
Tom Ruelle, Dirk Speelman, Dirk Geeraerts
Quantitative Lexicology and Variational Linguistics, University of Leuven
tom.ruelle@arts.kuleuven.be, dirk.speelman@arts.kuleuven.be,
dirk.geeraerts@arts.kuleuven.be

Aggregating and interpreting lexical alternation variables. Benefits of Weighted Multidimensional Scaling for lectal categorization

Research question

In this corpus-based sociolectometric study, we quantify the (lexical) difference between language varieties by aggregating a large number of lexical alternation variables with a technique that prevents variable-level details from being obscured by the aggregation step. This study consists of two parts: (a) finding an abundance of lexical alternation variables, and (b) carefully aggregating variables to find lectal patterns and to investigate the individual behavior of the variables. For step (a), we will fall back on a pre-made list of lexical alternation variables (Martin 2005). In the broader frame of this study, the lexical alternation variables are automatically generated on the basis of the similarity measures of semantic Vector Space Models (Turney & Pantel 2010; Peirsman et al. 2010). This aspect of the study will receive less attention in this paper, in which we focus on the aggregation procedure of step (b).

An inherent problem of aggregation techniques, as used in dialectometry (Goebel 1982; Nerbonne 2006) and sociolectometry, is that information of the individual variables is obscured or lost. This seems to be the price to pay for being able to make general claims, based on large amounts of data. Although a common metaphor is that the dialectometricist wants to see THE FOREST FOR THE TREES (see e.g. Szmrecsanyi, to appear), the loss of the “trees” (i.e. the linguistic variables) is problematic for our current purpose. We want to link the lectal patterns back to (types of) linguistic variables. Previous studies from both dialectometry and sociolectometry relied on manual scrutiny of complex in-between steps (Geeraerts et al. 1999; Soares da Silva 2010) or statistical comparison of aggregation solutions (Spruit et al. 2009) for assessing this link. The current study now proposes an aggregation method that grants more transparent access to the behavior of the individual variables.

Methodology

The cornerstone of the proposed method lies in the application of Weighted Multidimensional Scaling (WMDS, also referred to as “Individual Differences Multidimension Scaling”, see Cox & Cox (2001)). Traditionally, all variables are aggregated into one distance matrix that averages the behavior of every individual variable in each lect. Then, a dimension reduction technique takes this single distance matrix and identifies a coordinate for every lect on the retrieved dimensions, allowing for direct visualizations (Speelman et al. 2003) or further analysis (Szmrecsanyi 2011). WMDS, on the other hand, has no (theoretical) restrictions on the amount of distance

matrices that can be used. As a consequence, we are not forced to aggregate all variables into one distance matrix — exactly this caused the loss of information at the variable level in a traditional approach — and we can create a distance matrix for every single variable (or for pre-defined subgroups of variables). Using multiple distance matrices in a lectometric study is not new (see e.g. Spruit et al. 2009, mentioned above); the application of WMDS in a sociolectometric study, however, is innovative.

The outcome of a WMDS analysis consists of two parts. On the one hand, a single reduced space is returned, which is very similar to a typical MDS solution. Therefore, the WMDS approach is compatible with existing methodologies. On the other hand, WMDS also returns a *weight configuration space*, which gives information on the importance (weight) of every single input distance matrix for every dimension. This weighting coefficient is thus the key to an interpretation of the behavior of the variables.

Results and conclusions

In order to show the application of WMDS, we set out to verify the classification of “typical Belgian Dutch words” in the “Referentielijst Belgisch Nederlands” (Reference List Belgian Dutch, RBBN), described in Martin (2005). For a quick overview of the Dutch situation, see Section 4.1 of Geeraerts (2003). The verification here will be based on a very large corpus that combines spontaneous conversations, Usenet posts, popular and quality newspaper articles and official government announcements from both Belgium and The Netherlands. The incorporation of registers in this regionally patterned corpus will allow us to see the multivariate strength of a WMDS-based sociolectometric study.

The RBBN classified more than 4000 Belgian Dutch words manually into categories. Here, we will focus on the categories “colloquialisms” and “unique variants”. If we perform a traditional sociolectometric study (cf. Speelman et al. 2003), we see that texts in our corpus are divided in two dimensions on the basis of a strong regional difference and a slightly weaker register difference. From these results, one might assume that the “unique variants” and the “colloquialisms” are together responsible for the regional dispersion, but that the “colloquialisms” alone cause the register dimension. This assumption — which is basically the assumption that the categorization in the RBBN is accurate — can now be tested with WMDS.

In the WMDS approach, every variable on the RBBN (a combination of a Belgian Dutch word and its Netherlandic Dutch counterpart, hence a lexical alternation variable) is used to create its own distance matrix. These matrices are the input for the actual WMDS calculation. The first part of the WMDS outcome, the single, reduced space, shows the region and register variation on separate dimensions, just like the traditional approach. From the weight configuration space, however, we can see which variables “have more weight” on each of these dimension. It appears that the RBBN classification in “colloquialisms” and “unique variants” is distinctly related to the register and regional dimensions, yet not in a categorical fashion. From this, we conclude that the WMDS approach adds transparency and interpretability to a sociolectometric study. The comparison of the weight configuration space with the manual RBBN categorization adds trust in the accuracy of the methodology.

References

Cox, Trevor and Cox, Michael 2001. *Multidimensional Scaling*. Chapman & Hall.

- Geeraerts, D., Grondelaers, S. and Speelman, D. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding-en voetbaltermen*. Amsterdam: Meertens Instituut.
- Geeraerts, Dirk 2003. Cultural models of linguistic standardization. *Cognitive Models in Language and Thought. Ideology, Metaphors and Meanings* Dirven, René, Frank, Roslyn and Pütz, Martin (eds.), 25-68. Berlin: Mouton de Gruyter.
- Goebel, Hans 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Oesterreichische Akademie der Wissenschaften.
- Martin, Willy 2005. *Het Belgisch-Nederlands anders bekeken: het Referentiebestand Belgisch-Nederlands (RBBN)*. Tech. rept. Vrije Universiteit Amsterdam, Amsterdam, the Netherlands.
- Nerbonne, John 2006. Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21, 463-476.
- Peirsman, Yves, Geeraerts, Dirk and Speelman, Dirk 2010. The Automatic Identification of Lexical Variation between Language Varieties. *Natural Language Engineering* 16(4), 469-490.
- Soares da Silva, Augusto 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In *Advances in Cognitive Sociolinguistics*, Geeraerts, Dirk, Kristiansen, Gitte and Peirsman, Yves (eds.). Berlin/New York, Mouton de Gruyter.
- Speelman, Dirk, Grondelaers, Stefan and Geeraerts, Dirk 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317-337.
- Spruit, Marco Rene, Heeringa, Wilbert and Nerbonne, John 2009. Associations among Linguistic Levels. *Lingua* 119(11), 1624-1642.
- Szmrecsanyi, Benedikt 2011. Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1).
- Szmrecsanyi, Benedikt to appear. Aggregate data analysis in variationist linguistics. In *Research Methods in Language Variation and Change*, Krug, Manfred and Schlüter, Julia (eds.). Cambridge: Cambridge University Press.
- Turney, Peter and Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.
-