Degrees of similarities between Spanish and Portuguese varieties

Shuxian Pan
Universität Stuttgart
shuxian.pan@ ims.uni-stuttgart.de

Sabine Schulte im Walde
Universität Stuttgart
schulte@ ims.uni-stuttgart.de

Spanish and Portuguese originated as Romance languages in the Middle Ages. During the discovery and colonisation of Latin America, these languages spread and various varieties developed. Nonetheless, similarities between these varieties have rarely been examined in a large-scale data-driven manner. In this study, we provide a computational approach to compare language variants of Spanish and Portuguese in Europe and Latin America based on corpus data modelling and evaluation metrics.

We hypothesise that these degrees of similarities differ, due to shared and separate morphosyntactic language features that developed differently over time and regions. The selection of the morphosyntactic features is partially motivated by theoretical linguistic observations, including a) conjugations of verbs, b) positions of clitics, and c) pronoun types. All morphosyntactic features are extracted as n-grams and their frequencies in corpora across language varieties. To quantify differences in distributions, we calculate cosine distance, Kullback-Leibler divergence (Bizzoni et al., 2020) and χ^2 (Kilgarriff, 2001). In a more explorative series of analyses, we use z-score to rank n-grams of words and part-of-speech tags (Frassinelli et al., 2021). By evaluating the metric scores, we hope to confirm linguistic insights and in addition to empirically identify morphosyntactic features that may have been missed in previous observations. We use these metrics to determine not only how similar the varieties are in terms of certain features, but also whether any particular feature is more prominent in one variety than in the other. In a final step, we will use the metric scores to visualise the degrees of similarities between our two target languages and their varieties.

References: Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., & Teich, E. (2020): Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data–Driven Approach. *Frontiers in Artificial Intelligence* 3, 1–15. • Frassinelli, D., Lapesa, G., Alatrash, R., Schlechtweg, D., & Schulte im Walde, S. (2021): Regression Analysis of Lexical and Morpho-Syntactic Properties of Kiezdeutsch. *Proceedings of the 8th Workshop on NLP for Similar Languages, Varieties and Dialects*, 21–27. • Kilgarriff, A. (2001): Comparing Corpora. *International Journal of Corpus Linguistics* 6(1), 97–133.