Multidimensional scaling and linguistic theory

Martijn van der Klis and Jos Tellings*

December 10, 2020

word count: 9947

Abstract

This paper reports on the state-of-the-art in the application of multidimensional scaling (MDS) techniques to create semantic maps in linguistic research. MDS refers to a statistical technique that represents objects (lexical items, linguistic contexts, languages, etc.) as points in a space so that close similarity between the objects corresponds to close distances between the corresponding points in the representation. We focus on the recent trend to apply MDS to parallel corpus data in order to investigate a certain linguistic phenomenon from a cross-linguistic perspective.

We first introduce the mathematical foundations of MDS, intended for non-experts, so that readers understand notions such as 'eigenvalues', 'dimensionality reduction', 'stress values', etc. as they appear in linguistic MDS writing.

We then give an exhaustive overview of past research that employs MDS techniques in combination with parallel corpus data, and propose a set of terminology to succinctly describe the key parameters of a particular MDS application. We go over various research questions that have been answered with the aid of MDS maps, showing that the methodology covers topics in a spectrum ranging from classic typology (e.g. language classification) to formal linguistics (e.g. study of a phenomenon in a single language).

We finally identify two lines of future research that build on the insights of earlier MDS research described in the paper. First, we envisage the use of MDS in the investigation of cross-linguistic variation of compositional structures, an important area in variation research that has not been approached by parallel corpus work yet. Second, we discuss how MDS can be complemented and compared with other dimensionality reduction techniques that have seen little use in the linguistic domain so far.

1 Introduction

Multidimensional scaling (henceforth MDS) is a statistical technique that represents objects in a dataset as points in a multidimensional space so that close similarity between objects in the dataset corresponds to close distances between the corresponding points in the representation. Typically, MDS reduces a dataset that has variation in a large number of dimensions, to a representation in only two or three dimensions. MDS can therefore be seen as a dimensionality reduction technique, which makes it possible to graphically represent a highly complex

*Author names are in alphabetical order.

dataset as a 2D or 3D scatterplot. We will call such a visualization obtained by means of MDS an **MDS map**.

MDS as a statistical and visualization tool has been used in various fields of science (see e.g. Ding 2018). Recently, researchers in linguistic typology started using MDS as a method to chart cross-linguistic variation in semantic maps. One can distinguish two ways in which MDS has been employed for the creation of these maps. First, MDS maps can be seen as a step in the development of *classical semantic maps* (van der Auwera 2013; Georgakopoulos and Polis 2018). Whereas classical semantic maps are graph-like structures that are compiled from linguistic data by hand, MDS has been used to generate semantic maps directly from linguistic data. Used in this way, MDS maps are thus a computationally extendable and mathematically rigorous development of classical semantic maps. Second, and more recently, MDS has been applied in linguistics in a way that is unrelated to classical semantic maps. In this case, MDS is employed to map data from parallel corpora, with the goal to investigate a certain linguistic phenomenon from a cross-linguistic perspective. In this paper we will focus on this latter trend in the use of MDS in linguistics, although the link between MDS and classical maps will be discussed in passing as we follow the chronological development of MDS maps in section 3.[1]

We stress that an MDS map does not stand for a single concept: MDS is a technique that can be used to generate various kinds of maps, and they show different things and have different functions in the context of linguistic argumentation. In order to be clear what a given MDS map represents, we will discuss different sorts of MDS maps on the basis of three parameters: input data (what sort of linguistic data have been used as input for the MDS algorithm?), similarity measure (how is similarity between primitives defined?), and output (what do the data points on the output map represent?). We would like to push this forward as a concise way to provide the essential information for understanding how a given MDS map was construed.

Before we proceed to discuss several MDS maps along these parameters, we first describe the mathematical foundations of MDS (section 2). This helps to understand the fundamental role of similarity in the construction of the maps, and familiarizes the reader with some essential terminology (eigenvalues, stress factors, dimensionality) that is needed to understand the central concepts of multidimensional scaling. Then, in section 3, we review various MDS maps that can be found in the literature since Croft and Poole (2008). We look at the various types of linguistic input data, and explain how these MDS maps were constructed.

Section 4 covers how MDS maps can be interpreted by analyzing the dimensions of the map and the clustering of points, both by informal inspection and with the help of statistical tools. We also describe how this interpretation process links up to linguistic theory, by reviewing the types of research questions that MDS maps have been used to answer (section 4.3).

In section 5, we indicate promising future developments of MDS in the linguistic domain. Section 6 concludes.

[1] This paper is only concerned with MDS used as a method to create semantic maps. Other uses, such as the creation of areal maps in dialectometry (Wieling and Nerbonne 2015) will not be discussed. Moreover, this paper does not aim to provide a practical user guide about software packages that implement MDS, see for instance Croft and Timm (2013) or Borg and Groenen (2005, Appendix A) for this purpose.

2 Mathematical background on classical scaling

This section is a gentle introduction to some of the mathematical concepts behind MDS. It is intended for readers who do not have a background on matrix algebra, but want to understand notions such as 'eigenvalues' and 'stress factor' that are used in the linguistic MDS literature. Much more thorough expositions on the mathematics behind MDS are available, for example Borg and Groenen (2005).

Although in the linguistic literature the label 'multidimensional scaling' is typically used without further qualification, MDS actually stands for a family of methods and procedures consisting of numerous variants that have been developed for different applications. Here, we will introduce in some detail the version of MDS that is usually known as *classic scaling* or *classic MDS*, or more fully as *classic metric Torgerson scaling*, named after the work of Torgerson (1952). We opt for this variant for expository reasons – it is the conceptually simplest model, and contains the core concepts needed to understand multidimensional scaling and its related technical concepts.

Classical scaling is one of three MDS algorithms that have been used in linguistic applications of MDS, the other two being an iterative procedure known as SMACOF, which searches for a best fitting solution, and an algorithm known as Optimal Classification (OC) MDS. These other two algorithms will be introduced briefly in section 2.4, and illustrated further as we discuss the work in which they are used in section 3. For brevity of reference, we will continue the terminological abuse by referring to classical scaling simply as 'MDS' in this section.

The main component of MDS is a process called *eigendecomposition*. This is a more generic process that is also used in other statistical techniques, such as *Principal Component Analysis* (Jolliffe and Cadima 2016). What is specific about MDS is that it uses as input for eigendecomposition a set of similarity or dissimilarity data between objects. We will start at a more general level and describe the mathematical principles underlying eigendecomposition (§2.1), and then zoom in on some mathematical specifics of MDS, and the similarity data used as input (§2.2).

2.1 Matrix algebra and eigendecomposition

MDS is based on matrix algebra. Matrices can be added and multiplied, just like numbers can. Matrix addition and the multiplication of a matrix by a number (also known as 'scalar multiplication') are straightforward, as the following (arbitrary) examples illustrate:

$$\begin{pmatrix} 4 & 1 \\ -2 & 6 \end{pmatrix} + \begin{pmatrix} -3 & 7 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ -1 & 5 \end{pmatrix} \qquad \text{[matrix addition]}$$

$$3\begin{pmatrix} 3 & -4 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 9 & -12 \\ 0 & 6 \end{pmatrix} \qquad \text{[scalar multiplication]}$$

More significant is how two matrices are multiplied. We have a good intuition of what multiplication of numbers means. Likewise, matrix multiplication can be interpreted geometrically. This is easiest when we multiply a $n \times n$ matrix by a vector of length n. A vector is an arrow in n-dimensional space, so it has a length and a direction. It can be written as a matrix with n rows and 1 column (or 1 row and n columns). An example of matrix multiplication (with arbitrarily chosen numbers) is given below:[2]

$$\begin{pmatrix} -2 & 2 \\ -3 & 5 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} -12 \\ -22 \end{pmatrix} \tag{1}$$

Writing the matrix as $\mathbf{A}$, and the input and output vectors as v and w, we can represent the above equation as $\mathbf{A}v = w$. We can understand the multiplication by $\mathbf{A}$ as a geometric transformation: it acts like a function that maps input vector v to output vector w. The matrix $\mathbf{A}$ can be chosen in such a way that multiplication by $\mathbf{A}$ effects in a rotation, scaling, reflection, etc. of the input vector.

A special case arises when $\mathbf{A}v = \lambda v$, i.e. the result of applying $\mathbf{A}$ to v results in a vector with the same or opposite direction, only scaled by a factor λ (every coordinate of v is multiplied by the number λ). If this happens, v is called an **eigenvector** of $\mathbf{A}$, and λ its corresponding **eigenvalue**.[3] Matrix $\mathbf{A}$, as used in equation (1), has eigenvalues $\lambda_1 = 4$ and $\lambda_2 = -1$. The corresponding eigenvectors $v_1 = (\frac{1}{2}, \frac{3}{2})$ ($\mathbf{A}v_1 = (2, 6)$ has the same direction, but stretched by a factor 4) and $v_2 = (2, 1)$ ($\mathbf{A}v_2 = (-2, -1)$ has opposite direction, and same length) are displayed in Figure 1.[4]

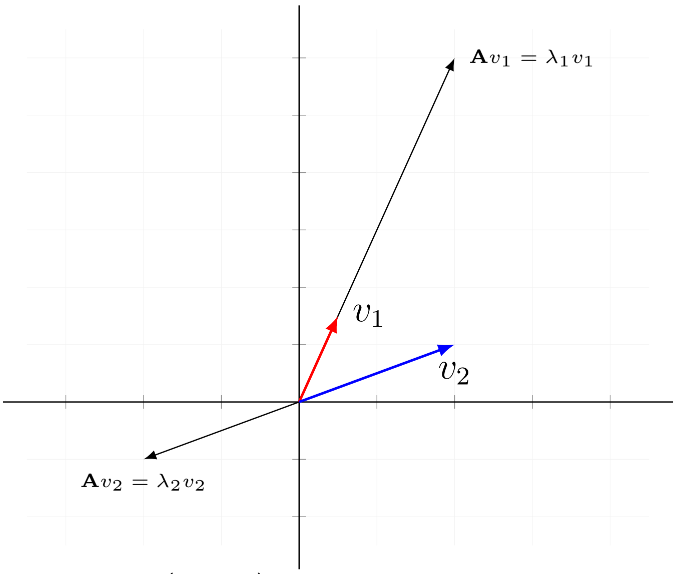

Figure 1: Eigenvectors of $\mathbf{A} = \begin{pmatrix} -2 & 2 \\ -3 & 5 \end{pmatrix}$ with a positive eigenvalue $\lambda_1 = 4$ and a negative eigenvalue $\lambda_2 = -1$

Eigenvectors and eigenvalues have many applications in mathematics and statistics, but for linguistic purposes it is easiest to interpret them in the context of data description. Suppose that the dataset is two-dimensional, so that it can be written as a matrix (for example

[2]Matrix multiplication is only possible if the two matrices have suitable sizes: a $m \times n$ matrix can be multiplied with a $n \times k$ matrix to yield a $m \times k$ matrix. The values in the resulting matrix are determined by *dot products* of rows in the first matrix, and columns in the second matrix (see any linear algebra textbook for definitions). In the example, the coordinates of the output vector are found by $-2 \times 4 + 2 \times -2 = -12$ and $-3 \times 4 + 5 \times -2 = -22$.

[3]For traditional reasons eigenvalues are denoted by λ. This has nothing to do with the lambda-operator in semantics.

[4]Note that there are infinitely many eigenvectors: all multiples of $(\frac{1}{2}, \frac{3}{2})$, such as $(1, 3)$, $(2, 6)$, $(100, 300)$ etc. are eigenvectors for $\lambda_1 = 4$. What is relevant is the number of eigenvectors that are *linearly independent* (cannot be written as combinations of scalar multiples of each other). For a $n \times n$ matrix, there are at most n linearly independent eigenvectors.

individuals for rows, and observations for columns). Then the eigenvectors of that matrix can be informally thought of as the dimensions along which most variation in the dataset occurs. The eigenvalue corresponding to an eigenvector indicates the relative significance of that eigenvector's dimension in describing the data.

Eigenvectors and eigenvalues have a further special property: for most matrices $\mathbf{A}$ – and in particular *symmetric* matrices,[5] which will show up in the setting of MDS – it is possible to reconstruct the matrix $\mathbf{A}$ by only using the eigenvectors/values. This process is called **eigendecomposition**, which is to say that $\mathbf{A}$ can be written as a product of three matrices, as follows:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

Here, $\mathbf{Q}$ contains the eigenvectors of $\mathbf{A}$ as its columns, and $\mathbf{\Lambda}$ (capital Greek letter lambda) is a diagonal matrix containing the eigenvalues of $\mathbf{A}$, which means that all its entries are 0 except for the ones on the diagonal, which contain the eigenvalues $\lambda_1, \dots, \lambda_n$ of $\mathbf{A}$:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix}$$

$\mathbf{Q}^{-1}$ is the *inverse* of $\mathbf{Q}$, which is to say that the product $\mathbf{Q}\mathbf{Q}^{-1}$ is the unit matrix, the diagonal matrix with ones on its diagonal.[6]

For our example matrix $\mathbf{A}$ from Figure 1, the eigendecomposition is as follows:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{3}{5} & -\frac{1}{5} \end{pmatrix}.$$

In general, applying eigendecomposition to a data matrix reveals the most important dimensions in the data (eigenvectors, from $\mathbf{Q}$), and the relative importance of those dimensions (eigenvalues, from $\mathbf{\Lambda}$).

2.2 MDS: similarity data

The input data for MDS are (dis)similarity data. Similarity between two objects i and j is represented as a numerical value $a_{i,j}$. Because the similarity between i and j is equal to the similarity between j and i ($a_{i,j} = a_{j,i}$), a (dis)similarity matrix is always a square symmetric matrix:

$$\begin{pmatrix} 0 & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ a_{2,1} & 0 & a_{2,3} & \dots & a_{2,n} \\ a_{3,1} & a_{3,2} & 0 & & \\ \vdots & \vdots & & \ddots & \\ a_{n,1} & a_{n,2} & & & 0 \end{pmatrix}$$

If the value $a_{i,j}$ increases as objects i and j become more similar, we speak of a similarity matrix. If the value decreases as the objects become more similar, we speak of a dissimilarity

[5] A matrix is symmetric if the value $a_{i,j}$ at row i and column j is the same as the value $a_{j,i}$ at row j and column i, for any i and j. A real-valued symmetric matrix has the property that it always has real-valued eigenvalues and eigenvectors.

[6] The counterpart in the domain of numbers is that the 'inverse' of the number a is $\frac{1}{a}$, because the product $a \times \frac{1}{a}$ is the unit number 1.

matrix. One can think of a matrix of driving distances between cities as a natural example of a dissimilarity matrix.

In order to apply MDS to linguistic data, these data must come in the form of a (dis)similarity matrix. It may at this point not be clear how linguistic data, such as translations or native speaker judgments, can be represented in such a way. Concrete examples of how linguistic data are turned into a similarity matrix will be discussed in section 3.

The steps in the classic scaling algorithm are as follows (Borg and Groenen 2005, §12.1):

1. Start with a matrix of dissimilarities $\mathbf{\Delta}$.
2. Apply an operation of *double centering* to the matrix of squared dissimilarities $\mathbf{\Delta}^2$. This does not affect the relative dissimilarities, but results in a matrix $\mathbf{B}$ in which the values are centered around the origin (the rows and columns add up to zero).
3. Eigendecompose $\mathbf{B}$ as $\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$.
4. Select the largest n eigenvalues from $\mathbf{\Lambda}$. Each of them corresponds with a column in $\mathbf{Q}$. The coordinates of the points in the reduced n-dimensional space are then found by keeping the n columns corresponding to the chosen eigenvalues, and removing the other columns.[7]

Because the matrix $\mathbf{B}$ is always a symmetric matrix (the original $\mathbf{\Delta}$, being a dissimilarity matrix, was also symmetric), a mathematical result ensures that the eigendecomposition of $\mathbf{B}$ results in a matrix $\mathbf{Q}$ that is *orthogonal*. This means that the inverse of $\mathbf{Q}$ is simply the *transpose* of $\mathbf{Q}$ (obtained by turning the columns into rows), written here as $\mathbf{Q}'$.

2.3 Stress and dimensionality selection

Another technical term that is sometimes used in the linguistic MDS literature is **stress**. Stress is a measure of the difference between the MDS output and the original dissimilarity data. It is a badness-of-fit measure, because a larger stress value indicates a worse fit. The basic stress measure is known as *Kruskal stress* and is based on the sum of squared deviations between the original dissimilarity data and the found coordinates in the output representation. There are various other stress measures – see Borg and Groenen (2005, §11) for full mathematical details of these and other fit measures, or Ding (2018) for a shorter exposition.

Step four in the above procedure also involves dimensionality selection. MDS is a dimensionality reduction technique, but the number of dimensions in the final MDS output is something the researcher chooses. When two or three dimensions are chosen, a 2D or 3D visualization can be readily obtained. Alternatively, a visualization can be made by selecting two or three dimensions from a higher-dimensional MDS output (for example, for a 5-dimensional MDS output, 2D maps can be generated for dimensions 1 and 2, or 2 and 4, etc.).

Stress can be used to help determine the optimal dimensionality of the MDS output. One easy general procedure is to generate MDS outputs of increasing dimensionality $2, 3, \ldots, n$, and calculate the stress value corresponding to each one. By comparing these (decreasing) stress values, the 'optimal' dimensionality k can be determined as the one for which the stress value do not decrease much anymore for higher dimensions than k. This method is known as the 'elbow method', after the shape of the line plot in a graphic representation of stress values for different dimensions (see Levshina 2016, Fig. 4 for an example of such a plot). Borg and Groenen (2005, §3.5) provide much more details on interpreting stress values.

[7]For a small numerical example of this procedure, see Borg and Groenen (2005, p. 263).

Instead of computing a single stress value by summing deviations for all points, it is also possible to use deviations for individual points to measure the badness of fit of individual points in the map. This can be used to detect potential outliers in the dataset, as in Levshina (2020, p. 15). Hence, stress values are a more flexible method in analyzing the fit of an MDS output than eigenvalues, which are associated with a dimension as a whole.

2.4 Other MDS algorithms

As mentioned above, classical scaling is only one of various MDS algorithms available. We will briefly mention two other algorithms that have been used in the linguistic MDS literature.

A method developed in de Leeuw and Heiser (1977) and subsequent work is based on an iterative method for minimizing a function, known as *majorization*. The corresponding MDS algorithm that minimizes stress by majorization is known as SMACOF (Scaling by MAjorizing a COmplicated Function) (Borg and Groenen 2005, §8; de Leeuw and Mair 2009). In practice, this implementation of MDS is more commonly used than classical scaling, as it is a more flexible method, and readily available, for example implemented in the software package R (see Borg, Groenen, and Mair 2018, sections 9 and 10 for a discussion of the benefits and potential issues of SMACOF).

The algorithm referred to as *Optimal Classification* (OC) by Croft and Timm (2013), belongs to a model that is known variously as an unfolding model, an ideal point model, or a preference model (see Borg and Groenen 2005, §14; Ding 2013, p. 247; Groenen and Borg 2014, p. 100; Borg, Groenen, and Mair 2018, §8 for more details). Unfolding is often classified as a technique closely related to MDS, but distinct from it with respect to the sort of input data. Unfolding is intended to map *preference data*, for example voting behavior of a number of individuals (Poole 2005). Both the individuals and their choice preferences are represented in the output visualization. The linguistic equivalent of this, as explicated by Croft and Poole (2008, §3), is to interpret the possibility to express function f by form s, as a (binary) preference for s by an individual f. Unfolding was mostly used in the earlier linguistic MDS literature, that was interested in the automatic generation of classical semantic maps, as discussed in section 3.1.

In the next section we will go over several examples from the literature in which the three MDS algorithms – classical scaling, SMACOF, and OC – have been used.

3 A typology of MDS maps

We will separate our discussion of MDS maps into two parts. This section will talk about the construction of the maps: which input data have been used, and what parameters were used in the generating of the MDS map? In other words, we attempt to provide a typology of MDS maps. We postpone interpretation of MDS maps, and how that links up to linguistic theory, until section 4.

The discussion in this section will be chronological, starting with a brief overview of MDS maps that aim to recreate classical maps (§3.1), and a related type of MDS map in which the points represent sentence contexts (§3.2). Then we will cover in more detail the recent trend of creating MDS maps on the basis of parallel corpus data (§3.3 and §3.4).

3.1 Recreating classical maps

A first type of MDS map is one that aims to recreate classical semantic maps. This was one of the early motivations of applying MDS in the linguistic domain: MDS was introduced because "the semantic map model is in need of a sound mathematical basis" (Croft 2007, p. 83). This was a methodological advancement, because MDS provided a way to automatize the process of building classical semantic maps, and made it possible to deal with large-scale sets of data that cannot be analyzed manually.

This type of MDS map is often based on questionnaire data: a number of sentence contexts that have been selected or designed by the researcher in order to investigate a particular domain (e.g. the tense/aspect questionnaire in Dahl 1985, or the performative questionnaire in de Wit et al. 2018). The questionnaire is applied by native speakers or fieldworkers in a number of languages, and the data obtained from these questionnaires serve as input for MDS.

In particular, the input data for these maps consists of specifications (Yes/No) for forms in various languages about whether or not that form can convey an abstract function. Two functions count as more similar when they share a higher number of forms that express that function. An example is given in Figure 2, which displays an MDS map for indefinites from Croft and Poole (2008), based on data from Haspelmath (1997). The construction of the MDS map in Figure 2 is summarized in the box below it.

Figure 4 from Croft and Poole (2008)

Figure 2: MDS map for indefinite functions from Croft and Poole (2008).

MDS-based classical semantic map

MDS algorithm: Optimal Classification / unfolding
Input for MDS: matrix of functions and forms, with Y(es) if that form conveys that function, and N(o) if it does not.

	function_1	function_2	...	function_n
form_1	Y	N	...	N
form_2	N	Y	...	N
$\vdots$	$\vdots$	$\vdots$		$\vdots$
form_k	Y	Y	...	N

Measure of similarity: the similarity between two functions is measured by the number of forms that co-express them: $\text{d}(\text{function}_i, \text{function}_j) = \frac{\#\text{Y's in common}}{k}$. This way a $n \times n$ dissimilarity matrix is obtained.
Output map: dots on the map represent abstract functions (the function_i's), while distance on the map represents similarity between functions.

Figure 2 aimed to recreate Haspelmath's (1997) classical semantic map of indefinites. Unlike in classical semantic maps, the distance between points is meaningful: points that are closer to each other are to be considered more similar. On the other hand, the dimensions have numerical values, but these do not have a direct linguistic interpretation. The dots on the MDS map may be connected in order to add the graph structure of the classical map (although this structure is not a result of the MDS algorithm), see Croft and Poole (2008, Fig. 6).

The similarity of this type of MDS maps to classical semantic maps entails that they are subject to some of the same shortcomings that classical maps have. For example, an issue discussed in the literature on classical maps is that the abstract functions that are used as nodes in a classical map ought to be theory-neutral and comparable across languages, i.e. should be *comparative concepts* (Haspelmath 2003). It is not always easy to make sure your data satisfy this property, and this problem persists for MDS-based classical maps.

Note that the points on the map in Figure 2 are multilingual abstractions, since they represent abstract functions that are positioned in the two-dimensional space based on how forms in various languages express these functions. However, a monolingual map can be created by adding *cutting lines* to the map that indicate how language-specific forms realize the functions on the map. In Figure 3 this is illustrated for Romanian. For example, the cutting line that is labeled *ori-* separates the functions on the map that the Romanian form *ori-* 'any' can convey (i.e. *free choice* and *comparative*) from functions that it cannot convey. Cutting lines work in this setting because of the binary nature of the input data, but cannot be used for other types of MDS input data.

This way, this type of MDS maps allows for the same two perspectives as classical semantic maps do, as described in Georgakopoulos and Polis (2018, p. 9): translational equivalents are visible in the MDS map as a whole, and designations of a particular meaning intralinguistically appear in language-specific maps.

Besides the work of Croft and Poole (2008), other domains for which MDS maps of this type have been made include Slavic tense (Clancy 2006), and causatives (Levshina 2020, §2). The latter study is noteworthy because it contains three-dimensional MDS maps that are construed on the basis of data from language grammars (Levshina 2020, Figures 4, 5).

Figure 5 from Croft and Poole (2008)

Figure 3: The map from Figure 2 with cutting lines for Romanian.

3.2 Incorporating sentence contexts

A variant of the type of MDS map described above appears in Croft and Poole's (2008) reanalysis of data from Dahl (1985). While a map such as the one in Figure 2 is based on forms (indefinite pronouns) and idealized functions, it does not include the data on which it was decided that a certain form may express a certain function. These data typically come in the form of sentence contexts that purport to show that form x can be used to express function y. Croft and Poole's map of Dahl's data does include these underlying data (sentence contexts from a questionnaire), but is otherwise conceptually similar to the maps discussed above in that it also involves an interpretation of the contexts in terms of abstract functions by the researcher.

The map, displayed in Figure 4, is based on Dahl's (1985) questionnaire on tense-aspect constructions in various languages. In this questionnaire, applied to various languages by native speakers or fieldworkers, informants were asked to translate sentences in context (such as 'He WRITE a letter' in the context where you saw someone engaging in an activity yesterday, Dahl 1985, p. 198). The constructions cross-cut languages, and include for example 'English *simple present*', 'French *imparfait*', 'Zulu *narrative past*', etc. Croft and Poole assigned each of the 250 sentence contexts to a prototype ('perfective', 'habitual', etc.). The contexts appear on the map as dots with a label for their prototype (such as the label V for 'perfective'). As a result, a single label appears several times on the map.

Lastly, the lines on the map in Figure 4 (imperfective/perfective and past/future) are added post-hoc by Croft and Poole as an interpretation of the two dimensions of the MDS map. In section 4.1, we return to the qualitative and quantitative assessment of the significance of MDS dimensions in more detail.

Croft and Poole's map of Dahl's tense-aspect data

MDS algorithm: Optimal Classification / unfolding
Input for MDS: matrix of sentence contexts and constructions.

	sentence context$_1$ code: V	sentence context$_2$ code: U	…	sentence context$_{250}$ code: r
construction$_1$	Y	N	…	N
construction$_2$	N	Y	…	Y
⋮	⋮	⋮		⋮
construction$_{1107}$	Y	Y	…	N

Measure of similarity: as above
Output map: dots on the map are sentence contexts, represented by their prototype code

Figure 8 from Croft and Poole (2008)

Figure 4: MDS map of Dahl's tense-aspect data (Croft and Poole 2008, Fig. 8)

MDS maps of a similar nature include the ones in de Wit et al. (2018), who use a questionnaire on aspectual constructions in performative contexts. Hartmann et al. (2014) apply MDS to map microroles (verb-specific semantic roles) from 25 languages. Similarity between two microroles is based on co-expression tendencies between the two (see their p. 469 for details on the similarity measure).

Map coloring In the same way that cutting lines were used to display information about a specific language in a multilingual map (recall Figure 3), MDS maps that map individual contexts also have a means to display language-specific data. This is done by changing the appearance of the dots on the map in order to indicate language-specific constructions (e.g. by using colors or symbols), a process we will refer to as **map coloring**. For example, Figure 5, taken from Hartmann et al. (2014, Fig. 5), shows the same map four times, but in each case the dots are represented in a different way, reflecting the constructions used in the four languages (the meaning of the contour lines on the map will be discussed in section 4.2).

Figure 5 from Hartmann et al. (2014)

Figure 5: Map coloring in Hartmann et al. (2014)

Map coloring is in important technique in MDS maps, as it allows to see language-specific patterns and cross-linguitic patterns in the same visualization. We will come back to map coloring in the next sections for other types of MDS maps.

3.3 Maps of parallel corpus data

Besides using questionnaire data, a second important source of data for linguistic MDS analyses is texts that have been translated in various languages, forming a *parallel corpus*. Wälchli and Cysouw (2012, p. 674) refer to this as *primary data typology*, contrasting it with analyses based on higher level sources such as reference grammars. Parallel corpora overcome some issues of data collection with classical maps: there is no dependency on existing comparable concepts, and using corpus data also allows to include frequency as a factor.

Examples of parallel corpora that have been used in MDS analyses include Bible corpora (Wälchli 2010, 2018; Wälchli and Cysouw 2012), translation corpora of novels (Verkerk 2014; van der Klis et al. 2020), Europarl (translated proceedings of the European parliament; van der Klis et al. 2017), and a corpus of subtitles (Levshina 2015, 2020). The choice of a parallel corpus is important, as it has been pointed out that a parallel corpus can be a limited source of data in that it may only provide a genre-specific perspective, might lack specific forms, and overuse prototypical forms (Levshina 2020). Also, just as with questionnaire data, translation effects may arise.

Once a suitable parallel corpus is selected, the relevant construction that the researcher is interested in must be extracted and annotated. For example, Wälchli and Cysouw (2012) extracted 360 clauses describing motion event from translations of the Gospel of Mark in 101 languages ('doculects' in their terminology) (see Wälchli 2010 for a similar study with a different sample from the Gospel of Mark).

Unlike the maps in section 3.2, in the setting of parallel corpora, a context corresponds with a sequence of translations. A toy example would be ⟨*book*, *libre*, *Buch*⟩ for the English, French, and German occurrences of that noun in a sentence from a parallel corpus. Similarity between contexts is then measured by a **distance function** applied to two such sequences. Choosing a suitable distance function is another parameter that goes into the design of an MDS research study. The most typical distance function used is the (relative) **Hamming distance**: a context is represented as a sequence of translations, and the distance between two sequences of n objects is defined as the number of objects that differ (compared pointwise) divided by n. For example the distance between $\langle A, B, C, D, E\rangle$ and $\langle A, B, X, D, Z\rangle$ is 2/5 because two of the five positions differ (the 3rd and the 5th).

Other distance functions are possible, such as the Levenshtein distance function that has been used in a number of (non-MDS related) applications in linguistics (see e.g. Greenhill 2011). Another plausible option is to define a distance function *ad hoc*, for example one that weighs certain components heavier than others. We are not aware of work in the linguistic MDS literature that explores different choices of distance function, and their qualitative effects on the resulting MDS output (but see section 5.1).

Parallel corpus MDS

Input for MDS: matrix of languages and contexts.

	language$_1$	language$_2$	...	language$_n$
context$_1$	translation$_{1,1}$	translation$_{2,1}$	...	translation$_{n,1}$
context$_2$			...	
⋮	⋮	⋮		⋮
context$_k$	translation$_{1,k}$		...	translation$_{n,k}$

Measure of similarity: similarity between two contexts is determined by relative Hamming distance function
Output map: dots on the map are contexts

There are several recent studies in which MDS has been applied to parallel corpus data. Here, we give a short overview of which kind of datasets have been used. In section 4.3, we return to most of these studies in more detail, to show their potential in answering research questions in both classical typology and formal semantics.

Wälchli (2018) investigates temporal adverbial clauses headed by words such as UNTIL, BEFORE, WHILE. Using a methodology similar to that of Wälchli and Cysouw (2012), he builds an MDS map representing contexts from the New Testament parallel corpus from 72 languages.

Verkerk (2014) uses a parallel corpus built from translations of three different novels in 16 Indo-European languages in order to investigate the encoding of motion events. This results in a 3D MDS map, but instead of computing Hamming distance between contexts (as in Wälchli and Cysouw's case above), distances are computed between languages. Hence, the dots in Verkerk's (2014, p. 349) MDS map represent languages, and not individual contexts.

Dahl and Wälchli (2016) studies perfects and the related category of iamatives (forms like English *already*). They create an MDS map in which the points represents *grams* (a word, suffix, or construction in a particular language with a specific meaning and/or function). They interpret the MDS space as a 'grammatical space'. Using NT Bible translations from 1107 languages, the similarity between two grams (for example English *Present Perfect* and Swahili *-me-*) is determined on the basis of how similar their distributions are across the text.

Beekhuizen et al. (2017) use parallel texts in an investigation of indefinite pronouns. Other than most studies on parallel corpora, these authors use the Optimal Classification algorithm for MDS (see section 3.1 above).

De Swart et al. (2012) apply MDS to occurrences of two Greek prepositions on the basis of a four-language sample of a parallel corpus of NT Gospels. The approach, including the similarity measure used, is similar to Wälchli (2010). They use a special variant of map coloring which they call "semantic overlays": they only display the points (i.e. occurrences of a preposition) that correspond with a given semantic feature. This way they can interpret if the poles of a given dimension correspond to the positive and negative value of a semantic feature.

Levshina (2015, 2016, 2020), in a series of papers, applies MDS by stress majorization (see section 2.4 above) in the domain of causatives. Levshina (2015) studies analytic causatives in 18 European languages with a parallel corpus of film subtitles constructed by Levshina. The procedure was similar to that of Wälchli and Cysouw (2012), but the annotated features for each causative constructions were assigned different weights (Levshina 2015, p. 498). Levshina (2016) is a similar study with the same corpus, but focuses on verbs of letting in 11 languages.

3.4 Translation Mining

van der Klis et al. (2017) developed a variant of the basic methodology from Wälchli and Cysouw (2012), which they dub *Translation Mining*. Instead of comparing translations by the lexical items that were chosen, they compare translations on the basis of a grammatical feature, namely the tense form used. So, for Wälchli and Cysouw, when comparing two constructions w_1 and w_2 in the same language, they count as equivalent if they are the same lexical item ($w_1 = w_2$). For van der Klis et al. (2017), on the other hand, w_1 and w_2 count as equivalent if they use the same tense form ($\text{Tense}(w_1) = \text{Tense}(w_2)$), but w_1 and w_2 need not be the same lexical item. In both cases, similarity of contexts is determined by means of the relative Hamming distance.

A consequence of this methodological step is that after the relevant data are extracted from the parallel corpus, they also need to be annotated for the grammatical feature in question, the step of 'tense attribution' in van der Klis et al. (2017). These authors have developed a software tool *TimeAlign*[8] to facilitate the process of annotation of parallel corpora.

In an extension of the 2017 study, van der Klis et al. (2020) investigate cross-linguistic variation of the Perfect in West-European languages, where small caps indicate a cross-linguistic tense category comprising language-specific forms such as the English *Present Perfect*, the French *Passé Composé*, etc. (these tense categories are purely form-based, e.g. auxiliary+participle). The parallel corpus used in this work contains translations of the French novel *L'Étranger* by Albert Camus (cf. de Swart 2007), and the MDS maps are created by the SMACOF algorithm.

[8]Source code for TimeAlign is available via `https://github.com/UUDigitalHumanitieslab/timealign`.

A slightly different version of map coloring is used in this line of work: colors correspond to cross-linguistic tense categories, and not language-specific tense forms (so, for example blue represents Perfect). With this method, differences in tense use between languages can be identified. This is illustrated in Figure 6: the same map is shown 7 times, but with colorings for the different languages in the corpus (blue for Perfect and green for Past). The stepwise reduction of the blue area (i.e. reduction of Perfect use) is the visual representation of what van der Klis et al. (2020) call a 'subset relation' between West-European languages' use of the Perfect. There is a core use for which all languages use their counterpart of the Perfect (blue), and then there is a scale from languages that use the Perfect in only the core contexts (modern Greek) to languages that use it more widely (French, Italian). Further interpretation of the cut-off points between pairs of languages feeds a cross-linguistic semantic analysis of the Perfect. Hence, MDS analysis is used to reveal a richer cross-linguistic variation in the domain of the Perfect than was previously assumed in the literature (see van der Klis et al. 2020 for further details).

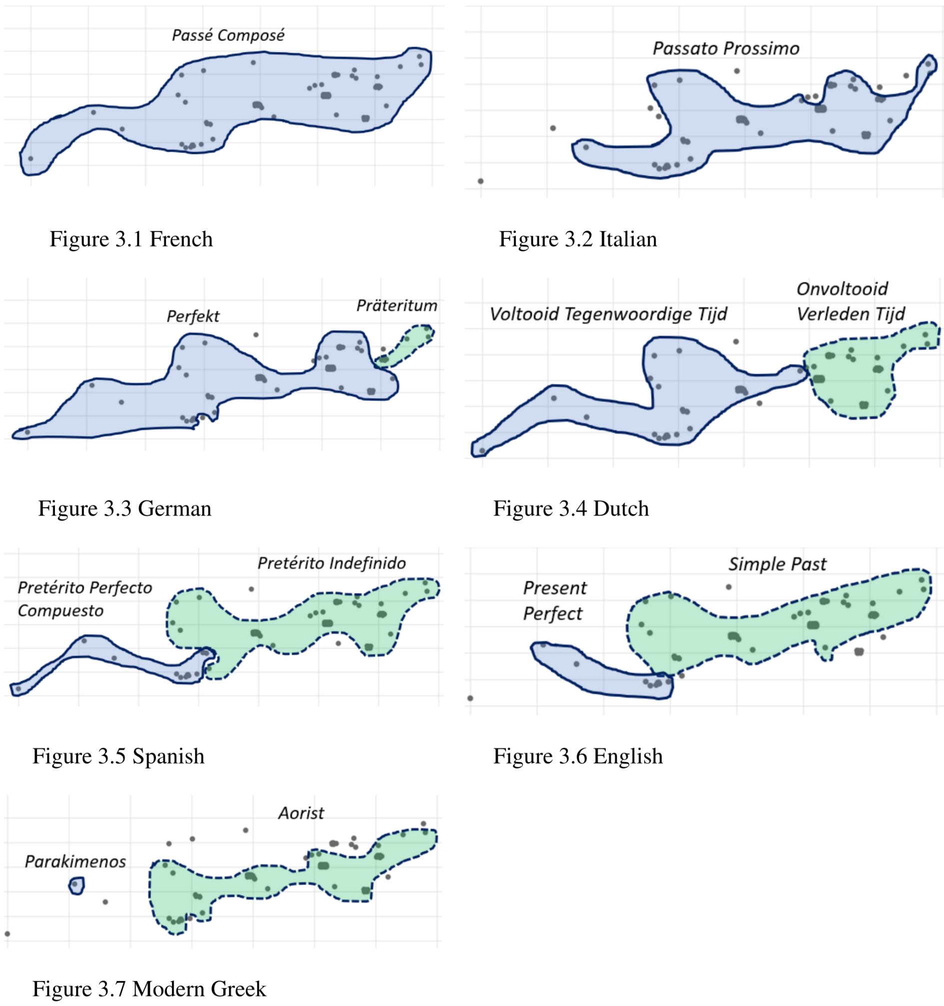

Figure 6: Subset maps from van der Klis et al. (2020)

This study on the Perfect gave rise to a line of (ongoing) work in which Translation Mining is applied in other domains. Bremmers et al. (2020) study definite determiners in German and Mandarin using a corpus of translations of *Harry Potter and the Philosopher's Stone* by J.K. Rowling. Tellings (2020) investigates variation in the domain of conditionals, see more on this in section 5.1 below.

Having provided a typology of MDS maps in this section, in the next section, we turn to the interpretation of MDS maps.

4 Map interpretation and links to linguistic theory

Broadly speaking, there are two ways to analyze MDS maps. First, one can attempt to assign a linguistic interpretation to the dimensions of the map. We will call this process **dimension interpretation**, and will discuss this in §4.1. Second, one can attempt to assign a linguistic interpretation to groups of points that cluster together on the map, a strategy that we refer to as **cluster interpretation** (§4.2). Note that dimension interpretation and cluster interpretation are not completely independent, as it is typically the case that when two clusters are separated on a map, they are also on opposing poles of one of the dimensions in the map.

§4.3 closes this section by linking interpretation of MDS maps to linguistic theory. We show how MDS is used both in classical typology and formal semantics.

4.1 Dimension interpretation

Recall that the dimensions in an MDS solution do not have an intrinsic linguistic meaning, but are the outcome of the algorithm.[9] However, a typical desideratum of MDS studies is to assign an interpretation to the dimensions so that it provides a qualitative assessment of the distribution of points on the map. For example, in Figure 4 the two axes are interpreted as an imperfective/perfective axis and a past/future axis.

As an example, Wälchli and Cysouw (2012) use eigenvalue analysis to find that 30 dimensions are relevant to describe their motion verb data. This is rather high for linguistic MDS studies, and is taken by the authors to be illustrative for the high degree of complexity of the variation in the domain of motion verbs (p. 689). Instead of assigning a single interpretative label to each dimension, the authors separately interpret the negative and positive 'pole' of a dimension. For example, the most important dimension (dimension 1) is analyzed as distinguishing 'come/arrive' contexts (negative pole) from 'go/depart' contexts (positive pole) (see their Table 4). As an example of how 2D maps are created for a high-dimensional MDS analysis, Figure 7 shows 2D maps plotting Dimension 1 and Dimension 10. Labels are displayed in regions of the map corresponding with the poles of Dimension 1 (x-axis).[10] As before, Figure 7 applies map coloring in order to indicate language-specific patterns on a multilingual map (Figures 7a and 7b display the same distribution of dots, but the coloring reflects Spanish and English, respectively).

Dimension interpretation is often done by visual inspection of MDS maps, but more rigorous approaches using statistical tools have also been proposed. Levshina (2020) uses a linear regression analysis to identify which of the semantic variables are the most strongly correlated with the placement of contexts in the MDS map. The procedure annotates the individual contexts of the MDS map with binary classifications (e.g. in the domain of causatives, one

[9] This relates to a general problem of visualizations that MDS maps are also subject to: they always show some structure in the data, even if this structure is only an artefact of the method applied (Cysouw 2008, p. 50). In this light, we should also view Zwarts's (2008) comment that the resulting dimensions in MDS maps not necessarily reflect semantic dimensions.

[10] The labels in Figure 7 correspond to *regions* on the map, not clusters. See Wälchli and Cysouw (2012, p. 690) for details on this rather subtle distinction.

Figure 3 from Wälchli and Cysouw (2012)

(a) Spanish coloring

Figure 3 from Wälchli and Cysouw (2012)

(b) English coloring

Figure 7: MDS maps from Wälchli and Cysouw (2012, Fig. 3)

could annotate for contexts being *intentional or not?* or *factitive or permissive?*). A regression analysis then correlates these variables with the positioning of a context on a single dimension. In other words, the method indicates which semantic phenomena best explain the cross-linguistic variation modeled by the MDS map.

4.2 Cluster interpretation and cluster analysis

A group of points that appear close together in a cluster in an MDS map is analytically relevant, because the proximity of the points indicates that the corresponding contexts are similar in a linguistically relevant way (and contrast with points outside the cluster). Clusters in a map can be identified either by informal inspection of the map, or with the help of statistical or algorithmic tools. For example, the contour lines in Figure 5 are obtained from a probabilistic method, see Hartmann et al. (2014, 471ff.) for details. Once the clusters are identified, cluster interpretation is the process of inspecting the contexts from the dataset corresponding to the points in the cluster, and finding some linguistic commonality between them. For example, Hartmann et al. (2014, p. 470), in their MDS map of semantic roles, recognize clusters of agent-like roles and patient-like roles.

The procedure above consists of cluster identification and interpretation *after* MDS has been applied to the dataset. An alternative is to identify clusters directly from the original dataset, and run MDS *parallel* to it. The direct identification of clusters from the distance matrix (or a transformation thereof) is known as **cluster analysis**. The resulting attribution of clusters to individual points can then be fed back to the MDS map as an additional layer of labelling. This procedure potentially facilitates the interpretation of the semantic dimensions at stake. Below, we describe two forms of cluster analysis that have been applied in combination with MDS.

4.2.1 k-means clustering

k-means clustering aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster. k-means clustering can be applied to a distance matrix to find k clusters consisting of similar data points. k-medoids clustering is a special case in which the center of each cluster is an actual data point; in k-means clustering, this need not necessarily be so.

In Wälchli (2018), k-medoids clustering (in particular, the Partitioning Around Medoids algorithm) is applied to cross-linguistic lexical variation in the expression of adverbial clauses. With k set to 3, AS.LONG.AS, UNTIL, and BEFORE appear as three different semantic clusters. This result confirms earlier typological analyses in this domain, but without taking these functions as a point of departure, but rather as a result of cross-linguistic lexical variation. With $k = 5$, two additional clusters appear: WHILE and FÖRRÄN (from Modern Swedish *förrän*). Figure 8 shows the MDS map with additional labels for the identified clusters.

Figure 2 from Wälchli (2018)

Figure 8: On the left: MDS map with added labeling for English and cluster analysis through the Partitioning Around Medoids algorithm. On the right: assignment of clusters to individual contexts by the Partitioning Around Medoids algorithm with $k = 5$. Adapted from Figure 2 in Wälchli (2018, p. 157).

A post-hoc analysis reveals that the optimal solution is with three clusters, and thus disregards WHILE and FÖRRÄN as meaningful clusters. From this result, one can infer that there are very few languages that have a separate lexical entry for FÖRRÄN as Modern Swedish does. Instead, languages in general have the same marker for FÖRRÄN and UNTIL. For English, the MDS map shows that there is a homogeneous distribution of *till* and *until* in these two clusters. A similar point can be made for WHILE, that has a separate lexical marker in English, but which is cross-linguistically usually expressed with the same marker that expresses AS.LONG.AS.

4.2.2 Hierarchical cluster analysis

Hierarchical cluster analysis aims to build a hierarchy of clusters. The default, agglomerative variant takes a bottom-up approach in which each observation starts in its own cluster, and pairs of clusters are iteratively merged while minimizing distance. The result is usually represented as a dendrogram.

In Levshina (2020), this type of cluster analysis is used to identify the semantic functions of causative constructions. Levshina annotated a typologically diverse sample of corpus subtitles and molded the parallel corpus data into the data structure posed in section 3.2 above. Hierarchical cluster analysis, as shown in Figure 9 below, then allows her to find seven clusters, that can serve as the input for a semantic map. Using Regier et al.'s (2013) method to induce edges, Levshina ends up with a fully data-driven classical semantic map.

Figure 1 from Levshina (2020)

Figure 9: Hierarchical cluster analysis on 18 causation contexts. The red rectangles delimit the seven identified clusters. Adapted from Figure 1 in Levshina (2020, p. 7)

.

Alternatively, not individual constructions, but rather languages as a whole are used as starting nodes of the hierarchical cluster analysis (e.g. in Hartmann et al. (2014, p. 475) and Levshina (2016, p. 106)). This move allows to generate hypotheses about genealogy or language contact. Recently, Neighbor-Nets has been put forward as a related method that also operates on the language level and has similar aims (Bryant and Moulton 2004). Neighbor-Nets has been successfully applied to parallel corpus data (e.g. in Dahl 2014; von Waldenfels 2014; Verkerk 2014, 2017).

Cluster analysis and dimension analysis are interpretation methods for the map itself, but the aim of MDS studies in linguistics is to answer some larger questions relating to linguistic theory. We now move to describing which part MDS maps play in the process of linguistic argumentation.

4.3 Links with linguistic theory

In this section we discuss how multidimensional scaling as a data visualization technique stands in relation to two approaches to the study of language, which we will label here as "classical typology" and "formal linguistics". The scare quotes indicate that the labels neither do justice to the large number of underlying assumptions that both approaches have (see e.g. Hawkins 1988), nor to the various variant and intermediate positions that exist.

MDS in the continuum from typology to formal linguistics We will adopt the following idealized definitions of the two approaches. (Classical) typology is a form of inquiry in which large-sample linguistic comparison is applied in order to reveal limits of cross-linguistic variation in the form of (implicational, restricted, biconditional, ...) universals of language. Formal linguistics (which is not restricted to generative linguistics) is an approach that, on the basis of data from a single or a small number of languages, provides an in-depth abstract analysis of a given phenomenon that leads to an account that is deductive in the sense that it makes falsifiable predictions. We do not aim to review the debate here of how these two approaches relate to each other, and to what extent there is a conflict between them (see e.g. Croft 2007; Cinque 2007; Haspelmath 2010; Newmeyer 2010 for differing opinions).

An important observation is that the application of MDS within linguistics is not restricted to classical typology only, but is also amenable to the methods of formal linguistics. We will show this by going over the studies cited in section 3 again, this time highlighting the theoretical contribution the authors had in mind by using MDS. We will see that MDS applications cover a continuum ranging from classic typological studies to formal linguistic studies.

On the typological side, a number of studies is primarily interested in research questions about language classification.

Verkerk (2014) is a clear example of this, given that her aims are strictly about language classification, and her MDS maps are unusual among the studies discussed here in that they locate languages rather than semantic functions or linguistic contexts. Verkerk's aim is to check the validity of the "strict dichotomy between satellite-framed and verb-framed languages" (p. 326) proposed by Talmy (2000). She compares MDS to other "aggregation methods" such as Neighbor-Nets (see above). From her MDS analysis, she concludes that a strict dichotomy cannot predict the attested variation, which gives rise to the potential identification of new language classes (Verkerk 2014, p. 351).

Dahl and Wälchli (2016) is an example of a large-sample MDS study (1107 languages). It addresses the question if two grams, Perfects and iamatives, whose prototypical examples form two areally distinct language groups (European vs. Southeast Asian), form two distinct clusters, or rather a continuum. The conclusion is that although certain areal groups can be identified as clusters in the MDS map, the distribution of grams forms a continuum.

Hartmann et al. (2014) investigate the clustering of semantic microroles in a classic scaling MDS map. On the basis of this, a metric is computed that classifies languages on the basis of pairwise similarity of microrole coding strategy. By this means, a hierarchical typology is constructed of the 25 languages in the study.

More towards formal linguistics is de Wit et al. (2018), who aim to investigate aspectual properties of performatives. They argue that, cross-linguistically, languages use the same aspectual category for performatives as they do for other constructions that have a similar epistemic property (see their §2 for details). They use an MDS study to show that aspectual categories indeed pattern this way, by identifying clusters on the map that can be separated on the basis of epistemic properties. This study can thus be argued to occupy somewhat of a middle ground: it is a typological study that investigates cross-linguistic patterns, but also aims to identify epistemic properties of performative and other speech acts.

In a similar position is Wälchli and Cysouw (2012), who employ MDS maps to represent the extent of variation in the domain of motion verbs (101 languages). Being a pioneering study that introduced the use of MDS in combination with parallel texts, many of their points are of a methodological nature, for example the need to use parallel corpus MDS methodology in domains with a high degree of typological variation. When we look at the claims that relate to linguistic theory proper, the authors apply detailed dimension and cluster interpretation on their MDS map to make several typological and language-specific claims about the cross-linguistic variation of motion verbs. By inspecting the linguistic contexts behind the motion verbs, the authors propose a new category type 'narrative *come*' (p. 696), showing that the distribution of motion verbs also has a discourse component.

The study by van der Klis et al. (2020) discussed in section 3.4 looks at a much smaller sample (seven European languages). However, this sample is sufficient to identify a subset relation in the use of the PERFECT, rather than a hitherto assumed dichotomy between strict and liberal PERFECT languages. This observation forms the starting point for a formal linguistic analysis of the contexts in which pairs of languages differ with respect to PERFECT use.

De Swart et al. (2012) represents a more radical departure from the typological studies discussed above in that it is primarily interested in a phenomenon in a single language – the semantics of the prepositions ἀπό (*apo*) and ἐκ (*ek*) in Ancient Greek. The authors use a parallel corpus MDS study to measure the semantic similarity between the two prepositions, stating explicitly that they want to investigate how the (broad-sample) MDS methodology "can be applied to a small language sample" (p. 163). By an analysis of the semantic features of the clusters on the map, they come to a better understanding of the semantic role of both prepositions.

Similarly, Bremmers et al. (2020) are primarily interested in a phenomenon in a single language: how is the formal distinction between weak and strong definites operational in Mandarin? A small-sample MDS study, with only three languages (English, German, and Mandarin Chinese), shows that contrary to earlier predictions, Mandarin bare nominals and demonstratives do not map directly on German contracted (weak definites) and uncontracted forms (strong definites). This discovery then forms the starting point of a formal linguistic analysis.

In summary, we have seen how the methodology of MDS has been used with linguistic applications ranging from language classification to single-language formal analysis. This wide-ranging use of MDS thus goes some way toward reconciling the sometimes perceived contrast that "typology starts with crosslinguistic comparison, while the structuralist/generative approach proceeds 'one language at a time' " (Croft 2007, p. 85), as we have discussed examples of work in formal linguistics that started out with cross-linguistic comparative data as MDS input.

The different approaches to applying MDS can also be appreciated by specifying the position that MDS maps take within the analytic process or process of argumentation. The classic typological papers use MDS maps to visualize cross-linguistic variation itself, and the dimensional/clustering patterns in the maps are the main theoretical interest, as this provides information about language classification. By contrast, the more formally oriented approaches have MDS maps in an earlier position within the analytic process: they use MDS to identify empirical distinctions that are relevant for building an analysis of the phenomenon in question. The MDS stage is then followed up by a formal analysis that proceeds in a manner that is fairly typical for the approach of formal linguistics.

We thus arrive at a more nuanced view than how Georgakopoulos and Polis (2018, p. 18) put it. They claim that whereas classical semantic maps are an *explanans* (they are a result of earlier analyses), MDS maps can be seen as an *explanandum*, i.e. they are visualizations of data that are not the end product, but the starting point of further linguistic analysis. As the above discussion has made clear, this may typically be the case for formal linguistic applications of MDS, but we contend it is not necessarily so for typological applications: the interpretation of clusters and dimensions of the MDS maps reflect the main topic of interest.

MDS and formal paradigms When MDS is used as a tool to identify data patterns that are the starting point for formal linguistic analysis, does this commit the researcher to a particular formal paradigm? One should be careful here in distinguishing the theoretical basis for creating semantic maps, and the theoretical paradigm for a subsequent formal analysis.

A number of MDS papers are explicit about their assumptions regarding the theoretical basis of semantic map methodology. Starting in Wälchli (2010, §2) and Wälchli and Cysouw (2012, §3), and later adopted in other MDS studies (e.g. de Swart et al. 2012, p. 167), a combination of *exemplar semantics* and *similarity semantics* has been proposed. This means that exemplars (individual occurrences) are compared instead of abstract concepts, and that similarity is a more basic notion than identity. The two are linked by Haiman's Isomorphism Hypothesis ("recurrent identity of form will always reflect some perceived similarity in communicative function"; Haiman 1985). This theoretical basis underlies MDS maps in which points represent individual contexts (see §3.2).

The exemplar approach to semantic maps is associated with a form-to-meaning direction of analysis. This not only holds for studies in the domain of lexical semantics, but also for example for van der Klis et al. (2020), in which Perfects are identified on the basis of their form (an auxiliary plus a participle), and not on one of the pre-conceptualized meanings of the Perfect from single-language analyses of that tense form (e.g. Portner 2003). This contrasts with a meaning-to-form approach, which is typical for classical semantic maps and some of the early MDS maps (see §3.1) that do not start with exemplars but with abstract functions.

These different theoretical foundations for building semantic maps should not be confused with theoretical assumptions that may be made relating to a formal analysis that is made on the basis of data from MDS maps. Although MDS methodology and the resulting maps crucially rely on a notion of similarity between linguistic objects, it does not follow that conclusions drawn about the semantic content of these objects must be based on similarity rather than identity.

A case in point is van der Klis et al. (2020), who argue that variation in the domain of the Perfect is to be described in terms of dynamic semantics, compositional semantics, lexical semantics, and others. So, for them, using a similarity-based statistical technique to create maps does not prevent them from an analysis in terms of well-established paradigms from the tradition of formal linguistics.

In conclusion, this section addressed the question whether MDS, in addition to a means to reveal descriptive patterns in complex multi-dimensional datasets, can be a valuable tool for the more formally oriented linguist. We have argued that this is the case: cross-linguistic comparison can be the starting point to, and the empirical core of, formal linguistic studies. Hence, multidimensional scaling on data from parallel corpora should be part of the (formal) linguist's toolkit. At the same time, this use of parallel corpus data in formal linguistics is still a fairly recent advance in need of further development. In the next section we will discuss some potential directions of future work, which we hope will further integrate the use of parallel corpus data in developing (formal) linguistic theory, as well as to point to alternatives for an analysis through MDS.

5 Future directions

In this section, we point at two possible future directions for applying MDS in linguistic research. First, we describe how we can use MDS when compositionality comes into play (§5.1).

So far, we have seen applications of MDS that only compare single lexical or grammatical features, but in most semantic domains, we see an interplay of variables. A compositional approach is therefore necessitated.

Second, we cover some alternatives for MDS as our method of choice (§5.2). Recently, a new set of dimensionality reduction methods have surfaced that assing more weight to local rather than global variation. We show how these methods can yield different perspectives on the datasets at hand.

5.1 Lexical-compositional step

In most of the MDS work reviewed in this paper, the methodology has been applied to word-size or phrase-size units (motion verbs, tense forms, causatives, etc.), and comparison has been made on the basis of one parameter. As a next step in the application of MDS techniques in semantic variation research, we envisage the application of this method to sentence-size constructions, for which comparison would be made on the basis of multiple parameters.

In abstract terms, consider a complex construction A whose meaning is compositionally determined by component expressions B and C:

$$\begin{array}{c} \mathrm{A} \\ \diagup\;\diagdown \\ \mathrm{B}\quad\mathrm{C} \end{array}$$

One can study variation for B and C separately, and then make predictions for what variation for A looks like. Alternatively, one can take construction A as primary data, and annotate various grammatical properties of A, including properties that relate to B and C. Then an MDS solution can be computed that takes these various parameters into consideration. This can be done either by a distance function that is a weighted average of distance measures for the various parameters (as described in Tellings 2020), or by *multi-mode* or *multi-way* MDS, extensions of MDS that take into consideration multiple similarity measures for each pair of objects (de Leeuw and Mair 2009). We are not aware of the use of these MDS extensions in the linguistic domain, but it promises to provide a way to take advantage of the MDS methodology for studying variation in the meaning of complex constructions. In addition, it would allow for studying variation in meaning composition, which is one of the aims of semantic cross-linguistic research (von Fintel and Matthewson 2008).

A first area for which this approach has been applied is conditional sentences (Tellings 2020). Conditionals were extracted from the Europarl parallel corpus, which due to its register contains a high number of conditionals. The data were annotated for the tense forms in *if*-clause and main clause, as well as the modal structure of the conditional. Preliminary results show that English two-past conditionals (i.e., with past perfect in the *if*-clause) are mostly stable in translations, i.e. are mostly translated with the counterparts of the past perfect in Dutch and Spanish. On the other hand, the English simple past in conditionals, which can have either a modal or temporal interpretation, shows much more variation across languages.

5.2 Alternatives to MDS

MDS is but one of a range of dimensionality reduction methods. This range of methods is usually subdivided into methods that attempt to retain global structure of the data (like MDS) and those that attempt to retain local structure of the data (e.g. Local Linear Embedding, LLE; Roweis and Saul 2000). Finally, there are methods that aim to operate at both the global and the local level. t-Distributed Stochastic Neighbor Embedding, abbreviated t-SNE, is an example of the latter (van der Maaten and Hinton 2008). In this section, we briefly compare these three kinds of solutions and show their applications in linguistics.

Generally, the difference between global-first (or *full spectral*) and local-first (*sparse spectral*) dimensionality reduction algorithms is exemplified using an artificial dataset called the Swiss roll, pictured on the left side of Figure 10 below. In a true Swiss roll, a sponge cake is rolled up to create a distinctive swirl effect. Similarly, the data in this set are curved when taking a three-dimensional perspective, but flat from a two-dimensional perspective.

Figure from web source

Figure 10: On the left: an artificial Swiss roll dataset. On the right: comparison of Euclidic and geodesic distance.

The right part of Figure 10 shows two ways of measuring distance between points. On a global level, points A and R are regarded close together, as their Euclidean distance is low. However, in the original sponge cake, A and R would be rather far away, and only end up close after rolling up. When reducing the manifold to two dimensions, one preferably wants to arrive at a solution that retains the adjacency of the points in a two-dimensional display. Dimensionality reduction methods that aim to retain local structure therefore attach more weight to geodesic distance instead: as a distance measure, they use the shortest path in terms of nearby points. In turn, A and R are far apart.

For the dataset pictured on the left side of Figure 10, MDS will map distant data points in the three-dimensional manifold to nearby points in the Cartesian plane. In turn, as shown in Figure 11 below, MDS produces a rather similar two-dimensional output to the three-dimensional input data. Thus, MDS fails to identify the underlying two-dimensional structure of the Swiss roll manifold.

LLE rather intends to display local differences. As a consequence, LLE produces a low-dimensional solution that preserves the neighborhood of the manifold. For the Swiss roll dataset, as shown in Figure 11 below. A downside to LLE is that the method has a general tendency to crowd points at the center of the map, which prevents gaps from forming between potential clusters (van der Maaten and Hinton 2008, pp. 6–7).

In the t-SNE solution in Figure 11, we again see that the local structure of the data is kept intact in the two-dimensional solution, but some of the curvature of the input data is also still apparent. A downside to t-SNE is that it comes with some parameters (most importantly perplexity) that require manual tuning (Wattenberg et al. 2016).

Figure 11 seems to point into the direction that retaining local structure yields better solutions regardless. However, whether or not the Swiss roll problem plays a role is very much dependent on the dataset at hand. Moreover, there are arguments in favor of using MDS. For example, an MDS map can be interpreted at the global level (by interpreting dimensions, see section 4.1 above), while LLE and t-SNE generally aim at showing local structure in the data, Figure from web source

Figure 11: Comparison of performance of three dimensionality reduction methods on an artificial Swiss roll dataset displayed on the left.

and therefore are more suitable to identify clusters in the dataset (see section 4.2).

While MDS prevails as the main method used in (typological) linguistics, recent research has shown applications of t-SNE. For example, Asgari and Schütze (2017) apply t-SNE to marking of tense cross-linguistically. They show how grammatical markers, e.g. past tense marking with *ti* in Seychellois Creole, can function as pivots to find all past-referring contexts in the parallel corpus of Bible translations. Iteratively selecting more pivots, e.g. Fijian *qai*, then allows to discern different sub-types of past-referring contexts: *qai* is used as a past tense marker in narrative progression, but not in progressive nor modal contexts. Applying t-SNE on formal similarity in the parallel corpus as a whole then neatly shows clustering of these aforementioned functions in the domain of past reference.

While we are unaware of implementations of LLE to (re)generate semantic maps, this section, along with sections 4.1 and 4.2 above, has convincingly shown multiple visualizations are often required to arrive at a full interpretation of the dataset (cf. Cysouw 2008, p. 50).

6 Conclusion

This paper reviewed how multi-dimensional scaling is used to create semantic maps in linguistic typology and cross-linguistic semantics. We have seen that MDS stands for a collection of algorithms that can reduce the dimensionality of a highly complex dataset, and represent this visually. Starting with a notion of similarity between linguistic objects, applying MDS results in a visualization of both cross-linguistic variation and single-language patterns, which then can be used to answer a variety of linguistic research questions.

What makes reading the MDS literature in linguistics potentially difficult is that there is so much variation with respect to various parameters. These include the particular MDS algorithm that is used, the type of linguistic data used as input, the similarity measure between primitives, what the points on the map represent, how clusters and dimensions are interpreted, and the place that MDS maps occupy in the process of linguistic argumentation. By identifying and explaining these parameters in this paper, and introducing useful terminology for describing MDS studies (*map coloring*, *dimension interpretation*, *cluster interpretation*, etc.), we hope to have provided the means to make existing MDS-based work in linguistics more accessible.

At the same time, we hope this paper will prompt future MDS studies. We suggested two directions for future work in particular. First, the use of MDS in a setting in which multiple semantic features are at play in a compositional way, so that the MDS methodology can contribute to the study of cross-linguistic variation of compositional structures. Second, we discussed how MDS can be complemented and compared with other dimensionality reduction techniques.

References

Asgari, Ehsaneddin and Hinrich Schütze (2017). “Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Vol. 2. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 113–124. DOI: 10.18653/v1/D17-1011.

Beekhuizen, Barend, Julia Watson, and Suzanne Stevenson (2017). “Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns”. In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, pp. 112–117.

Borg, Ingwer and Patrick J. F. Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media. DOI: 10.1007/0-387-28981-X.

Borg, Ingwer, Patrick J. F. Groenen, and Patrick Mair (2018). *Applied multidimensional scaling and unfolding*. Springer. DOI: 10.1007/978-3-319-73471-2.

Bremmers, David, Jianan Liu, Martijn van der Klis, and Bert Le Bruyn (2020). “Translation Mining: definiteness across languages. A reply to Jenks (2018)”. Submitted.

Bryant, David and Vincent Moulton (2004). “Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks”. In: *Molecular Biology and Evolution* 21.2, pp. 255–265. ISSN: 0737-4038. DOI: 10.1093/molbev/msh018.

Cinque, Guglielmo (2007). “A note on linguistic theory and typology”. In: *Linguistic Typology* 11.1, pp. 93–106. DOI: 10.1515/LINGTY.2007.008.

Clancy, Steven J. (2006). “The topology of Slavic case: Semantic maps and multidimensional scaling”. In: *Glossos* 7.1, pp. 1–28.

Croft, William (2007). “Typology and linguistic theory in the past decade: A personal view”. In: *Linguistic Typology* 11.1, pp. 79–91. DOI: 10.1515/LINGTY.2007.007.

Croft, William and Keith T. Poole (2008). “Inferring universals from grammatical variation: Multidimensional scaling for typological analysis”. In: *Theoretical Linguistics* 34.1, pp. 1–37. DOI: 10.1515/THLI.2008.001.

Croft, William and Jason Timm (2013). “Using Optimal Classification for multidimensional scaling analysis of linguistic data”. URL: http://www.unm.edu/~wcroft/MDSfiles/MDSforLinguists-UserGuide.pdf.

Cysouw, Michael (2008). “Generalizing language comparison”. In: *Theoretical Linguistics* 34.1, pp. 47–51. ISSN: 03014428. DOI: 10.1515/THLI.2008.003.

Dahl, Östen (1985). *Tense and aspect systems*. Basil Blackwell.

Dahl, Östen (2014). “The perfect map: Investigating the cross-linguistic distribution of TAME categories in a parallel corpus”. In: *Aggregating Dialectology, Typology, and Register Analysis*. Berlin, Boston: De Gruyter Mouton, pp. 268–289. DOI: 10.1515/9783110317558.268.

Dahl, Östen and Bernhard Wälchli (2016). “Perfects and iamitives: two gram types in one grammatical space”. In: *Letras de Hoje* 51.3, pp. 325–348. ISSN: 1984-7726. DOI: 10.15448/1984-7726.2016.3.25454.

de Leeuw, Jan and Willem J. Heiser (1977). “Convergence of correction matrix algorithms for multidimensional scaling”. In: *Geometric representations of relational data*. Ed. by J. C. Lingoes. Mathesis Press, pp. 735–752.

de Leeuw, Jan and Patrick Mair (2009). "Multidimensional scaling using majorization: SMACOF in R". In: *Journal of Statistical Software* 31.3. DOI: 10.18637/jss.v031.i03.

de Swart, Henriëtte (2007). "A cross-linguistic discourse analysis of the perfect". In: *Journal of Pragmatics* 39.12, pp. 2273–2307. DOI: 10.1016/j.pragma.2006.11.006.

de Wit, Astrid, Frank Brisard, and Michael Meeuwis (2018). "The epistemic import of aspectual constructions: The case of performatives". In: *Language and Cognition* 10.2, pp. 234–265. DOI: https://doi.org/10.1017/langcog.2017.26.

De Swart, Peter, Hanne M. Eckhoff, and Olga Thomason (2012). "A Source of Variation: A Corpus-Based Study of the Choice between ἀπό and ἐκ in the NT Greek Gospels". In: *Journal of Greek Linguistics* 12.1, pp. 161–187. DOI: 10.1163/156658412X649760.

Ding, Cody S. (2013). "Multidimensional scaling". In: *The Oxford handbook of quantitative methods in psychology*. Ed. by Todd D. Little. Oxford University Press, pp. 235–256. DOI: 10.1093/oxfordhb/9780199934898.013.0012.

Ding, Cody S. (2018). *Fundamentals of applied multidimensional scaling for educational and psychological research*. Springer. DOI: 10.1007/978-3-319-78172-3.

Georgakopoulos, Thanasis and Stéphane Polis (2018). "The semantic map model: State of the art and future avenues for linguistic research". In: *Language and Linguistics Compass* 12.2, pp. 1–33. ISSN: 1749818X. DOI: 10.1111/lnc3.12270.

Greenhill, Simon J. (2011). "Levenshtein distances fail to identify language relationships accurately". In: *Computational Linguistics* 37.4, pp. 689–698. DOI: 10.1162/COLI_a_00073.

Groenen, Patrick J. F. and Ingwer Borg (2014). "Past, present, and future of multidimensional scaling". In: *Visualization and verbalization of data*. Ed. by Jörg Blasius and Michael Greenacre. CRC Press, Taylor & Francis Group, pp. 95–117.

Haiman, John (1985). *Natural Syntax*. Cambridge University Press.

Hartmann, Iren, Martin Haspelmath, and Michael Cysouw (2014). "Identifying semantic role clusters and alignment types via microrole coexpression tendencies". In: *Studies in Language* 38.3, pp. 463–484. DOI: 10.1075/sl.38.3.02har.

Haspelmath, Martin (1997). *Indefinite pronouns*. Oxford University Press. DOI: 10.1093/oso/9780198235606.001.0001.

Haspelmath, Martin (2003). "The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison". In: *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Ed. by Michael Tomasello. Vol. II. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 211–242. ISBN: 9781315777443.

Haspelmath, Martin (2010). "Comparative concepts and descriptive categories in crosslinguistic studies". In: *Language* 86.3, pp. 663–687. DOI: 10.1353/lan.2010.0021.

Hawkins, John A. (1988). "On generative and typological approaches to universal grammar". In: *Lingua* 74.2-3, pp. 85–100. DOI: 10.1016/0024-3841(88)90055-1.

Jolliffe, Ian T. and Jorge Cadima (2016). "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, p. 20150202. DOI: 10.1098/rsta.2015.0202.

Levshina, Natalia (2015). "European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles". In: *Folia Linguistica* 49.2, pp. 487–520. ISSN: 0165-4004. DOI: 10.1515/flin-2015-0017.

Levshina, Natalia (2016). "Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles". In: *Languages in Contrast* 16.1, pp. 84–117. ISSN: 1387-6759. DOI: 10.1075/lic.16.1.04lev.

Levshina, Natalia (2020). "Semantic maps of causation: What different data types and methods can (and cannot) tell us". DOI: 10.5281/zenodo.3675342.

Newmeyer, Frederick J. (2010). “On comparative concepts and descriptive categories: A reply to Haspelmath”. In: *Language* 86.3, pp. 688–695. DOI: 10.1353/lan.2010.0000.

Poole, Keith T. (2005). *Spatial models of parliamentary voting*. Cambridge University Press. DOI: 10.1017/CBO9780511614644.

Portner, Paul (2003). “The (temporal) semantics and (modal) pragmatics of the perfect”. In: *Linguistics and Philosophy* 26.4, pp. 459–510. DOI: 10.1023/A:1024697112760.

Regier, Terry, Naveen Khetarpal, and Asifa Majid (2013). “Inferring semantic maps”. In: *Linguistic Typology* 17.1, pp. 89–105. ISSN: 1613-415X. DOI: 10.1515/lity-2013-0003.

Roweis, Sam T. and Lawrence K. Saul (2000). “Nonlinear Dimensionality Reduction by Locally Linear Embedding”. In: *Science* 290.5500, pp. 2323–2326. ISSN: 00368075. DOI: 10.1126/science.290.5500.2323.

Talmy, Leonard (2000). *Toward a Cognitive Semantics*. MIT Press.

Tellings, Jos (2020). “Translation mining in the domain of conditionals: first results”. Presentation at CLIN 30, Utrecht.

Torgerson, Warren S. (1952). “Multidimensional scaling: I. Theory and method”. In: *Psychometrika* 17.4, pp. 401–419. DOI: 10.1007/BF02288916.

van der Auwera, Johan (2013). “Semantic maps, for synchronic and diachronic typology”. In: *Synchrony and Diachrony: A dynamic interface*. Ed. by Anna Giacalone Ramat, Caterina Mauri, and Piera Molinelli. Amsterdam: John Benjamins Publishing Company, pp. 153–176. DOI: 10.1075/slcs.133.07auw.

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte de Swart (2017). “Mapping the PERFECT via Translation Mining”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 497–502.

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte de Swart (2020). “A multilingual corpus study of the competition between PAST and PERFECT in narrative discourse”. Submitted.

van der Maaten, Laurens and Geoffrey Hinton (2008). “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9, pp. 2579–2605. ISSN: 1532-4435. DOI: 10.1007/s10479-011-0841-3. arXiv: 1307.1662.

Verkerk, Annemarie (2014). “Where Alice fell into: Motion events from a parallel corpus”. In: *Aggregating Dialectology, Typology, and Register Analysis*. Ed. by Benedikt Szmrecsanyi and Bernhard Wälchli. Berlin, Boston: De Gruyter Mouton, pp. 324–354. DOI: 10.1515/9783110317558.324.

Verkerk, Annemarie (2017). “The goal-over-source principle in European languages”. In: *Space in Diachrony*. Ed. by Silvia Luraghi, Tatiana Nikitina, and Chiara Zanchi. Amsterdam: John Benjamins Publishing Company, pp. 1–40. DOI: 10.1075/slcs.188.01ver.

von Fintel, Kai and Lisa Matthewson (2008). “Universals in semantics”. In: *The Linguistic Review* 25.1-2, pp. 139–201. DOI: 10.1515/TLIR.2008.004.

von Waldenfels, Ruprecht (2014). “Explorations into variation across Slavic: Taking a bottom-up approach”. In: *Aggregating Dialectology, Typology, and Register Analysis*. Ed. by Benedikt Szmrecsanyi and Bernhard Wälchli. Berlin, Boston: De Gruyter Mouton, pp. 290–323. DOI: 10.1515/9783110317558.290.

Wälchli, Bernhard (2010). “Similarity semantics and building probabilistic semantic maps from parallel texts”. In: *Linguistic Discovery* 8.1, pp. 331–371. DOI: 10.1349/PS1.1537-0852.A.356.

Wälchli, Bernhard (2018). “‘As long as’, ‘until’ and ‘before’ clauses: Zooming in on linguistic diversity”. In: *Baltic Linguistics* 9, pp. 141–236. DOI: 10.32798/bl.372.

Wälchli, Bernhard and Michael Cysouw (2012). "Lexical typology through similarity semantics: Toward a semantic map of motion verbs". In: *Linguistics* 50.3, pp. 671–710. ISSN: 00243949. DOI: 10.1515/ling-2012-0021.

Wattenberg, Martin, Fernanda Viégas, and Ian Johnson (2016). "How to Use t-SNE Effectively". In: *Distill*. DOI: 10.23915/distill.00002.

Wieling, Martijn and John Nerbonne (2015). "Advances in dialectometry". In: *Annual Review of Linguistics* 1, pp. 243–264. DOI: 10.1146/annurev-linguist-030514-124930.

Zwarts, Joost (2008). "Commentary on Croft and Poole, Inferring universals from grammatical variation: Multidimensional scaling for typological analysis". In: *Theoretical Linguistics* 34.1, pp. 67–73. ISSN: 0301-4428. DOI: 10.1515/THLI.2008.006.