Evaluating Phonetically Weighted and Unweighted Distance Measures in Dialectometry

Alfred Lameli
Research Center Deutscher Sprachatlas, Marburg University
Pilgrimstein 16, 35032 Marburg
lameli@uni-marburg.de

Abstract

This paper compares phonetically weighted and unweighted string distance measures in dialectometry, examining how explicit phonetic modeling affects the quantitative representation of linguistic similarity. Using narrow IPA transcriptions from the German REDE corpus, we evaluate nine measures—Levenshtein distance, bigram and trigram overlap, cosine distance, Jaro-Winkler, Jaccard similarity, the Herrgen-Schmidt measure, and the Relative Identity Value—through correlational analysis, distributional comparison, stabilization testing, and multidimensional scaling. The phonetically weighted Herrgen-Schmidt measure consistently achieves the most balanced distance dispersion, earliest stabilization, and highest linguistic plausibility. Unweighted edit-based measures reproduce the same topological structure in compressed form; distributional and overlap-based metrics introduce systematic scale distortions through exaggeration or compression. These findings establish explicit phonetic weighting as a principled and analytically efficient extension of standard dialectometric procedures. Explicit phonetic weighting enhances resolution and interpretive precision without altering the underlying relational geometry of dialect classifications.

Keywords: Dialectometry, Edit Distance, String Similarity, German, Language Variation, Dialect

1. Introduction

Quantitative distance measures are central to dialectometry, providing a means of capturing and comparing systematic phonetic differences across dialect areas. Since the emergence of dialectometric approaches in the 1980s, a wide range of metrics has been applied to quantify linguistic similarity across dialect areas. Early studies employed unweighted procedures such as Goebl's (1982) Relative Identity Value (RIV, equivalent to the Simple Matching Coefficient), the Levenshtein distance for strings of equal or unequal length (Nerbonne and Heeringa, 1997, e.g.,), or the Hamming distance restricted to strings of identical length (Spruit, 2005, e.g.,). Subsequent work drew on overlap- and distributional measures, including the Jaccard coefficient (Scherrer and Stoeckle, 2016, e.g.,) and cosine-based distances derived from n-gram profiles (Birkenes and Fleischer, 2021, e.g.,).

What these procedures share is a treatment of phonetic differences as categorically uniform: every substitution, insertion, or deletion is assigned the same cost, regardless of the actual phonetic similarity of the segments involved. Frequency weighting—e.g., the Weighted Identity Value (GIW) by Goebl, 1982 or the LDϕ by Lameli et al., 2020—constitutes a partial exception, but it adjusts for token frequency rather than for phonetic magnitude. Explicit phonetic weighting, in which operation costs vary according to articulatory proximity, has remained marginal, with the study by Heeringa and Braun (2003) representing one of the few systematic applications.

This uniformity of cost assignment is problematic for phonetically transcribed data, where contrasts are inherently gradient. Articulatorily, the difference between [i] and [e] is substantially smaller than that between [i] and [a]; yet standard string metrics treat them as equivalent substitutions. Such flattening risks suppressing exactly the fine-grained phonetic distinctions that dialectometric analysis is designed to capture. Two independently developed phonetically weighted systems addressed this issue: Herrgen and Schmidt (1989) for dialect–standard comparisons and Almeida and Braun (1986) originally developed for transcription validation. Both assign graded segment scores along multiple articulatory dimensions.

Despite their conceptual appeal, such systems were rarely applied in large-scale dialectometry. Herrgen and Schmidt's metric remained limited to dialect–standard relations (see, e.g., Steiner, 1994, Kehrein, 2012, Vorberger, 2019, Lameli, 2025), whereas the Almeida-Braun system was adapted to dialectometric use (Heeringa and Braun, 2003), but without systematic comparison to unweighted procedures. The central question of whether, and to what extent, explicit phonetic weighting changes dialectometric outcomes has therefore remained unanswered.

The present study addresses this question by directly comparing a total of nine phonetically weighted or unweighted string distance measures. Using narrow IPA transcriptions from the REDE corpus (Schmidt et al., 2020ff.), our analysis proceeds in four steps: (1) we examine how the measures relate empirically through correlational and distribu-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4141–4151
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tional comparisons; (2) we evaluate their relative performance and robustness; (3) we assess stabilization behavior with increasing data size; and (4) we test linguistic plausibility through case studies and dialect classifications. Together, these four steps provide a basis for method selection that goes beyond correlation-based benchmarking.

2. Material and Methods

2.1. Data

Our empirical basis consists of three types of speech recordings from the REDE project (Schmidt et al., 2020ff.): (i) readings of Aesop's fable *The North Wind and the Sun*, (ii) translations of dialectal Wenker sentences (see Wenker, 2013) into Standard German, and (iii) translations of Standard German Wenker sentences into the respective local varieties. The three settings represent systematically different degrees of dialect intensity: while the reading and the standard-directed translation task elicit relatively standard-proximate speech, the dialect translation task is designed to elicit maximal dialect divergence. This deliberate variation ensures that the distance measures are tested across the full range of phonetic contrast available in the corpus, from near-standard to maximally dialectal. Each of the seven locations contributes recordings from multiple older speakers (> 60 years); the broader REDE corpus comprises 830 speakers across 148 locations, with up to six elicitation settings per site.

The sample covers the most divergent dialect areas (see Table 1) as delineated in Lameli (2013). Table 1 summarizes the field sites selected from the REDE corpus. Each location represents a distinct dialect area, so the analysis covers different German varieties. Each recording is available in a narrow phonetic IPA transcription, ensuring a high level of phonetic detail and comparability across the dataset. Figure 4 illustrates the range of phonetic transcription across locations, exemplified by the variant space of the functional word *und* ('and').

Location	Area
Oldenburg (OL)	Northern Low German
Prignitz (PR)	Brandenburgish
Merzig (MZG)	West German
Darmstadt (DA)	West Central Germ.
Zwickau (Z)	East Central Germ.
Loerrach (LOE)	West Upper Germ.
Traunstein (TS)	Bavarian

Table 1: Research locations from the REDE corpus, representing seven German dialect areas.

We use the complete reading text comprising 106 words per survey location. In addition, we take the first 200 words from each of the two translation tasks, yielding $\approx$ 500 tokens per location. Combined across all seven locations, this produces a total of 3,559 tokens spanning a wide range of dialectal and register variation. All tokens were measured against their Standard German equivalents, using the IPA reference transcription of the REDE project compiled from the HadiBOMP (Portele et al., 1995) and Verbmobil (Wahlster, 2000) projects (see Lipfert, 2024). For the dialectometric location comparisons reported in Section 3.6, the 500-token sets per location were used.

2.2. Distance Measures

The nine measures examined in this study fall into two broad methodological families. *Edit-based approaches*—Levenshtein distance and the Herrgen-Schmidt measure (including its PAM variant)—quantify the minimum number or cumulative weight of operations required to transform one string into another. *Overlap- and distributional measures*—bigram distance, trigram distance, cosine distance, Jaro-Winkler, Jaccard index, and the Relative Identity Value—assess similarity through the proportion of shared substrings, characters, or positional matches. As will be shown in Section 3.1, this theoretical classification is only partially empirically grounded: n-gram measures behave more like edit distances in practice, and Jaro-Winkler aligns with overlap measures rather than with Levenshtein. The individual measures are described below.[1]

Levenshtein distance (LV) calculates the minimal number of operations—substitution, insertion, and deletion—required to transform one string into another (Levenshtein, 1965). Each operation is assigned equal cost, which makes the method flexible for strings of unequal length. Although Levenshtein distance assigns uniform costs to all edit operations, its application to IPA-encoded strings introduces a functional analogue to phonological sensitivity: substitutions, insertions, and transpositions between symbols typically correspond to phoneme-level alternations, so that the metric implicitly captures some degree of phonological structure even in the absence of explicit phonetic weighting. In this paper, we apply the Damerau–Levenshtein distance, which also allows for transpositions such as

[1] Throughout this paper, the terms overlap-based, distributional, and inventory-based are used to characterize measures that assess similarity through shared substrings, character sets, or frequency vectors rather than through sequential edit operations. Although these terms reflect genuine differences in computational design—n-gram overlap, set intersection, and vector angle are distinct operations—they converge empirically in ways that make a strict terminological boundary impractical. The labels are therefore used descriptively, referring to the dominant mode of comparison rather than to a mutually exclusive classification.

metathesis (Damerau, 1964).

n-gram distances treat strings not as sequences of individual symbols but as distributions of adjacent sound clusters of length n. Similarity is determined by the degree of overlap between the n-gram profiles of two strings. In this paper, we consider bi- and trigrams (N2, N3 with boundary markers $\#\ldots\#$), which capture local sequential dependencies of length two and three. In contrast to LV, n-gram approaches emphasize phonotactic structure rather than sequential edit cost. By evaluating overlapping sound clusters, they allow phonetically adjacent segments to contribute proportionally to similarity. Bigram and trigram profiles thus reflect phonotactic cohesion and thereby encode aspects of phonological well-formedness. However, unlike explicitly weighted systems, n-gram distances do not quantify phonetic magnitude; their sensitivity arises from distributional rather than articulatory proximity.

Cosine distance (COS), by contrast, also builds on n-grams (bigrams in this paper) but takes their frequency distribution into account. Each string is represented as a vector of n-gram counts, and the distance is defined as one minus the cosine of the angle between the two vectors. While pure n-gram distances capture the extent of overlap, COS additionally reflects how similar the patterns of distribution are, even when exact overlap is limited. High similarity corresponds to a small angle (low distance), while divergence in the distribution of n-grams results in a larger angle.

Jaro–Winkler (JW) scores the proportion and order of matching characters within a window and down-weights transpositions, with a prefix bonus (prefix scaling in this paper: $p = 0.1$; max prefix length $l = 4$) for early matches (Winkler, 1990). Unlike LV's edit-path framing, JW operates via matching windows and positional weighting. This makes it particularly sensitive to near-identical forms with small positional changes. In this paper, JW is transformed into a distance measure, defined as $1 - JW$.

Jaccard distance (JAC) compares two strings by treating them as sets of elements (Jaccard, 1902). Here, we use sets of unigrams. JAC is here defined as one minus the ratio of the intersection to the union of the sets. In this way, it measures the proportion of shared units relative to the total number of distinct units. Because sequential order is disregarded, JAC abstracts away from string structure and instead highlights the extent of overlap at the level of segmental inventory.

Relative Identity Value (RIV) was introduced to dialectometry by Goebl (1982) as a re-labeling of the *Simple Matching Coefficient* (SMC). It is a similarity measure that calculates the proportion of identical positions in two strings of equal length. In this study, we transform the measure into a distance measure, defined as $1 - RIV$.

The *Herrgen-Schmidt distance* (HS) is a phonetically weighted distance measure. Each segmental opposition is evaluated along articulatory and acoustic dimensions, and graded scores are assigned depending on the degree of difference to the segmental counterpart (Herrgen and Schmidt, 1989). For vowels, differences in height (close–mid–open) and frontness (front–central–back) are measured in half- or full-step increments (0.5, 1.0, or 1.5 points), with additional points assigned for rounding and nasality. Thus, [i] vs. [e] is rated as a 0.5-point difference, whereas [i] vs. [a] is scored 1.5 points and [a] vs. [ã] is 0.5 points. The maximum vowel opposition is capped at 3 points. For consonants, three dimensions are considered: place of articulation, manner of articulation, and voicing. Each full difference adds 1 point (or 0.5 for diacritics highlighting phonetic tendencies). For example, [t] vs. [d] differs in voicing only and receives 1 point, while [t] vs. [m] differs in place and manner and is scored 2 points; [g] vs. [g̊] is 0.5 points. The maximum consonant opposition is 2 points. The cumulative score across a word yields its phonetic distance. We apply two variants of this method: (a) a strictly phonetic version based solely on these weighting rules (HS), and (b) a reduced version excluding coarticulatory reduction phenomena such as schwa deletion, assimilation, or final devoicing in spontaneous standard speech, following Herrgen et al. (2001) (PAM; *phonetische Abstandsmessung* in German).

PAM occupies an ambiguous position in the comparative framework of this study. Its design makes PAM specifically suited to research contexts where near-standard coarticulatory variation is treated as phonetically irrelevant noise rather than as a signal of dialectal distance. Because PAM's measurement protocols are referenced against Standard German, it cannot be applied symmetrically to all pairwise location comparisons and is therefore excluded from the MDS analysis in Section 3.6. Where it does appear, its results should be interpreted as a methodological reference point rather than a general-purpose competitor to the remaining eight measures.

2.3. Procedure

We computed all non-edit-based distances with the *R* package stringdist (van der Loo, 2014), and re-implemented the HS metric in *R* following the specification in Herrgen and Schmidt (1989) and Herrgen et al. (2001). HS was originally designed to quantify the phonetic distance of a spoken form from its Standard German equivalent rather than pairwise distance between two dialect varieties. Its application here to cross-dialectal comparison is an extension of its original scope. For measurements

excluding coarticulation (PAM), measurement protocols that contrasts spoken language with standard German were provided to us by the REDE project (see the technical reports by Meinck and Limper, 2013 and Lipfert, 2024). All distance computations were performed automatically on the basis of manually produced narrow phonetic transcriptions; only PAM requires additional manual verification of coarticulatory phenomena.

3. Results

3.1. Overview

For every pair of measures, a statistical correlation between all word–word distance values were calculated using Spearman's ρ, which we choose as the more robust measure. Figure 1 shows the correlation matrix ($p < .001$). Darker colors mean that two measures give similar results; lighter colors mean they behave differently (note that the similarity measures are transformed into distance measures).

From a methodological point of view, the procedures differ in their underlying logic: edit-based approaches such as LV and HS quantify the number or weight of operations needed to transform one string into another, whereas overlap- or distributional measures such as n-gram similarities (N2, N3), JAC, COS, and JW rely on the proportion of shared substrings or tokens (see above). Empirically, however, these theoretical divisions are not cleanly preserved. LV, HS, and the n-gram measures form a compact block with very high correlations (up to $\rho = .94$ for N2–N3, and $\rho = .89$ for LV–N2). Although methodologically overlap-based, N2 and N3 align more closely with edit distances in practice, reflecting their shared sensitivity to local segmental changes. By contrast, JAC, COS, and JW constitute a second group (JAC–JW $\rho = .93$, COS–JW $\rho = .78$, COS–JAC $\rho = .77$). Notably, JW—which is an edit-based method—behaves here like an overlap measure, aligning strongly with JAC and COS rather than with LV.

This behavioral realignment of JW is attributable to its prefix-weighting mechanism: by assigning bonus credit to matching characters at the onset of strings, JW systematically inflates similarity for forms sharing initial segments, irrespective of differences elsewhere in the string. The measure thus behaves less as a global edit-cost estimator and more as an onset-sensitive overlap detector, which is a design feature that has phonological utility in record linkage tasks but introduces a structural bias when applied to phonetically variable dialect data.

PAM connects moderately to both groups (PAM–HS $\rho = .63$, PAM–LV $\rho = .59$, PAM–JAC $r = .43$), but without forming a clear attachment. RIV correlates highest with JW ($\rho = .74$), JAC ($\rho = .73$) and

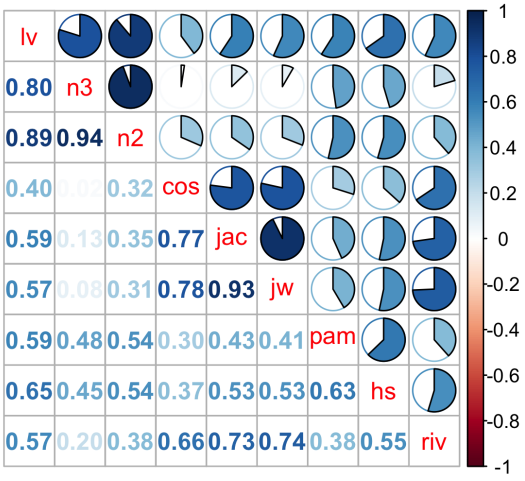

Figure 1: Correlation matrix of distance measures; darker colors indicate higher agreement.

COS ($\rho = .66$). Correlation with the edit/n-gram group is moderate (absolute $\rho \approx .38$–$.55$). Finally, HS emerges as a bridging measure, maintaining high correlations with LV ($\rho = .65$) and N2 ($\rho = .54$) but also moderate links to JAC and JW ($\rho = .53$).

Across contexts (reading, standard, dialect), overall agreement is lowest for dialect translations ($\bar{\rho}_{(dialect)} = 0.53$) and higher for reading ($\bar{\rho}_{(reading)} = 0.59$) and standard language ($\bar{\rho}_{(standard)} = 0.58$). This indicates that the greater heterogeneity of dialects slightly reduces convergence among methods.

3.2. Statistical Distributions

While the correlation analysis shows which methods are most strongly associated with each other, Figure 2 highlights how the central tendency and shape of the distributions differ across approaches. To this end the distant measures were *z*-standardized. A narrow, centered curve means the measure treats most words as moderately different. A wide or skewed curve means it stretches or compresses the differences.

The edit- and n-gram measures (HS, PAM, LV, N2, N3) display softly right-skewed distributions, indicating a smooth gradation of distance values with a predominance of smaller distances. By contrast, RIV, JAC, COS, and JW skew in the opposite direction, with a tendency toward larger distance values; JW is further distinguished by a bimodal distribution, reflecting its onset sensitivity rather than a continuous phonetic scale. LV and, to a lesser extent, HS approach a more symmetric configuration, suggesting that they distribute distances most evenly across the scale.

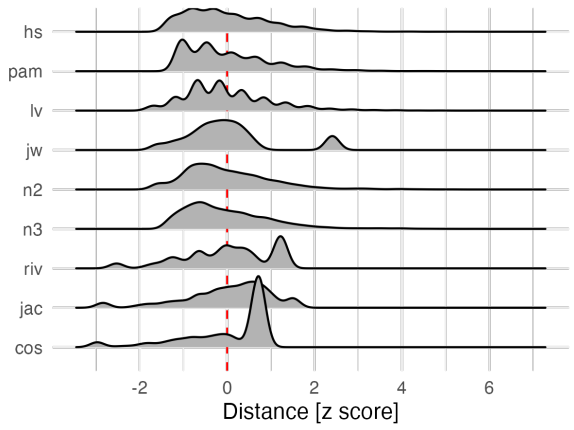

Figure 2: Distributions of standardized distance values for each measure.

The results show that the measures differ not only in the strength of their correlations but also in the shapes of their statistical distributions. Methodological distinctions are partly overridden by empirical grouping: overlap-based n-gram measures behave like edit distances, and JW aligns with overlap methods rather than with LV. A Friedman test on the pairwise distances confirmed significant variation across methods ($\chi^2(8) = 219.535, p < .001$), yet the associated effect size was extremely small (Kendall's $W = 0.008$).[2] This shows that the divergences are systematic but modest in magnitude.

3.3. Relative Performance and Robustness

To assess whether the relative ordering of measures is stable across items—i.e., whether each measure consistently produces larger or smaller distances than the others—we converted raw distances into within-pair ranks (rank 1 = smallest distance, rank 9 = largest). This rank transformation abstracts away from method-specific scaling and makes ordinal comparisons directly interpretable across heterogeneous items (i.e., the individual units over which the distances were calculated, e.g., word pairs, sounds and sound combinations). For each method, the average of these ranks (mean rank) indicates its typical position in the overall spectrum. Methods with low mean ranks tend to produce smaller distances on average, those with high mean ranks larger distances.

To test pairwise differences explicitly, we computed bootstrap differences of mean ranks (A–B) for every pair of methods ($B = 1000$). Figure 3 summarizes the mean differences and their 95% confidence intervals. The color intensity shows how often one measure produced higher or lower ranks than another. Red cells indicate that A has a higher mean rank (larger distances) than B, blue the opposite; bold values denote contrasts whose bootstrap confidence interval excludes zero (BH-adjusted[3]).

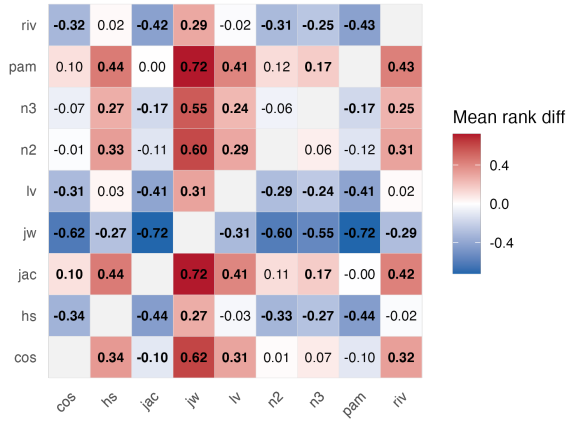

Figure 3: Heatmap of rank differences between measures based on bootstrap samples; stable color blocks indicate similar ranking behavior.

As can be seen from Figure 3, the differences in ranking are rather small, which indicates the robustness of the methods. Nevertheless, the heatmap reveals that RIV, LV, JW, HS consistently occupy lower positions across nearly all contrasts, setting them apart from the rest (PAM, N3, N2, JAC, COS), consistent with their tendency to produce smaller absolute distances as reported in Section 3.2. This suggests that the key result lies not in differences of robustness—which is uniformly high—but in the stable division into two distinct classes of distance measures.

3.4. Stabilization

Building on the assessment of robustness in the relative ranking of methods, we next turn to the question of stabilization over input size. Stabilization was assessed using the difference method (Diff) with the *R* package zoo (Zeileis and Grothendieck, 2005). The procedure compares the cumulative mean of each metric with its value a fixed number of words earlier. To this end, each metric was *z*-standardized over items. We evaluate windows $W \in \{100, 200, 400\}$ and thresholds $\theta \in \{0.0025, 0.005, 0.01\}$. A metric is declared stable at the earliest point where the absolute cumulative difference falls below θ for at least W. In the given case a window of $W = 200$ and a tolerance of $\theta = 0.005$ turned out to be the most powerful configuration in order to capture when the long-term

[2] Bootstrapped 95% confidence interval $[0.007, 0.010]$ with $B = 1000$.

[3] Multiple testing controlled with the procedure of Benjamini and Hochberg, 1995.

trend has leveled off and additional words no longer change the overall outcome substantially.

Metric	Stabilized after … words
HS	211
PAM	216
N2	217
JW	234
RIV	345
JAC	403
LV	431
N3	436
COS	–

Table 2: Estimated number of words required for each measure to reach a stable average distance. Smaller numbers = faster stabilization; (n = 3559).

The results in Table 2 show that especially the phonetically weighted measures stabilize very early: HS after about 211 words, PAM after 216 words, followed almost simultaneously by N2 (217) and closely by JW (234). The other measures reach stability somewhat later (345–436 words) while COS did not reach a stable point within the available data. This is consistent with the distributional findings reported in Section 3.2: COS, when applied to bigram vectors of short strings, is highly sensitive to minor changes in n-gram composition, resulting in high variance across items and a slow-converging cumulative mean. The stabilization points suggest that the phonetically weighted measures reach reliable estimates with comparatively little data, making them efficient indicators. By contrast, measures such as RIV, JAC, LV, and N3 require larger input sizes; N2 and JW, despite belonging to different methodological families, pattern with the phonetically weighted measures in their early stabilization.

3.5. Plausibility

Beyond efficiency, a distance measure must also behave in linguistically interpretable ways. To evaluate this, we take the variant space of the functional word *und* ('and') as an example. This word is particularly suited for this comparison because it is phonetically reduced under all dialects and registers, yielding a large and structurally diverse variant space (over 80 different variants in the REDE corpus). This diversity provides a demanding test case: a plausible distance measure should assign systematically increasing scores as variants deviate progressively from the standard form, while remaining sensitive to the gradient phonetic differences between variants.

Figure 4 visualizes, for each variant (v) and measure (m), the baseline-centered difference

$$\Delta_{v,m} = \overline{z}_{v,m} - \overline{z}_{[\text{ʊnt}],m}$$

where $\overline{z}_{v,m}$ denotes the z-standardized distance value of variant v for method m. The reference form [ʊnt] serves as the baseline, meaning that $\Delta_{v,m}$ expresses how much a given variant deviates—positively or negatively—from the standardized score assigned to the baseline form ($\Delta = 0$). Colors in Figure 4 encode these differences: blue indicates smaller additional distance relative to [ʊnt], red larger ones. Baseline centering thus removes method-specific offsets and highlights systematic tendencies as well as relative strengths and weaknesses in capturing phonetic relationships.

The edit-based measures (LV, n-grams) provide a conservative baseline. They order variants in a phonologically consistent way: deletion of [t] in [ʊn] yields modest increases (≈ 0.5 for LV; ≈ 0.25 for N2/N3), voicing in [ʊnd] is treated similarly, and [n̩] rises further (LV ≈ 1.0; *n*-grams up to 0.77). Yet they compress the dynamic range: large structural changes are only slightly costlier than small edits, and vowel quality differences remain invisible. Plausible as baselines, they are too coarse for fine-grained analysis.

HS scales differences more convincingly. Subtle changes such as [ʊ̜nt] hardly shift the score (≈ 0.2), deletion of [t] in [ʊn] produces a moderate increase (≈ 0.6), and [n̩] rises strongly (≈ 1.6). HS thus captures both minor and major shifts without inflating them, making it a very good balanced and phonetically plausible procedure for the given word.

PAM suppresses distances from reduction or coarticulation but amplifies categorical changes. [ʊn] or [ʊ̜nt] stay near the baseline (≈ 0.39, 0.45), and [ən] adds virtually no distance (≈ 0.00). By contrast, [ɔn] rises sharply (≈ 2.1). This asymmetry reflects its design: PAM is plausible if reduction is to be ignored, but underestimates variation when reductional detail matters.

COS and JAC emphasize inventory and distribution. Even [ʊn] yields high values (COS ≈ 1.1; JAC ≈ 1.5), and [ən] or [n̩] reach extremes (COS up to 3.8; JAC ≈ 3.2–3.4). Such exaggeration leads to implausibly large distances for moderate changes. Moreover, COS under $n \geq 2$ cannot represent one-segment forms, which disappear from the analysis.

JW is dominated by onset sensitivity. Variants with matching onsets remain low ([ʊn] ≈ 0.5; [ʊ̜nt] ≈ 0.3), but [n̩] jumps dramatically (≈ 3.9). This reflects algorithmic weighting, not phonetic similarity: JW is diagnostic for onset mismatches but implausible as a general phonetic measure.

RIV highlights positional identity. Scores remain close to baseline when alignment is preserved (≈ 1.3 for [ʊn], [ʊnd]) but drop steeply when structure is reorganized (≈ 3.8 for [n̩]). RIV thus distinguishes sharply between preserved and disrupted structure but over-penalizes breaks and masks gradient phonetic proximity.

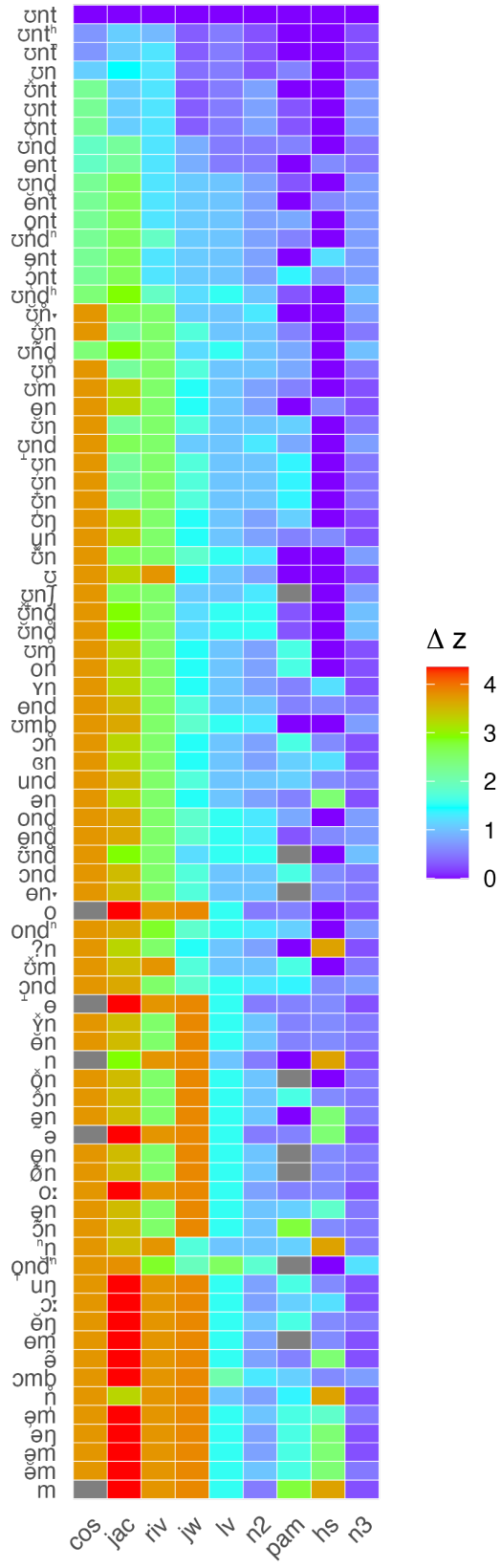

Figure 4: Average deviation (Δ) from the standard form [ʊnt] across dialect variants.

The boxplots in Figure 5 confirm these profiles. A wider box means more variation in the distances, that is the measure is more sensitive to fine phonetic detail. A narrow box means it compresses the differences. LV and n-grams cluster tightly below zero, treating almost all variants as near-standard. HS also centers near zero but with wider spread, reflecting its finer discrimination. PAM shows a broad distribution with low values for reductional variants and long positive whiskers for categorical changes. RIV stays near zero on average but produces extremes when alignment fails. JW has the widest spread, with a median around zero and a long positive tail, underscoring its onset bias. JAC and COS show positive medians and wide ranges, indicating their systematic emphasis on inventory divergence.

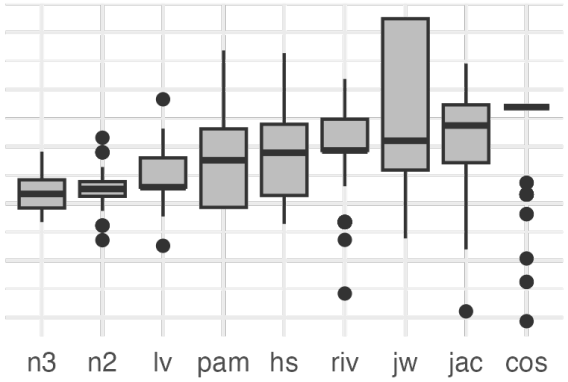

Figure 5: Standardized distances for dialect realizations of /ʊnd/ by measure.

Taken together, HS provides the most linguistically calibrated distance representation, while the remaining measures distort the distance scale through compression (LV, n-grams), suppression (PAM), exaggeration (JAC, COS), or structural bias (JW, RIV).

3.6. Dialectometric Comparison of Classifications

Finally, in order to examine the structural behavior of the distance measures a dialectometric comparison between all locations Multi-Dimensional Scaling (MDS) was applied to the distance matrices for each measure. For each location, the distance matrices aggregate tokens from all three recording conditions (reading, standard translation, dialect translation), producing a location-level distance profile that conflates register variation with dialectal divergence. The MDS analysis therefore does not aim to reproduce geographic dialect topology but rather to assess which measures produce internally consistent and structurally stable classifications of the data regardless of register, and which measures distort that structure through scale artifacts.

We report two-dimensional maps that preserve the relative differences as well as possible. For this

analysis PAM was left out because the available measurement protocols only report distances to the standard language. Figure 6 indicates how the dialect regions relate to each other visually; similar pronunciations appear closer together, different ones farther apart; measures that produce similar maps model dialect space consistently.

The configurations in Figure 6 reveal three distinct behavioral regimes rather than a single uniform topology. First, the phonetically weighted HS together with the unweighted edit-based metrics (LV, N2, N3) form a coherent core group. All four produce highly similar spatial arrangements, differing mainly in the degree of dispersion. HS spans the widest coordinate range ($Dim1 \approx -30\ldots60$; $Dim2 \approx -20\ldots40$), resulting in a clear and evenly proportioned configuration. This expansion reflects the effect of phonetic weighting: it enhances proportional contrast and distributes distances more uniformly without altering the relational geometry. LV ($STRESS_1 = 0.147$), N2 ($STRESS_1 = 0.147$l), and N3 ($STRESS_1 = 0.151$) occupy the same structure as HS ($STRESS_1 = 0.192$) but in a compressed range ($Dim1 \approx -15\ldots30$; $Dim2 \approx -10\ldots15$), showing that they capture comparable relations yet underestimate small phonetic differences.

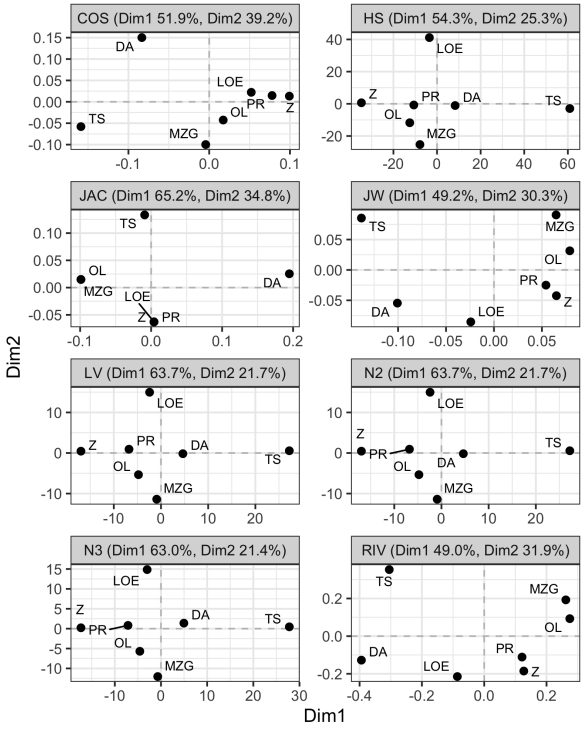

Figure 6: MDS maps visualizing dialect similarity: closer points = more similar pronunciations.

A second regime comprises COS ($STRESS_1 = 0.460$) and JAC ($STRESS_1 = 0.654$). Both exaggerate global contrasts and stretch the configuration outward, producing inflated distances between otherwise closely related points ($Dim1 \approx -0.15\ldots0.25$; $Dim2 \approx -0.1\ldots0.15$). This expansion stems from their reliance on frequency distributions and inventory overlap, which emphasize categorical divergence over fine-grained correspondence. The adjustment of the configuration to the raw data performs significantly worse here than the above-mentioned measures, based on the $STRESS_1$ values.

Finally, JW ($STRESS_1 = 0.145$) and RIV ($STRESS_1 = 0.163$) exhibit the opposite pattern: they contract the data into narrow clusters and differentiate only weakly ($Dim1 \approx -0.4\ldots0.3$; $Dim2 \approx -0.2\ldots0.3$). Despite their distinct computational logic, both minimize overall spread and display nearly parallel geometric behavior. Their restricted variance suggests limited sensitivity to intermediate phonetic gradience.

Against this background, the MDS analysis shows that only the core edit-based group (HS, LV, N2, N3) yields a consistent and interpretable topology of linguistic distances. The remaining measures diverge systematically. COS and JAC by amplifying contrast, JW and RIV by suppressing it. Phonetic weighting thus improves resolution and contrast balance within an otherwise stable structural core, whereas distributional and similarity-based metrics reshape the space more radically through scale distortion.

4. Discussion

The central question of this study—whether, and to what extent, explicit phonetic weighting changes dialectometric outcomes—can now be answered precisely: phonetic weighting does not alter the topological relations among dialect varieties, but it enhances the metric resolution with which those relations are represented. The difference between HS and unweighted edit distance is one of calibration, not of structure and consequently of degree, not of kind: HS amplifies existing contrasts rather than revealing new ones.

Three interpretive claims can be highlighted.

The measures share a topology but differ in resolution. The MDS analysis reveals that HS, LV, N2, and N3 produce nearly identical spatial configurations, differing only in the degree of spread. This is structurally relevant, because the pairwise correlations between HS and the unweighted measures are only moderate. Correlation, being sensitive to rank agreement across individual item pairs, does not fully capture agreement at the level of aggregate distance geometry. The nine measures therefore differ along two orthogonal dimensions: *topological faithfulness*—whether a measure preserves the re-

lational geometry of the dialect space—and *metric resolution*—how finely it differentiates among points within that space. The edit-based core group (HS, LV, N2, N3) is uniformly high on the former. HS additionally achieves higher resolution through its explicit phonetic weighting.

Phonetic weighting refines rather than redefines. Within the topologically faithful core, HS outperforms unweighted measures on three independent criteria: distributional balance, stabilization speed, and linguistic plausibility. All three advantages share one source: by assigning graded costs to segmental oppositions based on articulatory proximity, HS encodes the continuous, multidimensional structure of phonetic space rather than collapsing it onto a binary distinction between identity and change. Each edit operation carries information about the phonetic magnitude of the change, not merely about whether a change occurred. The result is less noise per token, a faster-converging cumulative mean, and a more proportionate distance scale. Crucially, however, phonetic weighting does not alter the relational geometry of the dialect space, nor does it introduce new groupings. It is best understood as a calibration of edit distance, that is an adjustment of the cost function that brings the metric into closer alignment with the gradient structure of phonetic contrast.

Non-edit-based measures impose structurally incompatible assumptions. The failure modes of COS, JAC, JW, and RIV are not random but systematic, and each is traceable to a specific design feature. COS and JAC treat strings as sets or frequency vectors, abstracting away from sequential order; applied to within-word phonetic comparison, this produces geometric distortion through exaggeration, confirmed by their high $STRESS_1$ values. JW and RIV distort in the opposite direction: JW through prefix-weighting that rewards onset matching irrespective of divergence elsewhere; RIV through a positional identity requirement that assigns maximal distance as soon as string alignment breaks down. Both collapse the dialect space into a narrow, poorly differentiated cluster. These are not merely suboptimal choices; they reflect assumptions about phonetic similarity that are structurally mismatched with gradient dialectal variation.

Methodological recommendations. (i) Edit-based measures should be preferred over distributional and overlap-based measures for phonetically transcribed data. Among edit-based measures, HS is to be preferred in case that narrow phonetic transcriptions are available, as it delivers higher resolution and faster stabilization at no cost to topological validity. The PAM variant is additionally appropriate in research designs where coarticulatory reduction constitutes noise rather than signal. (ii) The efficiency advantage of HS—stable estimates after approximately 200 tokens—suggests that investment in narrow phonetic transcription pays dividends in statistical efficiency: phonetically informed transcription combined with phonetically informed distance metrics reduces the data requirements for reliable dialectometric estimates. (iii) The study demonstrates that correlational evaluation alone is an insufficient criterion for measure selection; it must be complemented by distributional, stabilization, and geometric analyses, as moderate correlations may mask topological agreement, and high correlations may mask geometric distortion. The four-step design of this study demonstrates that no single criterion suffices: each step reveals a dimension of measure behavior invisible to the others.

5. Conclusion

This study has shown that explicit phonetic weighting provides a principled refinement of established distance metrics in dialectometry. By incorporating articulatory proximity into string distance computation, the Herrgen-Schmidt measure achieves clearer differentiation, earlier stabilization, and higher linguistic plausibility than unweighted or distributional approaches, without altering the underlying structural relations among dialect varieties.

Differences among measures are primarily scalar rather than structural. Edit-based procedures form a stable topological core, while overlap- and inventory-based metrics systematically distort scale through expansion or compression. Phonetic weighting thus bridges empirical adequacy and methodological rigor, extracting more information from less data.

Beyond its empirical findings, this study advances a broader methodological argument: the effectiveness of distance-based modeling depends not on algorithmic innovation but on the phonetic realism of the underlying cost assumptions. Future work should examine whether these advantages generalize to cross-dialectal comparisons, other language families, and spontaneous continuous speech. These are contexts, in which coarticulatory complexity and typological diversity may challenge the articulatory assumptions underlying the Herrgen-Schmidt weighting scheme.

6. Acknowledgements

This research is supported by the Academy of Science and Literature Mainz (grant REDE 0404), the German Research Foundation (DFG, grant 432304149) and the German Federal Ministry of Research, Technology and Space (BMFTR, grant

AnDy 16DKWN007). I am grateful to five anonymous reviewers for their valuable comments and discussion.

7. Ethical Considerations

The data used in this study originate from the REDE project and were collected in full compliance with applicable ethical and legal standards. The dataset does not contain personally identifiable information; all data were anonymized prior to analysis. Therefore, no personal data were processed within the meaning of relevant data protection regulations. The REDE project obtained the required ethical approvals before data collection. Where applicable, informed consent was secured in accordance with institutional and legal requirements. Data use complies with the project's defined access and licensing conditions.

8. Bibliographical References

Antonio Almeida and Angelika Braun. 1986. "Richtig" und "Falsch" in phonetischer Transkription. Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, pages 158–172.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Magnus Breder Birkenes and Jürg Fleischer. 2021. Syntactic vs. phonological areas: A quantitative perspective on hessian dialects. *Journal of Linguistic Geography*, 9(2):142–161.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Hans Goebl. 1982. *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Verlag der Österreichischen Akademie der Wissenschaften.

Wilbert Heeringa and Angelika Braun. 2003. The use of the Almeida-Braun system in the measurement of dutch dialect distances. *Computers and the Humanities*, 37:257–271.

Joachim Herrgen, Alfred Lameli, Stefan Rabanus, and Jürgen Erich Schmidt. 2001. Dialektalität als phonetische Distanz. ein Verfahren zur Messung standarddivergenter Sprechformen. Technical report, Forschungsinstitut für deutsche Sprache – Deutscher Sprachatlas, Philipps-Universität Marburg.

Joachim Herrgen and Jürgen Erich Schmidt. 1989. Dialektalitätsareale und Dialektabbau. In *Dialektgeographie und Dialektologie. Günter Bellmann zum 60. Geburtstag von seinen Schülern und Freunden*, pages 304–346, Marburg. Elwert.

Paul Jaccard. 1902. Lois de distribution florale dans la zone alpine. *Bull Soc Vaudoise Sci Nat*, 38:69–130.

Roland Kehrein. 2012. *Regionalsprachliche Spektren im Raum. Zur linguistischen Struktur der Vertikale*. Steiner.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Walter de Gruyter.

Alfred Lameli. 2025. *Gesprochenes Deutsch in den Regionen: Eine Standortbestimmung für die Bundesrepublik Deutschland*, pages 51–80. De Gruyter, Berlin, Boston.

Alfred Lameli, Elvira Glaser, and Philipp Stöckle. 2020. Drawing areal information from a corpus of noisy dialect data. *Journal of Linguistic Geography*, 8(1):31–48.

Vladimir I Levenshtein. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmissions*, 1(1):8–17.

Salome Lipfert. 2024. Arbeitsschritte und Messmethode des Programms phonetische Abstandsmessung (PAM). Technical report, Forschungszentrum Deutscher Sprachatlas, Projekt "Regionalsprache.de", AG Empirie.

Bettina Meinck and Juliane Limper. 2013. Anleitung zur phonetischen Abstandsmessung (PAM). Technical report, Forschungszentrum Deutscher Sprachatlas, Projekt "Regionalsprache.de", AG Empirie.

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

T. Portele, J. Krämer, and D. Stock. 1995. Symbolverarbeitung im sprachsynthesesystem hadifix. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 1995*, pages 97–104. Dresden: Techn. Univ., Inst. für Technische Akustik.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer. 2020ff. Regionalsprache.de (REDE III). Forschungsplattform zu den modernen Regionalsprachen des Deutschen. *Forschungszentrum Deutscher Sprachatlas, Marburg*.

Marco René Spruit. 2005. Classifying dutch dialects using a syntactic measure: The perceptual daan and blok dialect map revisited. *Linguistics in the Netherlands*, 22(1):179–190.

Christiane Steiner. 1994. *Sprachvariation in Mainz: quantitative und qualitative Analysen*. Steiner.

M.P.J. van der Loo. 2014. The stringdist package for approximate string matching. *The R Journal*, 6:111–122.

Lars Vorberger. 2019. *Regionalsprache in Hessen. Eine Untersuchung zu Sprachvariation und Sprachwandel im mittleren und südlichen Hessen*. Steiner.

Wolfgang Wahlster. 2000. *Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg.

Georg Wenker. 2013. *Schriften zum Sprachatlas des Deutschen Reichs: Gesamtausgabe*. Olms.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.

Achim Zeileis and Gabor Grothendieck. 2005. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.