Detecting (Social and) Geographical Varieties in Spoken Italian

.ical

References –
Auer, P. (2005). “Europe’s Sociolinguistic Unity, or: A Typology of European Dialect- Standard Constellations”. In N. Delbecque, J. van der Auwera & D. Geeraerts (Eds.), Perspectives on Variation: Sociolinguistic, Historical, Comparative. Berlin/New York: De Gruyter, 7–42. doi.org/10.1515/9783110909579.7.
Ballarè, S., Goria, E. & Mauri, C. (2022), Italiano parlato e variazione linguistica. Teoria e prassi nella costruzione del corpus KIParla. Bologna: Pàtron.
Berruto, G. (2005). “Dialect/standard convergence, mixing, and models of language contact: The case of Italy”. In P. Auer, F. Hinskens & P. Kerswill (Eds.), Dialect change. Convergence and divergence in European languages. Cambridge: Cambridge University Press, 81–97. doi.org/10.1017/CBO9780511486623.005.
Berruto, G. (2018). “The languages and dialects of Italy”. In W. Ayres-Bennett & J. Carruthers (Eds.), Manual of Romance Sociolinguistics. Berlin/New York: De Gruyter, 494–525. doi.org/10.1515/9783110365955-019.
Braunmüller, K., Höder, S. & Kühl, K. (Eds. 2014). Stability and divergence in language contact. Factors and mechanisms. Amsterdam/Philadelphia: John Benjamins.
Cerruti, M. & Tsiplakou, S. (Eds. 2020). Intermediate language varieties. Koinai and regional standards in Europe. Amsterdam/Philadelphia: John Benjamins.
Cerruti, M. & Vietti, A. (2022), “Identifying language varieties: Coexisting standards in spoken Italian”. In: Beaman, K. V. & Guy, R. G. (Eds.), The coherence of linguistic communities. Orderly heterogeneity and social meaning. New York: Routledge, 261-280
Costa, P. S., Santos, N. C., Cunha, P., Cotter, J. & Sousa, N. (2013), “The use of multiple correspondence analysis to explore associations between categories of qualitative variable in healthy ageing”. Journal of Aging Research, 2013, doi.org/10.1155/2013/302163.
D’Achille, P. (2002), “L’italiano regionale”. In: Cortelazzo, M., Marcato, C, De Blasi, N. & Clivio, G. (Eds.), I dialetti italiani. Storia, struttura, uso. Torino: UTET, 26-42.
Ghyselen, A. S., & De Vogelaer, G. (2018), “Seeking systematicity in variation: Theorethical and methodological considerations on the variety concept”. Frontiers in Psychology 2018 (9). doi.org/10.3389/fpsyg.2018.00385.
Vietti, A. (2019), “Phonological variation and change in Italian”. In: Oxford Research Encyclopedia of Linguistics. Oxford: Oxford University Press. doi.org/10.1093/acrefore/9780199384655.013.494.
Wu, T., Duchateau, J., Martens, J. P. & Van Compernolle, D. (2010), “Feature subset for improved native accent identification”. Speech Communication 52(2), 83-98. doi.org/10.1016/j.specom.2009.08.010.

Abstract –
The range of regional varieties of Italian, as well as that of most other European languages (Auer 2005; Braunmüller, Höder & Kühl 2014; Cerruti & Tsiplakou 2020), is currently affected by both convergence and stability of linguistic features (cf. Berruto 2005, 2018). Our study aims to investigate these dynamics by comparing different socio-geographically defined groups of speakers in terms of the actual clusterings of regional features in speech data. We focus on a set of 30 phonetic/phonological, morphological, syntactical and lexical features which includes both region-specific phenomena (e.g. Emilian suffix -ìzia, as in stupidìzia “stupidity”, vs. Italian suffixes -ità/-ezza, as in stupidità/stupidezza “stupidity”) and features distinguishing northern varieties from southern varieties (e.g. intervocalic [s]/[z], as in Northern [ˈkaːza] “home”, [ˈmɛːze] “month”, [ˈnaːzo] “nose”, etc. vs. Southern [ˈkaːsa] “home”, [ˈmɛːse] “month”, [ˈnaːso] “nose”, etc.); see, e.g., D’Achille 2002, Vietti 2019.
Our analysis is based on the KIPPasti corpus (Ballarè et al. 2022) that consists of around 40h of dinner table conversations collected in different Italian regions, involving speakers with different social characterization (in terms of gender, age, educational degree, and type of employment). More specifically, we consider a subcorpus composed of 16 hours of spoken interactions (8 h recorded in northern regions -namely Veneto and Emilia Romagna- and 8 h recorded in southern regions – namely Puglia and Campania) to focus on the comparison between northern and southern regional varieties.
In order to examine the influence of the aforementioned features in the identification of the hypothesized groups, we explore statistical analyses on multiple tiers. As commonly done in similar studies (Ghyselen & De Vogelaer 2018), the underlying idea is to reduce the number of dimensions used to describe the data in order to evaluate their contribution to discriminate the different varieties. To do so, we employ three different quantitative methods:
1) Correspondence analysis (CA), which allows for detection of clusters of features that behave alike and therefore can be considered as defining features of a specific variety (Costa et al. 2013);
2) Dimensionality reduction techniques, i.e. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), which, by selecting and aggregating dimensions that best approximate the full dataset, yield a representation where geometrical properties of the objects involved can be linguistically interpreted (Cerruti & Vietti 2022);
3) Support Vector Machine (SVM) classification (a powerful supervised algorithm based on the construction on maximum-margin hyperplanes that best separate the considered categories) which, through the Recursive Feature Elimination (RFE) procedure yield an estimate of the contribution of each dimension to the classification hypothesis (Wu et al. 2010).
The analysis shows (i) on the one hand, that regional differentiation is becoming less pronounced, partly due to the supra-regional use of features originally confined to distinct areas and partly because several region-specific features are now exclusive to elderly (and poorly educated) speakers, and (ii) on the other hand, that differences between regional varieties of Italian still persist,

as the socially unmarked usage slightly differs across regions.