Exploring the correlation between pronunciation and lexical variation in Tuscany

Simonetta Montemagni

Istituto di Linguistica Computazionale, Pisa (Italy)

It is a well-known fact that different types of features contribute to the linguistic distance between any two locations, which can differ for instance with respect to the word used to denote the same object or the pronunciation of a particular word. Yet, the correlation between different feature types in defining patterns of dialectal variation still represents a research area to be explored. If on the one hand the traditional notion of dialectal variation might suggest that a high degree of correlation should be expected among features pertaining to different linguistic levels, on the other hand theoretical studies suggest that patterns of dialectal variation based on different types of features do not necessarily coincide (Chambers and Trudgill, 1998). In traditional dialectology, there is no obvious way to approach the problem beyond fairly superficial and impressionistic observations. The situation changes if the same research question is addressed in the framework of dialectometric studies, where it is possible to measure dialect distances with respect to different linguistic levels and to compute whether and to what extent observed distances correlate. Pioneering work in this direction is represented by Nerbonne (2003), Gooskens and Heeringa (2006) and Spruit et al. (in press): the results of these studies show that this is a promising line of research.

The main goal of this work is to investigate the degree to which patterns of dialectal variation computed with respect to different linguistic levels correlate in the language varieties spoken in an Italian region, Tuscany. The study is carried out on the entire corpus of dialectal data of an online dialectal resource, ALT-Web (Cucurullo et al. 2006), and builds on the results of a dialectometric study focussing on pronunciational and lexical variation in Tuscany (Montemagni 2007). By exploiting a multi-level representation model of dialectal data, the linguistic distances among 224 locations have been measured with the Levenshtein distance with respect to different linguistic levels (pronunciation and lexicon). Correlational analyses have then been performed on the resulting distance matrices in order to estimate the degree of association between the different levels: in our case such a correlation does not appear to be particularly strong. The role of geography and other extralinguistic constraints (including sex, age, and socio-cultural status of informants) is also being investigated.

References

Chambers J.K., Trudgill P., 1998, *Dialectology* (2nd Edition), Cambridge University Press, Cambridge.

Cucurullo S., Montemagni S., Paoli M., Picchi E., Sassolini E., 2006, *Dialectal resources on-line: the ALT-Web experience*. In: *Proceedings of LREC-2006*, Genova (Italy), May 2006.

Gooskens C., Heeringa W., 2006, *The Relative Contribution of Pronunciation, Lexical and Prosodic Differences to the Perceived Distances between Norwegian dialects*. "Literary and Linguistic Computing", Vol. 21, n. 4 pp. 477-492.

Montemagni S., 2007, *Patters of phonetic variation in Tuscany: using dialectometric techniques on multi-level representations of dialectal data*, in P. Osenova (ed.) *Proceedings of the RANLP Workshop on Computational Phonology* Workshop at the conference *Recent Advances in Natural Language Phonology* Borovetz, 2007.

Nerbonne J., 2003, *Linguistic Variation and Computation*, in *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*, April, 2003. pp.3-10.

Spruit M.R., Heeringa W., Nerbonne J., in press, *Associations among Linguistic Levels*, accepted for publication in a Special Issue on Syntactic Databases of "Lingua", available at http://marco.info/pro/pub/shn2007dh.pdf.