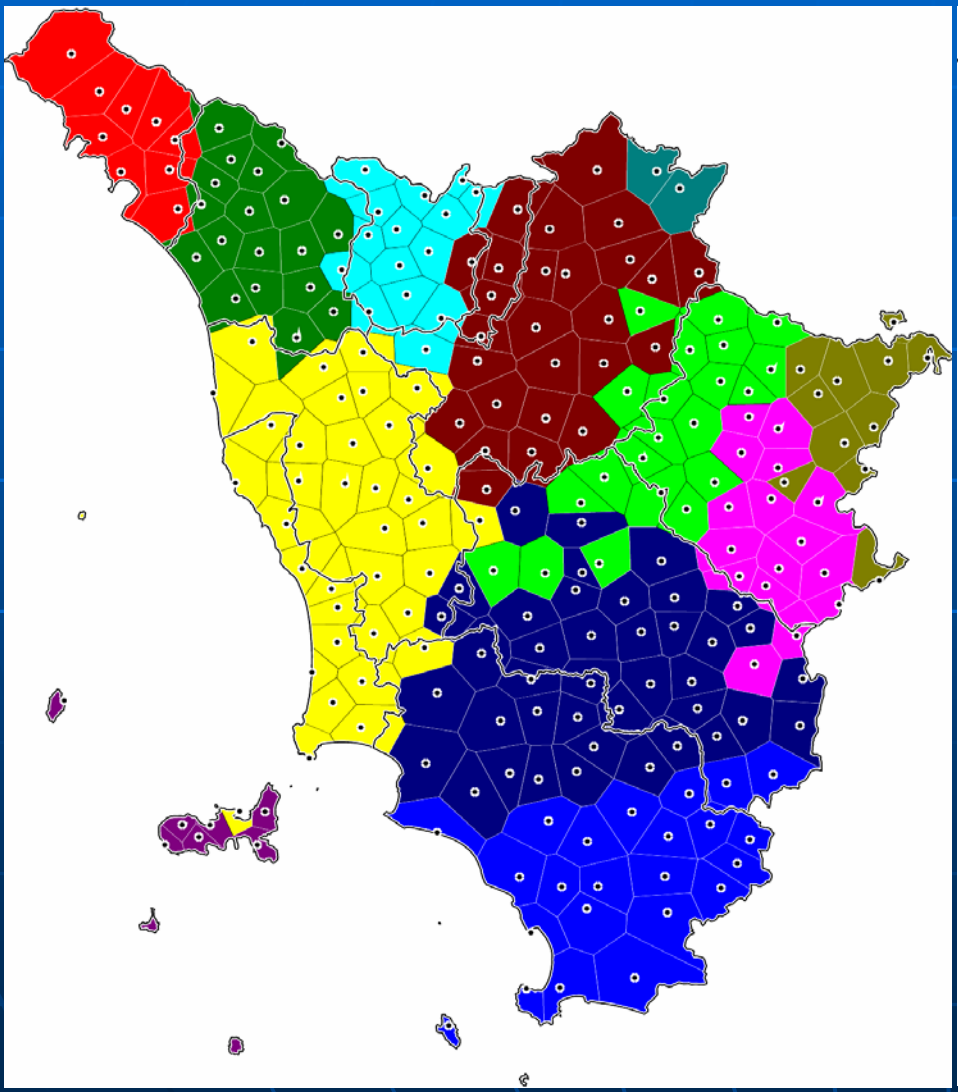
- Matrice delle distanze basata su rappresentazioni a tratti
  - 18 tratti
  - 78 foni
- Tecnica usata: clustering gerarchico agglomerativo
- Attestazioni fonetiche considerate: 93.076
- Cronbach Alpha: 0,985522
- Aree fonetiche identificate:
  - aree concentriche
  - nucleo centrale che dall'area fiorentina si spinge verso
    - l'area senese
    - il pistoiese e l'area lucchese
    - l'area pisano-livornese (con sporadici sbocchi costieri)
  - area di contorno all'interno della quale si distinguono nettamente i dialetti romagnoli, il lunigianese e alcuni dialetti aretini (Val di Chiana, Val Tiberina)

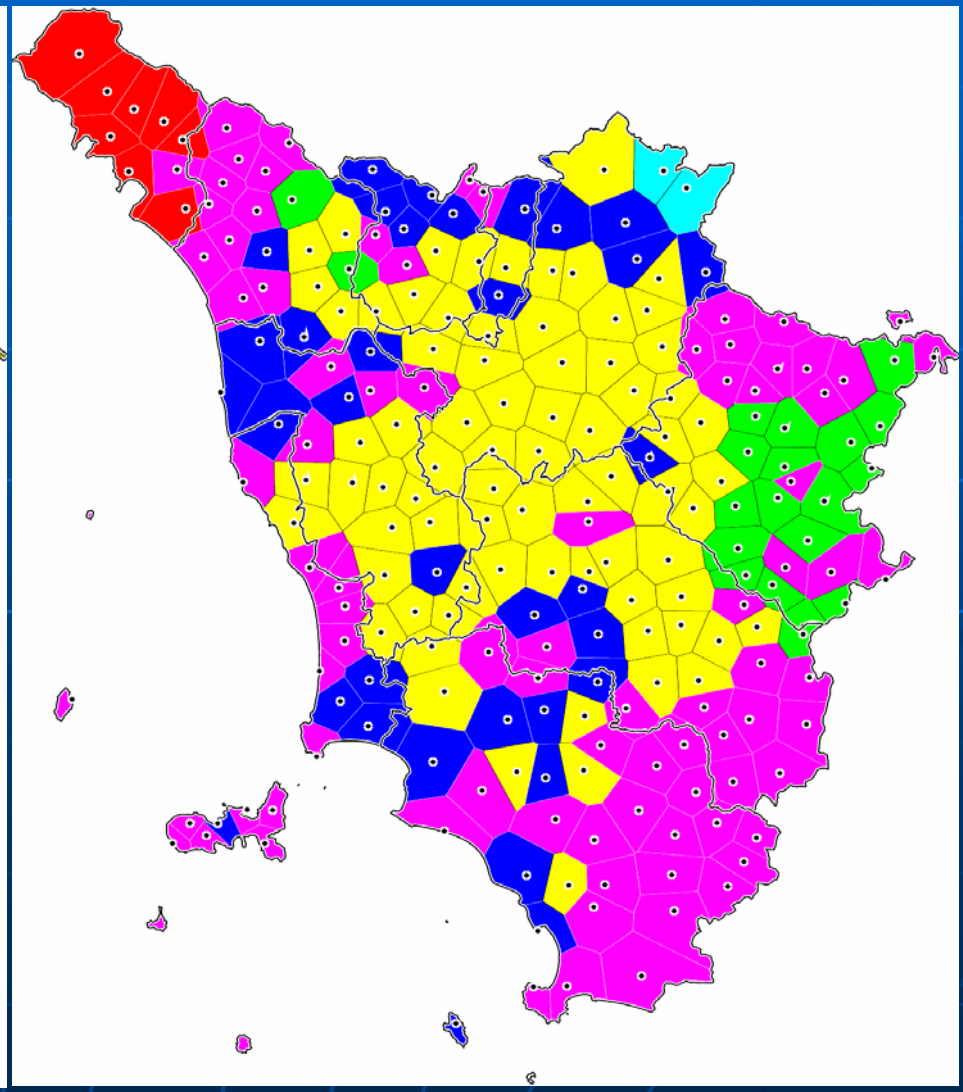


Aree fonetiche e lessicali a confronto: il problema

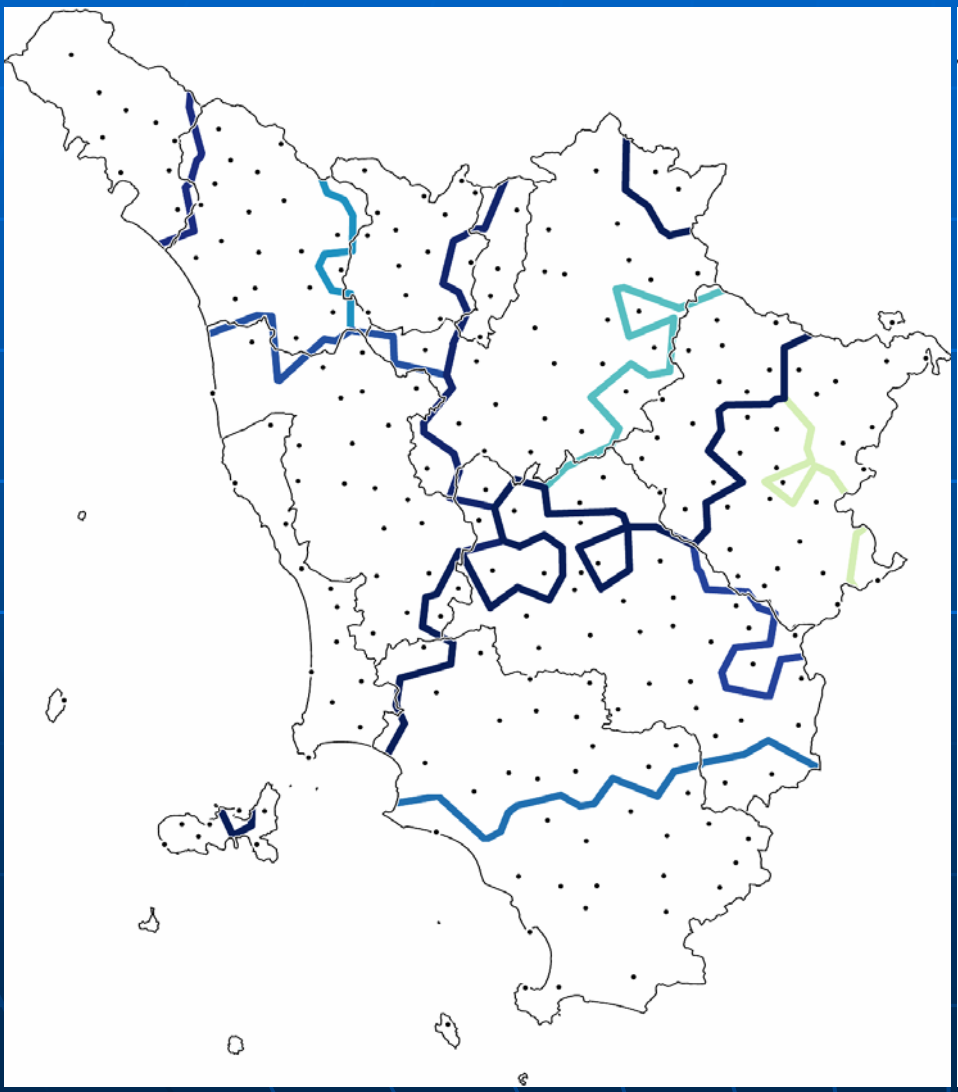
- Netta discrepanza tra la morfologia delle aree lessicali e fonetiche toscane rilevate
  - aree giustapposte vs aree concentriche
- Il lessico alla base di entrambe le classificazioni proposte
  - Aree lessicali: diverse lessicalizzazioni di uno stesso concetto
  - Aree fonetiche: diverse pronunce di uno stesso tipo lessicale
- Chambers and Trudgill (1998)
  - “It seems a safe assumption to rank lexical differences as more superficial than pronunciation differences because the former are more likely to be subject to self-conscious control or change by speakers than the latter.”
- Giannelli (2000) fornisce una classificazione dei dialetti toscani basata su diversità morfosintattiche, fonetiche, fonologiche e lessicali

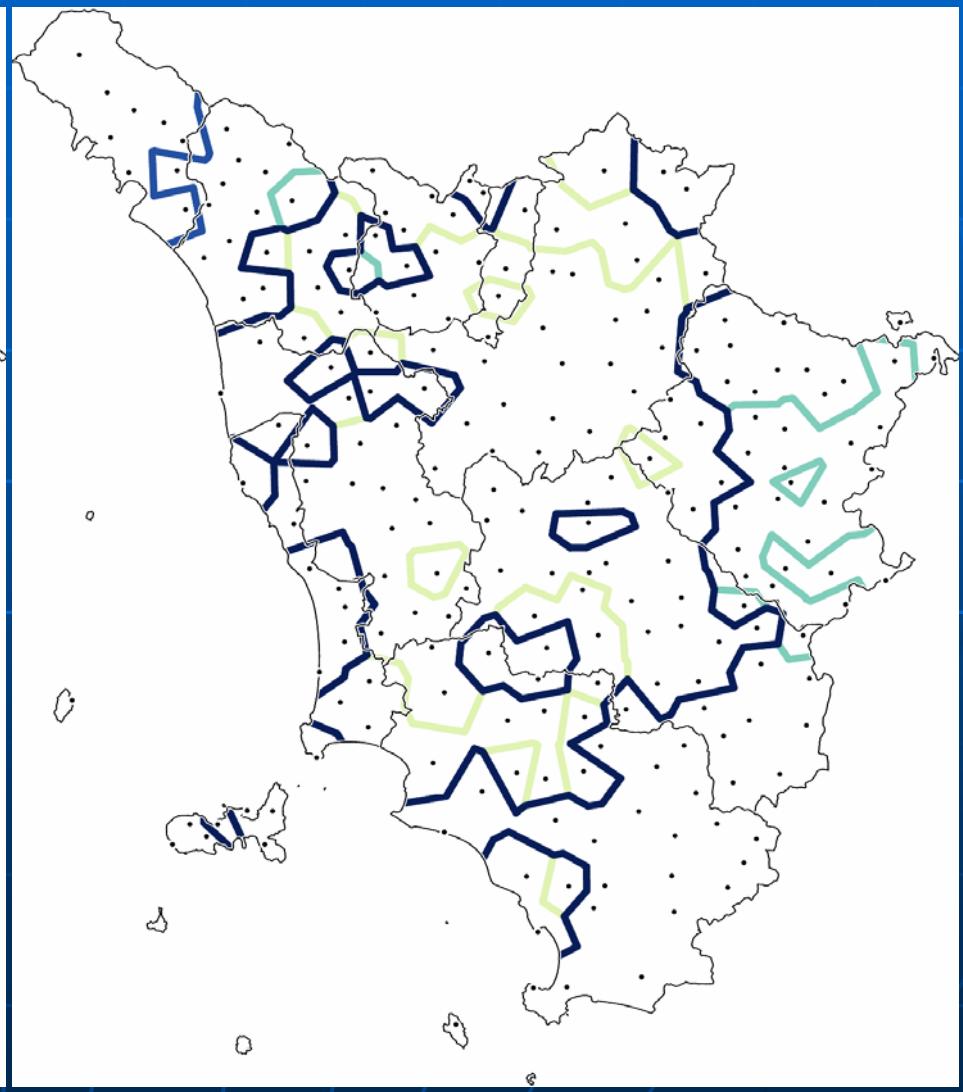
Aree fonetiche e lessicali a confronto: l'evidenza dell'ALT





Aree fonetiche e lessicali a confronto: l'evidenza dell'ALT





Conclusioni
e ulteriori direzioni di ricerca

- Applicazione di tecniche dialettologico-computazionali ai dati dell'ALT per lo studio degli schemi di variazione fonetica e lessicale in Toscana
- Risultati preliminari ma promettenti che confermano le analisi di Giacomelli e Giannelli
- Nuove prospettive di ricerca per quanto riguarda la correlazione tra variazione fonetica e lessicale. Ulteriori indagini in corso:
  - sulla misura della correlazione tra variazione fonetica e lessicale
  - sulla "instabilità" delle aree lessicali
- Estensione dell'applicazione di queste tecniche per lo studio della variazione linguistica in relazione a variabili extralinguistiche quali sesso, età e status socio-culturale