Title: Linguistic Distances in Dialectometric Intensity Estimation

Authors: Simon Pickl (University of Salzburg),
Aaron Spettl (Ulm University),
Simon Pröll (University of Augsburg),
Stephan Elspaß (University of Salzburg),
Werner König (University of Augsburg),
Volker Schmidt (Ulm University)

Lead author: Simon Pickl
Fachbereich Germanistik
Universität Salzburg
Erzabt-Klotz-Str. 1
5020 Salzburg
Austria

+43 (0)662 8044 4359

simon.pickl@sbg.ac.at

Short title: Linguistic Distances in Intensity Estimation

Linguistic Distances in Dialectometric Intensity Estimation

Abstract

Dialectometric intensity estimation as introduced in Rumpf et al. (2009) and Pickl & Rumpf (2011, 2012) is a method for the unsupervised generation of maps visualizing geolinguistic data on the level of linguistic variables. It also extracts spatial information for subsequent statistical analysis. However, as intensity estimation involves geographically conditioned smoothing, this method can lead to undesirable results. Geolinguistically relevant structures such as rivers, political borders or enclaves, for instance, are not taken into account and thus their manifestations in the distributions of linguistic variants are blurred. A possible solution to this problem, as suggested and put to the test in this paper, is to use linguistic distances rather than geographical (Euclidean) distances in the estimation. This methodological adjustment leads to maps which render geolinguistic distributions more faithfully, especially in areas that are deemed critical for the interpretation of the resulting maps and for subsequent statistical analyses of the results.

Acknowledgements

The authors' work is supported by the *Deutsche Forschungsgemeinschaft* through the research project *Neue Dialektometrie mit Methoden der stochastischen Bildanalyse* ('New Dialectometry Using Methods from Stochastic Image Analysis').

1. MOTIVATION

What exactly can dialect maps tell us about the language variation in a specific region? How reliable are they when it comes to considering individual data at particular locations? As shown in Rumpf, Pickl, Elspaß, König & Schmidt (2009, 2010) and Pickl & Rumpf (2011, 2012), the data contained in dialect atlases and similar geolinguistic data collections tend to have the disadvantage of limited reliability when viewed in detail. For instance, a single record at a single site that was uttered by an informant in a specific interview situation may or may not reflect common usage in the local dialect. A range of influencing factors can reduce the accuracy of an informant's answers, such as poor memory, observer effects, or personal background. Not all of these factors can be controlled entirely, especially as they are subject to a certain degree of randomness. The responses of an individual informant to a specific question may even differ from day to day. Methodological restrictions (such as observer effects) aside, this is also a manifestation of the fundamental probabilistic nature of language variation (cf. Cedergren & Sankoff, 1974; Pickl, 2013: 13, 41–42,205–207). Individual records of variants are thus little more than statistical samples. Even though the overall picture that a dialect map of a specific variable gives is in all likelihood a good representation of the actual geographical distribution of that variable, the individual details of such maps can be inaccurate. A further problem is that in many cases dialect atlases only show one or two records per site, but no information about their relative frequencies is given.

All analyses and examples in this paper are based on data from the *Sprachatlas von Bayerisch-Schwaben* (SBS), a dialect atlas covering an area of approx. 11,000 km^2 in Southern Germany, which is based on explorations carried out in the 1980s and 1990s. Featuring data from 272 record locations, the atlas's 14 volumes comprise approximately 2,700 individual maps. Fig. 1, an original point-symbol map taken from the SBS, features responses for the concept 'woodlouse'. Each symbol stands for one specific variant; the small

triangle symbol that is prevalent in the north-west, for instance, stands for the lexical variant *Maueresel* (or similar). The number of recorded variants per location ranges from zero (symbol "+") to three. Although at most of the locations only one variant is recorded, it seems likely that a different sample of informants at one of the sites – or even the same informants at another time – may have come up with different variants, e.g. with divergent variants that were recorded at neighbouring locations.

One way to deal with this uncertainty is to "aggregate the differences in many linguistic variables in order to strengthen their signals" (Nerbonne, 2010: 3822). This is the approach of traditional, aggregative dialectometry, which is useful for making global, overall structures in large corpora of dialect maps visible. The pivotal instrument of aggregative dialectometry is a distance (or similarity) matrix that is derived from the maps of a large number of variables. The matrix contains the overall (i.e. aggregated) relations among every possible pair of locations in the data collection, and can be used, for instance, to produce similarity maps that show degrees of similarity in relation to one location in question. By aggregating the differences between locations across linguistic variables, however, the distinctiveness of these variables and their variants' distributions is made void – the variation among these distributions collapses. For some research interests, however, it is essential to maintain and analyse exactly these differences, e.g. when one wishes to investigate which variants' distributions behave similarly or which are affected by certain extra-linguistic factors. Standard aggregative dialectometry is not suitable for this kind of study. Only recently, a number of works have appeared that try to overcome the problem of losing sight of individual variants when viewing the larger picture, using different methodological approaches (cf. e.g. Shackleton 2005, 2007, Nerbonne 2006, Grieve 2009, Grieve, Speelman & Geeraerts 2011, Wieling & Nerbonne 2011, Pickl 2013, Pröll, Pickl & Spettl (forthcoming)).

One such approach, which is also aimed at 'strengthening the signals' of individual variables, is so-called intensity estimation (cf. Rumpf, Pickl, Elspaß, König & Schmidt 2009, 2010 and Pickl & Rumpf 2011, 2012). In intensity estimation, it is not the information from other variables that is employed to stabilize the overall signal; instead, the information from nearby locations is used to stabilize the information for individual locations. The simple assumption that justifies such a course of action has been formulated by Nerbonne & Kleiweg (2007: 154) as the Fundamental Dialectological Postulate: "Geographically proximate varieties tend to be more similar than distant ones", or, put more generally by Waldo Tobler: "everything is related to everything else, but near things are more related than distant things" (Tobler's so-called "First Law of Geography"; Tobler, 1970: 236). Based on this principle, it is to a certain extent legitimate to infer the variants used at a site from the variants recorded at surrounding sites. In intensity estimation as introduced in Rumpf et al. (2009, 2010) and Pickl & Rumpf (2011, 2012), the geographical proximity of two sites is used as a measure of how well these two sites can 'speak for one another'. If, for example, site a has a record for variant x and neighbouring site b features variant y, then there is some probability that x is also used at b by some speakers with a certain frequency, especially if all other sites surrounding b also have x. Thus, all locations in an area under investigation are mutually dependent, near ones more so than distant ones, and hence each location can be assigned a specific 'intensity' for each variant, based on how densely the actual records for this variant are distributed in the location's vicinity.[1] In a probabilistic interpretation, a variant's intensity can also be seen as representing the variant's estimated relative frequency of occurrence, or its probability.

2. INTENSITY ESTIMATION

The equation for the calculation of the intensity $i_{x_1}(a_1)$ of variant x_1 at location a_1 based on geographical proximity is the following (for its derivation and explanation cf. Rumpf et al., 2009 or Pickl & Rumpf, 2011):[2]

$$i_{x_1}(a_1) = \frac{1}{\sum_{x \in X} \sum_{a \in A} K(d(a_1,a),h) \cdot w_x(a)} \cdot \sum_{a \in A} K(d(a_1,a),h) \cdot w_{x_1}(a) \quad (1)$$

X: the set of variants of the respective variable

A: the set of record locations in the area under investigation

$w_x(a)$: the 'weight' of variant x at site a, meaning the proportion of records for x in all records for the respective variable at a

$d(a_1, a_2)$: the geographical distance between the sites a_1 and a_2

h: the so-called 'bandwidth' of the intensity estimation, a parameter that defines how quickly the influence between two sites decreases with increasing distance

(There are several algorithms that optimize the bandwidth automatically.[3])

$K(d, h)$: the so-called kernel, another parameter that defines how the actual influence between sites at a certain distance is calculated

(Often, the two-dimensional normal distribution is used for K. In this study, the so-called K_3-kernel, a very similar kernel that has some advantages over the normal distribution is used; cf. Silverman, 1986: 76–77, 88–89.)

After applying this equation to every variant and every site, the results can be visualized as a set of maps, each with the intensity field of one variant (see Rumpf et al., 2009, 2010; Pickl & Rumpf, 2011 or 2012 for examples), or alternatively as one map that integrates the intensities of all variants. In order to do the latter, each location is assigned to the variant that

has the highest intensity locally, which is the 'dominant' local variant. The result is a graded area-class map, which features fuzzy variant areas that are delimited by the relative dominance of the respective variants, but graded in their shades to display the dominant variants' intensities and to illustrate the fact that these variants overlap in space to a certain extent.

Fig. 2 is an example of such a graded area-class map. The different colours stand for different variants; their local intensity or dominance is represented by different colour shades. The orange lines delimit areas of dominance, which implies that they do not represent clear-cut isoglosses but centres of transition zones: in the lighter-shaded zones, less frequent non-dominant variants are also present.

Area-class maps of this kind are not only abstracted visualizations of point-symbol data, they are also the basis for further analyses. Rumpf et al. (2010) performed analyses to investigate whether variables with similar intensity distributions are also related on an extra-linguistic level, and found, for instance, that lexical variables from the semantic field *crop* tend to have similar geolinguistic configurations. This points to similar patterns of spatial diffusion that led to these distributions. Certain characteristic measures that are derived from the intensity information can further be used in statistical testing (cf. Pickl, 2013: 125–140; Pröll, 2013: 149–153). Moreover, the resulting (fuzzy) isoglosses can be used to validate presumed dialect borders (cf. Pickl, 2013: 141–157). For further applications of intensity maps cf. Meschenmoser & Pröll (2012 a, b).

3. PROBLEM

In some of the applications of intensity estimation as depicted in the preceding section, the way in which the data is abstracted towards an overall distribution pattern can be problematic.

As intensity estimation is basically a form of geographically conditioned smoothing, some basic characteristics of individual maps can be levelled out. If, for example, one or more of the locations in the area under investigation are language islands, this fact will have no consequences for area-class maps produced with intensity estimation. The reason is that even locations that consistently have variants diverging from their neighbours will be treated as outliers and 'smoothed over'. The same problem holds for dialect borders. Any structure that entails significant dialect differences will not have any relevance for individual area-class maps, as the majority of features are not taken into account. Therefore, it frequently happens that differences in individual variables that coincide with structural dialect borders are straightened, shifted, or otherwise blurred. All kinds of geographical structures that are systematic in the sense that they show up on a lot of individual maps will not be recognized or rendered as such by geographically-informed intensity estimation.

In Fig. 2, this problem becomes visible (or rather not visible) in the cases of several larger cities and towns whose records differ from the variants recorded in the surrounding countryside, e.g. Augsburg (122), Günzburg (96) or Memmingen (205) (cf. Fig. 1). While it is not surprising that densely populated places of urban character behave linguistically differently from rural areas, this difference is not reproduced in the area-class map: All these locations are smoothed out; they become indistinguishable parts of larger areas. The same is true for the river Lech (cf. Fig. 1, Fig. 3), which is a well-known and well accounted for strong dialect border in this region; also in the case of 'woodlouse' the Lech shows a clear separating effect on the level of linguistic variants (cf. Fig. 1), dividing the line-shaped symbol in the east from the triangle and rectangle symbols in the west. However, the corresponding isogloss in Fig. 2 has become fuzzy like all the other isoglosses; what is more, the isogloss itself is shifted significantly to the east (especially in the south). In these instances, intensity estimation does not do justice to the data. Relying entirely on geographical distances, intensity estimation is so to speak less informed than we are, because we have

knowledge about linguistically relevant geographical structures that go beyond pure geometry.

A possible solution to this problem lies in a combination of traditional, i.e. aggregative dialectometry and intensity estimation. While the latter allows us to process individual maps quantitatively, the former has the advantage of taking all the available data into account to draw a much more general picture. A straightforward integration of the two concepts is to use linguistic distances instead of geographical distances in intensity estimation. The idea behind this approach is that geographical space is not the only determinant of geolinguistic processes – other conditions, like accessibility, traffic routes, terrain structure, attractiveness of places etc. also shape the way in which people from different areas interact and the way in which dialects come into contact. Several studies have investigated to what extent dialect similarity is related to variables like travel distance, population density, migration and others (cf. e.g. Trudgill, 1974; Gooskens, 2004; Inoue 2004, 2006; Szmrecsanyi, 2012). All of them find that using more sophisticated distance measures based on such extra-linguistic variables in addition to geographical distances leads to a better representation of dialect similarity. The exact impact of these variables on dialect similarities or distances is, however, of minor importance for our study. Their influence is already condensed in the form of dialect distances (probably including a certain amount of random effects). If we use measurable dialect distances instead of geographical distances, all these effects are therefore virtually automatically included, and isolated language islands as well as dialect borders are represented in the model by higher linguistic distances (either in relation to geographically surrounding locations or to locations 'across the border', respectively). Linguistic distances render exactly those effects that actually did influence variant distributions in the past. Hence they provide what could be seen as a model of linguistic space (as opposed to Euclidean space; cf. Pickl 2013, 62–63). Linguistic space can be understood as the network of mutual relations between local dialects that is constituted by pairwise contact probabilities, and which

establishes a framework in which dialectal accommodations and diffusion of features take place. This network is shaped also, but not exclusively, by Euclidean distances. As the effect of geographical distances on this linguistic network is rendered indirectly through the extent to which dialect similarities are influenced by them, geographical distances as such can be disregarded on the whole throughout the intensity estimation introduced in this article.

4. LINGUISTIC DISTANCE MEASURES IN DIALECTOMETRY

There are various implementations of linguistic distance; one of the most popular is Goebl's Relative Distance Value RDV_{jk} (cf. e.g. Goebl, 2010), which is suitable especially for nominal-scale data such as lexical variants. Simply put, RDV_{jk} is the percentage of variables for which two locations j and k have different variants. The exact value is calculated using the following formula:

$$\mathrm{RDV}_{jk} = 100 \cdot \frac{\sum_{i=1}^{p} (\mathrm{COD}_{jk})_i}{\sum_{i=1}^{p} (\mathrm{COI}_{jk})_i + \sum_{i=1}^{p} (\mathrm{COD}_{jk})_i} \tag{2}$$

In this notation, $(\mathrm{COD}_{jk})_i$ is a so-called co-difference function that returns 0 if the locations j and k have the same variant for variable i and 1 if they have different variants for variable i. The co-identity function $(\mathrm{COI}_{jk})_i$ does the opposite, returning 1 for identical values and 0 for different values. Thus RDV_{jk} is the fraction of different answers at the location j and k within p variables. By using the sum of CODs and COIs instead of the total p, it is ensured that variables with missing values for the locations in question are disregarded.

One fundamental restriction of RDV is that it is only defined for unequivocal answers, i.e. for the case that for each variable and each location there is only one variant in the data.[4] As modern dialectological atlases and other geolinguistic data sources often feature a range of

different answers at a location, RDV is applicable to this kind of data format only if a substantial portion of the data is discarded previous to analysis.[5] For this kind of data, an adaption of the RDV (or a completely different approach) is necessary. Speelman, Grondelaers & Geeraerts (2003: 320–321) (cf. also Speelman & Geeraerts 2008) present what they call "city block distance" (Speelman et al., 2003: 320) or "profile-based dissimilarity" (Speelman & Geeraerts, 2008: 227–228), a distance measure between linguistic profiles, i.e. percentages of answers that add up to 100 %. This distance is defined for each pair of locations and for each variable as the average of the differences between variants' percentages. For instance, if the variable is 'jeans' and at one location 70 % of the answers belong to the variant *jeans* and 30 % to the variant *spijkerbroek*, while at the other location 97 % belong to *jeans* and 3 % belong to *spijkerbroek*, then the difference is 27 % for *jeans* and also 27 % for *spijkerbroek*. The sum is 54 %, but we have counted the same difference twice, so we divide by 2 and receive a distance of 27 % or 0.27. Subsequently, the individual distances for variables can be aggregated to overall distances by arithmetic averaging. Again, this measure is especially suitable for nominal-scale and therefore particularly for lexical data. Even for phonetic data, a nominal-scale approach can be appropriate (Pröll, 2013: 48–51).

Phonetic information, however, can also be treated as interval-scale data. The so-called Levenshtein distance is a very popular phonetic distance measure. It is a string edit distance between two strings of phonetic (typically IPA) transcriptions. The Levenshtein distance measure counts insertions, deletions and (mis)matches of characters between two transcription strings. For this purpose the two strings have to be aligned previously by following certain rules. In contrast to RDV, it is suitable for the calculation of graded distances even between individual word pronunciations. Still, they are usually aggregated in order to obtain more solid overall distances. For a detailed account of the Levenshtein distance measure cf. Heeringa (2004: 121–143).

5. IMPLEMENTATION AND APPLICATION

Regardless of which distance measure is chosen, its implementation in intensity estimation is rather straightforward. Firstly, the linguistic distances between all location pairs have to be calculated, based on the specified measure and the corpus of maps to be investigated. Then equation (1) has to be adapted in that the geographical distance d has to be replaced by a linguistic distance d_{ling}; apart from that the equation remains unchanged.

In this section, this procedure is applied to data from the SBS, more specifically to a lexical sub-corpus of maps comprising 736 variables. The linguistic distance used is based on this subcorpus and calculated following the Speelman et al. (2003) approach. It is therefore a purely lexical distance. For an individual variable X with variants $\{x_1;\ldots;x_n\} \in X$, the lexical distance between locations a_i and a_j is defined as follows:

$$d_X(a_i, a_j) = \frac{1}{2}\sum_{x \in X}\left|w_x(a_i) - w_x(a_j)\right| \quad (3)$$

For a corpus of maps $\mathbb{M}$ with variables $\{X_1;\ldots;X_n\} \in \mathbb{M}$, the overall lexical distance between locations a_i and a_j is:

$$d_{\mathbb{M}}(a_i, a_j) = \frac{1}{\sum_{X \in \mathbb{M}} n_X(a_i, a_j)} \sum_{X \in \mathbb{M}} d_X\left(a_i, a_j\right) \cdot n_X(a_i, a_j) \quad (4)$$

In this equation, $n_X(a_i, a_j)$ is 1 if both locations a_i and a_j have a record for X and 0 if at least one of them has no record for X. Thus if two locations cannot be compared reasonably with respect to a specific variable, this variable is excluded.

$d_{\mathbb{M}}(a_i, a_j)$ is now calculated for all possible pairs of locations. With a total of 272 locations, this yields a matrix with 36,856 individual values. It is to be expected that these values are correlated to a high degree with the respective geographical distances. At the same

time, it is desirable that the correlation is not too high, so that there actually is an improvement in the representation of linguistically relevant relations between locations. In the current study, the Pearson correlation coefficient is 0.80, which corresponds to an explained variance of 64 % using a linear regression model. As the corresponding scatter plot (Fig. 4) reveals, however, the relation seems to be logarithmic rather than linear.

This is in line with the findings of a number of other studies investigating the relation between linguistic and geographical distances, e.g. Heeringa & Nerbonne (2001), Nerbonne & Heeringa (2007: 288–289), Nerbonne (2010), Szmrecsanyi (2012). A logarithmic regression analysis returns 69.3 % of explained variance. The circumstance that the difference first rises rather steeply and then gradually goes towards a ceiling is a pattern that is commonly found in linguistic distances and is easily explained:[6] If the spatial lexical replacement rate, i.e. the proportion of variants that change on average if a certain distance is crossed, is relatively homogeneous, then a certain number of variants is usually different at a location that is at a certain distance. If the distance is doubled, the percentage of different variants is not doubled, because some of the already different variables may change again, not affecting the overall difference. Thus the farther away a location lies, the smaller the role is that distance plays, even if this may sound paradoxical. The reason is that dialects that are separated by a large distance are already relatively different, so there is not much leeway for them to differ much more.

Still, this interpretation only explains the logarithmic shape of the scatter plot; it does not tell us anything about an improvement brought about by linguistic distances. About 30 % of the values of linguistic distance in relation to geographical distance cannot be accounted for by logarithmic relation. These 30 % can be due to random fluctuations or additional effects. While naturally a certain amount of randomness is to be expected, the scatter plot reveals certain hints about what else makes linguistic distances different from geographical ones. One

such hint is the fact that the distribution of dots in Fig. 4 has more than one condensation area. While the majority scatters around the regression line, there is a second, rather hard to distinguish concentration area forming a long stretch above the major concentration. These dots, which have been highlighted in the visualization on the right in Fig. 4, represent those location pairs that are separated by the river Lech (cf. Fig. 3), a recognized dialect barrier, and therefore have an increased linguistic distance in relation to their geographical distance. On the whole, these pairs contribute to the logarithmic shape of the distribution, but they add a significant effect that cannot be accounted for by geographical distance. This effect is even stronger if other parts of the map corpus – not the lexical subcorpus – are used for the calculation of the linguistic distance. Fig. 5 shows the respective scatter plots of distances built on phonetic and morphological sub-corpora, as well as on the whole corpus. Again, the location pairs separated by the river Lech have been highlighted in the right-hand visualization.

In these figures, up to three agglomeration areas can be discerned. These concentrations are best visible for phonetic distances. One of them is apparently caused by the separating effect of the river Lech. What the exact meaning of other agglomerations is remains as yet unclear, but it is very likely that they represent some kind of geolinguistic condensation areas. Other such peculiarities are the pairs that appear as outliers in the scatter plot; these pairs can be outliers for various reasons, for instance if they include dialect islands. These additional features that go beyond pure geographical information promise to make linguistic distance a more suitable distance measure for intensity estimation.

As mentioned above, the procedure for implementing linguistic distances in intensity estimation is simply to substitute the geographical distance d in equations (3) and (4) for the linguistic distance. The result for the data underlying Fig. 1 and 2 (Map 63 from vol. 8 of the SBS) is shown in Fig. 6.

Compared to Fig. 2, the differences are immediately clear. The shapes of the isoglosses are more jagged, and the colour shades representing intensity values are distributed less equally across the area. Also the fringe of lighter colours that accompanies each isogloss in Fig. 2 is less regular here. What is also noteworthy is that the isogloss running from north to south now follows the river Lech almost exactly (the only exception is location 199 (Landsberg am Lech), a district capital). Also, structures that do not appear in Fig. 2 do show up here, for instance individual locations that have variants that diverge from the variants that are prevalent in their surroundings, such as the towns Memmingen (205) or Neu-Ulm (109). Most of the locations that 'behave' differently from their surroundings are larger towns or cities, which is also true for Günzburg (96), Schongau (269), Königsbrunn (156), Augsburg (122) and Augsburg's borough Lechhausen (123). Königsbrunn is a peculiar case – being a relatively young colony, founded only in the 19th century, its settlers came mostly from the northwest of Bavarian Swabia. This explains why Königsbrunn features the same variant as the northwest of the area under investigation – coloured in blue –, like on many other maps (cf. Pickl, 2013: 172). This specialty is lost on practically all area-class maps based on geographical distance but is rendered clearly on linguistic-distance-based area-class maps like Fig. 6. Generally speaking, while linguistic-distance-based maps do an equally good job at abstracting away from the original data, more details are preserved.

Despite the clear improvement in these points, some problems that arise from the use of linguistic distance also have to be addressed.

One of these problems is that deficits in the data layer at individual locations (e.g. due to a large number of missing records at a site or a generally unreliable informant) will lead to flaws in the resulting maps. With our test corpus of maps, this effect only rarely led to visible problems.

Another, more fundamental and theoretical problem is that the reliance on linguistic distances equals an exclusive reliance on the data that are the basis for these distances. In other words, if the linguistic distances are calculated from lexical data only, it is plausible that they are suitable for drawing lexical maps, but probably not for drawing phonetic maps. This leads to the general question of how the dataset that is used to calculate the distances should be chosen. Would it, for instance, be 'better' to use maps from all linguistic levels for the distances, or to establish an individual distance matrix for each of the levels? And which criteria should we use to define what is 'better'?

Finally, we use linguistic distances to construct a network of locations which is then used for the estimation of linguistic distributions. At first glance, this may appear to involve a circularity problem, because linguistic distances obtained from a geolinguistic dataset are used for the estimation of underlying distributions in individual maps of the same dataset. Yet, there is no circular reference, because the linguistic distances are computed using the original weights of variants. As the proportion of the information referring to one variable is relatively small in relation to the whole dataset, only a very small fraction of the data is used twice in each estimation. This is not a problem in practice and there is no theoretical reason why the data of the considered map should be excluded from linguistic distance computations.

In the following section, the second of these problems, being quantifiable, will be discussed. The central question will be, however, whether linguistic distances have a measurable advantage over geographical distances in the generation of area-class maps.

6. VALIDATION

The question of whether the linguistically based implementation of intensity estimation is 'better' than the geographically based intensity estimation cannot be answered conclusively, but certain quality criteria can be defined and then compared.

In our case, we have chosen to analyse the accuracy that is achieved using the two distance measures in 'predicting' individual records. By accuracy of prediction we mean the frequency with which intensity estimation yields intensities at a location that favour the variant(s) actually recorded there, if the estimation is performed without taking this location's records into account. Concretely, we perform intensity estimation for each map as often as there are locations on the map (in our case 272), each time with one of the locations dismissed. This technique is known as leave-one-out cross-validation. It tells us how well the intensity estimation 'predicts' the actual record. For each location and each variable, we assign a score that expresses the distance between actual records and intensities inferred. A well-performing implementation of intensity estimation will yield low scores in this kind of leave-one-out cross-validation, all the while providing a reasonable degree of smoothing.

For our study, we have defined a score that compares the dominant variant estimated for a location with the record(s) found in the raw data. This means that we establish how well the visible division into variant areas reflects the raw data. Specifically, we assign a score to individual locations each variable that quantifies the difference between the estimated dominant variant and the actual records at the location. It is defined as 1 minus the local weight of the dominant variant (cf. Section 2), which can take on only the values 0, 1/3, 1/2, 2/3, or 1 with our data. The score is therefore 0 if the dominant variant equals the only variant at that location for that variable in the raw data, between 0 and 1 if the dominant variant equals one of two or more variants at that location for that variable in the raw data, and 1 if the dominant variant is a variant that is not attested at that location for that variable in the raw

data. The lower the score, the better the areal division of the intensity map reflects the raw linguistic data.

We calculated average scores for intensity maps based on geographical and linguistic distances, using two different kernels ($K_{\text{Gauß}}$, K_3) and two different bandwidth algorithms (LCV, *CL*). In direct comparison, linguistic distances yield better results in most cases (cf. Fig. 7). The best overall result is attained with a combination of $K_{\text{Gauß}}$, LCV and linguistic distance. Note that likelihood-cross-validation (LCV, cf. Silverman, 1986: 52–53) chooses the bandwidth such that the predicted densities match the weights of the variants (cf. Section 2) best. Leave-one-out cross-validation is a special case of LCV and its idea is therefore the same, with the only difference that (in our case) dominant variants are compared, not the estimated density values themselves. Nonetheless, it is to be expected that LCV yields almost optimal (i.e. low) average scores. *CL* as a cost-curve approach that tries to balance map complexity vs. map fidelity cannot compete with LCV in this validation approach, but it often produces graded area-class maps that are 'nicer' to look at (more smoothing for high-complexity maps, less smoothing for very homogeneous maps). Especially in combination with $K_{\text{Gauß}}$, *CL* clearly does not provide the best results for both geographical and linguistic distances.

A more detailed perspective can be provided by calculating average scores for each location separately. This is done using once linguistic and once geographic distances. The resulting values can then be mapped in combined maps such that each location is assigned a colour value depending on which of the two scores is lowest, i.e. which of the two methods of estimation gets closer to the recorded data (cf. Fig. 8). In these maps, the blue locations are the ones where better results are obtained using the linguistic distance; the red locations are the ones where geographical distance is better. Generally, linguistic distances yield better results for sensitive areas, as exemplified by the regions around Augsburg and along the river

Lech, even in the map for $K_{\text{Gauß}}$ and *CL* (Fig. 8b), where geographical distances do better globally.

While these findings suggest that the use of linguistic distances does lead to improvements in the generation of area-class maps with intensity estimation in most scenarios, the question remains of how the map corpus used for the calculation of distances should be chosen in the first place.

Generally, two possibilities seem plausible: 1) The use of the subcorpus of maps that represents the linguistic level in which the individual maps are situated (in our case the lexical subcorpus), or 2) the use of the entire map corpus comprising maps for variables from all linguistic levels. The latter could be helpful if the subcorpus to be analysed is very small, or if it is assumed that subcorpora representing other linguistic levels have certain relevance for the distributions in the subcorpus in question. What is not suggested, however, is to exclude the maps to be analysed using intensity estimation from the calculation of distances and to use maps from subcorpora representing other levels only. It is simply not plausible that for instance a morphological dataset should make a better statement about relations that are relevant for lexical items than a dataset containing other lexical information.

To get an initial empirical idea of how well the individual subcorpora are suited for intensity estimation based on linguistic distances, we calculated the average scores for the three subcorpora (lexical, morphological and phonetic data) and the entire map corpus comprising all these levels based on distances calculated from exactly the same maps that are contained in these test corpora (cf. Fig. 9). We used $K_{\text{Gauß}}$ and LCV as parameters as these yielded the best results with the lexical subcorpus, as discussed above.

Somewhat surprisingly, the morphological and phonological subcorpora achieve much better results than the lexical subcorpus (and hence also than the entire corpus). This is

probably due to the fact that there is a tendency in morphological and phonological maps toward larger and more solid variant areas, which makes them less susceptible to intensity estimation induced smoothing (cf. Pröll, 2013: 151).

In the next step, we calculated the average scores for the lexical subcorpus, each time with a different distance measure based on the lexical subcorpus, on the morphological subcorpus, on the phonological subcorpus, and on the entire map corpus (cf. Fig. 10).

As is to be expected, the best results are achieved with distances calculated from the lexical corpus and the complete corpus, but the use of the morphological or the phonological maps for the calculation of distances leads only to a minimal deterioration. This suggests that the choice of corpus for distance extraction is not crucial, especially as the magnitude of its effect lies well within the range of effects caused by smoothing parameter selection (cf. Fig. 7). This finding is corroborated by a look at the average scores for all combinations of test corpora and corpora used for distance extraction (cf. Fig. 11); again, the quality of the estimation depends on the test corpus more than on the distance corpus. It should be emphasized, however, that the base corpus for the extraction of distances should not be too small, as this would reduce the accuracy of intensity estimation because there would not be enough information about the relations between sites.

7. SUMMARY

The use of intensity estimation for the drawing of graded area-class maps is a means to deal with uncertainty in geolinguistic data, to provide a quantitative account of the geographical configuration of an individual variable's map, and to provide visual abstractions from pointwise data for better and quicker inspection. This article presents a method to improve the accuracy of the results of intensity estimation by utilizing linguistic instead of

geographical distances. Visual inspection of the results obtained with this new approach shows that significant geographical structures like dialect barriers (such as, in this, case, a river) or dialect islands (in this case towns or a colony) are rendered much more faithfully when using linguistic rather than geographical distances. These findings are corroborated by leave-one-out cross-validation, which shows that with most parameter settings and especially in 'critical' regions of the area (i.e. regions where geographical structures influence the distribution of dialectal variants), linguistic distances lead to better results. The best overall results for linguistic distances are attained in combination with LCV bandwidth optimization and the two-dimensional normal distribution kernel. In some, especially less 'interesting' regions, different parameter settings can lead to a better prediction of left-out records. As this study is restricted to a certain dialect area in the south of Germany and to a specific data set (the SBS), it is clear that in a different region and with other data, the results could be in favour of other parameter settings or even of geographical distances. We hope to have shown, however, that the use of linguistic distances can be an improvement of intensity estimation, thus honing the results of the estimation, which are useful for subsequent statistical analyses, as well as the resulting maps, which provide unsupervised visualisations of geolinguistic data.

End Notes

[1] In this regard, the resulting maps of intensity estimation bear some resemblance to those obtained using methods that measure spatial autocorrelation, which have become increasingly popular in recent times (cf. Grieve 2009, Grieve, Speelman & Geeraerts 2011, and Lameli 2013: 98–102 in an aggregative perspective). Especially local spatial autocorrelation measures such as Getis-Ord G_i* yield results that resemble those of intensity estimation at least superficially, but the two underlying methods work quite differently. While methods of spatial autocorrelation like Getis-Ord G_i* are aimed at identifying areas of statistically significant spatial clusters (hot spots), intensity estimation is used to infer an underlying probability distribution from a number of observances.

[2] The denominator in this equation normalizes local sums of intensities to 1.

[3] In this study, two algorithms are used: Likelihood-Cross-Validation (LCV) and a cost-curve approach (*CL*) that optimizes certain characteristics of the resulting maps (cf. Pickl, 2013: 110–113).

[4] Cf. Pröll (2013: 18–19) for a discussion of this problem.

[5] For instance Bauer (2009: 172), using Goebl's approach, reduces the dataset to contain only one variant per variable and site.

[6] Séguy (1971) gave the first account of this phenomenon; Nerbonne (2010: 3821) thus suggests using the term *Séguy's curve*. Stanford (2012) examines the applicability of this relation for short geographical distances.

References

Bauer, Roland. (2009). *Dialektometrische Einsichten. Sprachklassifikatorische Oberflächenmuster und Tiefenstrukturen im lombardo-venedischen Dialektraum und in der Rätoromania.* San Martin de Tor: Istitut Ladin Micurà de Rü.

Cedergren, Henrietta J. & Sankoff, David. (1974). Variable Rules: Performance as a Statistical Reflection of Competence. *Language* 50/2: 333–355.

Goebl, Hans. (2010). Dialectometry and quantitative mapping. In A. Lameli, R. Kehrein, & S. Rabanus (eds), *Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping.* Berlin/New York: de Gruyter. 433–457.

Gooskens, Charlotte. (2004). Norwegian dialect distances geographically explained. In B.-L. Gunnarsson, L. Bergström, G. Eklund, S. Fridell, L. H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren, & M. Thelander (eds), *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLaVE 2.* Uppsala: Universitetstryckeriet. 195–206.

Grieve, Jack. (2009). *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English.* Dissertation, Northern Arizona University.

Grieve, Jack, Speelman, Dirk, & Geeraerts, Dirk. (2011). A Statistical Method for the Identification and Aggregation of Regional Linguistic Variation. *Language Variation and Change* 23, 193–221.

Hansen, Sandra, Schwarz, Christian, Stoeckle, Philipp, & Streck, Tobias (eds). (2012). *Dialectological and folk dialectological concepts of space. Current Methods and Perspectives in Sociolinguistic Research on Dialect Change.* Berlin/New York: de Gruyter.

Heeringa, Wilbert. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen: Rijksuniversiteit Groningen.

Heeringa, Wilbert, & Nerbonne, John. (2001). Dialect Areas and Dialect Continua. *Language Variation and Change* 13: 375–400.

Inoue, Fumio. (2004). Multivariate Analysis, Geographical Gravity Centers and the History of the Standard Japanese Forms. *Area and Culture Studies* 68: 15–36.

Inoue, Fumio. (2006). Geographical Distance Center and Rate of Diffusion of Standard Japanese. In A. Timuška (ed.), *Proceedings of the 4th International Congress of Dialectologists and Geolinguists.* Riga: Latvijas Universitāte. 239–247.

Lameli, Alfred. (2013). *Strukturen im Sprachraum. Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland.* Berlin/Boston: de Gruyter.

Meschenmoser, Daniel, & Pröll, Simon. (2012a). Using fuzzy clustering to reveal recurring spatial patterns in corpora of dialect maps. *International Journal of Corpus Linguistics* 17/2: 176–197.

Meschenmoser, Daniel, & Pröll, Simon. (2012b). Automatic detection of radial structures in dialect maps: determining diffusion centers. *Dialectologia et Geolinguistica* 20: 71–83.

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

Nerbonne, John. (2010). Measuring the Diffusion of Linguistic Change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365/1559: 3821–3828.

Nerbonne, John, & Heeringa, Wilbert. (2007). Geographic distributions of linguistic variation reflect dynamics of differentiation. In S. Featherston & W. Sternefeld (eds), *Roots. Linguistics in Search of its Evidential Base.* Berlin/New York: De Gruyter Mouton. 267–297.

Nerbonne, John, & Kleiweg, Peter. (2007). Toward a Dialectological Yardstick. *Journal of Quantitative Linguistics* 14: 148–166.

Pickl, Simon. (2013). *Probabilistische Geolinguistik. Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben.* Stuttgart: Steiner.

Pickl, Simon, & Rumpf, Jonas. (2011). Automatische Strukturanalyse von Sprachkarten. Ein neues statistisches Verfahren. In E. Glaser, J. E. Schmidt, & N. Frey (eds), *Dynamik des Dialekts – Wandel und Variation. Akten des 3. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD).* Stuttgart: Steiner. 267–285.

Pickl, Simon, & Rumpf, Jonas. (2012). Dialectometric Concepts of Space. Towards a Variant-Based Dialectometry. In Hansen et al. (2012). 199–214.

Pröll, Simon. (2013). *Raumvariation zwischen Muster und Zufall. Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben.* PhD thesis, University of Augsburg.

Pröll, Simon, Pickl, Simon, & Spettl, Aaron (forthcoming). Latente Strukturen in geolinguistischen Korpora. In M. Elmentaler, M. Hundt, J. E. Schmidt (eds), *Deutsche Dialekte – Konzepte, Probleme, Handlungsfelder. Akten des 4. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD).* Stuttgart: Steiner.

Rumpf, Jonas, Pickl, Simon, Elspaß, Stephan, König, Werner, & Schmidt, Volker. (2009). Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik* 76/3: 280–308.

Rumpf, Jonas, Pickl, Simon, Elspaß, Stephan, König, Werner, & Schmidt, Volker. (2010). Quantification and Statistical Analysis of Structural Similarities in Dialectological Area-Class Maps. *Dialectologia et Geolinguistica* 18: 73–98.

SBS: *Sprachatlas von Bayerisch-Schwaben.* (1996–2009). Edited by Werner König. 14 volumes. Heidelberg: Winter.

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

Silverman, Bernard W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman & Hall.

Shackleton, Robert G. (2005). English-American speech relationships: a quantitative approach. *Journal of English Linguistics* 33: 99–160.

Shackleton, Robert G. (2007). Phonetic variation in the traditional English dialects: a computational analysis. *Journal of English Linguistics* 35: 30–102.

Speelman, Dirk, & Geeraerts, Dirk. (2008). The Role of Concept Characteristics in Lexical Dialectometry. *International Journal of Humanities and Arts Computing* 2/1–2: 221–242.

Speelman, Dirk, Grondelaers, Stefan, & Geeraerts, Dirk. (2003). Profile-Based Linguistic Uniformity as a Generic Method for Comparing Language Varieties. *Computers and the Humanities* 37: 317–337.

Stanford, James N. (2012). One size fits all? Dialectometry in a small clan-based indigenous society. *Language Variation and Change* 24/2: 247–278.

Szmrecsanyi, Benedikt. (2012). Geography is overrated. In Hansen et al. (2012). 215–231.

Tobler, Waldo R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46/2: 234–240.

Trudgill, Peter. (1974). Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society* 3/2: 215–246.

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

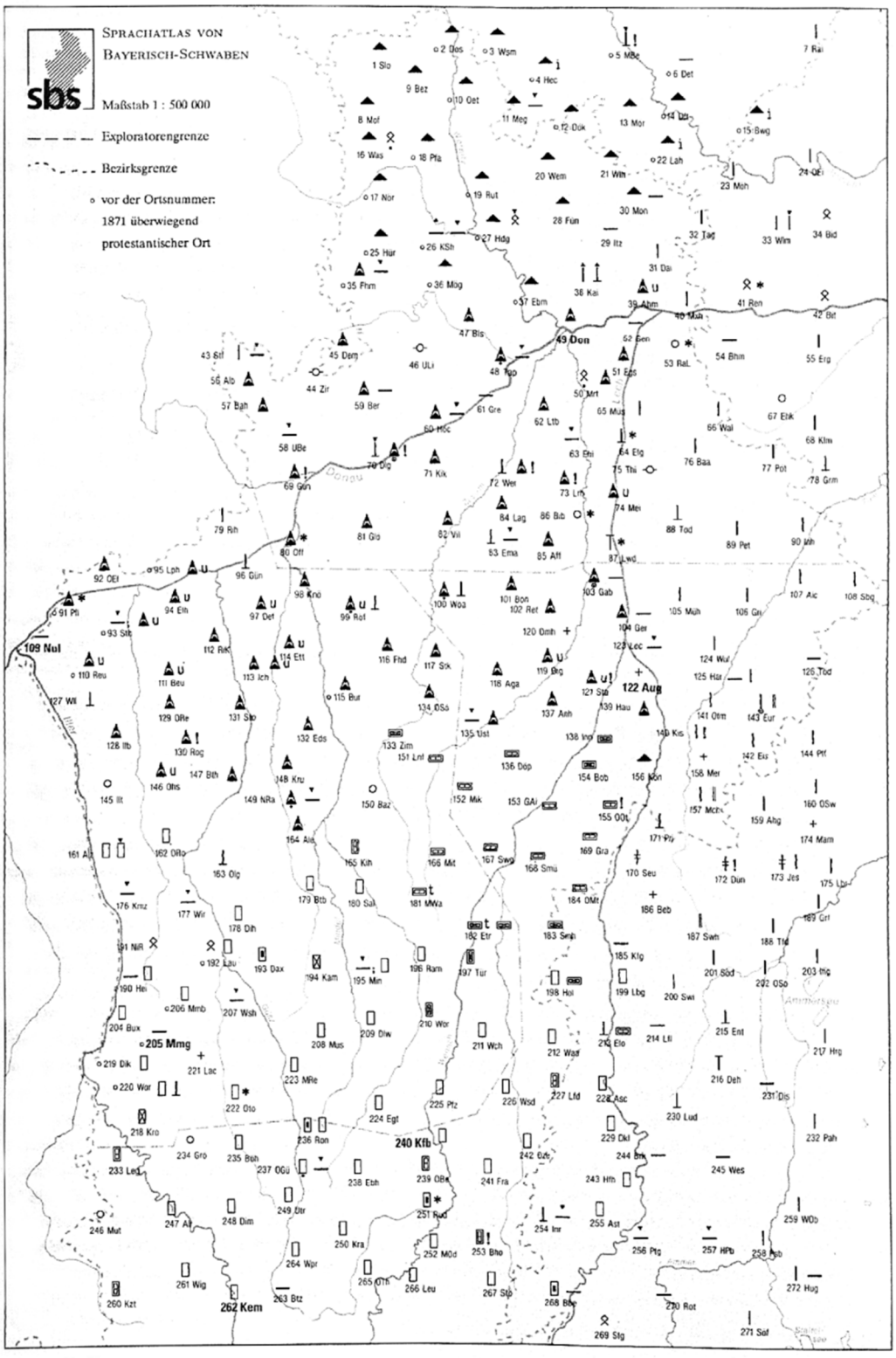

Fig. 1a: Point-symbol map

Karte 63 (Frage 258.14)

Kellerassel (Porcellio scaber)

Dreiecksymbole: Komposita mit »-esel«/»-essel«/»-isel« m.

▲ »Mauer-« (z.B. *mâorēsl*)

▲ j Vokal *i* im Grundwort (z.B. *mārīsl*)

▲ »Tunk-« (z.B. *dǭŋgēsl, dǫŋkhēsl*)

▲ u Vokal nicht gesenkt (z.B. *duŋkhēsl*)

Strichsymbole: »Assel« und verwandte Worttypen

— »(-)Assel« f. (z.B. *ǫsl, kxerasl, aflə*)

| »(-)Nassel« f. (z.B. *nǫsl, nasl*)

| »Nässel« [wohl f. Pl.] (z.B. *nēfl, nęflə*)

⊥ »(-)Rassel« f. (z.B. *rǫsl, raflə* Pl.)

T »Dassel« f. (216 Deh *daflə*)

ǂ »Natzel« f. (z.B. *nādsl, nādsləx* Pl.)

˙ über Strichsymbolen: Bestimmungswort »Mauer-«

Rechtecksymbole: Bildungen mit »Mauch-«/»Mauk-« u. ä.

□ »(-)Mauchen« Pl. (z.B. *môuxə*, 225 Pfz *môũxə*)

▣ »Maucher« m. (*môuxər* Sg. = Pl.)

▯ »Mauche(r)le^in^« n. (z.B. *môuxərlə, môuxələ*)

▯ »Maucheler« m. (z.B. *môuxələr*)

▯ »Mauken« Pl. (z.B. *môukə, môukxə*)

▭ »Mäuchele^in^« n. (z.B. *mêixələ*)

▭ »Mäucheler« m. (z.B. *mêihələr, mêixələr*)

t nach Rechtecksymbolen: Stamm »Mäucht-«

˙ über allen Symboltypen: Bestimmungswort »Keller-«

☿ Seltenheiten:

16 Was: *sibədâosədvīəslər* »Siebentausendfüßler«

27 Hdg: *khęlrkhēvər* »Kellerkäfer«

34 Bid: *dâosədvêislə* »Tausendfüßler«

41 Ren: *nēslkhēvə* »Nässelkäfer«

42 Bit: *rusn* »Russen« Pl.

50 Mrt: *rīəsl* »?Rüssel«

191 NiR: *bōdəkxēəvr* »Bodenkäfer« cher Sammelbegriff

192 Lau: *maltə, maltənə* Pl. »Malte« f.

269 Stg: *rufə* »Russen« Pl.

Weitere Zeichen:

- unter dem Symbol: Beleg als „älter", „richtiger" u.ä. qualifiziert
- ° unter dem Symbol: Beleg als E, „sehr alt" u.ä. qualifiziert
- ; zwischen den Symbolen: semantische Differenzierung; vgl. Belegliste
- ○ Wort unbekannt
- -○- Wort und Sache unbekannt
- \+ nicht gefragt
- ! Hinweis auf Belegliste
- * Hinweis auf Kommentar

Fig. 1b: Legend

Fig. 1: Map 63 from vol. 8 of the SBS, showing the distribution of the variants for 'woodlouse'.

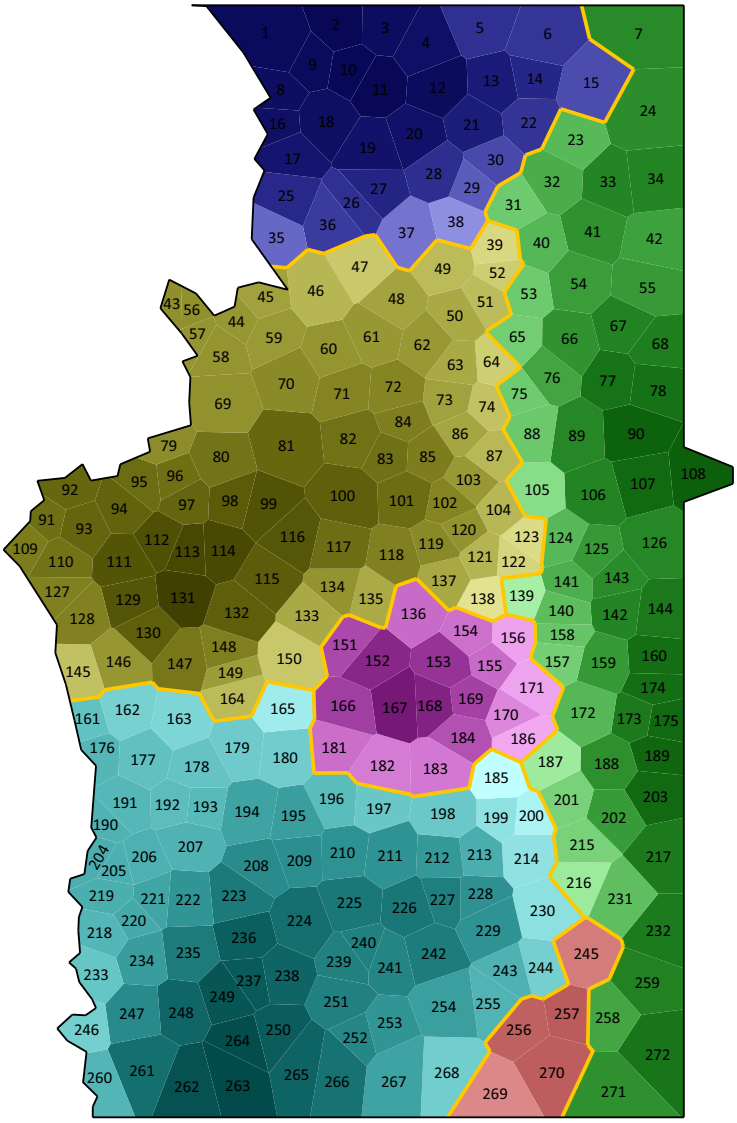

Fig. 2: Map 63 from vol. 8 of the SBS, intensity estimation, geographical distance, Level 2, K_3, h = 22 km (*CL*).[1]

[1] The parameter *Level* specifies the degree of abstraction from the raw data, i.e. what criteria were used to categorize the individual records into variants (cf. Pickl, 2013: 72–78). Level 2 is a medium setting and is used throughout this paper.

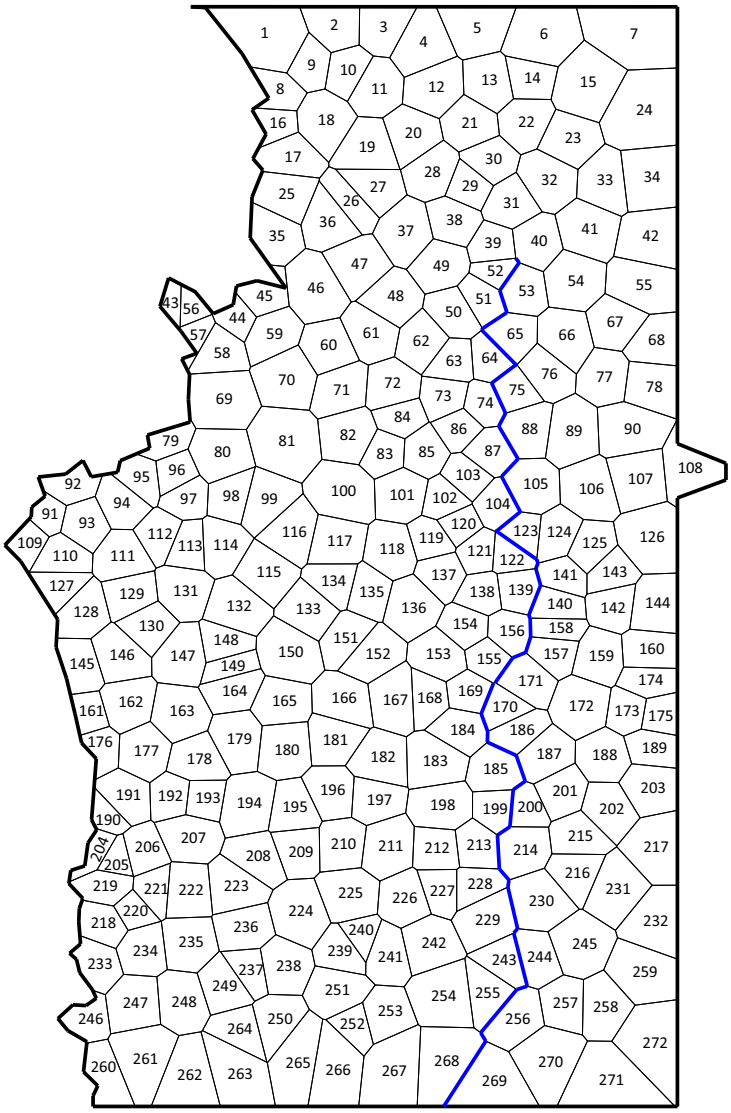

Fig. 3: Shape of the river Lech in the area under investigation in Voronoi rendering.

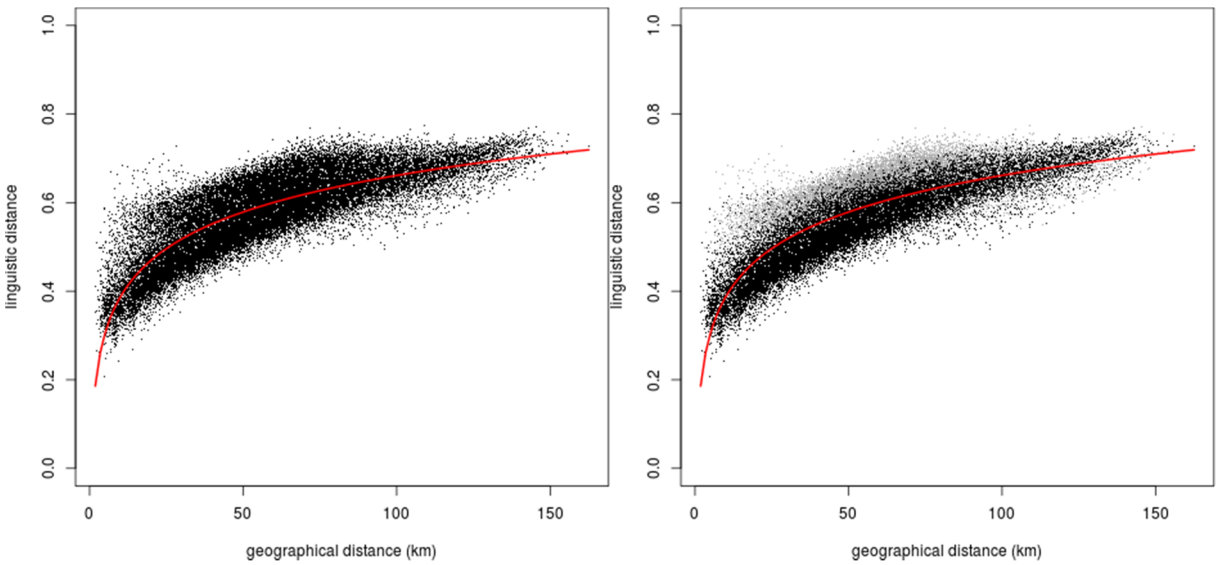

Fig. 4: Scatter plot of geographical vs. lexical distances and logarithmic regression curve. In the visualization on the right, the location pairs that are separated by the river Lech are highlighted.

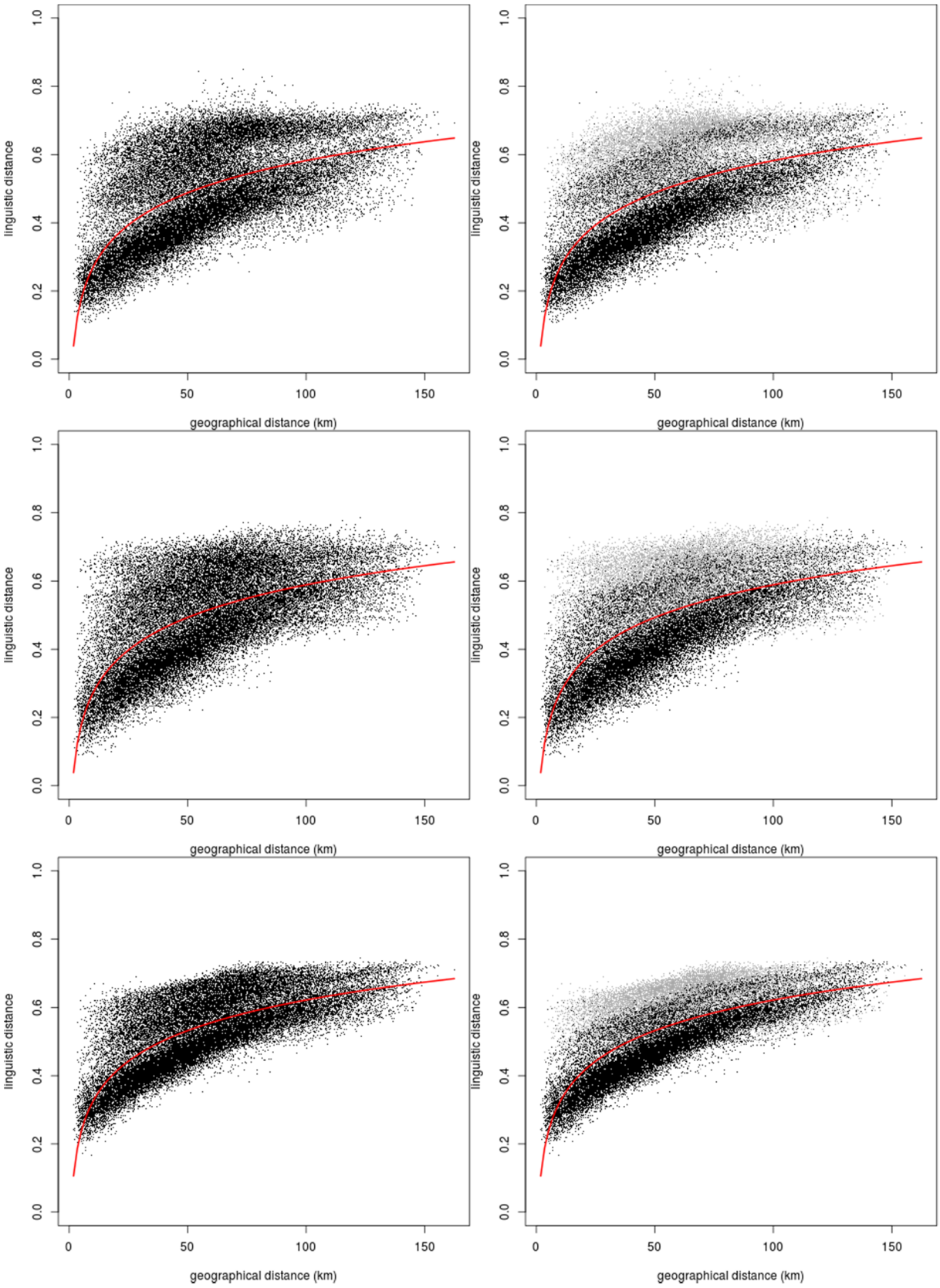

Fig. 5: Scatter plots of geographical vs. linguistic (from top to bottom: morphological, phonetic, all) distances and logarithmic regression curves. In the visualizations on the right, the location pairs that are separated by the river Lech are highlighted.

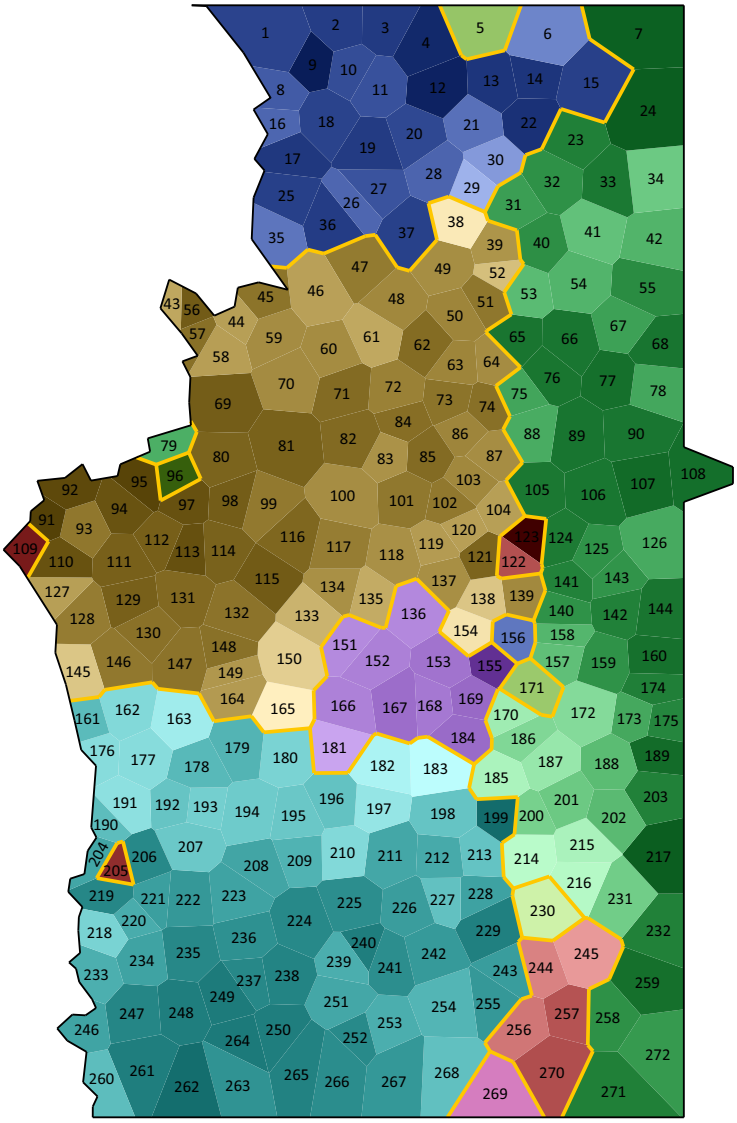

Fig. 6: Map 63 for vol. 8 of the SBS, intensity estimation, lexical distance, Level 2, K_3, $h = 0.55$ (CL).

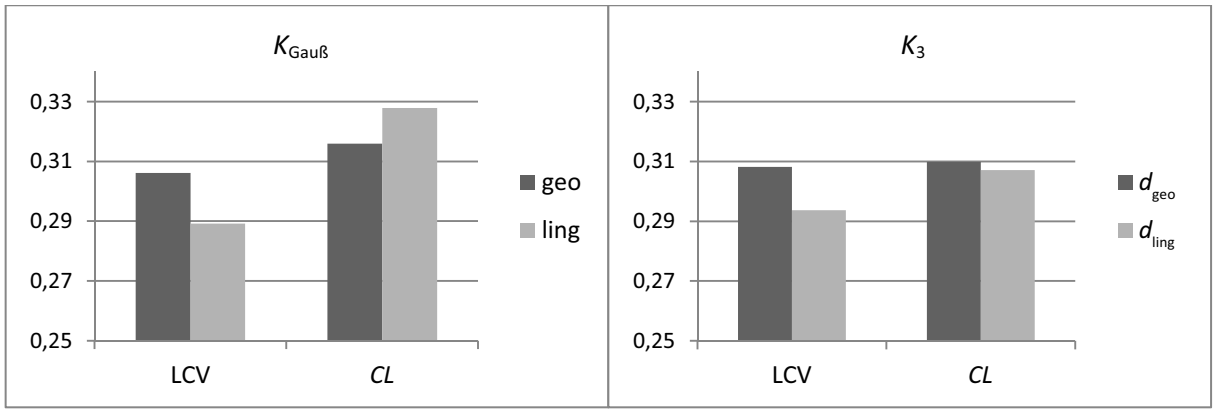

Fig. 7: Average scores for lexical and geographical distances in an application of intensity estimation to 736 lexical maps across all locations and all variables (lower is better).

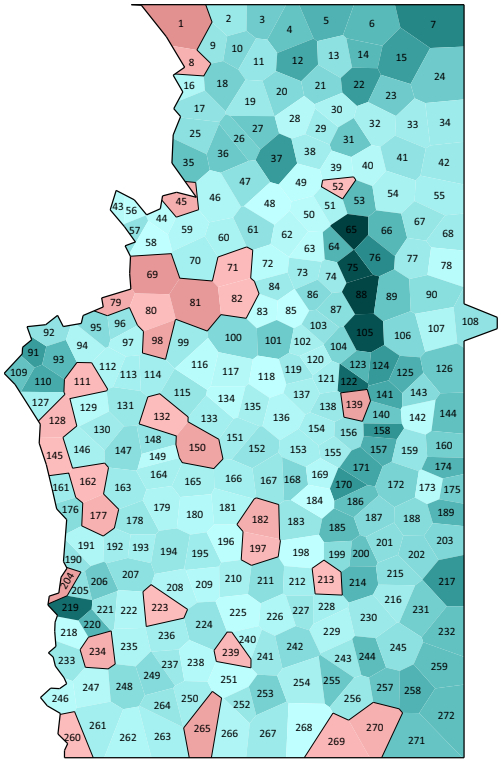

Fig. 8a: $K_{\text{Gauß}}$, LCV

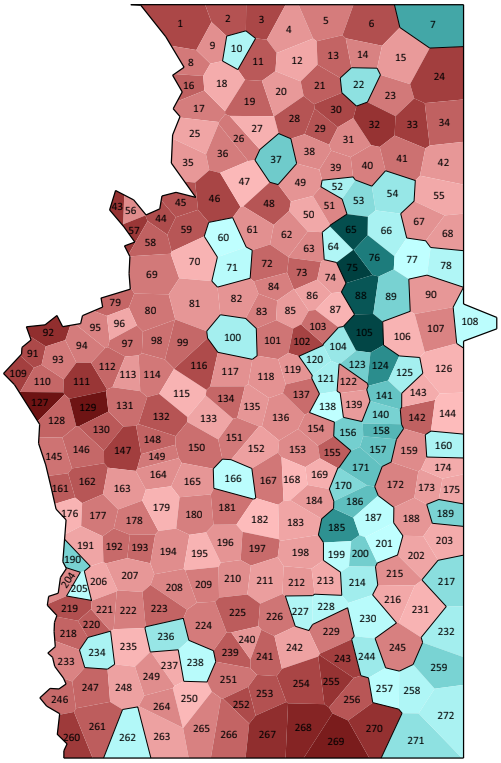

Fig. 8b: $K_{\text{Gauß}}$, *CL*

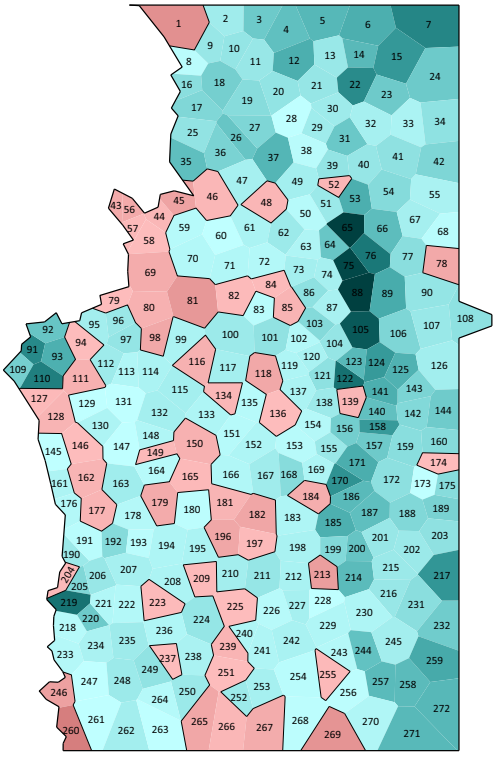

Fig. 8c: K_3, LCV

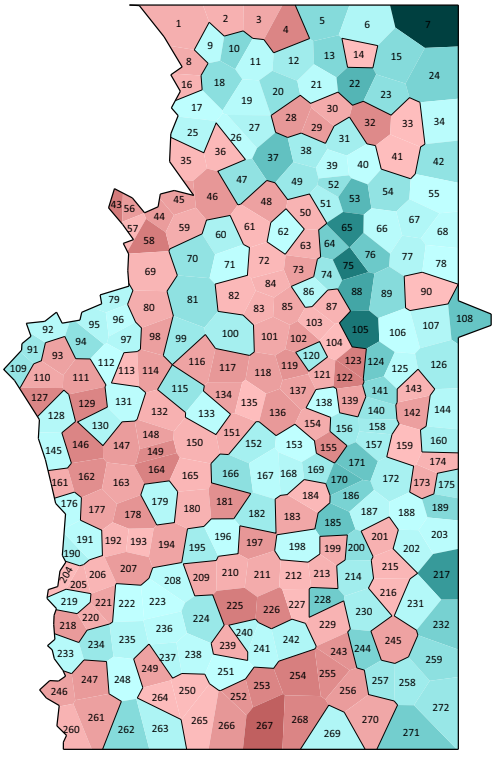

Fig. 8d: K_3, *CL*

Fig. 8: Average scores for lexical (blue) and geographical (red) distances in an application of intensity estimation to 736 lexical maps across all locations. The respective lower value (linguistic or geographical) was colour-coded at the individual locations, the colour intensity being higher for scores with a larger advantage over the respective other one.

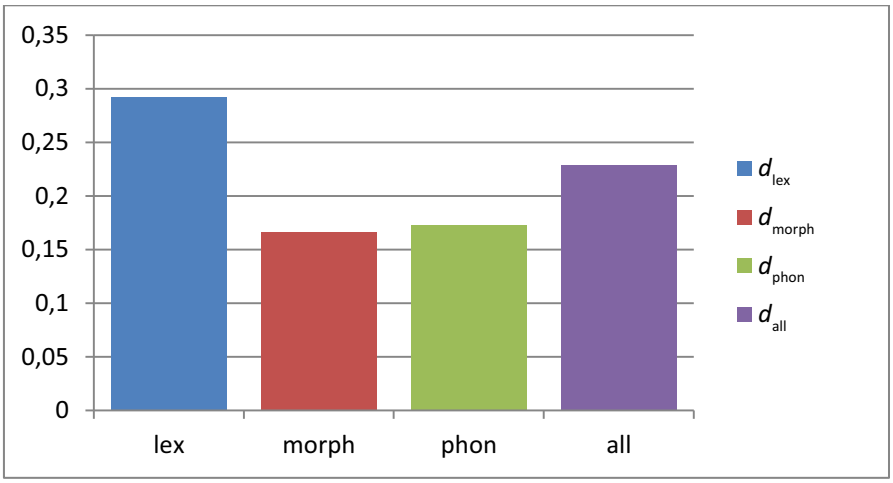

Fig. 9: Average scores for linguistic distances (d_{lex} etc.) in an application of intensity estimation to the respective test corpus (lex etc.) across all locations and all variables (lower is better).

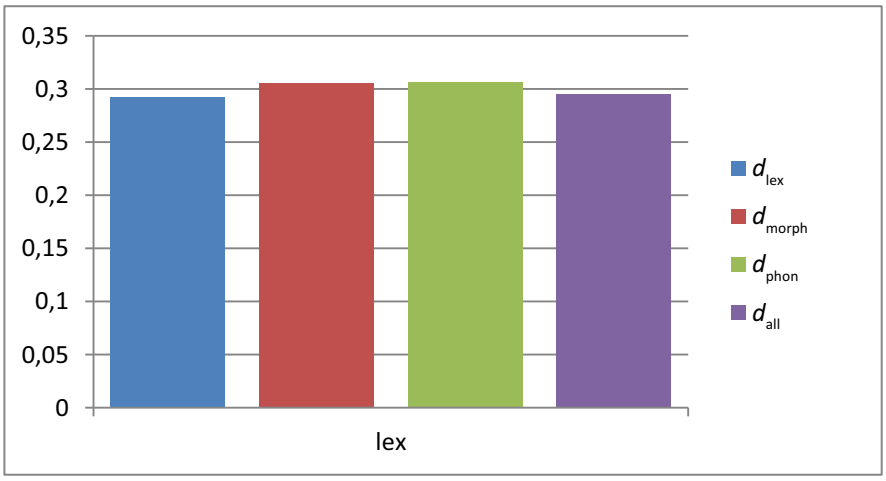

Fig. 10: Average scores for linguistic distances (d_{lex} etc.) in an application of intensity estimation to the lexical subcorpus (lex) across all locations and all variables (lower is better).

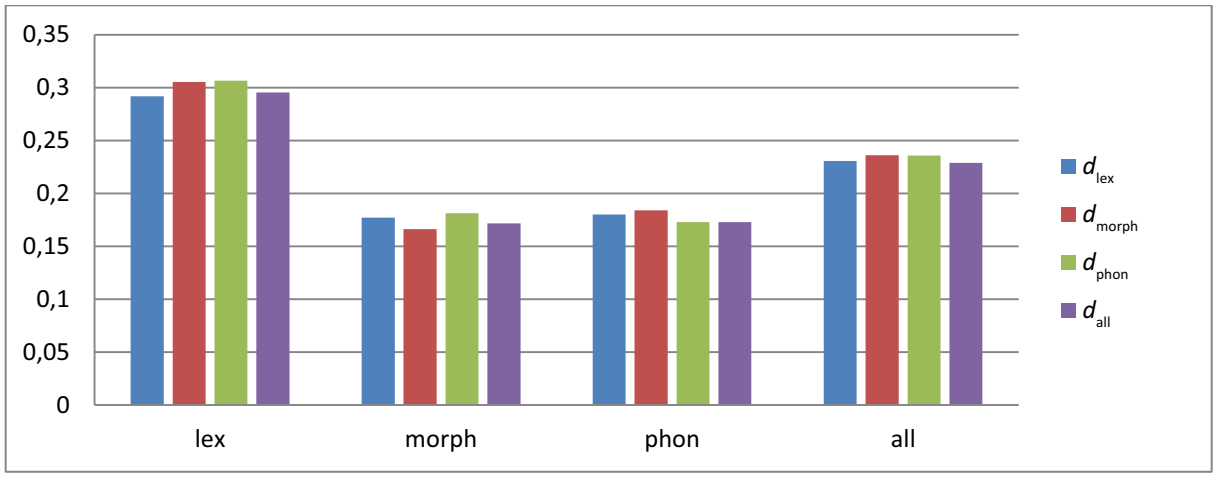

Fig. 11: Average scores for linguistic distances (d_{lex} etc.) in an application of intensity estimation to the all test corpora (lex etc.) across all locations and all variables (lower is better).