Transparant aggregation of linguistic variables with Individual Differences Scaling

Tom Ruette and Dirk Speelman

1 Introduction

The goal of the current paper is to introduce an existing, but as far as we know not yet linguistically applied Multidimensional Scaling (MDS) method called *Individual Differences Scaling* (henceforth, INDSCAL) to the field of variationist aggregation studies. In such aggregation studies, e.g. Seguy (1971), Goebl (1982), Geeraerts et al. (1999) or Nerbonne (2006), many linguistic variables are considered simultaneously to reveal a structure among the language varieties that are being described by the input variables. Although the focus of aggregation studies is specifically on the structure of language varieties, their explanatory power is reduced by the fact that the behavior of the underlying linguistic variables is completely obscured (Horan, 1969).[1] Technically speaking, the loss of the behavior of the variables is due to the fact that all the variables are averaged out by means of an aggregating distance metric to form a single varieties × varieties distance matrix. This step is tangibly explained by Szmrecsanyi (2011, Fig. 1). Therefore, we propose to use INDSCAL as a different aggregation technique that can take a distance matrix for every (group of) variable(s) as its input, and which is able to reveal both the structure of the varieties and the structure of the (groups of) variables. Intuitively, one could say that the aggregation of the variables is postponed and moved into the MDS phase of the analysis.

To account for individual differences is an emerging trend in aggregation studies. Heeringa (2004, p. 268–270) already used complex calculations to retrieve which individual variables are strongly represented in the clusters that emerge from the aggregation. More in line with INDSCAL, the recent work of e.g. Spruit et al. (2009) shows how dialectometrists are already producing separate distance matrices for different types of variables, but without an overarching method (except for basic correlation measures) to bring the separate results together. A similar approach can be found in Cysouw et al. (2008), where the distance matrices for individual features are compared to the aggregated solution. The benefit of INDSCAL over these more post-hoc approaches is that INDSCAL is an integrated branch of the well-established Multidimensional Scaling method, that is already commonly used in aggregation studies. This link to the existing framework of aggregation studies is also its main benefit over the use of bipartite spectral graph partitioning (Wieling and Nerbonne, 2011) or Generalized Additive Models with mixed-effects (Wieling et al., 2011)

This paper is structured as follows. We explain INDSCAL in Section 2 by introducing the specific terminology of the method, interpreting an example analysis and giving some mathematical properties. To show the value of INDSCAL for variationist aggregation studies, we apply the method to a dataset that was gathered by Geeraerts et al. (1999) to show lexical convergence between two national varieties of Dutch. We revisit the data and findings from Geeraerts et al. (1999) in Section 3. In Section 4 we apply INDSCAL to that dataset and we show how the INDSCAL analysis confirms and extends the previous findings. Finally, we conclude the paper by summing up further possible applications of INDSCAL in Section 5.

2 Individual Differences Scaling

Individual Differences Scaling, abbreviated as INDSCAL, is a fairly standard Multidimensional Scaling (MDS) technique that is described in most MDS textbooks, e.g. Cox and Cox (2001) or Borg and Groenen (2005). These textbooks usually begin by introducing *two-way* MDS, where two-way refers to the fact that the dissimilarity input matrix has two dimensions (rows and columns), representing the proximities between pairs of objects. Because the rows and the columns carry the same objects, this is called *one-mode* input. Usually, these proximities are averages of proximities from multiple sources, e.g. test subjects or object characteristics. However, one of the objections against two-way one-mode MDS, made already by Horan (1969) and many people since, is that the averaging and aggregation of many sources into a single distance matrix is sometimes not acceptable, because it does no justice to the individual differences between the sources[2]. Therefore, Carroll and Chang (1970) proposed the *Individual Differences Scaling* method, abbreviated as INDSCAL. INDSCAL is a type of *three-way* MDS, and can take several objects × objects matrices as its input, thus objects × objects × sources. Because there are two types of input, i.e. objects and sources, this is called a *two-mode* input. Typically, it is used to show the individual differences between a number of judges (sources) who have rated the objects under investigation. Well-known examples are *The Whisky Tasting experiment* and *The Body Part study*, both of which will be discussed further on.

Let it be clear, however, that this method still assumes considerable similarity between the sources, just as is required for a two-way MDS. Indeed, if there is not at least some consensus among all the sources, aggregation makes no sense. INDSCAL does allow for somewhat more variation between the sources than a two-way one-mode MDS, but not to the extent that the sources do not share some underlying characteristics, sensitivities or judgmental processes, which can become the dimensions of the MDS solution (Arabie et al., 1987, p. 21). In other words, individual differences scaling as a method can be applied to datasets that are somewhere between two extremes. The negative extreme is that aggregation is not possible because the sources that we want to aggregate are too different from each other. The positive extreme is that aggregation is the obvious thing to do, because the sources behave all very similar and there is no need to account for the behavior of the individual sources.

Before we can look into an example output of INDSCAL, we need to introduce some terminology. The input of an INDSCAL analysis is an array of proximity matrices[3]. Every proximity matrix gives the (dis)similarities between all pairs of objects, according to a source that quantifies and estimates these proximities. In the Whisky Tasting experiment, n whisky experts are asked to compare all possible pairs of whiskies. At the end of the experiment, there are n proximity matrices, and every matrix represents the judgements of a single whisky expert. The output of an INDSCAL analysis then consists of two parts: the *Group Stimulus Space* and the *Configuration Weights*. The Group Stimulus Space (also called *Stimulus Space*, *Group Space*, *Object Space* or *Common Space*) shows the low-dimensional solution for the objects (e.g. whiskies) that is characteristic of the entire group of sources (e.g. all whisky experts together). This solution can be interpreted in the same way as the solution of a two-way MDS. The Configuration Weights (also called *Source Weights*) indicate the importance attributed to each dimension of the Group Stimulus Space by each source (e.g. the whisky experts). Although there is quite some mathematical complexity behind these Configuration Weights (Arabie et al. (1987, p. 17–25), Borg and Groenen (2005, Chapter 22)), the gist of the approach is the following. A Configuration Weight of 1 means that the source (e.g. whisky expert) agrees with the distinction made on the respective dimension of the Group Stimulus Space. If the Configuration Weight is smaller than 1, the respective dimension is shrunken in the source's perception, effectively giving less importance or weight to the distinction that is made by the dimension. If the Configuration Weight is larger than 1, the respective dimension is stretched in the source's perception, and thus the respective distinction should be given more importance. However, it is not allowed to interpret the Configuration Weights as percentages relative to some baseline or as probabilities, e.g. source A gives twice as much importance to this dimension as source B (Arabie et al., 1987, p. 23). Obviously, in order to benefit from the explanatory power of the Configuration Weights, the interpretation of the Group Stimulus Space should be based on meaningful dimensions. A non-dimensional interpretation (Borg and Groenen, 2005, Chapter 4), i.e. an interpretation of the results that is not directly linked to the dimensions in the algorithm's output, is not suited for INDSCAL. By matrix multiplying the Group Stimulus Space with the Configuration Weights of a specific source (see Equation 2), the *Private Object Space* for the source is generated, which can be regarded as the view of the source, if he were forced to accomodate his judgements along the consensus dimensions of the Group Stimulus Space (e.g. the judgement of a single whisky expert squeezed into the dimensions of the Group Stimulus Space). Note that the Private Object Space of a certain source is not necessarily equivalent to the two-way MDS of that source's distance matrix.

Let us introduce the interpretation of the output of the method with the frequently cited Body Parts example of Jacobowitz (1973). The goal of Jacobowitz (1973) is to discover how people conceptualize the human body, and whether this conceptualization is different for children and adults. To find this out, he asked 15 children and 15 adults to give similarity ratings for a number of body parts. The results of the INDSCAL analysis, performed by Takane et al. (1977), of his 30 dissimilarity matrices are visually presented in Fig. 1. The three dimensional Group Stimulus Space in Fig. 1a can be dimensionally interpreted just as one would do with the solution of a two-way MDS. The Group Stimulus Space represents the conceptual dimensions with which all individuals can agree to a certain degree; the degree with which they agree is captured in the Configuration Weights. On the first dimension of the Group Stimulus Space (vertical), a distinction between the head and the limbs is made. Dimension two (horizontal) distinguishes the legs from the arms. And the third dimension (depth) expresses a whole-part relationship with a cline from the full body at the front over head, leg and arm in the middle, to ear, toe and finger at the back. The cubes of the Configuration Weights plot in Fig. 1b answer the question *how much importance do children and adults give to the distinctions made by the consencus dimension of the Group Stimulus Space?* In Fig. 1b, the zero-value of all the Configuration Weights, which indicates a spot where none of the distinctions from the Group Stimulus Space are deemed important, is plotted in the left bottom corner at the back; the Configuration Weights of the adults are indicated with a black cube, whereas the Configuration Weights of the children are indicated with a white cube. It now becomes immediately clear that the adults and children value the distinction of Dimension 2 (left to right) differently. The adults are generally closer to the origin of Dimension 2, indicating that they give little importance to this dimension, which distinguished the arms from the legs. Children, however, have higher Configuration Weights for Dimension 2, and this means that they make a more pronounced distinction between arms and legs.

Intuitively, the rationale behind INDSCAL is easy: every source is represented in its own distance matrix; INDSCAL aggregates these distance matrices for the aggregate perspective (yielding the Group Stimulus Space); and finally, the aggregate perspective is compared to the individual input matrices (yielding the Configuration Weights). Mathematically speaking, however, three-way Multidimensional Scaling is fairly complex and its development knows many branches and competing approaches. Instead of giving a detailed account of its developmental history, the mathematical properties of the different approaches and a review of the available implementations, we refer the reader to the literature in Arabie et al. (1987), Cox and Cox (2001, Chapter 10) and Borg and Groenen (2005, Chapter 22). For our analyses, we stick to the INDSCAL approach of Carroll and Chang (1970) and we use an implementation of INDSCAL in the *SMACOF* package for *R* by de Leeuw and Mair (2009). The package actually offers a specific way of finding the optimal lower-dimensional MDS solution, called *Scaling by MAjorizing a COmplicated Function*, abbreviated as SMACOF, first proposed by de Leeuw (1977) and described in Cox and Cox (2001, Section 11.2) and Borg and Groenen (2005, Chapter 8). The SMACOF approach in the *R* package is applied to all sorts of metric and non-metric branches of multidimensional scaling, including the INDSCAL approach to three-way MDS. We will make use of the out-of-the-box implementation for our example analysis of a variationist dataset, taken from Geeraerts et al. (1999).

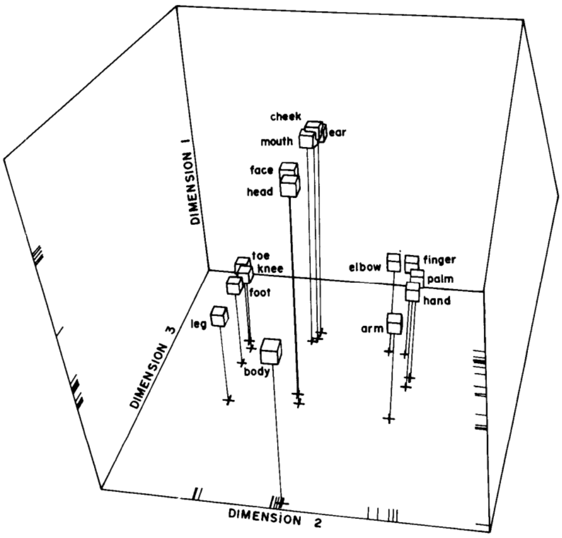
Before we deal with the example analysis, we give the high level mathematical properties of INDSCAL. In general, the problem for (metric) INDSCAL three-way MDS consists of representing the given dissimilarity δ_{ijk} between objects i and j as seen by individual k by a Euclidian distance d_{ijk}. Given a solution $\mathbf{G}$ and Configuration Weights $\mathbf{W}_k$, that distance can be calculated as follows:

$$\begin{aligned} d_{ijk}(\mathbf{GW}_k) &= [\sum_{a=1}^{m}(w_{aak}g_{ia} - w_{aak}g_{ja})^2]^{\frac{1}{2}} \\ &= [\sum_{a=1}^{m}(w_{aak}^2(g_{ia} - g_{ja})^2]^{\frac{1}{2}} \end{aligned} \quad (1)$$

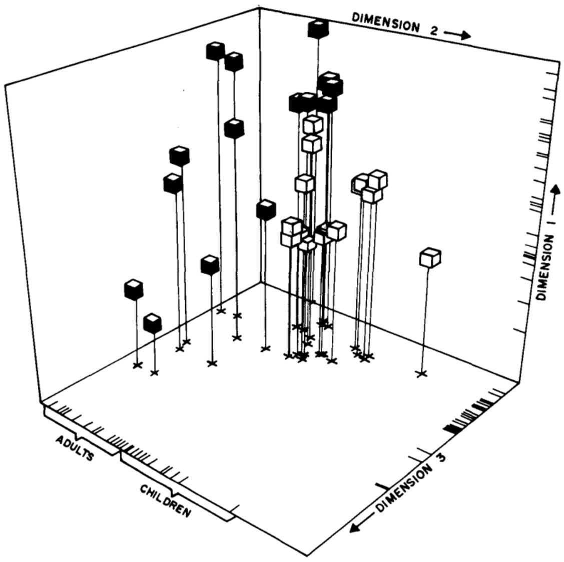
where $i, j = 1, \ldots, n$; $k = 1, \ldots, K$; $a = 1, \ldots, m$; $\mathbf{W}_k$ is an $m \times m$ diagonal matrix of nonnegative weights w_{aak} for every dimension a for individual k; and $\mathbf{G}$ is the matrix of coordinates of the Group Stimulus Space $\mathbf{G}$. For an individual k, its individual configuration $\mathbf{X}_k$ should be:

$$\mathbf{X}_k = \mathbf{GW}_k \quad (2)$$

To find the optimal solution for $\mathbf{X}_k$ and $\mathbf{G}$, the difference between the given individual dissimilarities and the distances, summed over the calculated individual



(a) Group Stimulus Space: similarity of body parts according to 30 raters



(b) Configuration Weights: individual differences between 30 body parts similarity raters

Fig. 1: Example of INDSCAL output, based on Jacobowitz (1973) and taken from Takane et al. (1977) (Permission pending)

configurations needs to be as small as possible. This difference is called *stress*, represented by σ. The stress that needs to minimized can be written as follows:

$$\sigma(\mathbf{X}_1,\ldots,\mathbf{X}_k) = \sum_{k=1}^{K}\sum_{i<j}(\delta_{ijk} - d_{ij}(\mathbf{X}_k))^2 \tag{3}$$

The approach to minimize this total stress is called *majorization* (de Leeuw, 1977). We refer the reader to Borg and Groenen (2005, Chapter 8) for a detailed elaboration of majorization, its application in MDS, and the non-metric flavour of INDSCAL.

3 The dataset of Geeraerts et al. (1999)

The goal of the current paper is to show how INDSCAL can be applied to linguistic aggregation studies. Therefore, we will perform INDSCAL on a dataset that has been compiled and analysed in Geeraerts et al. (1999). One of the goals of the monograph was to empirically investigate whether there is diachronic convergence or divergence between 1950 and 1990 in the lexicon of Dutch as spoken in Belgium and The Netherlands. With this goal in mind, a list of 32 concepts from two lexical fields, i.e. *Football* and *Clothes*, was manually collected. For every concept, words that name this concept were listed and counted in the corpus-materials introduced below. As an example, the concept OFF-SIDE can be named in Dutch with the words *buitenspel* or *offside*, or the concept DRESS can be named with *jurk*, *japon* or *kleed*. The distance between two varieties is measured by means of observed preferences for choosing a certain word to name a concept. The actual distance metric is introduced below, but intuitively one could say that if both varieties prefer *buitenspel* over *offside* to name the concept OFF-SIDE, they are closer together then if one of the varieties prefers *buitenspel* and the other variety prefers *offside*.

As Geeraerts et al. (1999) want to study the diachronic movement of two national varieties of Dutch empirically, these words have to be attested in actual language material, representative for the two variational dimensions (temporal dimension and national dimension). Therefore, they collected magazines and newspapers from Belgium and The Netherlands (national dimension) which were written around 1950, 1970 and 1990 (temporal dimension). In this corpus, the occurrences of all the words that name the football and dress concepts were recorded and brought together in a table, of which a sample can be found in Table 1[4]. The complete table can be found in Geeraerts et al. (1999, Appendix 1). In this table, the concept is identified with a descriptive English name at the beginning of the line in small caps, the actual word that names the concept follows at the second position in italics. After that, the frequencies with which this word occurs in the national-temporal specific subsets of the data: N50 refers to Netherlandic material from 1950, B90 refers to Belgian material from 1990, etc.

CONCEPT	*variant*	N50	B50	N70	B70	N90	B90
KICK-OFF	*aftrap*	2	8	8	22	14	66
	kick-off	0	3	1	6	0	2
OFF-SIDE	*buitenspel*	17	9	21	28	18	5
	off-side	7	13	3	2	2	1
FOUL	*foul*	0	17	0	2	0	0
	fout	9	18	1	47	0	9
	overtreding	12	0	16	26	49	20

Table 1: Some examples from the frequency table in Geeraerts et al. (1999)

The findings of Geeraerts et al. (1999) concerning the evolution of two national varieties of Dutch point in the direction of convergence during a period of forty years.

For both lexical fields, there seems to be an increasing convergence between Belgian Dutch and Netherlandic Dutch along the three measure points 1950, 1970 and 1990. In the analysis of Geeraerts et al. (1999), a systematic analysis of the behaviour of the individual concepts and of how this relates to the overall patterns is missing. Therefore, we propose the INDSCAL analysis.

4 INDSCAL analysis of Geeraerts et al. (1999) data

The INDSCAL method can be applied to the lexical convergence and divergence study of Geeraerts et al. (1999) if we substitute the individuals (e.g. that rated the similarities between body parts) by the lexical fields or the individual concepts, and if we substitute the objects (e.g. the body parts) by the national varieties at the three measuring points[5]. The Group Stimulus Space will then show the lectal dimensions along which the national varieties are distributed, and the Configuration Weights will inform us about the importance that the lexical fields or concepts give to these lectal dimensions. Given the input data, we expect to find a structure of the varieties along a diachronic dimension and a national dimension, present in the variation of the concepts or lexical fields.

To show the application possibilities of INDSCAL, we will perform two analyses. In a first analysis, we consider the distances between the varieties per lexical field — thus a dissimilarity matrix for *Clothes* and a dissimilarity matrix for *Football* — so that the three-way input is variety × variety × lexical field. As there are only two lexical fields, the Configuration Weights of this analysis will be easy to interpret. This first example will actually be overly simplistic because there are only two sources. Technically, it is advised to have at least five sources in an INDSCAL analysis, as will be the case in the more realistic example with many individual concepts, introduced in Section 4.3. However, we chose to first illustrate the procedure with a case that is as simple as possible. The second analysis looks at the distances between the varieties for every concept in the *Football* lexical field. For every concept, a distance matrix is constructed, so that the three-way input is variety × variety × concept. As there are 15 concepts in the *Football* lexical field, we will have a more complicated Configuration Weights scatterplot which will give us the opportunity of showing the more advanced interpretations one could make. Although we could show the analysis for the *Clothes* field and the analysis that considers all concepts from both fields, we restrict ourselves to the above cases to keep this paper from becoming too repetitive, as our goal is primarily to introduce a new method.

4.1. Distance metric

To construct a distance matrix of the varieties — or rather an array of distance matrices —, the lexical distances per lexical field or concept need to be measured on the basis of the attested frequencies in Geeraerts et al. (1999). In the original 1999 study, the lexical similarity between two varieties V_1 and V_2 for a certain concept L is captured by the Uniformity metric $U_L(V_1, V_2)$ in Equation 4. In a later study (Speelman et al., 2003), the City-Block distance $D_{CB,L}(V_1, V_2)$ of Equation 7 was used.

$$U_L(V_1, V_2) = \sum_{i=1}^{n} \min(R_{V_1,L}(x_i), R_{V_2,L}(x_i)) \tag{4}$$

This Uniformity metric is a similarity metric and related to the City-Block distance metric D_{CB} as follows: $U_L(V_1, V_2) = 1 - D_{CB,L}(V_1, V_2)$. As an MDS method relies on distances as its input, we will use the City-Block distance metric presented in Speelman et al. (2003, Section 2.2 and 2.3). For completeness, we repeat the details of this distance metric below. The main advantage of the proposed Uniformity similarity metric and City-Block distance metric is that it takes the level of the concept into account. Instead of aggregating over the frequencies of all individual words, the overlap in relative preferences for choosing a specific word to name a concept are aggregated. The advantages of this onomasiological semantic control have been shown in Speelman et al. (2003).

Now, we revisit the details of the distance metric. Given two subcorpora V_1 and V_2 that represent two of the varieties under scrutiny, a concept L (e.g. FOUL) and x_1 to x_n the list of words (e.g. {*foul, fout, overtreding*}) that can refer to the concept L, then we define the absolute frequency F of the usage of x_i for L in V_j with:

$$F_{V_j,L}(x_i) \tag{5}$$

Subsequently, we introduce the relative frequency R:

$$R_{V_j,L}(x_i) = \frac{F_{V_j,L}(x_i)}{\sum_{k=1}^{n}(F_{V_j,L}(x_k))} \tag{6}$$

Now we can define the lexical City-Block distance D_{CB} between V_1 and V_2 on the basis of concept L as follows (the division by two is for normalization, mapping the results to the interval [0,1]):

$$D_{CB,L}(V_1, V_2) = \frac{1}{2}\sum_{i=1}^{n} |R_{V_1,L}(x_i) - R_{V_2,L}(x_i)| \tag{7}$$

The City-Block distance is a straightforward descriptive dissimilarity measure that assumes the absolute frequencies in the sample-based profile to be large enough to be good estimates for the relative frequencies. If however the samples are rather small, the relative frequencies become unreliable, and a supplementary control is needed. For this we measure the confidence of there being an actual difference between two profiles with the Log Likelihood Ratio test described in Dunning (1993). This time, unlike with D_{CB}, we look at the absolute frequencies in the profiles we compare. When we compare a profile in one language variety to the profile for the same concept in a second language variety, we use a Log Likelihood Ratio test to test the hypothesis that both samples are drawn from the same population. We use the p-value from the Log Likelihood Ratio test as a filter for D_{CB}. We set the dissimilarity between subcorpora at zero if $p > 0.05$, and we use D_{CB} if $p < 0.05$.[6] The argument for setting D_{CB} to zero if the two samples appear to be drawn from the same population (a single language variety), i.e. if the $p > 0.05$, is that there is no statistical evidence that the two samples come from a different population, and thus their lexical distance should be zero.

To calculate the dissimilarity between subcorpora on the basis of many concepts, e.g. all concepts from the lexical field *Football*, we just sum the dissimilarities for the individual concepts. In other words, given a set of concepts L_1 to L_m, then the global dissimilarity D between two subcorpora V_1 and V_2 on the basis of L_1 up to L_m can be calculated as:

$$D_{CB}(V_1, V_2) = \sum_{i=1}^{m} D_{L_i}(V_1, V_2) W(L_i) \quad (8)$$

The W in the formula is a weighting factor. We use weights to ensure that concepts which have a relatively higher frequency (summed over the size of the two subcorpora that are being compared[7]) also have a greater impact on the distance measurement. The sum of all weights is one. In other words, in the case of a weighted calculation, concepts that are more common in everyday life and language are treated as more important.

However, the W only applies when multiple concepts are being aggregated into a single distance matrix. In the case of the *Football* and *Clothes* example below, the W weighs the football concepts in the construction of the *Football* distance matrix, and similarly for the *Clothes* distance matrix, but the *Football* lexical field is considered to be equally important as the *Clothes* lexical field. In the case of the *Football* example further down, where every concept of the *Football* lexical field is the basis for a separate distance matrix, Equation 8 does not even come into play, and all concepts are considered equally important. The conceptual weighting is in that case completely absent. Although we would like to include the conceptual weighting in the INDSCAL approach in future research, the situation as presented in this paper is equivalent to the U metric of Geeraerts et al. (1999, p. 41). An approach that incorporates the conceptual weighting would be equivalent to the (unweighted) U' metric (Geeraerts et al., 1999, p. 42). Although there are out-of-the-box possibilities to incorporate the conceptual weighting, we stick to the unweighted approach for now, in order to keep the introduction of INDSCAL as straightforward as possible in this article.

4.2. Football and Clothes

As a first example, we will look at the structure of the language varieties according to the two lexical fields *Football* and *Clothes.* The focus lies on the behavior of the lexical fields, and not on the individual concepts. Therefore, we aggregate the lexical distances of all concepts per lexical field using the City-Block distance metric, introduced above. This gives us an array of two distance matrices — one for *Football* and one for *Clothes* — that we can use as input for the (non-metric) INDSCAL analysis[8]. The analysis produced the Group Stimulus Space in Fig. 2 and the scatterplot of Configuration Weights in Fig. 3. The low stress value of 1% indicates that there is not much difference between the input distances and the two-dimensional representation. To aid the interpretation, we manually added lines to the Group Stimulus Space that indicate the diachronic connection between the subcorpora. Note, that INDSCAL allows us to create higher dimensional representations, too.

The interpretation of the dimensions of the Group Stimulus Space is straightforward: Dimension 1 distinguishes the subcorpora of measuring point 1950 from the subcorpora of 1970 and 1990. Dimension 2 represents the distinction between Belgian and Netherlandic subcorpora. In general, one can observe the convergence of Belgian and Netherlandic Dutch terminology for *Football* and *Clothes* very clearly, as the national subcorpora are farther apart for 1950 than for 1990. A reader familiar to two-way MDS or Factor Analysis might suggest to slightly rotate the Group Stimulus Space counterclockwise, so that the dimensions become clearer. However, in the case of three-way MDS, this kind of arbitrary rotations is prohibited, as these rotations

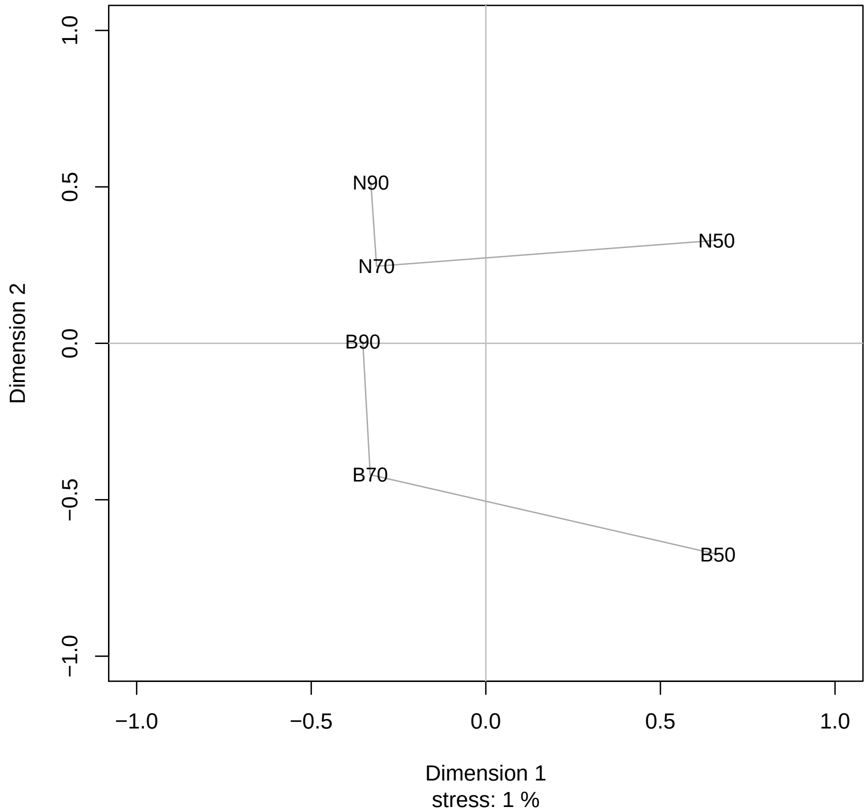


Fig. 2: Group Stimulus Space for the combination of the *Clothes* and *Football* field

would have an influence on the Configuration Weights. Given that the solution which comes out of the INDSCAL analysis is already the best possible representation of the data, any change in the Group Stimulus Space and consequently in the Configuration Weights would reduce the goodness-of-fit.

The interpretation of the scatterplot of the Configuration Weights needs some more guidance. Remember that in a two-dimensional solution a Configuration Weight coordinate of $(1, 1)$ means that the source completely agrees with the proposed Group Stimulus Space. A Configuration Weight of less than 1 means that the source would shrink the dimension and thus de-emphasize the distinction made along that dimension. A Configuration Weight of more than 1 means that the source would stretch the dimension and thus give more importance to the distinction made along that dimension. The scatterplot of Configuration Weights in Fig. 3 shows immediately that the two lexical fields behave differently. The *Football* terminology has a Configuration Weight smaller than 1 for Dimension 1 and this indicates that *Football* terminology gives less weight to the temporal dimension. The dissimilarities between the 1950s and the 1990s are considered to be smaller, which means that *Football* terminology has not changed much over the decades. The Configuration

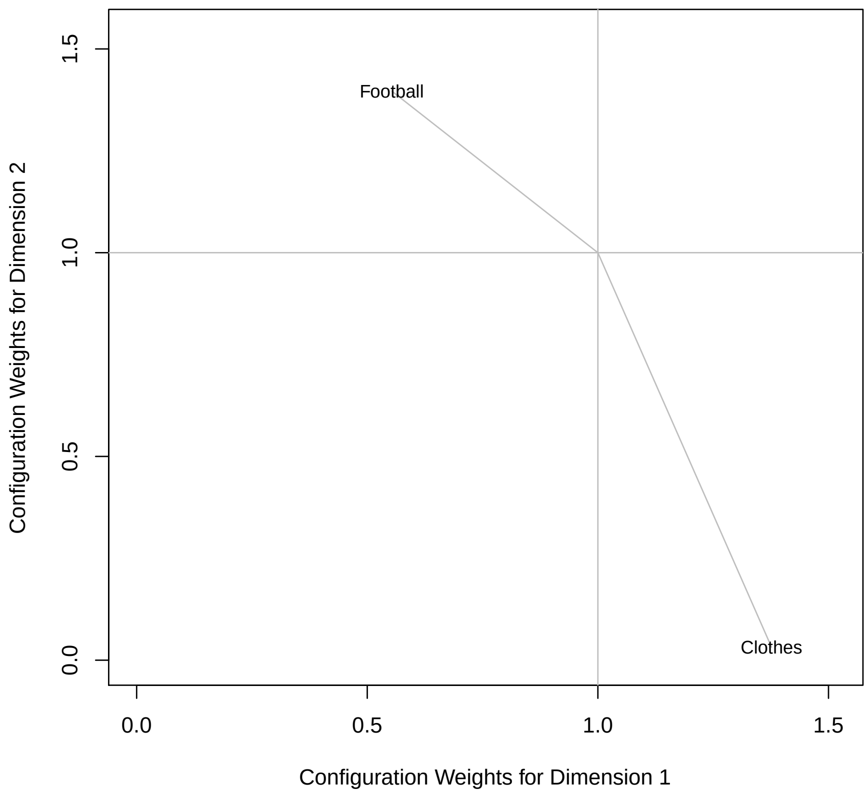


Fig. 3: Configuration Weights for the combination of the *Clothes* and *Football* field

Weight of the *Football* field for Dimension 2 is larger than 1. This means that the national dissimilarities are increased. In other words, *Football* terminology is more different in Belgium and The Netherlands than the consensus of the Group Stimulus Space suggests. The interpretation for the *Clothes* field are exactly opposite: the first dimension is emphasized and the second dimension is de-emphasized. It is important to point out that, although these differences are clearly present in the visualizations, INDSCAL is merely an exploratory technique, designed to assist the researcher in formulating hypothesis. Any strong conclusions should be based on further confirmatory statistics.

4.3. Football

As a second example, we will now perform a more detailed analysis of the *Football* lexical field. This time, we will consider every concept as a single source. As Geeraerts et al. (1999) came up with 15 concepts in the *Football* field, an array of 15 distance matrices, generated with Equation 7 will be the input of the INDSCAL analysis. A two dimensional solution yielded an outcome with an unacceptably high stress value.

Therefore, we calculated a three dimensional solution to obtain the stress value of 15%, which is just acceptable. In order not to overcomplicate the introduction of INDSCAL, we will only discuss the first two dimensions. This is in fact common practice in the MDS framework.

The analysis produced the Group Stimulus Space in Fig. 4 and the scatterplot of Configuration Weights in Fig. 7. Note now that we are only using the first part of the City-Block distance metric that was introduced above: only Equation 7 is needed to construct the distance matrix for a single concept, and Equation 8 is not applied here, effectively removing the conceptual weighting W.

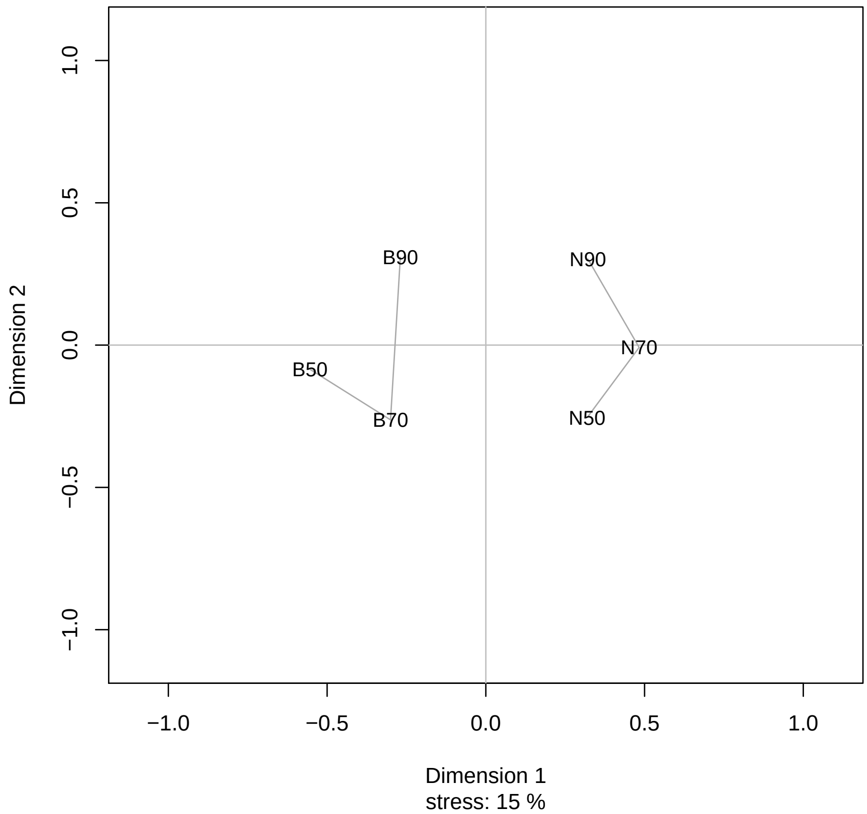


Fig. 4: Group Stimulus Space for the *Football* field

The Group Stimulus Space in Fig. 4 splits the Belgian and Netherlandic subcorpora on the first dimension. The second dimension sorts the subcorpora diachronically. The promotion of the national dimension to the first dimension, in comparison to the lexical field example above, is not suprising. In Fig. 3 we already saw that the *Football* domain values the national distinction much stronger than the temporal distinction. Although the diachronic order is perfect for the Netherlandic subcorpora, the position of the Belgian subcorpus from the 1950s is off. This strange position may be due to the high stress value of the analysis. However, introducing the residuals plot in Fig. 5

— a plot that is somewhat comparable to a Shepard plot in two-way MDS —, we see no extreme deviations between the input dissimilarities (x-axis) and the calculated distances in the Group Stimulus Space (y-axis) for the concepts, which indicates a very accurate lower dimensional representation. Moreover, introducing the stress decomposition plot in Fig. 6 — a plot that shows which objects are responsable for most of the stress —, we find that the 1950s Belgian subcorpus was actually one of the subcorpora with the lowest amount of stress.

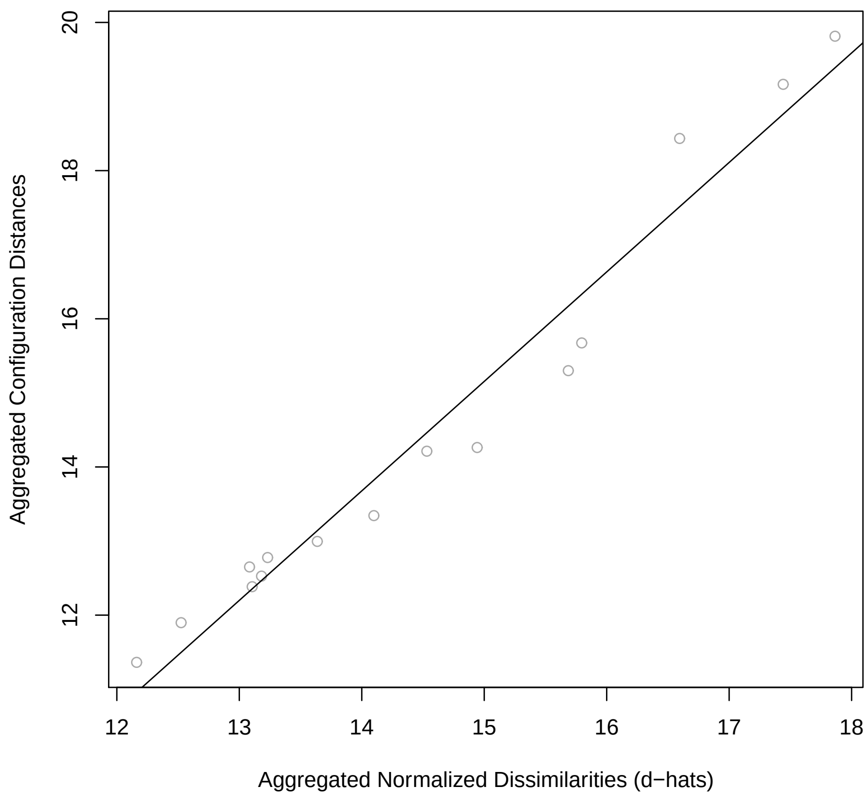


Fig. 5: Residuals for the *Football* field

If the strange position of the 1950s Belgian subcorpus is thus not an artifact of INDSCAL, one could propose a not too far-fetched interpretation. The clear alignment of Belgian football terminology between 1950 and 1970 with the Netherlandic terms of the 1950 seems plausible in the light of the Belgian language policy that was followed during the 60s, stating that Belgian speakers should embrace the Netherlandic norm. Previous researchers have hypothesized that this language policy could cause a certain *retardation* effect on Belgian Dutch: before the Netherlandic (N50) norm is accepted in Belgium (from B50 to B70), the Netherlandic situation changed already (N70). Although the hypothesis seems to have some visual support in our analysis, Geeraerts et al. (1999, p. 69) do not find statistically significant proof for this.

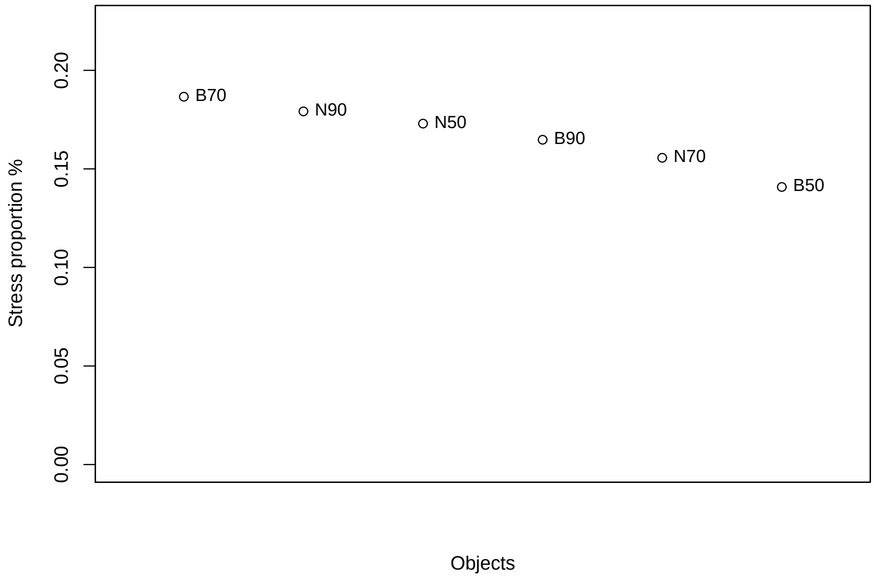


Fig. 6: Stress decomposition for the *Football* field

The interpretation of the scatterplot of Configuration Weights in Fig. 7 for individual concepts is now more challenging than in the lexical field example. The concepts form a fan around the $(1,1)$ point, which indicates that the concepts do not all agree with the proposed Group Stimulus Space. In itself, this shows the importance of performing INDSCAL: whereas a typical aggregation (and two-way MDS analysis) would have assumed that all concepts behave similarly, the three-way MDS analysis makes the diverging behavior of the individual concepts explicit. Figure 7 can be interpreted on the basis of the four quadrants that are formed by the horizontal and vertical grey line at $x = 1$ and $y = 1$. The first quadrant (left upper corner) contains the concepts with a Configuration Weight smaller than 1 for the first dimension and a Configuration Weight larger than 1 for the second dimension. These concepts thus alter the Group Stimulus Space so that the diachronic evolution is emphasized and the national distinction is made smaller. The second quadrant (right upper corner) contains concepts that have a Configuration weight larger than 1 for boh Dimension 1 and 2. These concepts thus merely stretch up the Group Stimulus Space. In the third quadrant (right bottom corner), the concepts have a Configuration Weight larger than 1 for the first dimension and smaller than 1 for the second dimension. The behavior of these concepts is thus the opposite of the concepts in the first quadrant: they promote the national distinction and they de-emphasize the diachronic evoluation. The fourth quadrant contains only one concept, which has Configuration Weights smaller than 1 for both dimensions. Mirroring the second quadrant, this concept merely shrinks the Group Stimulus Space.

The two most interesting quadrants are the first and the third. To get a visual grip on what is happening in these quadrants, we produce the Private Object Spaces

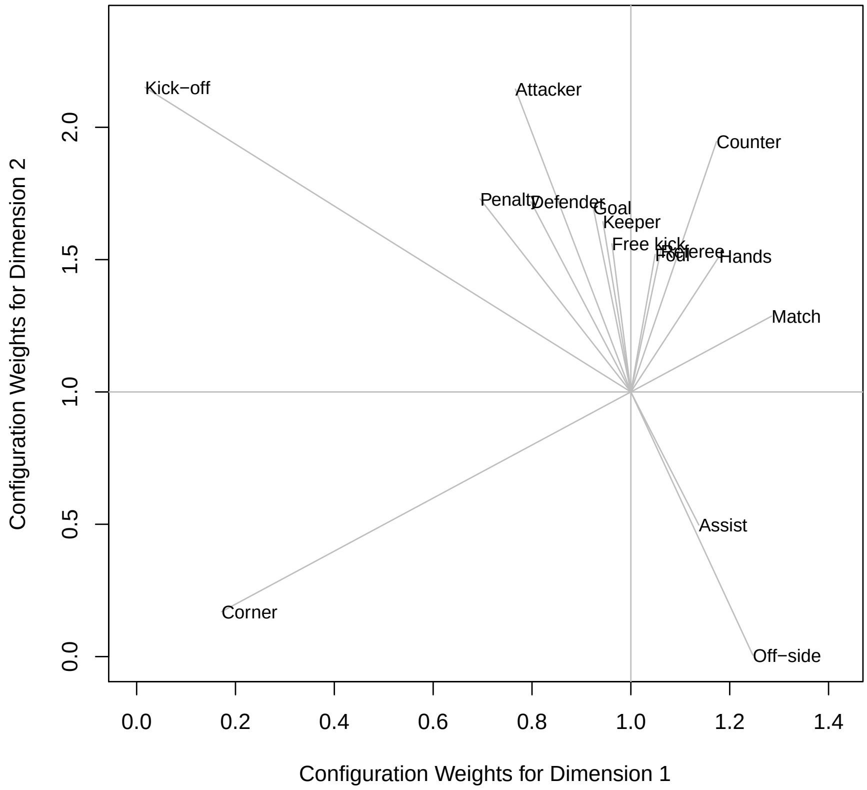


Fig. 7: Configuration Weights for the *Football* field

for the most extreme concepts (KICK-OFF and OFF-SIDE) in Fig. 8 and Fig. 9 to see how they differ from the Group Stimulus Space. As mentioned above, the Private Object Space shows the view of a single source if he were forced to accomodate his judgements along the consensus dimensions of the Group Stimulus Space. In the case of individual concepts, this shows how sensitive a single concept is to the national and temporal dimension from Fig. 4. The Private Object Space for a single concept is constructed by (matrix) multiplying the Group Stimulus Space with the Configuration Weights of the source. The Private Object Space does not give any more information than the numerical information of the Group Stimulus Space and the Configuration Weights. However, as INDSCAL is an exploratory method, any visual representation that makes the data more accessible than pure numbers adds to the appeal of the method. Note also, that the Private Object Space of a specific concept is not (necessarily) equivalent to the two-way MDS solution of the distance matrix of that concept.

The result of this multiplication for the KICK-OFF concept can be found in Fig. 8. As expected, Dimension 1 (national distinction) is practically completely removed, putting all the subcorpora on top of each other. For clarity, the subcorpora are from

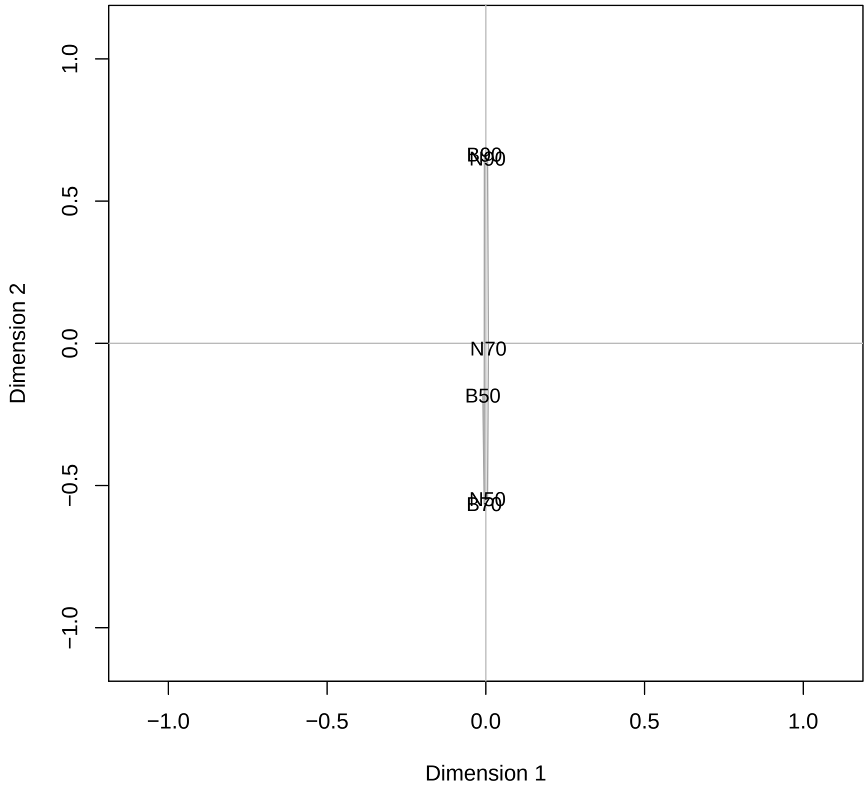


Fig. 8: Private Object Space for KICK-OFF

bottom to top: B70, N50, B50, N70, N90, B90. Only Dimension 2 (temporal evolution) is present, putting the subcorpora of the 1950s and 1970s at the bottom and the 1990s at the top. The Private Object Space visualizes that KICK-OFF is a concept that changed over time (especially after 1970), but the change was parallel in Belgium and the Netherlands.

The Private Object Space for the OFF-SIDE concept is presented in Fig. 9. As expected, Dimension 2 (temporal evoluation) is completely flattened and only the first dimension (national distinction) remains. For clarity, the subcorpora from left to right: B50, B70, B90, N50, N90, N70. Here, the Private Object Space very clearly shows that OFF-SIDE has not changed over time, but that there is a clear difference in naming the concept between Belgium and the Netherlands.

5 Conclusion

To conclude this methodological paper, we would like to address three questions of the proposed method: (1) can we find more examples of this method, (2) what are the formal requirements for INDSCAL, and (3) how can this method be applied to

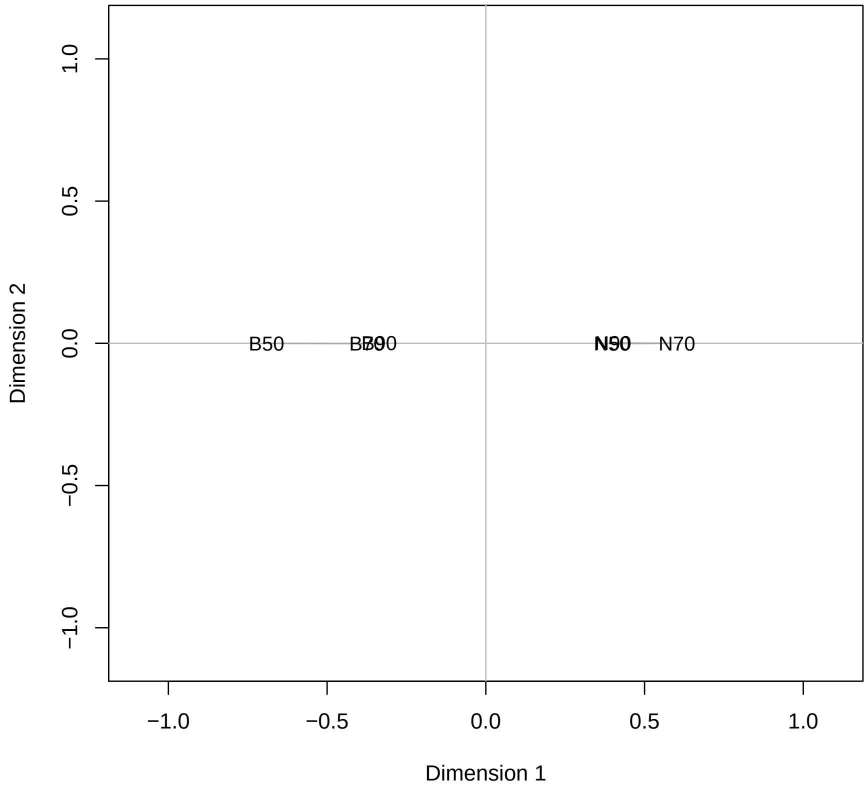


Fig. 9: Private Object Space for OFF-SIDE

other linguistic problems?

First, the proposed method has been mainly applied in psychological research, and even there it has only been applied by a handful of scholars since it was developed in the 70s. Despite its seeming unpopularity, one does find publications in psychological and sociological domains that employ three-way MDS almost every year and three-way MDS is covered in every statistics book that addresses Multidimensional Scaling. It is however never covered in books that focus on the use of (exploratory) statistics for linguistics. The shortage of studies employing three-way MDS is probably due to the lack of a widespread implementations in popular statistical packages. Luckily, this problem is now overcome. In the now commonly used statistical program *R*, the *SMACOF* package by de Leeuw and Mair (2009) is the first *R* implementation of three-way MDS. Although the package is still undergoing development, its distribution is stable.

Second, there is an important assumption that underlies the application of INDSCAL. As noted in Arabie et al. (1987, p. 21), 'the benefit of fitting the INDSCAL model is the inherently unique orientation that usually results for the object space [Group Stimulus Space]', but 'this uniqueness suggests that the dimensions of an INDSCAL object space should correspond to “fundamental” perceptual or judgmental [...] processes whose importance, strength, or salience may differ from source to source’. For sources such as experts and judges, or humans in general, this assumption of comparable sources is certainly defendable. But can we claim that linguistic phenomena comply with the given assumption? As MDS is an exploratory technique, it is acceptable to apply it to linguistic datasets, even if not all assumptions are undisputably met. However, one should keep this assumption in mind when interpreting the visualizations and when devising datasets to be submitted to an MDS analysis.

Third, the application of INDSCAL in aggregation studies should not be restricted to lexical variation. One of the examples in Arabie et al. (1987, p. 26) takes different experimental setups as the sources, which indicates that INDSCAL might be valuable in the perception and attitude field. Closer to the lexical analysis above, recent publications in the domain of dialectometry (see Wieling and Nerbonne (2011) for an overview of efforts), have shown an interest in identifying the behaviour of different types of variables that were part of aggregate studies. And research on historical developments of linguistic structures could take time periods as sources (e.g. Hilpert, 2011). INDSCAL would be a valuable exploratory tool for these problems.

With the above introduction and example application of INDSCAL, we have shown the relevance and advantages of a psychometrical method in the field of linguistics. Its wide application possibilities make INDSCAL a very rich method for exploring complex three-way, two-mode datasets. The main advantage of the method is that the aggregation over the sources shows the average pattern that is common to all sources, while still allowing the researcher to investigate the differences among the sources.

Notes

1. The linguistic studies that use Factor Analysis or Principal Components Analysis, e.g. Biber (1988), do have access to the behavior of the underlying variables, but they are fundamentally different from the type of aggregation studies that we aim at here. The research we aim at is distance-based, with elaborate distance metrics to measure the distances between measure points; in Factor Analysis and Principal Components Analysis, the input is two-way, two-mode and the processing to a two-way, one-mode matrix is based on correlation measures.
2. In fact, this is where the origin of Cronbach’s α lies: to check if there is enough similarity between the sources so that taking their average is not a too drastic reduction of the variance in the sources.
3. For INDSCAL, these proximity matrices should be square, symmetric, two-way, one-mode distance matrices.
4. It is of course not so, that all the variables in Geeraerts et al. (1999) are necessarily alternations between Dutch and English.
5. It would be impossible to apply Factor Analysis or Principal Components Analysis to that dataset because (a) we would not be able to use the specific distance metric of Section 4.1. and (b) the dataset does not comply with the strict statistical requirements of Factor Analysis or Principal Components Analysis.
6. Because a proper functioning of the Log Likelihood Ratio test requires frequencies not to be too small, we introduced a frequency threshold. If the frequency of the concept was lower than 30 for the two varieties that are being compared, that concept was excluded from the comparison.

7. The size of the two subcorpora is not the actual amount of words in the two subcorpora, but the sum of the frequencies of all variables in these two subcorpora.
8. As the input dissimilarities are City-Block distances, we could have considered using the City-Block model for three-way MDS (Heiser, 1989). However, in this paper we chose to restrict the demonstration of INDSCAL to the out-of-the-box settings of the method.

References

Arabie, P., J. D. Carroll, and W. S. DeSarbo (1987). *Three-way scaling and clustering*. Number 65 in Quantitative Applications in the Social Sciences. SAGE Publications.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Borg, I. and P. Groenen (2005). *Modern Multidimensional Scaling* (2 ed.). New York: Springer.

Carroll, J. and J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrica 35*, 283–319.

Cox, T. and M. Cox (2001). *Multidimensional Scaling*. Chapman & Hall.

Cysouw, M., M. Albu, and A. Dress (2008). Analyzing feature consistency using similarity matrices. *Sprachtypologie und Universalienforschung 61*(3), 263–279.

de Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. Barra, F. Brodeau, G. Romier, and B. van Cutsen (Eds.), *Recent Developments in Statistics*, pp. 133–145. Amsterdam: North Holland.

de Leeuw, J. and P. Mair (2009). Multidimensional Scaling using Majorization: SMACOF in R. *Journal of Statistical Software 31*(3), 1–30.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1), 61–74.

Geeraerts, D., S. Grondelaers, and D. Speelman (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.

Goebl, H. (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Oesterreichische Akademie der Wissenschaften.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph. D. thesis, Rijksuniversiteit Groningen.

Heiser, W. (1989). The city-block model for three-way multidimensional scaling. In R. Coppi and S. Bolasco (Eds.), *Analysis of Multiway Data Matrices*. Amsterdam: North-Holland.

Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics 16*(4), 435–461.

Horan, C. (1969). Multidimensional Scaling: combining observations when individuals have different perceptual structures. *Psychometrica 34*(2), 139–165.

Jacobowitz, D. (1973). *Development of semantic structures.* Ph. D. thesis, University of North Carolina at Chapel Hill.

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing 21*, 463–476.

Seguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane 35*, 335–357.

Speelman, D., S. Grondelaers, and D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities 37*, 317–337.

Spruit, M. R., W. Heeringa, and J. Nerbonne (2009). Associations among linguistic levels. *Lingua 119*(11), 1624–1642.

Szmrecsanyi, B. (2011). Corpus-based dialectometry: a methodological sketch. *Corpora 6*(1), 45–76.

Takane, Y., F. Young, and J. de Leeuw (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika 42*(1), 7–67.

Wieling, M. and J. Nerbonne (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language 25*(3), 700–715.

Wieling, M., J. Nerbonne, and H. Baayen (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE 6*(9), e23613.