A corpus-based study of lexical uniformity in the standardization of Italian

One of the fundamental aspects that defines the standardization phase of many European languages, is the development towards more uniformity and less variability (Milroy, 2001). Although it is still debated which degree of variability is acceptable in order to consider a linguistic situation to be standardized (Soares da Silva, 2010), lexicon-oriented quantitative standardization research has devised specific measures to grasp these dynamics (Daems, Heylen, & Geeraerts, 2015).

The Italian situation is not different from other European languages, in that nation building efforts after the political unification in 1861 led to the programmatic reduction of lexical variation. But the peculiarity of Standard Italian is the historical overabundance of formal variants of the same word, sometimes called *allotropi* (D'Achille, 2010), whose single etymological base developed along different paths and whose reflexes eventually (re-)entered the Italian language from different sources and in different periods (borrowings, Latinisms and analogical formations). Well-known examples are the doublets '*gioco/giuoco*' (IŎCUS "game") and '*veduto/visto*' (VĬSUM "seen"). Earlier corpus-based investigations have mainly focused on a limited number of alternations and have only briefly touched on the sociolinguistic distribution of these variants (Thornton, 2012).

The goal of this study is to frame this phenomenon more explicitly in previous quantitative standardization research, and to scale up the analysis by looking at 5 sets of roughly 10 lexical variables that exemplify a particular alternation type (eg.: absence/presence of: orthographical rendition of syntactic gemination, mobile diphthongs, raised vowels in Latin prefix *'re-'*, etc.). The frequencies of each variant will be extracted from the DiaCORIS (28 mln tokens), a diachronic corpus of Italian texts which covers a period from 1861 to 2001, and that includes multiple written genres sampled from different areas in Italy. The data will be analyzed by building different mixed-effects logistic regression models per alternation set and compare them.

Given the evident reduction of formal lexical variability in the Italian language over the past century and a half, the central research question to be answered is whether the influence of typical sociolinguistic dimensions can help explain the specific dynamics of this process of reduction. Which areas of the peninsula lead this development to more uniformity, and which lag behind? Do variants associated with literature supplant more informal variants, or did the opposite happen? As to the internal factors that played a role, we also ask whether this reduction is associated with the loss or acquisition of certain fixed lexical patterns that involve these variants.

References

D'Achille, P. (2010). *L'italiano contemporaneo* (2nd [2003]). Bologna: Il Mulino.

Daems, J., Heylen, K., & Geeraerts, D. (2015). Wat dragen we vandaag: een hemd met blazer of een shirt met jasje? *Taal En Tongval*, *67*(2), 307–342.

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, *5*(4), 530–555.

Soares da Silva, A. (2010). Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in Cognitive Sociolinguistics* (pp. 41–84). Berlin, Boston: De Gruyter Mouton.

Thornton, A. M. (2012). Reduction and maintenance of overabundance. A case study on Italian verb paradigms. *Word Structure*, *5*(2), 183–207.