Exploring the correlation between pronunciation and lexical variation in Tuscany

Simonetta Montemagni

Istituto di Linguistica Computazionale – CNR

simonetta.montemagni@ilc.cnr.it

Outline

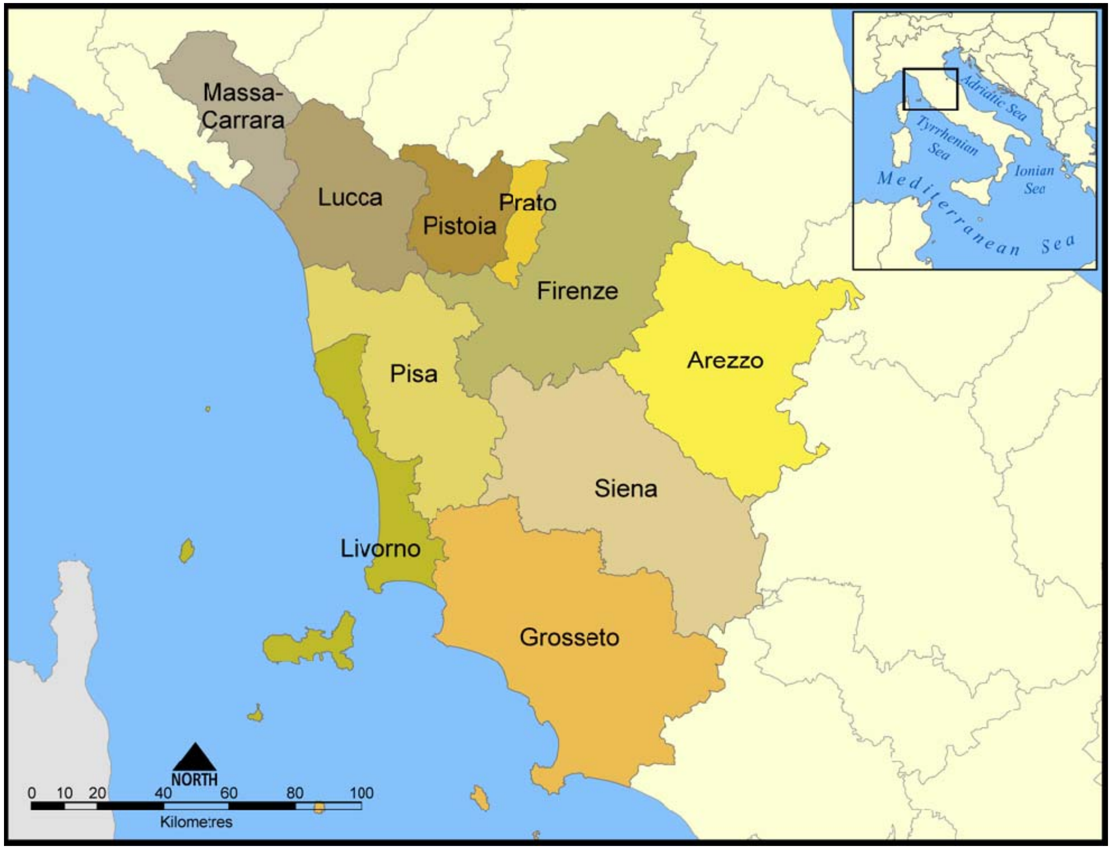
- Research questions
- Tuscany
- The data source
- Induction of patterns of linguistic variation
  - Pronunciation variation
  - Morpho-lexical variation
- Correlations
  - Between linguistic levels
  - With geography
- Behind patterns of pronunciation variation: work in progress
- Conclusions

Research questions

1. Whether and to what extent are observed patterns of pronunciation and lexical variation associated with one another?
   - in traditional dialectology, no obvious way to answer this question beyond fairly superficial and impressionistic observations
   - in the framework of dialectometric studies, dialect distances can be measured with respect to different linguistic levels and their correlation can also be investigated
     - Nerbonne (2003), Gooskens and Heeringa (2006) and Spruit et al. (in press)

2. Whether and to what extent do pronunciation and lexical distances correlate with geographic distance? If this turns out to be the case, are they expected to correlate in the same way?

Tuscany



- Special status of Tuscany in the puzzle of Italian dialects
- Compromise between northern and central-southern dialects
- Source of Italian language
- Not easy linguistic characterization
  - very few features common to all and only Tuscan dialects
  - elements of differentiation present at all levels of linguistic description

The data source

- The *Atlante Lessicale Toscano* (ALT) available as an on-line resource (*ALT-Web*) at http://serverdbt.ilc.cnr.it/ALTWEB/
  - dialectal data have both a **diatopic** and **diastratic** characterisation
- ALT interviews carried out
  - in **224** localities of Tuscany
  - with **2,193** informants selected wrt parameters ranging from age, socio-economic status to education and culture
- Field workers employed a questionnaire of **745** target items, designed to elicit variation mainly in **vocabulary**, semantics and **pronunciation**
  - Data collection: 1973-1986
- More than **350.000 geo-referentiated answers** were collected which were integrated with **additional material** emerged during the interviews (about 30.000 dialectal items)
  - corresponding to more than 84,000 different dialectal items

The data source: representation model of dialectal data

- **Multi-level representation model**
  - ***phonetic transcription***
    - the phonetic alphabet used in the ALT project was a geographically specialized version of the “Carta dei Dialetti Italiani” (CDI) transcription system
  - ***basic orthographic transcription***
    - to help the non-expert user to understand the phonetically transcribed form
    - to account for the variety of attested phonetic realizations
  - ***normalized representation***
    - abstracting away from within-Tuscany vital phonetic variation
    - **NOT** dealing with
      - morphological variation (neither inflectional nor derivational)
      - no longer productive phonetic processes

Induction of patterns of pronunciation and lexical variation:
building the data set

- focus on the levels of
  - **phonetic transcription** and
  - **normalised representation**
- multi-level representation exploited in different ways
  - the **alignment** of the representation levels used to automatically extract all phonetic realizations attested in Tuscany for the same abstract normalized word form
    - pronunciation distances calculated wrt these data **testify productive phonetic processes only**, without interference from any other linguistic description level (e.g. morphology)
  - patterns of **pronunciation and lexical variation** can be studied with respect to **different representation levels** of the same dialectal data

Measuring pronunciation distances in Tuscany (1)

- The experimental data set
  - normalised forms (NF) attested as answers to ALT questionnaire items, having at least 2 different phonetic variants attested in at least 2 different locations
- The distance between the pronunciations of corresponding words in different locations calculated on the basis of the Levenshtein distance (LD)
  - no normalisation by the length of compared words
- Two different experiments carried out on the selected data set, with LD operating respectively on
  - **phone-based** representations
    - two phones are equal or different: rough measure
  - **feature-based** representations automatically generated on the basis of a system of 18 features, identified starting from the ALT phonetic transcription system
    - more sensitive representation accounting for phone similarities
- Used software: Peter Kleiweg's RUG/L04 dialectometric package

Measuring pronunciation distances in Tuscany (2)

- Some numbers
  - 9,553 different normalised forms
  - 34,074 different phonetic variants
  - 221,705 geo-referentiated phonetic variants
- Cronbach α = 0.99 in both experiments
- Comparing the distances identified on the basis of phone-based and feature-based representations
  - Pearson's correlation coefficient: r=0.99
    - feature-based representations do not lead to much improved analyses
    - the rough measure working on phone-based representation appears to be reliable when working with large data sets
- Pronunciation distances between 224 locations
  - the distance between two locations equal to the average of LDs calculated for individual word pairs
  - missing pronunciations ignored

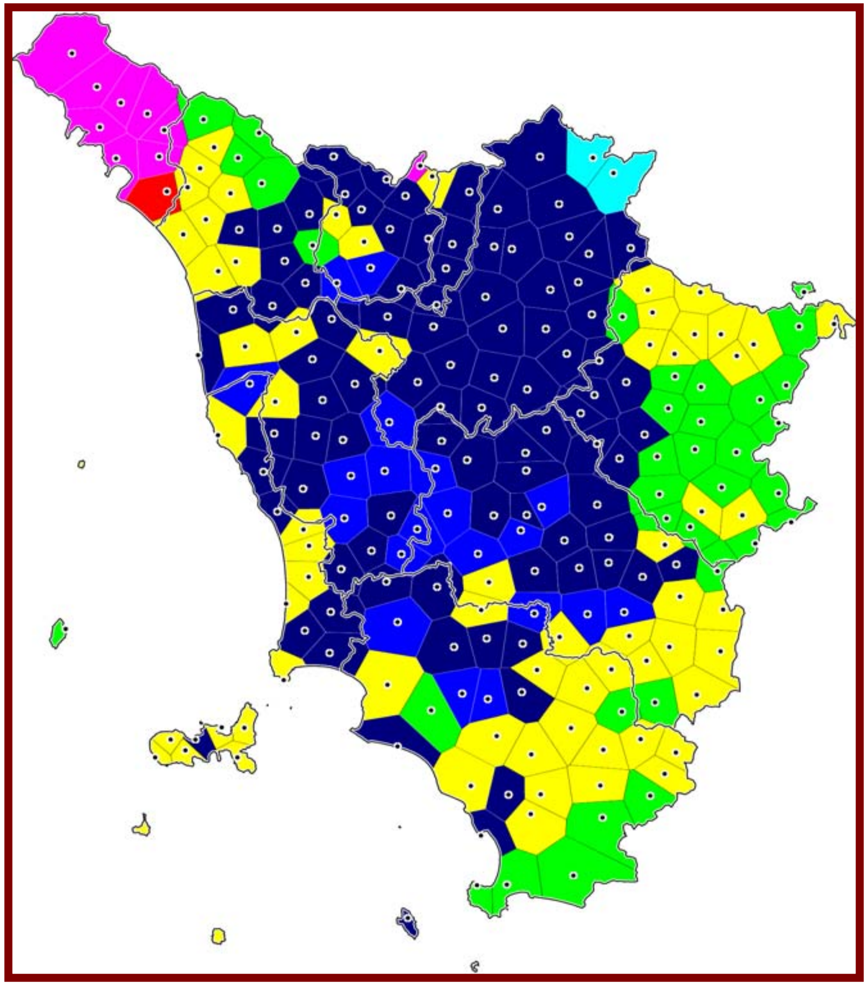
Measuring morpho-lexical distances in Tuscany (1)

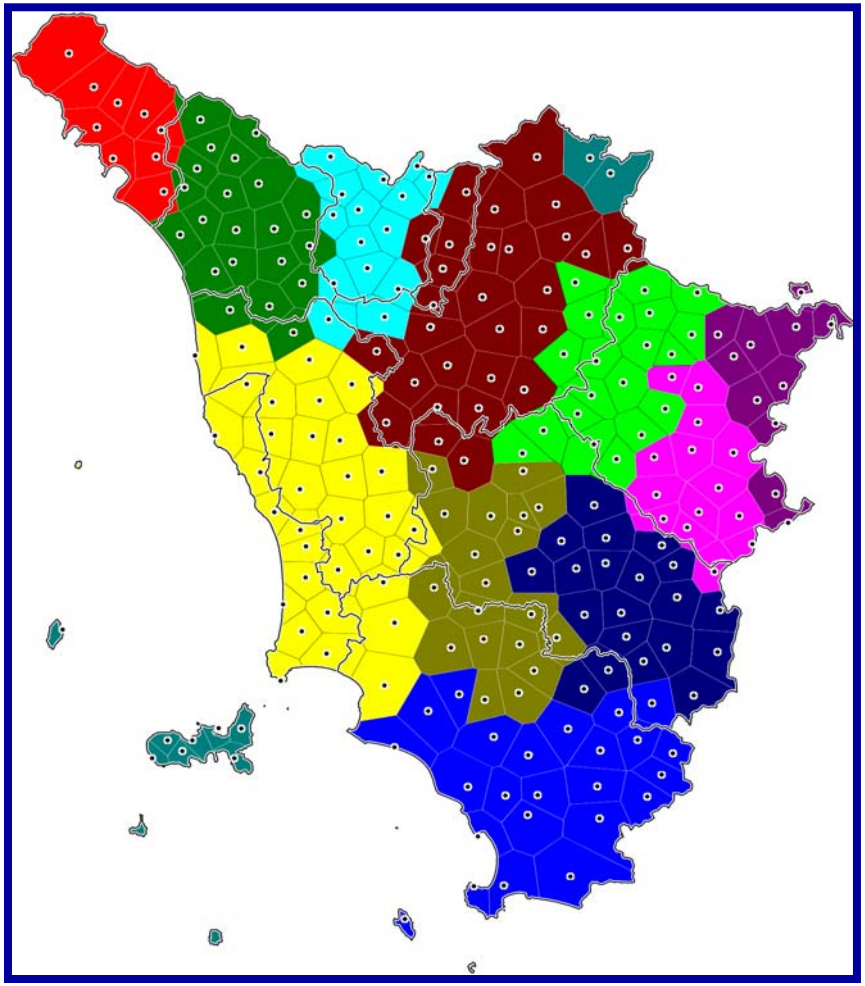
- Experimental data set
  - normalised answers to all onomasiological questions of the ALT questionnaire
- Lexical distances measured through LD against normalised representations
  - the partial similarity of related lexical items accounted for in the measure of lexical distance. The resulting measure of lexical distance reflects
    - patterns of morphological (both inflectional and derivational) variation
      - *schiacciata* vs *schiacciate*
      - *schiacciatina* vs *schiacciatello* vs *schiacciata unta*
    - but not only
      - *empitella* vs *epitella* vs *lempitella* vs *lepitella* vs *mepitella* vs *nempitella* vs *nepitella*
  - normalisation of the distance by the length of compared words

Measuring morpho-lexical distances in Tuscany (2)

- Some numbers
  - 460 questionnaire items
  - 32,448 different normalised answers
  - 227,555 geo-referentiated normalised answers
- Cronbach α = 0.97
- Morpho-lexical distances between 224 locations
  - the distance between two locations equal to the average of LDs calculated for individual word pairs
  - missing answers ignored

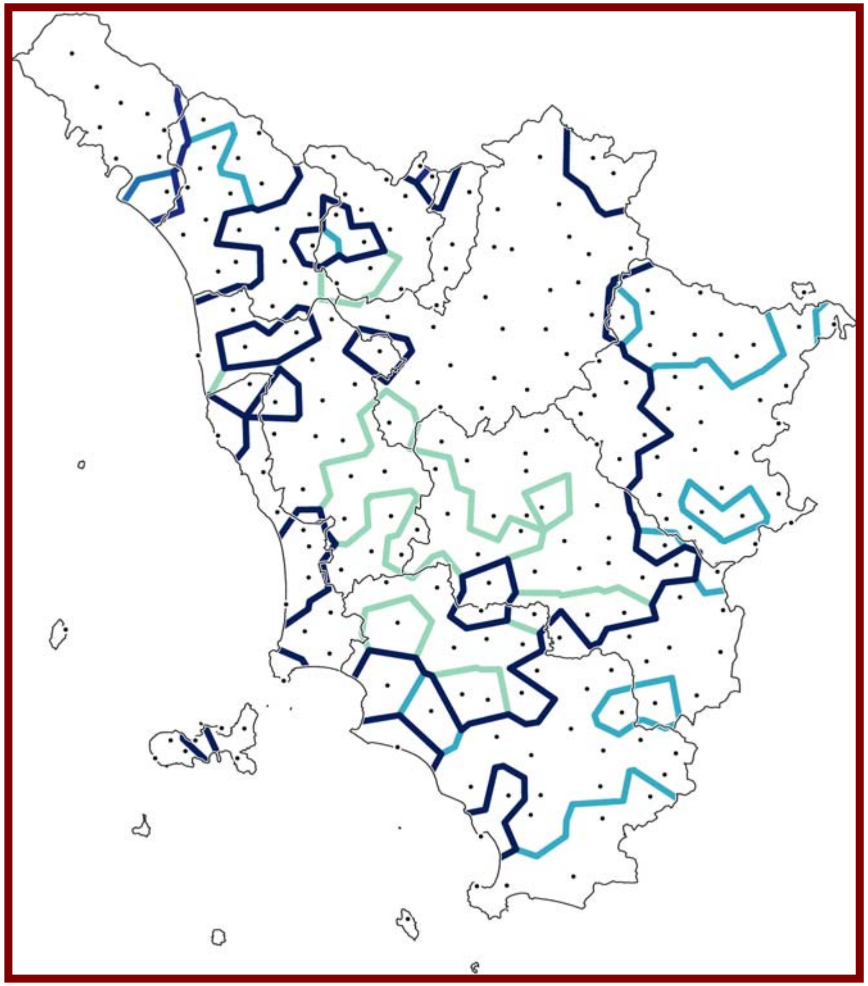
Pronunciation vs morpho-lexical variation: *clustering*

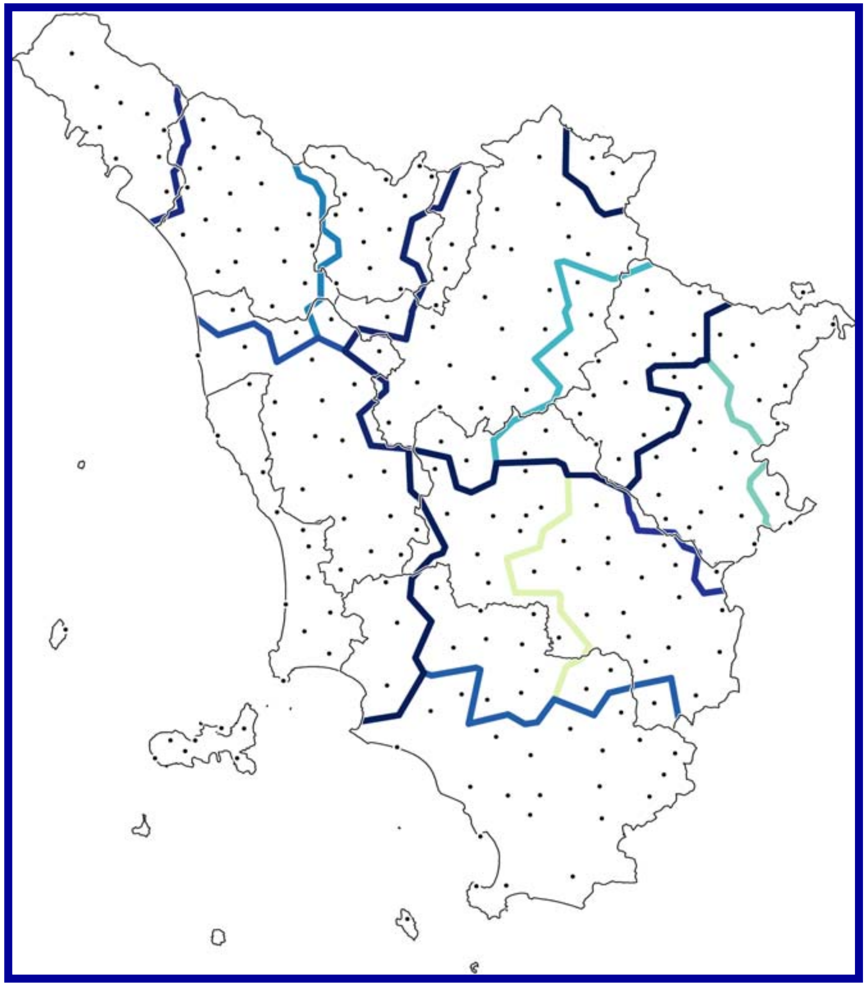




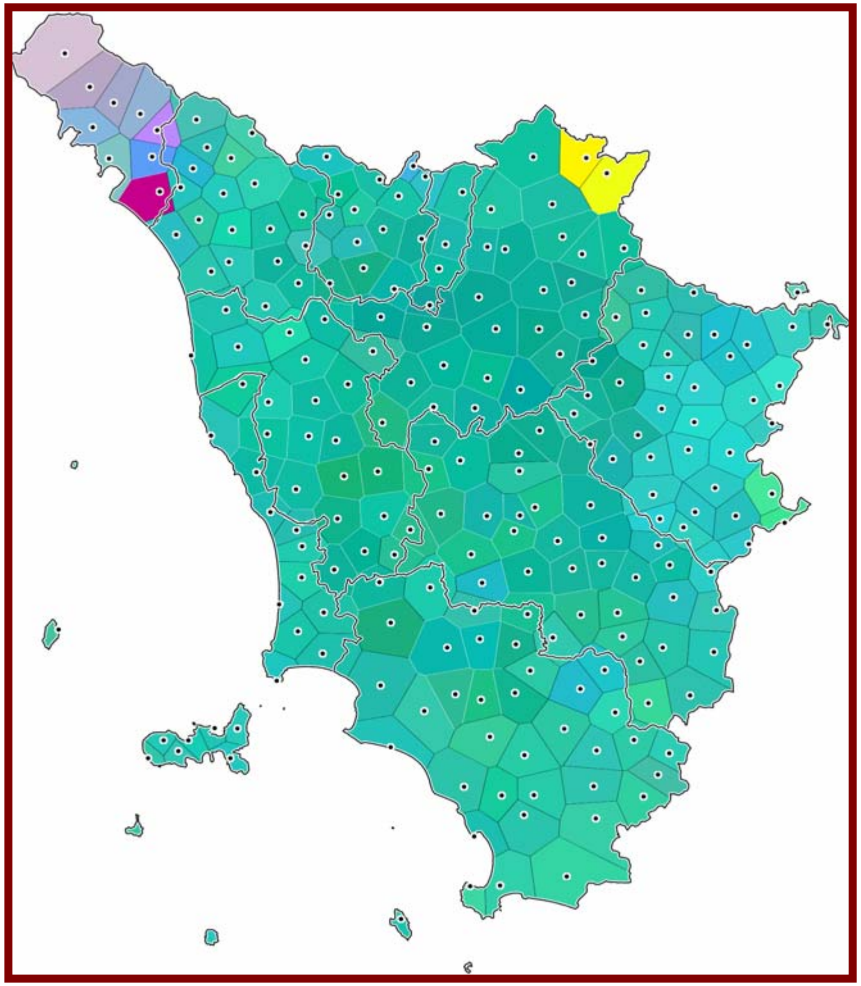
Pronunciation vs morpho-lexical variation: *cluster composite map*

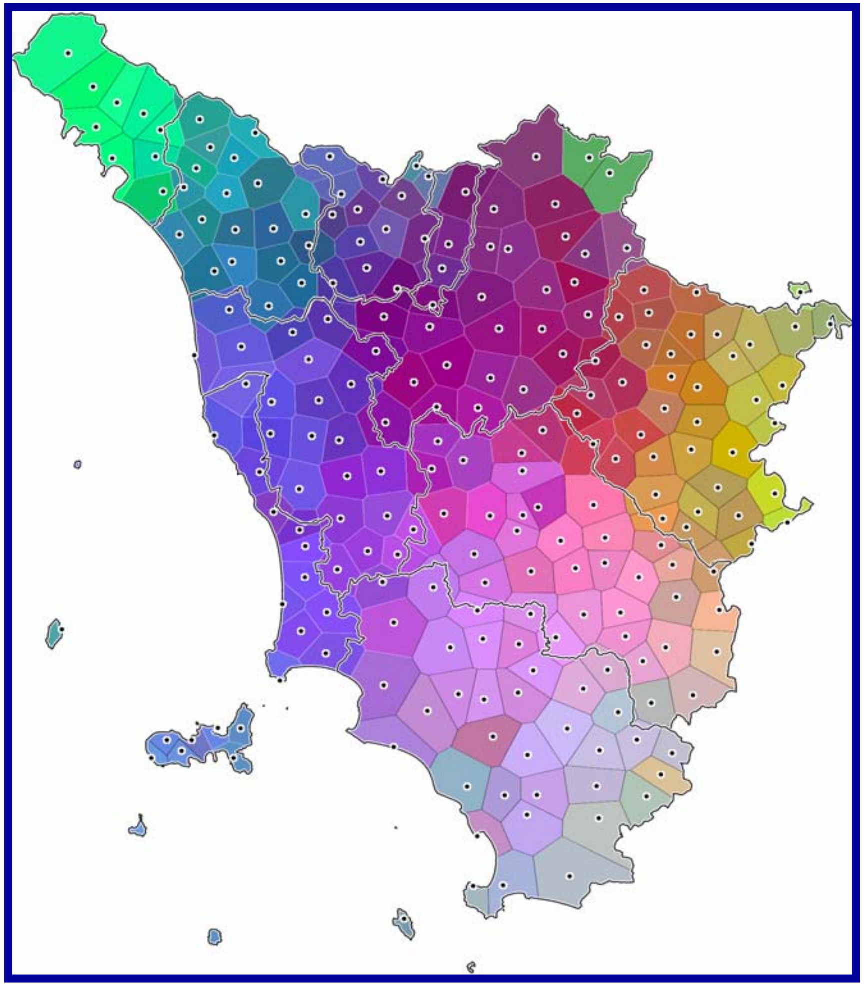
(Kleiweg *et al.* 2004)



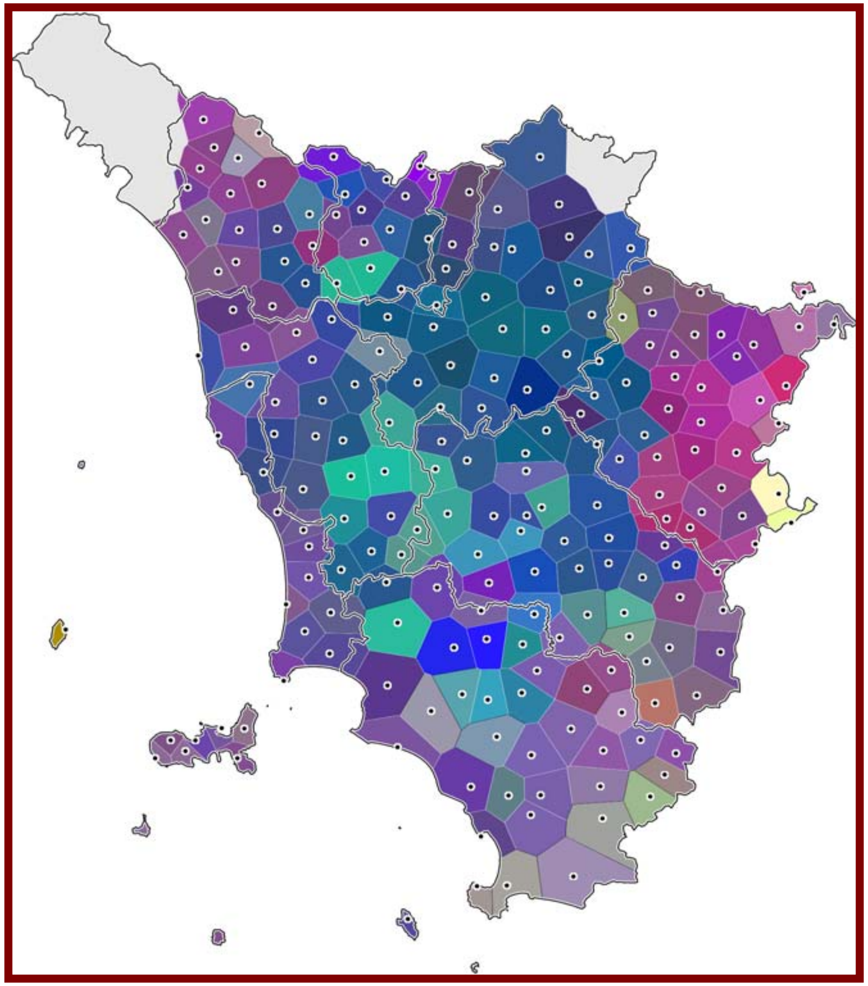


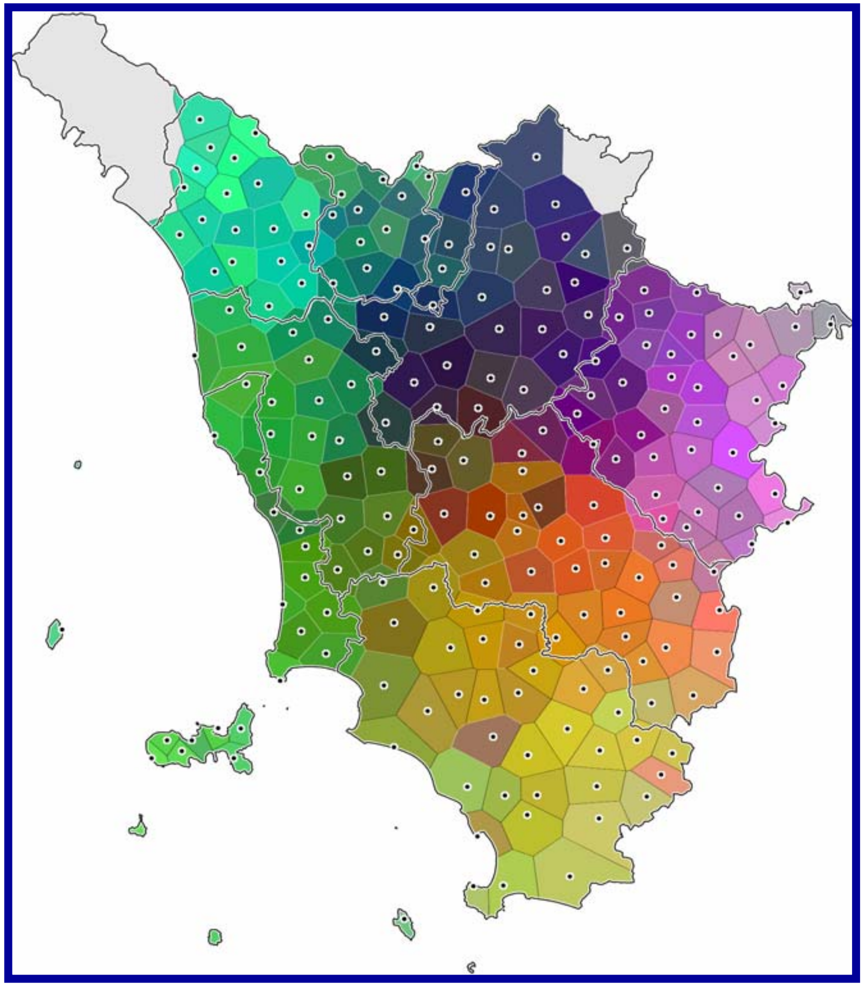
Pronunciation vs morpho-lexical variation: *MDS map*





Pronunciation vs morpho-lexical variation: *MDS map* without non-Tuscan dialects





Correlation between pronunciation vs morpho-lexical distances

- Pronunciation distance: LD calculated against phone-based representations
- For all correlations coefficients: $p < 0.0001$

Pronunciation vs morpho-lexical distances	r	r^2 * 100
All Tuscany (224 locations)	0.6582	43%
Tuscany without non-Tuscan dialects (213 locations)	0.4125	17%

- in line with Chambers and Trudgill (1998) assumption that lexical differences do not necessarily coincide with pronunciation differences
- not reflected in the analysis of the main scholar of Tuscan dialects (Giannelli 1976, 2000)

Correlation of pronunciation and morpho-lexical distances with geography

- Geographic distance calculated with ll2dst (RuG/L04) starting from the longitude- latitude coordinates
- For all correlations coefficients: $p < 0.0001$

Morpho-lexical distances	**r**	**r^2 * 100**
All Tuscany (224 loc)	0.5417	29%
Tuscany without non-Tuscan dialects (213 loc)	0.5306	28%

Pronunciation distances	**r**	**r^2 * 100**
All Tuscany (224 loc)	0.2422	5.8%
Tuscany without non-Tuscan dialects (213 loc)	0.0906	0.8%

Correlations between linguistic and geographic distances

- Different correlations observed in the literature
  - Dutch (Nerbonne et al. 1996): r=0.67
  - Norwegian (Gooskens 2004): r=0.22
  - explained in terms of differences in geography
- In Tuscany the correlation between linguistic and geographic distances appears to vary significantly across the different linguistic levels
  - considerably lower in the case of pronunciation distances
  - cannot be accounted for in terms of geography!
- What lies behind the low correlation between pronunciation and geographic distances in Tuscany?

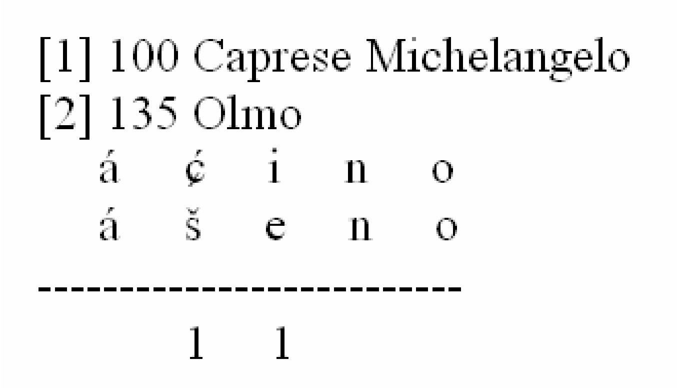
Behind identified patterns of pronunciation variation:
work in progress

- Following Kondrak (2002) and Prokic (2007), extraction of regular sound correspondences from aligned word pairs
- Focus on the aligned phonetic variants of 519 NFs selected on the basis of extra-linguistic criteria:
  - Geographical coverage: => 100 localities (out of 224)
  - Variation range: between 34 and 5
- Experimental data set
  - All Tuscany: 5,218 phonetic variants corresponding to 89,715 geo-referentiated items
  - Without non-Tuscan dialects: 3,911 phonetic variants corresponding to 86,809 geo-referentiated items
- Attested phonetic variants were aligned using RUG/L04

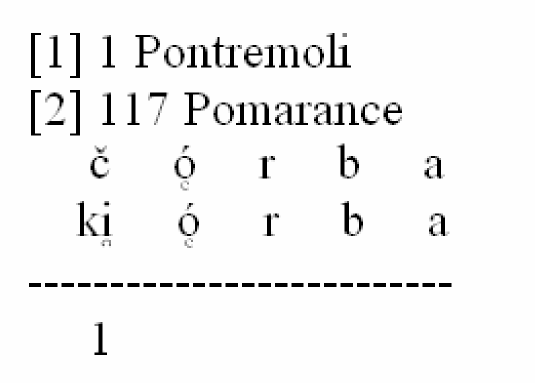
Behind identified patterns of pronunciation variation:

work in progress

- Extraction of regular sound correspondences: examples



```
[1] 100 Caprese Michelangelo
[2] 135 Olmo
    á   ḉ   i   n   o
    á   š   e   n   o
--------------------------
        1   1
```



```
[1] 1 Pontremoli
[2] 117 Pomarance
    č   ó̜   r   b   a
   ki̯   ó̜   r   b   a
--------------------------
    1
```

- Alignments were induced by enforcing the syllabicity constraint
  - only vowels may match with vowels, consonants with consonants, [j] and [w] with both
- Multiple alignments: only the first one is considered

Behind identified patterns of pronunciation variation:
work in progress

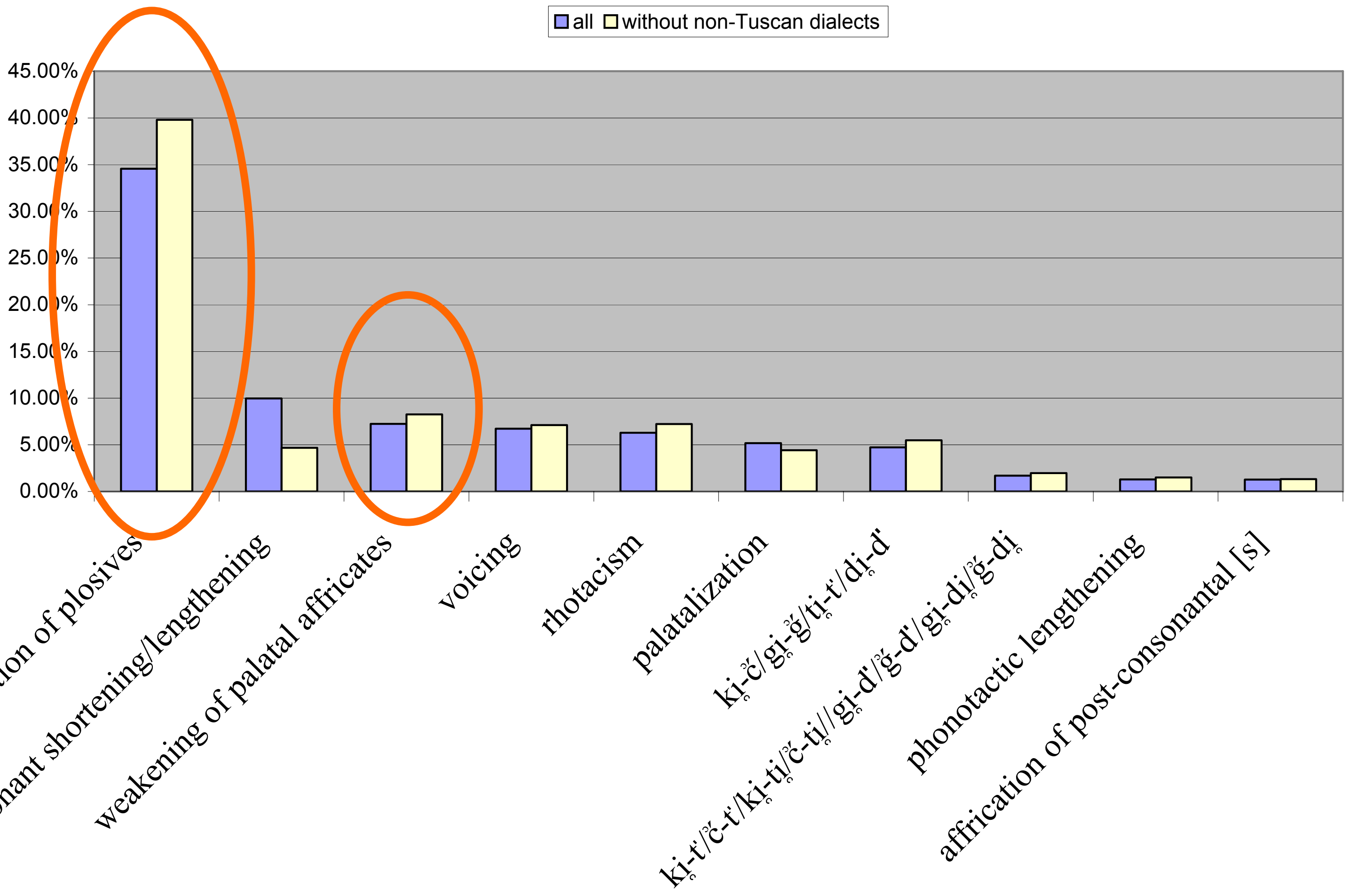
- Extraction from all aligned word pairs of both matching and non-matching phonetic segments
  - 25,132,756 segment pairs: all Tuscany
  - 21,963,192 segment pairs: without non-Tuscan dialects
- A coarse-grained classification of non-matching phonetic segments shows that consonants play a major role in Tuscan pronunciation variation

	All Tuscany		Without non-Tuscan dialects	
Vowels	1,449,840	29.72%	1,007,084	26.05%
Consonants	3,408,790	**69.88%**	2,843,826	**73.55%**
Other	19,298	0.40%	15,433	0.40%
	4,877,928	100.00%	3,866,343	100.00%

Behind identified patterns of pronunciation variation:

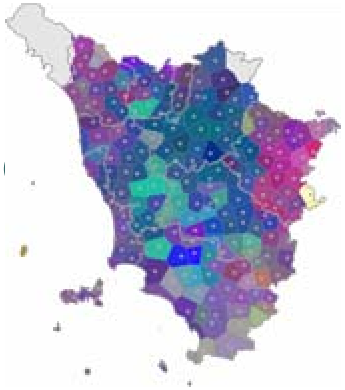
work in progress

A finer-grained classification of non-matching phonetic segments involving consonants



Behind identified patterns of pronunciation variation: work in progress

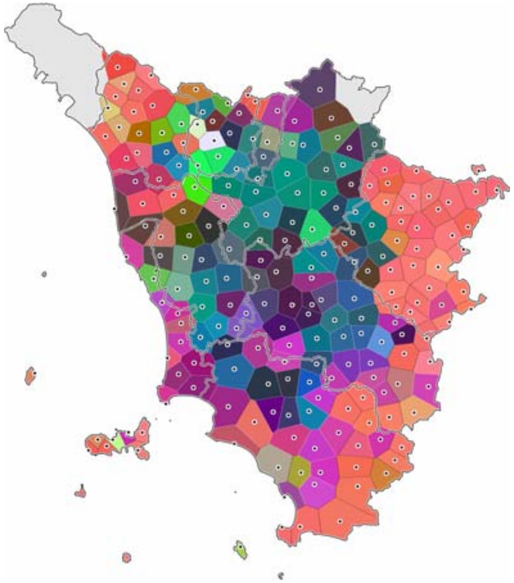
- a significant part of non-matching phonetic segments classified as spirantization phenomena
  - 42% in the case of whole Tuscany
    - 35% spirantization of plosives, e.g. /k t p/ > [h θ ϕ]
    - 7% weakening of palatal affricates, e.g. /tʃ/ > [ʃ]
  - 48% when we focus on Tuscan dialects only
    - 40% spirantization of plosives
    - 8% weakening of palatal affricates



Behind identified patterns of pronunciation variation:

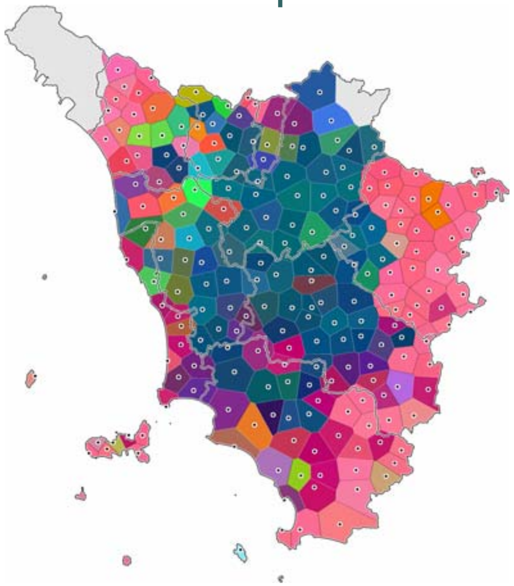
work in progress

Spirantization of plosives



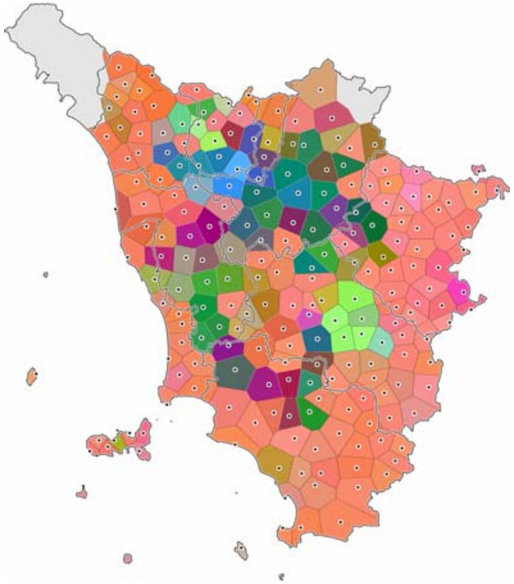
$r = 0.61$
$r^2 * 100 = 40\%$

Spirantization of voiceless plosives



$r = 0.60$
$r^2 * 100 = 35\%$

Spirantization of voiced plosives



$r = 0.31$
$r^2 * 100 = 10\%$

Correlation with overall pronunciation distances

Behind identified patterns of pronunciation variation:
work in progress

- Among the linguistic features playing a major role in determining identified pronunciation variation patterns there appears to be spirantization phenomena (so-called “Tuscan gorgia”)
    - Florence traditionally viewed as the epicenter
    - From Florence, the gorgia spreads its influence 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
    - It is present in Siena and further south whereas it does not appear in far southern Tuscany

- *Tuscan gorgia*: increasingly accepted as being a local and **innovative** (dating back the Middle Ages) natural phonetic phenomenon (consonantal weakening) spreading from the locally influential center of Florence in all directions
    - this can help to shed light on the reasons why pronunciation distances do not correlate with geographic distances: there are geographically remote areas which are linguistically similar

Conclusions

1. Whether and to what extent are observed patterns of pronunciation and lexical variation associated with one another?
   - the multi-level representation model of ALT data permitted to focus on pronunciation and morpho-lexical variation respectively, without any interference from any other level
     - orthogonal views of the same data set
   - pronunciation and morpho-lexical variation do not correlate perfectly: identified dialectal areas and continua differ significantly

2. Whether and to what extent do pronunciation and lexical distances correlate with geographic distance? If this turns out to be the case, are they expected to correlate in the same way?
   - asymmetric correlation wrt geography suggesting that pronunciation and morpho-lexical variation in Tuscany is regulated by different patterns of linguistic diffusion
     - morpho-lexical distances correlate significantly with geographic distance
     - pronunciation distances are not fully cumulative