Combining the frequency-based and string similarity metrics for measuring the language distance between non-standardized Italian and Croatian idioms of the 18th century

Ilia Afanasev
National Research University Higher School of Economics
szrnamerg@gmail.com

Olga Lyashevskaya
National Research University Higher School of Economics, Vinogradov Russian Language Institute RAS
olesar@hse.ru

Viacheslav Kozak
Institute for Linguistic Studies RAS
viacheslav.kozak@gmail.com

Keywords: computational dialectology, language distance, string similarity measures, frequency-based metrics

Modern linguistics witnesses the rise of the need to develop a trustworthy inventory of methods for computing the language distance (List, 2021) that allows one to reliably split a set of language idioms into clusters automatically. In historical linguistics, this task is usually complicated by a small amount of available data.

Dialectal studies have proposed different methods to deal with the issue on the material of modern languages. String similarity measures and frequency-based metrics proved themselves to be one of the most efficient (Gamallo et al., 2017; Gooskens, Heeringa, 2004). However, they were used on the prepared lists of words, mainly Swadesh lists (Gooskens, Heeringa, 2004) that have not been intended for studying raw texts from historical corpora. The question of whether these methods are going to retain their efficiency on historical data is the main subject of the research.

The output of each of these metrics may be used for two main purposes. The obvious one is to automatically cluster the idioms under consideration. The second purpose is to collect some insights on how the research output may be interpreted linguistically. Which n-grams contribute the most to each of the metrics? Which metric is the most efficient or explainable? Which one is the greatest at the simulation of an actual scientist's work? These are the questions that are to be addressed in the research. And the linguistic interpretability of the results is going to be the main instrument to evaluate the actual efficiency of the metrics.

The proposed method may be described as follows. The initial step is to split tokens of the analyzed texts into character n-grams and to sort them by their frequency rate for each of the supposed idioms. After that, two metrics are applied for the idioms in the dataset pairwise. First of these metrics is Damerau-Levenshtein distance, a string similarity measure that is designed to find the minimal number of operations (such as deletion/insertion, transposition, substitution) required to transform one string into another (Damerau, 1964; Levenshtein, 1966). For each n-gram, the minimal Damerau-Levenshtein distance is scored (which means that the vector of such distances to each n-gram of the opposite idiom is acquired, and then the minimal one is chosen). The second metric is DistRank (Gamallo et al., 2017), which takes the ranks in the frequency dictionary for corresponding n-grams within two idioms, subtracts one from another, and averages the absolute value of resulting expressions. The results of both DistRank and Damerau-Levenshtein distance measurements (both calculated pair-of-tokens-wise) are gathered into two lists of numbers. Each value in these lists is normalized (by splitting by overall number of values in list). Finally, the average value for each of these lists, and thus, the average value for each metric, is acquired. The experiments are conducted on the different fractions of the dataset to illustrate how the size of the sample affects the metrics.

The material under consideration is a set of manuscripts, representing a single text — the decree of the Venetian prince of Zadar Zuanne Moro from 1762. The text was composed in Italian and translated in two written variations of Croatian that differ on all linguistics levels, cf. the orthography of [j] in

imaiu/*imaû*, reflexes of PSl. *stj in *karĉanski*/*karščanski*, morphological forms *sudcem*/*sudcima*, or lexical pair *licencie*/*testir* (*licenza*).

In the conclusion, clustering, based on measuring the language distance between the three idioms, is visualized. The method allows one to automatically distinguish both close and distant languages and cluster non-standardized language varieties. The study contribute to the reconstruction of linguistic landscape, localization, and dating of manuscripts.

References

1. Damerau, F.J. (1964). *A technique for computer detection and correction of spelling errors.* Communications of the ACM vol. 7 no. 3, pp. 171–176.
2. Gamallo P., Campos J.R.P., Alegria I. (2021). *From language identification to language distance*. Physica A: Statistical Mechanics and its Applications vol. 484, pp. 152 – 162.
3. Gooskens C., Heeringa W. (2004). *Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data*. Language Variation and Change vol. 16 no. 3, pp. 189 – 207.
4. Levenshtein, V.I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady vol. 10 no. 8, pp. 707–710.
5. List, J.-M. (2021). *Computer-assisted approaches to historical language comparison.* Habilitation Thesis, Friedrich-Schiller-Universität, Jena.