

Synchronic patterns of Tuscan phonetic variation and diachronic change: evidence from a dialectometric study

Simonetta Montemagni*, **Martijn Wieling⁺**,
Bob de Jonge⁺ and **John Nerbonne⁺**

*Istituto di Linguistica Computazionale – CNR (Italy)

⁺Center for Language and Cognition Groningen, University of
Groningen (The Netherlands)

simonetta.montemagni@ilc.cnr.it

{m.b.wieling,r.de.jonge,j.nerbonne}@rug.nl



Novel: automating attention to phonological context

- Background on *gorgia Toscana*
- Data: *Atlante Lessicale Toscano* (<http://serverdbt.ilc.cnr.it/ALTWEB/>)
- Spectral clustering of bipartite graphs
 - Sound correspondences in context
- Radial spread from Florence,
 - Generalization of phonological context
 - Attention to current demographics
- Conclusions



When diachrony meets synchrony

- How diatopic linguistic variation can be used to shed light on diachronic phonetic processes
- Starting from a synchronic, dialectometric analysis of phonetic variation in a central Italian region - Tuscany - we investigate a controversial feature of Tuscan dialects
 - Spirantization, and specifically the so-called *Gorgia toscana*, whose earliest reference dates back to the beginning of the 16th century
- Method (graph-theoretic): spectral partitioning of bipartite graphs, used by Wieling and Nerbonne (2010, 2011) to cluster dialectal varieties and simultaneously determine the underlying linguistic basis (features)



The phenomenon of spirantization in Tuscany: what

- *Gorgia toscana*: popular term for voiceless stop spirantization intervocalically
 - Originally restricted to the shift from /k/ to /x/
 - Later extended to voiceless dental and bilabial stops /t p/
- Rapid spread of spirantization through the Tuscan consonants:
 - Spirantization of /k p t/ in non-intervocalic contexts
 - Voiced stops /b d g/ undergo similar processes
 - Affricates /tʃ/ and /dʒ/ also strongly affected by spirantization
- Focus here on spirantization of voiceless and voiced stops in different contexts
 - peculiar phenomenon of Tuscan dialects



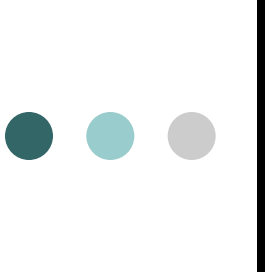
Spirantization in Tuscany: whence

- Tuscan *gorgia* increasingly accepted as being a local and **innovative** natural phenomenon (lenition, consonantal weakening) spreading from the influential center of Florence in all directions
 - Florence traditionally viewed as the epicenter
 - From Florence, the *gorgia* spreads along the entire Arno valley, losing strength nearer the coast
 - It is also present to some extent in the northwest and the northeast
 - The Apennines are the northern border of the phenomenon
 - Present in Siena and further south but not in far southern Tuscany
- Intervocalic voiceless spirantization is expanding not only geographically but also phonologically
 - Tuscan Spirantization no longer restricted to intervocalic voiceless stops
 - Extension of *gorgia* to voiced stops, fuelled by perceived prestige of *gorgia*-related phenomena amongst speakers in the region



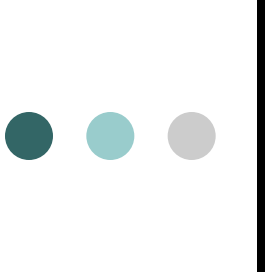
Data source

- *Atlante Lessicale Toscano* (<http://serverdbt.ilc.cnr.it/ALTWEB/>)
 - Regional linguistic atlas focusing on dialectal variation throughout Tuscany, a region where both Tuscan and non-Tuscan dialects are spoken
- ALT interviews carried out
 - In **224** localities of Tuscany
 - With **2,193** informants selected wrt socio-demographic parameters
 - On the basis of a questionnaire of **745** target items designed to elicit lexico-semantic variation
 - Data collection: 1973-1986
- Multi-level representation of dialectal data
 - Focus on phonetic transcription and normalized representation levels where the latter abstracts away from Tuscan phonetic variation
 - Alignment of representation levels exploited to automatically extract phonetic variants (PV) sharing the same normalized form (NF)



Building the experimental data set (1)

- ALT dialectal data used as a *corpus*
 - We did not start from a predefined set of questionnaire items specifically designed to investigate the geographic distribution of phonetic features, but rather from the set of the attested ALT lexical items, which were elicited from informants for quite different (mainly, lexico-semantic) purposes
 - By using atlas data as a corpus, the problem of inherently subjective feature selection is significantly reduced, thus providing a more “realistic” linguistic signal (Szmrecsanyi)
- But – by using atlas data as a corpus one main advantage ascribed to atlas-based studies, namely broad geographical coverage, can no longer be taken for granted
 - To overcome this potential problem, a minimal geographic coverage threshold was enforced in the selection of normalised forms used in this study



Building the experimental data set (2)

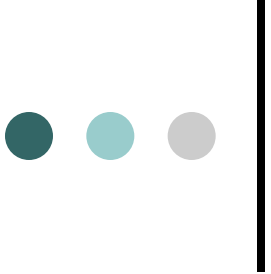
- Focus on
 - Tuscan dialects: 213 locations
 - Phonetic variants of 444 lexical types selected from the ALT dialectal corpus on the basis of
 - Geographical coverage: ≥ 100 locations
 - Phonetic variability: between 5 and 34 variants
 - Morpho-syntactic category: nouns and adjectives, both single words and multi-word expressions
 - -- for a total of 502.799 phonetic variant tokens
- Representativeness of the selected sample wrt the whole set of NFs having at least two PVs attested in at least two locations assayed using the correlation between overall phonetic distances and phonetic distances using only the selected sample
 - $r = 0.994$
- The experimental dataset also includes the phonetic realization of the selected NFs in a reference variety
 - Standard Italian

Methods: extracting sound correspondences

- Every variety attested at a given location is described in terms of the realizations of phonetic segments wrt standard Italian
 - Attested phonetic realizations encoded in terms of sound correspondences (SCs) linking the dialectal allophones to corresponding realizations in the standard (reference variety)
 - SCs generated with the Levenshtein algorithm using PMI-based segment distances (Wieling et al., 2009)

Italian	a	l	b	i	k	ɔ	k:	a
Montecatini Val di Cecina	a	r	b	i	h	ɔ	k:	a

- context-free vs. context-sensitive representation of sound correspondences
 - /l/:[r] vs V/l/C:V[r]C
 - /k/:[h] vs V/k/V:V[h]V

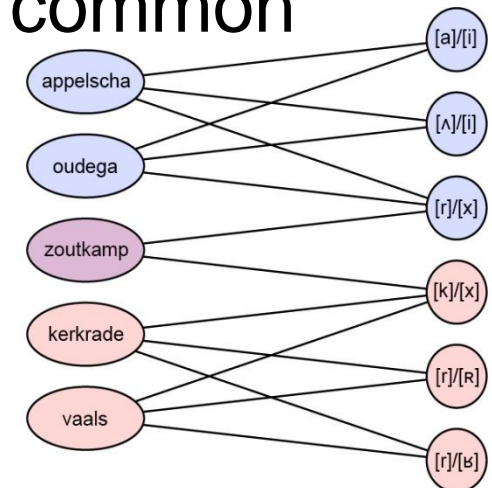
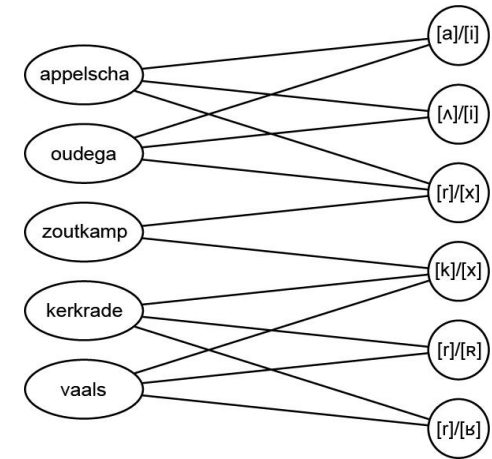


Extracting sound correspondences involving spirantization

- Focus on
 - Most frequent phonetic variants of each selected normalized form attested in a given location
 - Phonetic correspondences involving both identical and non-identical segments
 - With a stop on the reference (standard) side
 - With either an occlusive or a spirantized (including absent) realization on the allophonic (dialectal) side
- For a total of 16 context-free & 84 context-sensitive sound correspondences
- Construction of a variety x sound-correspondence matrix with normalized frequencies
 - SC frequencies normalized by dividing by the number of words, as not all words are attested in every variety

Clustering SCs & varieties simultaneously

- From a site \times feature matrix
- Create a bipartite graph (right)
- Eigenvalues of (Laplacian) graph's spectrum effectively cluster sites (based on common features) and features (based on common sites)
- Hierarchical version used here





Verifying the most important features in clusters

③ We measure representativeness & distinctiveness

○ Feature f is *representative* of cluster c

- $Rep(f, c) = \frac{|sites\ in\ c\ with\ f|}{|sites\ in\ c|}$

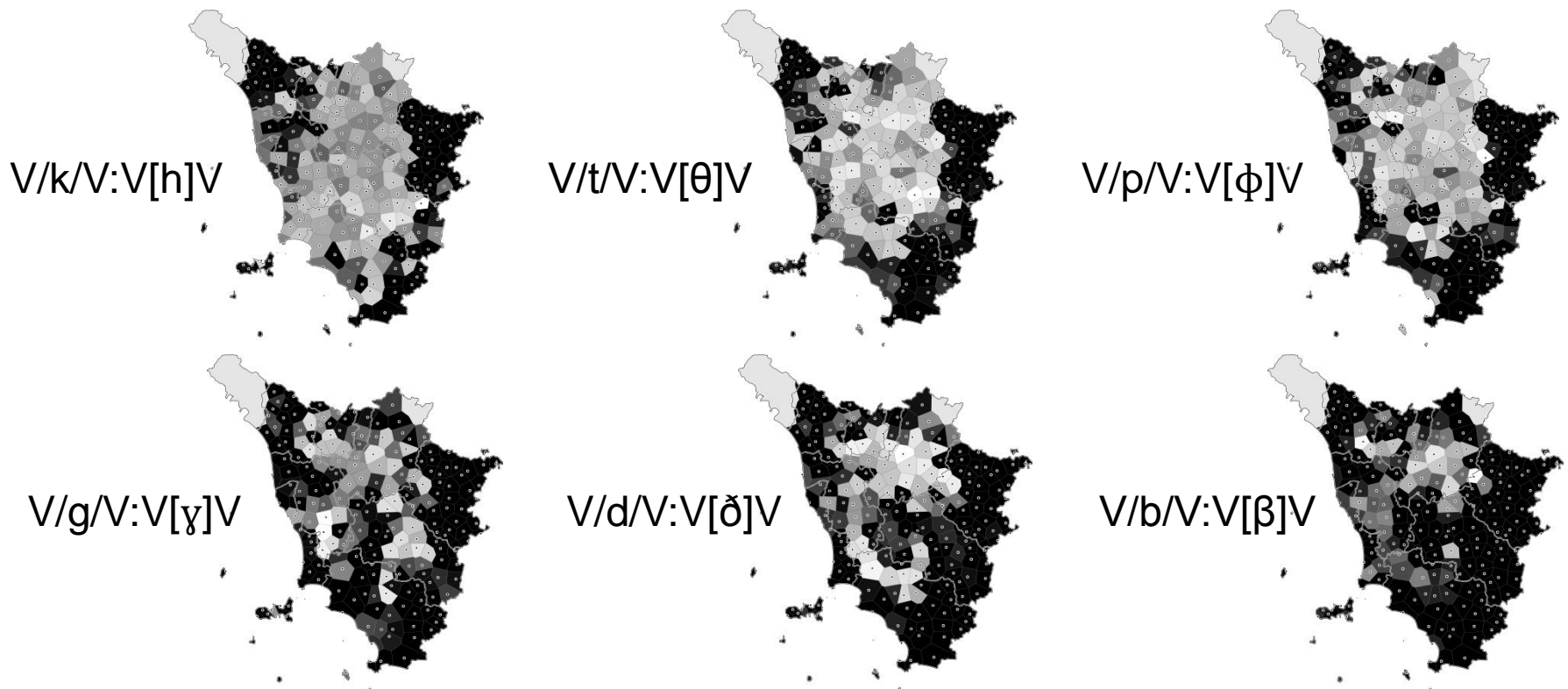
○ Feature f *distinguishes* cluster c

- $Distinct(f, c) = \frac{|sites\ in\ c\ with\ f|}{|sites\ with\ f|}$ (w. correction for chance occurrence)

○ *Importance* ~ mean of representativeness and distinctiveness

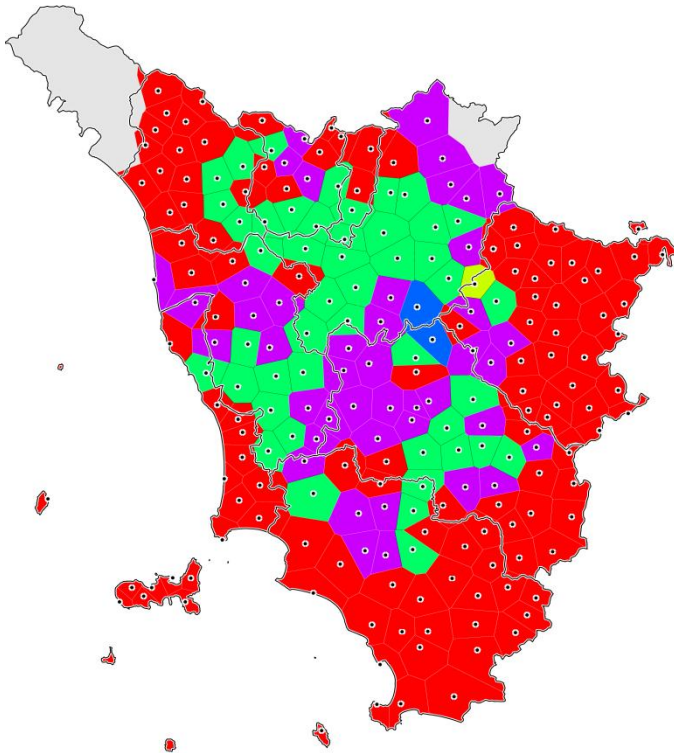
Areal distribution of sound correspondences

- SCs involving voiceless and voiced stops and their spirantized counterpart in intervocalic context

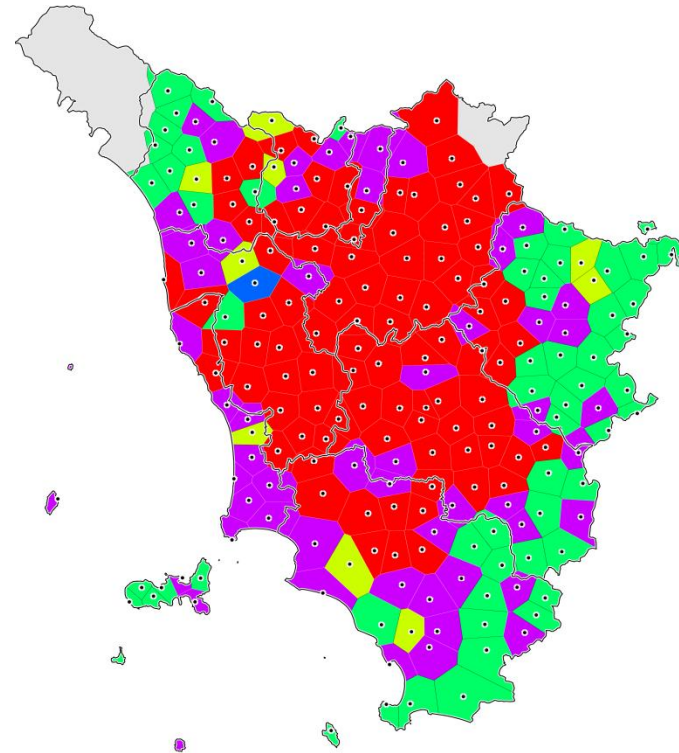


Clustering of Tuscan varieties

- Geographic clustering of Tuscan wrt spirantization



with contextualised SCs



without contextualised SCs

Features underlying Tuscan clusters

Representativeness=1
Distinctiveness=1

Ranked sound correspondences

V/g/V:V[ɣ]V (0.319115)

V/d/V:V[ð]V (0.280644)

/k/C:[h]C (0.210480)

/k/V:[h]V (0.126210)

V/b/C:V[β]C (0.112370)

Ranked sound correspondences

V/t/V:V[θ]V (0.191697)

/p/V:[ϕ]V (0.163595)

V/p/V:V[ϕ]V (0.152429)

V/p/C:V[ϕ]C (0.144144)

/t/C:[θ]C (0.130167)

V/t/C:V[θ]C (0.130073)

/p/B:[ϕ]B (0.127868)

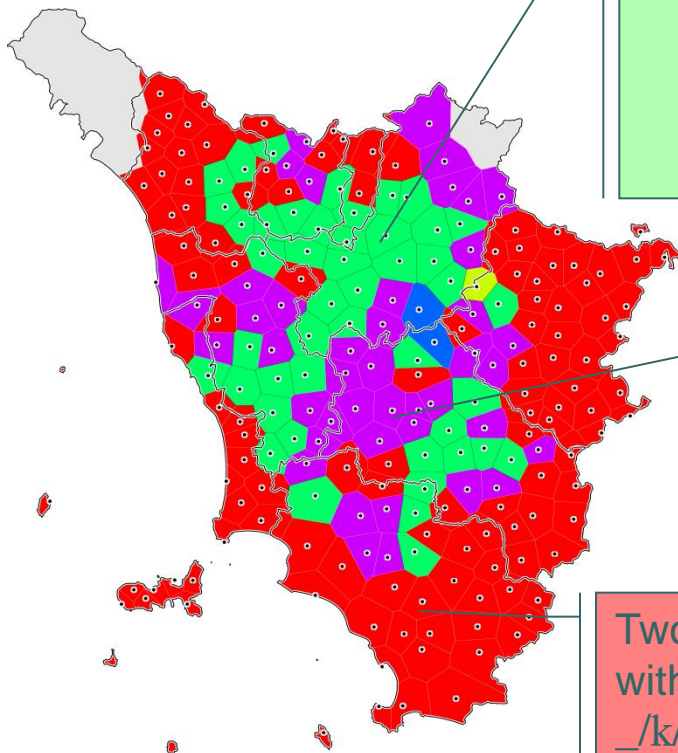
V/k/V:V[h]V (0.112285)

Two SCs only
with spirantization

/k/V:[x]V (0.133616)

V/k/V:V[x]V (0.116877)

with contextualised SCs



Features underlying Tuscan clusters

Ranked sound correspondences

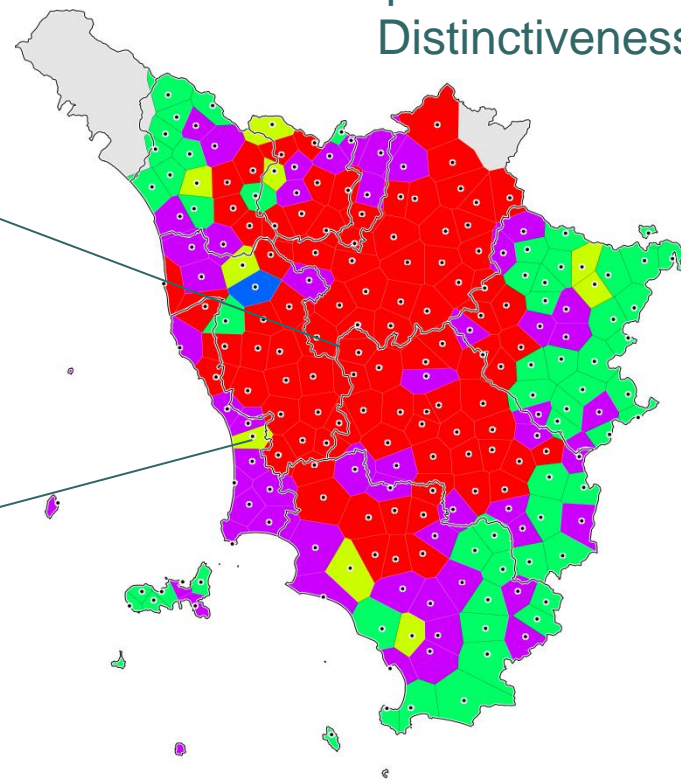
/t:[h] (0.500465)
/d:[ð] (0.484426)
/t:[θ] (0.448604)
/p:[ϕ] (0.421344)
/b:[β] (0.421309)
/g:[ɣ] (0.404903)
/k:[h] (0.258726)
/t:[ø] (0.177900)

Ranked sound correspondences

/k:[x] (0.197257)

No spirantized SCs underlying the marginal clusters (green and purple)

Representativeness=1
Distinctiveness=1



SCs without context



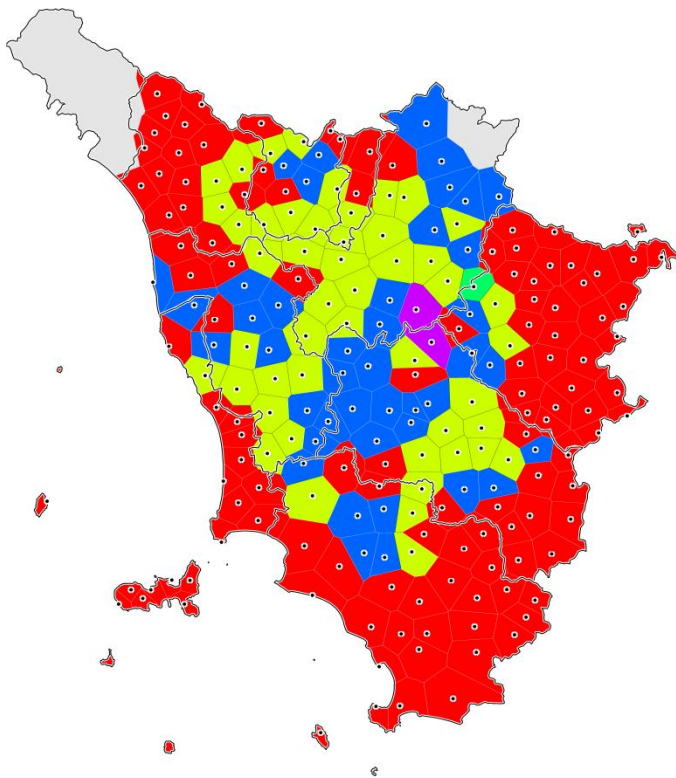
Results: role of context

- Results show that context information plays a central role
 - Sound changes are recognized to be conditioned by phonetic context, as we saw in the case of Tuscan *gorgia*
 - Contextualised SCs enable the detection of an articulated and linguistically well-founded diffusion, both at the level of regional coherence and the underlying linguistic features
- Using contextualised SCs we were able to “reconstruct” the spreading of spirantization phenomena
 - Geographically: across Tuscany starting from Florence
 - Phonologically: through the consonantal phonology by originally involving the velar stop /k/, then /p t/ up to the voiced stops /b d g/
- Without context information a more static picture emerges with a single cluster characterized by spirantization

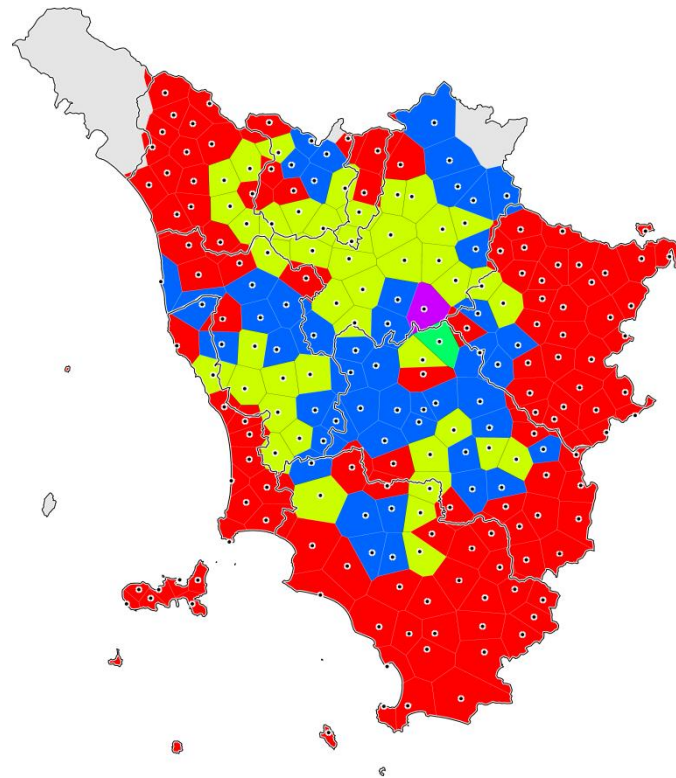
Geographical results: old vs young speakers

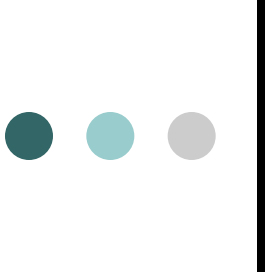
- Geographic clustering of Tuscan wrt spirantization using contextualised SCs

Old speakers (born in 1930 or earlier)



Young speakers (born after 1930)





Linguistic results: old vs young speakers

- The main differences in age groups involve underlying features
 - **Same** typology of features underlying the major clusters
 - **Different importance** assigned to individual features, reflected both in the ranking and the score assigned to each SC

Old vs Young speakers

- **Lower vs higher** salience assigned to most innovative SCs
 - **Core spirantization cluster:** SCs involving voiced stops /g d b/
 - **External spirantization cluster:** SCs involving /p t/

- Minor differences across age groups, mainly at the level of feature salience
 - ALT data elicited on the basis of a questionnaire focused on lexico-semantic variation
 - careless, informal, emotive pronunciation rarely testified in ALT data

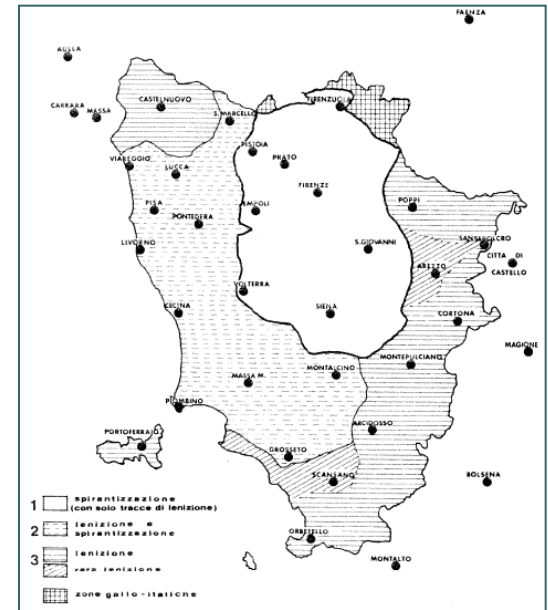
Discussion

- Results are in line with the primary texts on the topic of *Gorgia Toscana*

- Giannelli and Savoia (1978, 1980)
- Hajek (1996)

- Spirantization arose in Florence and spread to other areas

- Intervocalic voiceless spirantization (or Tuscan *Gorgia*) expanded in different respects
 - geographically
 - phonologically
 - demographically (age-based analysis)
- Spirantization in Tuscany is still a native feature which is quite resistant to standardization





Conclusion

- The method of spectral partitioning of bipartite graphs when applied to synchronic dialectal data can effectively be used to investigate diachronic phonetic processes
 - Case study carried out on Tuscan dialects, in particular on the phenomenon of spirantization with a specific view to the so-called *gorgia toscana*
- A careful analysis of the sound correspondences involved in spirantization provides truly valuable information for the reconstruction of the diachronic process of spirantization
 - geographically
 - phonologically
 - demographically
- On the phonological side
 - Crucial role played by contextual information