

Article

The Structure of Lexical Variation in Austrian Varieties: A Quantitative Analysis

Philipp Stöckle ¹  and Markus Pluschkovits ^{2,*} 

¹ Austrian Centre for Digital Humanities, Austrian Academie of Sciences, 1010 Vienna, Austria; philipp.stoeckle@oeaw.ac.at

² Institute for German Studies, University of Vienna, 1010 Vienna, Austria

* Correspondence: markus.pluschkovits@univie.ac.at

Abstract

This paper investigates variation on the level of lexis in Austrian German across registers with an apparent-time approach, drawing on data from a large-scale online survey of lexical variation conducted across Austria between 2021 and 2023, comprising 45 lexical concepts elicited from over 1900 respondents. We utilize Shannon entropy as a proxy to model lexical diversity across different registers in young and old speakers and investigate the influence of area, age group, phenomenon and register. We use ANOVA, multi-dimensional scaling and generalized additive modelling and find that the lexical concept under investigation exerts by far the strongest influence on lexical diversity, followed by a small but significant regional effect—one that, however, disappears when modelled at the level of individual localities rather than aggregated regions. While area influences lexical diversity, we do not find evidence of clearly distinguished lexical isoglosses in our data but rather broad east–west distinctions. While we find more diversity in older speakers in their nonstandard, we observe more pronounced east–west distinctions in younger speakers. We interpret this as a levelling of standard and nonstandard forms in younger speakers but along broader regional lines. This is corroborated by the reduced diversity in younger speakers in nonstandard varieties but slightly higher diversity in standard registers, pointing towards a levelling process.

Keywords: lexical variation; apparent-time; Shannon entropy; multi-dimensional scaling; generalized additive model; areal variation

1. Introduction

Lexicon—and lexical variation in particular—occupies a distinctive position within the overall architecture of language. Traditionally, lexical phenomena have been documented in dictionaries that target diverse segments of a language’s vocabulary, and lexical variation has long played a central role in dialect atlases, alongside phonology and morphology. Compared to other linguistic levels, lexical variation exhibits several distinctive characteristics. On the one hand, the lexicon of a language in its full extent is difficult to delimit and grasp empirically, which means that it can only ever be investigated in selected excerpts. On the other hand, lexical phenomena are particularly accessible to non-linguists (cf. Kleene, 2020, Figure 25; Stöckle, 2014, pp. 446–452), making lexical variation a frequent topic of public discourse and metalinguistic reflection. This dual status—methodologically challenging yet socially salient—sets lexical variation apart from other domains of variationist research.



Academic Editor: John Nerbonne

Received: 3 February 2026

Revised: 22 June 2026

Accepted: 26 June 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In German-language variationist linguistics, research has traditionally focused less on lexical variation than on other system levels, despite the long-standing documentary traditions of dictionaries and linguistic atlases (cf. Schmidt et al., 2019, pp. 43–45). Lexical variation has thus often been treated somewhat marginally in quantitative variationist studies. In recent years, however, there has been a noticeable increase in empirical research on lexical variation, particularly in the context of the Special Research Program “German in Austria. Variation–Contact–Perception” (‘Deutsch in Österreich. Variation–Kontakt–Perzeption’)¹ (e.g., Lenz et al., 2021, 2023, in press). The present contribution positions itself within this emerging line of work.

The empirical basis of this quantitative investigation consists of data from the first round of the Atlas of Lexical Variation in Austria in the 21st Century (LexAT21) survey (cf. Lenz et al., in press). The dataset comprises 45 variables corresponding to lexical concepts, collected via an online survey conducted across Austria between June 2021 and May 2023. Lexical variation was elicited using linguistic picture stimuli (see also Section 3.1). In addition, several extra-linguistic variables were collected, including the place of participation and respondents’ age. Different registers were also taken into account, allowing for the analysis of register-related variation.

The aim of this contribution is to provide a quantitative analysis of these data, with particular focus on the following research questions:

Which factors exert the strongest influence on lexical variation?

Are there systematic differences by age and register that could, in an apparent-time perspective, be interpreted as evidence of ongoing change?

Do the quantitative analyses yield spatial patterns that can be interpreted in dialect-geographical terms?

Following a brief overview of related research (Section 2), the dataset is introduced in more detail (Section 3). Section 4 presents the quantitative analyses, employing a range of methods, including entropy as a measure of variation, analysis of variance, multidimensional scaling (MDS), and generalized additive modelling (GAM). Finally, Section 5 summarizes the results and discusses them in relation to the research questions outlined above.

2. Related Research

Naturally, variation constitutes one of the central concepts in variationist linguistics. Most existing studies, however, focus on the nature of linguistic variants, their relationship to extra-linguistic factors, their social evaluation by speakers, or their relevance for spatial patterns. With regard to variation on the lexical level, there is a tradition of investigating spatial patterns in lexical variation by means of atlases, initially stemming from a dialectological interest.

An early example of such dialectal atlas work in German is represented by the *Deutsche Wortatlas* by Mitzka and Schmitt (1951–1980), a dialectal atlas focused on lexical variation. The *Deutsche Wortatlas* saw itself in the tradition of the seminal *Sprachatlas des Deutschen Reiches* by Georg Wenker. The *Deutsche Wortatlas* is based on over 48,000 linguistic questionnaires covering 200 linguistic variables collected between 1939 and 1942 (see Wiegand & Harras, 1971, p. 7). It aimed at describing dialectal variation and utilized teachers to record the dialectal variants for given localities—its interest was therefore not so much in the speech of individuals, but in variants considered representative for the dialect of a given location. The maps in the *Deutsche Wortatlas* record not just lexical but also spelling variants, and offer point symbols for each location but also isoglosses, indicating where certain variants are dominant (although those areas might be far from homogenous). The *Deutsche Wortatlas* remains an important reference work to this day.

In contrast to the *Deutsche Wortatlas*, the *Wortatlas der deutschen Umgangssprachen* (Eichhoff, 1977–2000) targeted variation not in dialect or standard, but rather within “everyday language”. Smaller in scale than the *Deutsche Wortatlas*, the first round of surveys in the 1970s encompassed 402 localities (see Eichhoff, 1997, p. 187) followed by supplementary surveys. Overall, these resulted in four volumes of maps investigating primarily lexical variation but also including material on phonologic, morphologic and syntactic variation. In contrast to the *Deutsche Wortatlas*, the aim of the *Wortatlas der deutschen Umgangssprachen* was to record common language and not exclusively base dialect. This was due to the fact that such base dialect is conceived as one extreme of a spectrum which is not everyone’s common language—at least not for younger, more urban speakers, which Eichhoff considered to be users of such “everyday language”. Eichhoff assumed that e.g., for Austria, dialect was still the more commonly spoken and usual register (Eichhoff, 1997, p. 189). Additionally, there was the motivation to record German varieties as they are actually spoken for teaching purposes (see Eichhoff, 1997, FN 13).

The *Atlas zur deutschen Alltagssprache* (‘Atlas of Colloquial German’, henceforth AdA, Elspaß & Möller, 2003) similarly targets not just exclusively base dialect but rather varieties “one would normally hear in your town” (Möller, 2021, p. 198). This long-standing project, at the time of writing in its 14th survey round, has published over 550 maps (see Möller, 2021, p. 197) mostly documenting lexical variation but also maps on different phonetic and grammatical phenomena. Similarly to both the *Deutsche Wortatlas* and the *Wortatlas der deutschen Umgangssprachen*, the aim is to document the language use in a locality instead of an individual. The *Atlas zur deutschen Alltagssprache* covers different registers, depending on what is considered typical for a given location. With over 30.000 answers for the latest completed survey round, the *Atlas zur deutschen Alltagssprache* is extremely well sourced and covers Germany, Austria and the German-speaking part of Switzerland.

In addition to these research projects, there has been somewhat of a renewed interest in lexical variation in the German-speaking world, as evidenced by, e.g., the recently published *Dialektatlas* (Leemann et al., 2025) covering dialectal change in German-speaking Switzerland. The *Variantenwörterbuch des Deutschen* (Ammon et al., 2016) and the *Zentrum für Digitale Lexikographie der deutschen Sprache*² indicate that this interest is also no longer just confined to spoken or nonstandard language. LexAT21 is very much part of this current interest in lexical variation, although it appears to be the first focusing not just on variation between localities, but between registers of an individual. LexAT21 aims to represent not just what might be considered typical for a location by speakers, but rather on the registers of the speakers in those localities themselves.

This degree of variation itself, that is, the number and distribution of variants available within a given speech community, register, or individual repertoire, has as of yet received comparatively little systematic attention.

Yet the extent of variation can be highly informative with respect to language dynamics. A high degree of variation may indicate ongoing change, for instance when several variants of a variable are in competition. At the same time, variation need not always be transitional: particularly in the lexical domain, stable variation may reflect a differentiated and semantically rich inventory of designations for specific lexical concepts. Conversely, low degrees of variation can be interpreted as a sign of linguistic levelling, where one variant has become dominant at the expense of others, as may occur in processes of dialect levelling or increased standardization.

In research on German and its varieties, such a degree of variation has only been addressed in a limited number of studies. In a study on variation in Swiss German (morpho-)syntax, Stöckle (2016) introduces the “Variation Index”, which operationalizes

variation as the degree of agreement among speakers at a given location with respect to the dominant variant. Lower agreement is interpreted as higher variation. Building on this approach, [Wirtz et al. \(2025\)](#) propose the concept of “Variation Intensity”, which extends the measure by taking into account not only the frequency of the dominant variant, but the full distribution of all variants and their relative frequencies. Their study is based on 15 variables drawn from the AdA, comprising five lexical, five morphological, and five phonetic variables. In addition, the authors incorporate geographic location and an age distinction, which allows them to investigate change in variation intensity (CVI) in an apparent-time perspective.

A different approach to measuring variation is pursued by [Stöckle et al. \(in press\)](#), who employ Shannon entropy to systematically compare variation in historical lexical data from the *Dictionary of Bavarian Dialects in Austria* (Wörterbuch der bairischen Mundarten in Österreich, WBÖ; cf. [Stöckle, 2021](#)) with contemporary data from the LexAT21 project (cf. [Lenz et al., in press](#)), which also forms the empirical basis of the present study. While Variation Intensity can be interpreted as the probability that two randomly selected responses represent different variants, Shannon entropy quantifies the average informational content of a distribution and thus captures how evenly variants are distributed. Shannon entropy is also used as a central measure in the present study; its methodological implementation is discussed in more detail in Section 4.1.

Beyond the degree of variation, the present study also investigates the spatial patterns that emerge when lexical variation is quantified. This line of research has a long and well-established tradition under the umbrella term dialectometry, encompassing a wide range of quantitative methods for identifying and visualizing spatial structures in linguistic data.

Several dialectometric studies have explicitly addressed lexical variation in nonstandard varieties of German. For Swiss German, [Scherrer and Stöckle \(2016\)](#) conduct a dialectometric analysis based on atlas data, comparing lexical, phonological, and morphological variation using data from the *Linguistic Atlas of German-Speaking Switzerland* (*Sprachatlas der deutschen Schweiz*), and syntactic variation based on the *Syntactic Atlas of German-Speaking Switzerland*. To identify spatial structures, they apply hierarchical cluster analysis to distance matrices computed using the Jaccard distance measure. Their comparison of the resulting spatial patterns shows substantial similarities between lexicon, phonology, and morphology, whereas syntax exhibits more divergent structures. Importantly, they also demonstrate that lexical variation yields clearly delineated spatial patterns that are readily interpretable in dialect-geographical terms.

The identification of spatial patterns in lexical data is likewise central to the work of [Pickl \(2013\)](#) and [Pröll \(2015\)](#), whose analyses are based on data from the *Linguistic Atlas of Bavarian Swabia* (*Sprachatlas von Bayerisch-Schwaben*). Their methodological approach involves an initial smoothing of individual variables or maps using kernel density estimation (cf. [Rumpf et al., 2009](#)). These smoothed maps are then subjected to factor analysis, with the aim of reducing the overall variation in the dataset to a small number of underlying factors (cf. [Pickl, 2013](#), p. 119). The resulting factors can either be mapped individually or combined into a composite map, in which each location—or, following dialectometric convention, each polygon—is coloured according to the dominant factor and its intensity (cf. [Pröll, 2015](#), p. 193). This procedure yields spatial representations that can likewise be interpreted from a dialect-geographical perspective. Similar to the atlas-based analyses by [Scherrer and Stöckle \(2016\)](#), the studies by [Pickl \(2013\)](#) and [Pröll \(2015\)](#) reveal clearly structured spatial patterns for lexical variation.

Using similar quantitative techniques, [Pickl et al. \(2019\)](#) analyse data from AdA, with a strong emphasis on lexical variables. In addition to an analysis of the entire German-speaking area covered by the AdA, the authors place a particular focus on Aus-

tria. Applying factor analysis with a three-factor solution, they identify a comparatively clear spatial tripartition of Austria. Notably, the boundaries of the resulting linguistic regions largely coincide with the administrative borders of the Austrian federal states, suggesting a close alignment between lexical variation patterns and established regional divisions (cf. [Pickl et al., 2019](#), p. 51).

Another method frequently employed in dialectometry, but so far not applied to lexical variation in German varieties, is multidimensional scaling (MDS) (cf. [Backhaus et al., 2015](#), pp. 350–354). MDS is a dimensionality-reduction technique that, like hierarchical cluster analysis, operates on aggregated distance measures. It projects distances (or similarities) between objects—typically locations—into a low-dimensional space in such a way that their spatial configuration reflects the original distances as closely as possible. In linguistic geography, two or three dimensions are commonly used, the latter being particularly well suited for mapping onto the RGB colour space. The resulting colour values can then be used to visualize locations or polygons on a map, making spatial structures immediately apparent. First introduced to linguistics by [Nerbonne et al. \(1999\)](#) (cf. also [Wieling & Nerbonne, 2015](#), p. 245), MDS has since been applied in numerous dialectometric studies. In the present study, the method is employed to identify and visualize spatial patterns in the lexical data.

A different approach is represented by the use of generalized additive models (GAMs). Such GAMs allow for non-linear relations between predictor and response variable in any number of dimensions as a wiggly surface (or hypersurface, see [Wieling et al., 2014](#), pp. 678–679). Essentially, this means that the influence of geography can be modelled by a combined smooth consisting of latitude and longitude of a locality, meaning that we can utilize coordinates of a locality as a predictor value. GAMs have become more and more common in linguistic research and have been used productively by, e.g., [Wirtz and Pfenninger \(2025\)](#) or [Ko et al. \(2014\)](#).

3. Materials

3.1. The LexAT21 Dataset

The data used in our analysis were collected as part of the atlas of lexical variation in Austria in the 21st century (LexAT21), a project of the Special Research Programme “German in Austria: Variation—Contract—Perception”, funded by the Austrian Science Fund FWF (Grant Number FWF F060). LexAT21 aims to document and research lexical variation of contemporary German as spoken in Austria. The basis for the maps and analysis of LexAT21 is semi-regular online surveys on language use, with the first survey round additionally serving as the data basis for the present study. The data are available online (see [Lenz, 2025](#)).

LexAT21 survey round one was conducted between July 2021 and May 2023. The survey was orchestrated purely online, utilizing LimeSurvey ([LimeSurvey Project Team & Schmitz, 2026](#)). Overall, the first survey round collected 1925 relevant completed surveys (see [Lenz, 2025](#)). Surveys were considered relevant if the respondents stated they were representative speakers of a locality in Austria and if the survey was fully completed. Partially completed surveys are not considered in the analysis. The survey consists of three parts: a short initial questionnaire on demographic data of the informant, the linguistic portion in which 45 linguistic variables were queried and the final section collecting further metadata on the informant.

Within the first part of the survey, the informants specified a location for which their language use was characteristic, and indicated whether or not this location was in Austria. While those informants who chose non-Austrian locations could partake in the survey, their data are not considered. Overall, the first survey round includes data from

over 500 localities in Austria. The distribution of the informants across these locations is however not balanced, with more informants stemming from the east (especially Vienna). Furthermore, informants had to select which glottonym they considered more appropriate for their nonstandard register: they could choose between “Dialekt (Mundart)” (‘Dialect’) or “Umgangssprache (Alltagssprache)” (‘Colloquial Language (Everyday Language)’). The question (“Sprechen Sie selbst einen Dialekt (Mundart), der sich vom Hochdeutschen unterscheidet?” ‘Do you speak a dialect that is different from high German?’) utilized common conceptions of these registers and this self-identification required no linguistic knowledge. The answer to this question had an influence on the linguistic portion of the survey: if informants identified as speakers of dialect, the answer boxes for the nonstandard register in the linguistic portion were labelled as “Dialekt (Umgangssprache)”, otherwise they were labelled as “Umgangssprache (Alltagssprache)”. From survey round two on, informants could also identify as only speaking a standard variety of German.

The linguistic portion constituted the majority of the survey. The informants were asked to name 45 concepts (targeted were 40 nouns, four verbs and one adjective) both in their standard variety and their nonstandard variety, with the option to also list further denotations independent of register. The 45 concepts were chosen based on variables from the *Deutsche Wortatlas*. The name of the nonstandard variety depended on what register they identified with (as detailed above), whereas the name of the standard register was randomly chosen from a pool of three (“Ihr Hochdeutsch”, “Ihr bestes Hochdeutsch” and “Ihr österreichisches Hochdeutsch”, ‘Your High German’, ‘Your Best High German’ and ‘Your Austrian High German’; see [Koppensteiner \(2023\)](#) for a discussion of these terms, which have been collapsed in the analyses to foreground differences between standard and nonstandard varieties more broadly). Each variable was presented as a picture stimulus accompanied by a written stimulus formulated as a question (see Figure 1 for an example).

Wie nennen Sie das abgebildete Kleidungsstück?

Bitte mindestens eine Antwort ausfüllen

In Ihrem Dialekt (Mundart)	In Ihrem besten Hochdeutsch	Verwenden Sie noch weitere Bezeichnungen?

Figure 1. Stimulus for the variable CAP. The verbal stimulus reads “What do you call the garment pictured”, with answer boxes for the “In your dialect”, “in your best high German” and “Do you use any further terms?”.

The picture stimuli were chosen with permissive licenses (CC-0), which allows for replication of the survey. All verbal stimuli were slight variations on the basic question of “how do you call . . .”. For verbs, informants were asked to answer in a full sentence. At least one of the boxes had to be filled in order to progress to the next question.

The final part of the survey was dedicated to further metadata on the informants, with a focus on their prior mobility. Informants were asked to detail their residency history to detect possible influences in their language use and compare the residency data to the locality the informants chose to be representative of, which sometimes showed

significant discrepancies (e.g., informants claiming their language use was representative of a locality they just recently moved to). In a first pre-processing step, those incongruities were resolved.

Further processing of the data was done in Baserow (Baserow, 2026). The initial LimeSurvey results were exported in Baserow and then manually classified according to lemma. This classification—and not the raw strings—is the basis of our present analysis, as we focus on lexical variation and not the phonetic variation that is reflected in spelling variations of identical lemma. The entire data we base our analysis on are available online (see Lenz, 2025).

3.2. Data Preparation

We prepared the data by classifying all responses by speaker age into one of three groups: the group of young speakers, including speakers below age 31; the group of middle-aged speakers, which we excluded from the analysis and consisted of speakers between the ages of 31 and 50; and speakers over 60, which were classified as old. Table 1 indicates the number of speakers and responses (i.e., lexical variants) per group:

Table 1. Distribution of speakers and responses across age groups.

Age Group	Speakers	Responses
Young	660	52,875
Middle	709	58,371
Old	554	45,102

To achieve greater discriminatory power with respect to age distribution and to ensure a sufficient age difference of at least one generation while maintaining adequately sized groups, the subsequent analyses compare only the young and old speaker groups, excluding the middle age category.

Register. The LexAT21 data originally distinguishes five register types: dialect (“Dialekt (Mundart)”), colloquial speech (“Umgangssprache (Alltagssprache)”), and three standard-oriented varieties (Your High German “Ihr Hochdeutsch”, Your Austrian High German “Ihr österreichisches Hochdeutsch”, Your best High German “Ihr bestes Hochdeutsch”). Since our primary interest lies in distinguishing between standard and nonstandard registers, we collapsed these into two categories: standard (comprising all three High German variants) and nonstandard (dialect and colloquial speech), following the classification already established in the LexAT21 framework.

Region. Several options exist for geographical classification. The original data record the survey location for each data point. However, the distribution of respondents across locations is highly uneven, with many locations represented by only one or very few speakers. This poses a problem for quantitative comparison, as an age-group analysis at the location level would require excluding numerous sites due to insufficient or incomparable data. We therefore adopted the network of so-called major regions (*Großregionen*) used in the *Dictionary of Bavarian Dialects in Austria* (WBÖ) (cf. Stöckle, 2021, p. 18). This system comprises 29 regions covering the Bavarian-speaking areas of Austria and has served as the primary geographical classification scheme in both lexicographic work and numerous analyses based on WBÖ data (e.g., Stöckle et al., 2021; Stöckle & Vergeiner, 2025). Since Vorarlberg is the only Alemannic-speaking federal state in Austria and thus falls outside the WBÖ’s coverage area, it was added as an additional major region, yielding a total of 30 regions for geographical classification. Responses from individual survey locations were assigned to their respective major regions, reducing the 539 distinct locations to 30 aggregated regions for spatial analysis. Additionally, we pursue a more granular

approach utilizing localities instead of major regions in Section 4.3 by utilizing generalized additive models. For this analysis, each locality was assigned geocoordinates utilizing the API of Google Maps (v3). We manually corrected misassigned localities.

Phenomenon. Finally, the phenomenon itself was included as the sole linguistic factor. As noted above, the LexAT21 dataset contains 45 phenomena corresponding to lexical concepts (e.g., ROSE HIP, GOOSEBERRY, TO SMOKE, etc.), each of these representing a linguistic variable as almost every concept exhibits relevant variation in its denomination by the informants.

All data preparation as well as subsequent analysis was undertaken in R (version 4.3.2) (R Core Team, 2024).

4. Results

Various quantitative methods have become established in variationist linguistics for empirically identifying factors that influence language variation. However, most research has focused on the relationship between different extra-linguistic factors or on the relationship between extra-linguistic factors and linguistic variables. Measuring the extent of variation itself and relating it to other variables has been comparatively rarely addressed.

For our analyses, we therefore employ Shannon entropy as a measure of lexical diversity. The resulting entropy values then serve as the central dependent variable in a subsequent variance decomposition, which examines how entropy is distributed across the factors age, register, region, and phenomenon (Section 4.1). To identify spatial structures in the data, we use multidimensional scaling (MDS), a dimension-reduction technique frequently employed in dialectometry that—similar to hierarchical cluster analysis—operates on a distance matrix containing aggregated pairwise linguistic distances between survey locations (Section 4.2). Finally, we employ generalized additive models (GAMs) to model the spatial distribution of lexical variation while accounting for potential non-linear relationships between geographic coordinates and entropy values (Section 4.3).

4.1. Entropy-Based Analysis of Lexical Variation

One of the central questions of the present study is to what extent lexical variation is structured geographically and how it differs across speaker subgroups defined by socio-demographic parameters. Addressing this question requires a statistical measure that adequately captures variation. Following Stöckle et al. (in press), we conceptualize lexical variation as comprising both the number of variants attested for a given phenomenon and their relative frequencies. A suitable measure for this purpose is Shannon entropy (see Shannon, 1948; Cover & Thomas, 2006, pp. 13–14), originally developed in information theory, which quantifies the average information content associated with each act of selecting a signal from an information source. Entropy can be interpreted as a measure of uncertainty or diversity within a distribution and reflects the average number of binary decisions required to identify a symbol from a set of alternatives. It is conventionally denoted by the Greek letter eta (H) and measured in bits.

Applied to lexical variation, this can be illustrated as follows: if a given lexical concept is realized by only a single variant, entropy is $H = 0$ bits, since no decision is required to identify the form. If exactly two variants occur with equal frequency—analogue to a fair coin toss—entropy amounts to $H = 1$ bit. More generally, higher entropy values indicate a greater number of variants that are more evenly distributed, whereas lower entropy values point to the dominance of a single form. Entropy thus provides a principled way to quantify both the richness and balance of lexical variation. What entropy does not capture, however, is which variants occur—and this distinction is consequential for interpreting the analyses that follow. A location where 90% of speakers use *Hagebutte* and 10% use *Hetscherl* yields

the same entropy value as one where these proportions are reversed. The measure is, by design, indifferent to the concrete identity of variants: it quantifies the degree of uncertainty or competition within a distribution, not the direction of that competition. This sets entropy apart from the distance-based measures used in the MDS analysis below (Section 4.2), where the identity of shared or divergent variants directly determines how linguistically close two locations are judged to be. The two approaches are therefore complementary rather than redundant: entropy addresses the degree of variation within a group or location, while MDS addresses the spatial patterning of variant choices across locations.

To analyse lexical variation across speaker groups, we employed analysis of variance (ANOVA) to determine which factors contribute significantly to differences in lexical entropy. The main-effects model included age group (young vs. old), register (nonstandard vs. standard), major region, and lexical phenomenon as predictors. We used Type-II sums of squares, which are appropriate for unbalanced designs without interaction terms (cf. Langsrud, 2003).

To assess whether the effect of age differs across registers, we fitted a separate model including age group, register, and their interaction. For this model, we relied on Type-III sums of squares (cf. Backhaus et al., 2023, p. 202), which test each effect while controlling for all other effects, including the interaction term. This approach ensures correct estimation of main effects in the presence of a significant interaction. Effect sizes were quantified using partial eta-squared (η^2) with 95% confidence intervals. Following Cohen’s (1988) conventions, effect sizes were interpreted as negligible ($\eta^2 < 0.01$), small ($0.01 \leq \eta^2 < 0.06$), medium ($0.06 \leq \eta^2 < 0.14$), or large ($\eta^2 \geq 0.14$).

All analyses were conducted in R (R Core Team, 2024). ANOVAs were computed using the *car* package (Fox & Weisberg, 2019), effect sizes with the *effectsize* package (Ben-Shachar et al., 2020), MDS using the *MASS* package (Venables & Ripley, 2002) and GAMs using *mgcv* (Wood, 2025). Maps were created with *ggplot2* (Wickham, 2016) and *sf* (Pebesma, 2018).

We begin by reporting the main effects of the ANOVA. The Type-II analysis revealed significant effects for all predictors on normalized Shannon entropy (Table 2). Phenomenon explained the largest share of variance ($F(44, 2588) = 67.23, p < 0.001, \eta^2 = 0.53$), confirming that different lexical concepts exhibit inherently different patterns of variation. Region also showed a significant, though comparatively modest effect ($F(18, 2588) = 3.65, p < 0.001, \eta^2 = 0.03$), reflecting the dialectal structure of Austria. Both sociolinguistic variables of primary interest were statistically significant, but their effect sizes were negligible according to conventional benchmarks (Cohen, 1988): age group ($F(1, 2588) = 20.85, p < 0.001, \eta^2 = 0.008$) and register ($F(1, 2588) = 4.48, p = 0.034, \eta^2 = 0.002$). This discrepancy between statistical significance and practical effect size is unsurprising given the large sample size ($n \approx 2600$), which provides sufficient power to detect even very small effects; statistical significance in this context should therefore not be equated with practical importance.

Table 2. Type-II ANOVA results for main effects model predicting normalized Shannon entropy.

Factor	df	F	p	η^2	Interpretation
Phenomenon	44	67.23	<0.001	0.53	large
Region	18	3.65	<0.001	0.03	small
Age	1	20.85	<0.001	0.008	negligible
Register	1	4.48	0.034	0.002	negligible
Residuals	2588				

η^2 = partial eta-squared. Effect size interpretation follows Cohen (1988): negligible (<0.01), small (0.01–0.06), medium (0.06–0.14), large (≥ 0.14).

At the same time, the small effect sizes warrant interpretation in context rather than dismissal. In a model dominated by phenomenon-level variation ($\eta^2 = 0.53$), the sociolinguistic variables age and register cannot be expected to account for large portions of variance, since most of the variation in entropy is attributable to the lexical concept itself rather than to speaker characteristics. We therefore treat the age and register effects reported here as modest but consistent directional tendencies—younger speakers exhibited systematically lower entropy than older speakers, and the standard register showed lower entropy than nonstandard varieties—rather than as evidence of strong sociolinguistic differentiation in their own right.

To examine whether the age effect operates differently across registers, we fitted a separate model including the interaction between age and register. The Type-III ANOVA confirmed significant main effects for age ($F(1, 2649) = 10.58, p = 0.001$) and register ($F(1, 2649) = 8.86, p = 0.003$), as well as a significant interaction ($F(1, 2649) = 7.07, p = 0.008$; Table 3). As with the main effects model, the effect sizes are small ($\eta^2 < 0.01$ for all effects). The total variance explained by this model is correspondingly low ($R^2 = 0.005$), with 99.5% of the variance in entropy remaining unexplained. This is expected, however, since phenomenon and region—excluded from this model to avoid aliased coefficients—account for the vast majority of variance. This model is not intended as a comprehensive predictive account of entropy, nor should the interaction be read as indicating a sociolinguistically large effect; its purpose is the narrower one of establishing whether age and register interact at all once the dominant influence of phenomenon and region is set aside. On this narrower question, the interaction is statistically detectable ($p = 0.008$), and we treat it in what follows as a reliable, if modest, signal of register-specific age differentiation rather than as a substantial driver of variation in its own right.

Table 3. Type-III ANOVA results for the Age $\times$ Register interaction model.

Factor	df	F	p	η^2
Age	1	10.58	0.001	0.004
Register	1	8.86	0.003	0.003
Age $\times$ Register	1	7.07	0.008	0.003
Residuals	2649			

The descriptive statistics (Table 4) illustrate the nature of this interaction. In the nonstandard register, older speakers show higher entropy ($M = 0.56$) than younger speakers ($M = 0.49$), a difference of $\Delta = 0.07$. In the standard register, however, older and younger speakers show comparable entropy levels ($M = 0.49$ vs. $M = 0.50$), with virtually no age difference ($\Delta \approx 0.01$). This pattern indicates that the age effect on lexical diversity is concentrated in the nonstandard register, while the standard register remains stable across generations.

Table 4. Descriptive statistics: mean normalized entropy by age group and register.

Age	Register	M	SD	n
old	Nonstandard	0.56	0.34	664
old	Standard	0.49	0.35	598
young	Nonstandard	0.49	0.35	785
young	Standard	0.50	0.34	606

M = mean normalized Shannon entropy; SD = standard deviation; n = number of observations.

This interaction pattern is visualized in Figure 2, which shows the mean entropy values for each combination of age group and register. The non-parallel lines clearly illustrate

the differential age effect: while entropy decreases substantially from older to younger speakers in the nonstandard register, it remains virtually stable in the standard register.

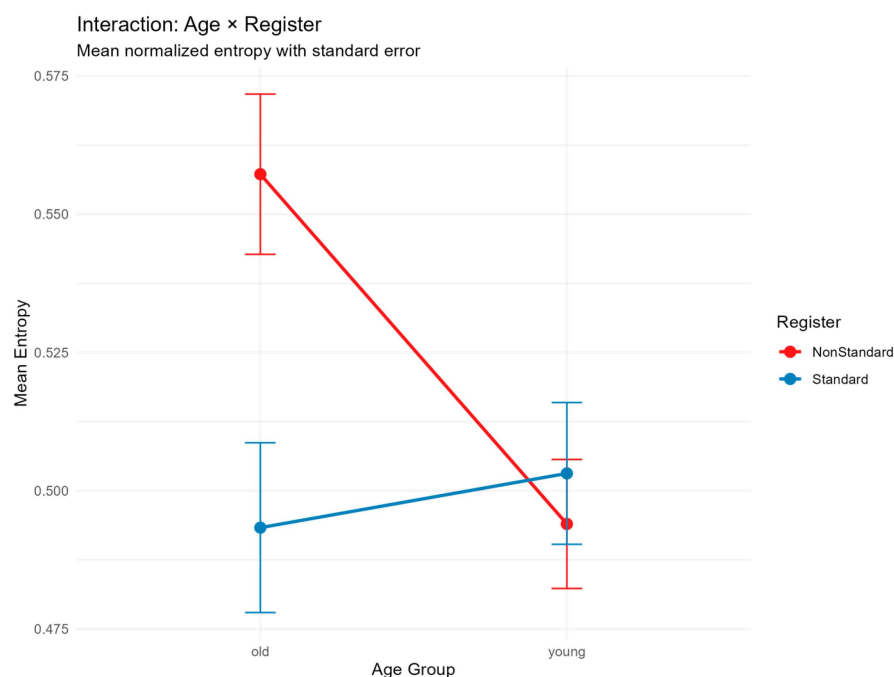


Figure 2. Interaction between age group and register on normalized Shannon entropy. Error bars represent standard errors.

The diverging pattern indicates that the age effect on lexical diversity is register-dependent: younger speakers show reduced entropy (i.e., less variation) in the nonstandard register, but not in the standard register.

The significant interaction can be interpreted from two complementary perspectives, each highlighting a different aspect of the ongoing language change. The first perspective focuses on lexical change within the nonstandard register. Table 5 presents the phenomena with the largest entropy differences between older and younger speakers in dialect and colloquial speech. Positive values indicate that older speakers show higher entropy (more lexical diversity), suggesting levelling among younger speakers.

Table 5. Top 10 phenomena with largest age effect in the nonstandard register.

Phenomenon	Entropy (Old)	Entropy (Young)	Difference
ROSE HIP	0.82	0.26	+0.56
GOOSEBERRY	0.66	0.31	+0.35
CAP	0.63	0.32	+0.31
CARROT	0.59	0.30	+0.29
MOLE	0.68	0.40	+0.28
GRANDFATHER	0.60	0.34	+0.26
GRANDMOTHER	0.59	0.35	+0.24
SPRING	0.73	0.50	+0.23
SHOELACE	0.50	0.72	−0.22
TOOTHACHE	0.23	0.45	−0.22

Difference = entropy difference (old – young). Positive values indicate reduced variation among younger speakers; negative values indicate lower variation among older speakers.

Lexical concepts such as ROSE HIP, GOOSEBERRY, CAP and CARROT exhibit particularly strong levelling effects. Older speakers maintain multiple competing nonstandard

variants, whereas younger speakers converge on a markedly smaller set of forms. This pattern is illustrated below using the concepts ROSE HIP and GOOSEBERRY.

As shown in Figure 3, *Hagebutte*—which corresponds to the Standard German form—is the dominant variant in both age groups (50.4% in the older and 94.0% in the younger group). Among younger speakers, however, alternative forms such as *Hetschepetsch* or *Hetscherl* occur only sporadically and remain marginal in comparison to the dominant variant. In contrast, the older speaker group not only displays a larger number of variants but also a more even distribution among them, resulting in substantially higher entropy values.

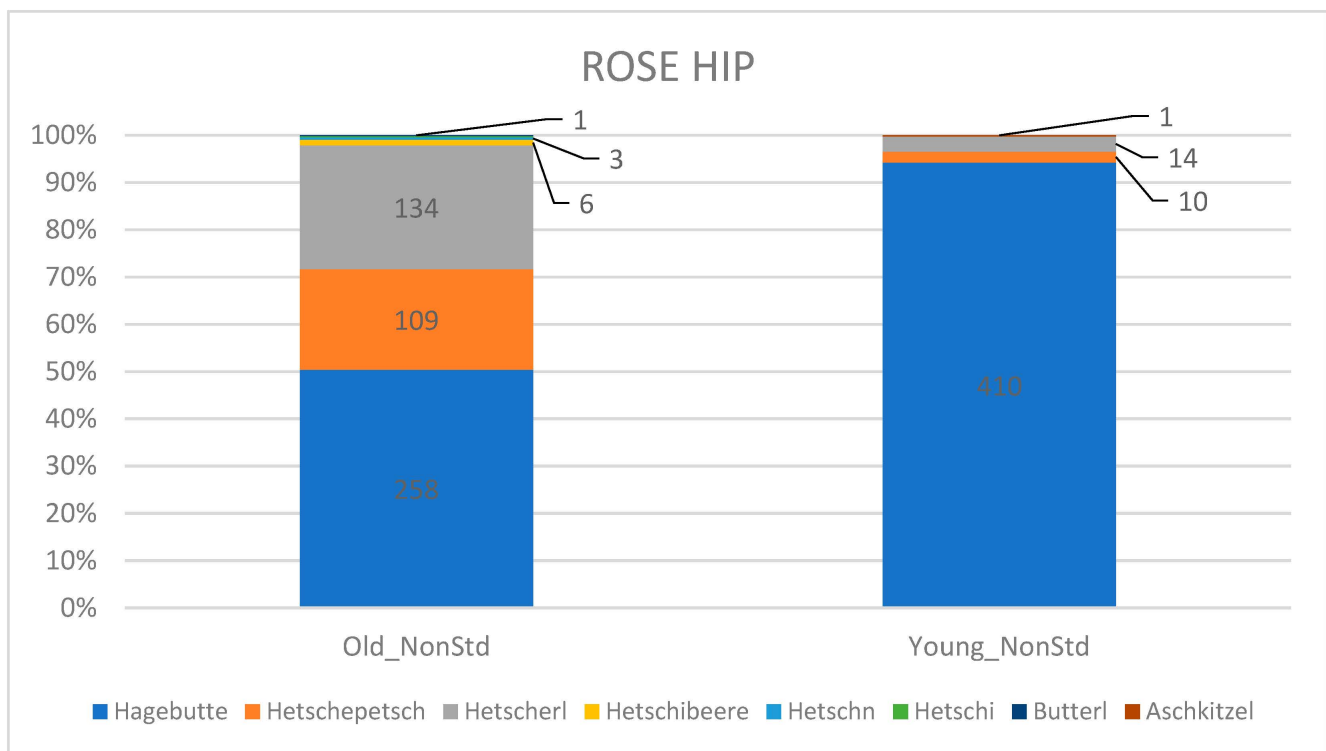


Figure 3. Lexical realizations for ROSE HIP in older and younger speakers (for nonstandard register).

A comparable pattern is observed for GOOSEBERRY (Figure 4). Although *Stachelbeere* is the dominant variant in both age groups, its relative frequency is substantially higher among younger speakers (93.2%) than among older speakers (70.0%). A similar age-related contrast can be seen for the second most frequent variant *Ogrosl*, a nonstandard lexeme that accounts for 30.1% of the responses in the older group but only 6.2% in the younger group. In addition, the older speakers exhibit a further variant, *Eiterpatzn* (three attestations, 0.6%), which is entirely absent from the younger group.

However, not all phenomena follow this levelling trajectory. As shown in Table 5, two concepts—SHOELACE and TOOTHACHE—exhibit the opposite pattern, with younger speakers displaying *higher* entropy than older speakers in the nonstandard register. This reversed pattern may reflect different dynamics: for TOOTHACHE, for instance, younger speakers may be adopting a broader range of informal or expressive variants alongside the standard form, rather than converging on a single dominant form. Alternatively, such cases may point to concepts undergoing active lexical innovation or borrowing in younger registers. These counterexamples underscore that lexical change in the nonstandard register is not uniformly directional: while levelling towards dominant—often standard-proximate—forms is the predominant pattern, some concepts show evidence of ongoing diversification. A more detailed account of these diverging cases remains a task for future research.

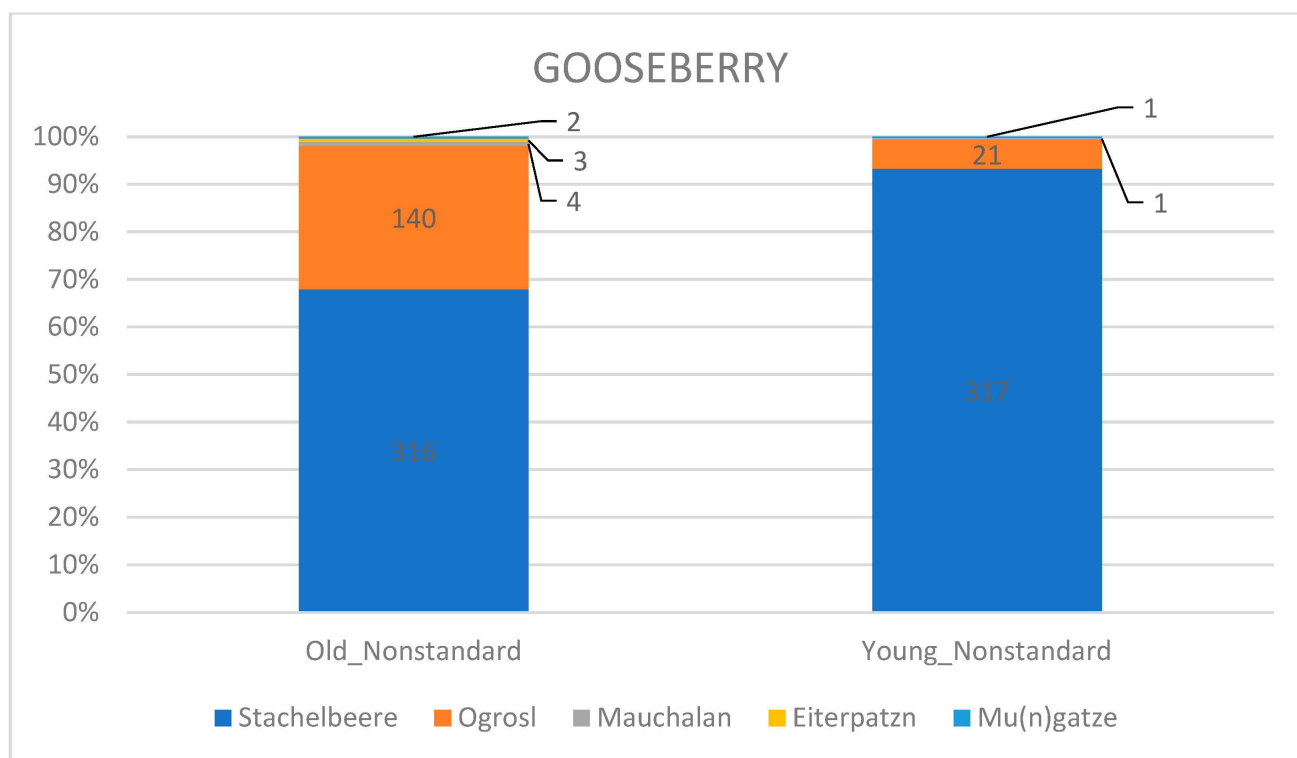


Figure 4. Lexical realizations for GOOSEBERRY in older and younger speakers (for nonstandard register).

The second analytical perspective focuses on the extent to which older speakers differentiate between registers. Table 6 lists the lexical concepts showing the largest entropy differences between nonstandard and standard usage among older speakers. Two distinct patterns can be identified. For concepts such as ROSE HIP, GOOSEBERRY, TO SMOKE, and MOLE, older speakers exhibit substantially greater variation in the nonstandard register. For other concepts, including HORSERADISH and PRUNE PLUM, the pattern is reversed, with higher entropy values in the standard register, as there is no competing nonstandard form (*Kren* is almost exclusively used for HORSERADISH, *Zwetschke* for PRUNE PLUM) but at least one competing variant in the standard (*Meerrettich* in the case of HORSERADISH, *Pflaume* for PRUNE PLUM).

Table 6. Top 10 phenomena with largest register effect among older speakers.

Phenomenon	Entropy (NonStd)	Entropy (Std)	Difference
ROSE HIP	0.82	0.02	+0.81
GOOSEBERRY	0.66	0.01	+0.65
TO SMOKE	0.66	0.02	+0.64
MOLE	0.68	0.07	+0.62
HORSERADISH	0.14	0.62	−0.48
WATERING CAN	0.52	0.08	+0.44
GOAT	0.76	0.32	+0.44
BAREFOOT	0.79	0.36	+0.43
PRUNE PLUM	0.17	0.58	−0.41
COLD	0.51	0.10	+0.41

Difference = entropy difference (nonstandard – standard). Positive values indicate more variation in nonstandard; negative values indicate more variation in standard.

Since ROSE HIP and GOOSEBERRY were already discussed in the previous sections, the findings for register effects are illustrated using the concepts TO SMOKE and HORSERADISH, which exemplify opposite directions of register-based entropy change.

Figure 5 shows the realizations of the lexical concept TO SMOKE. Although the standard variant *rauchen* is dominant in both registers, its relative frequency is markedly higher in the standard register (99.6%) than in the nonstandard register (70.0%). In the nonstandard register, *rauchen* co-occurs with the dialectal form *tschicken*, which accounts for 23.4% of the responses, as well as a small number of additional minor variants. By contrast, usage in the standard register is almost categorical, with only two attestations (0.4%) of *(durch)ziehen*. Consequently, variation for this lexical concept is virtually absent in the standard register, whereas the nonstandard register retains a comparatively rich and internally differentiated set of variants.

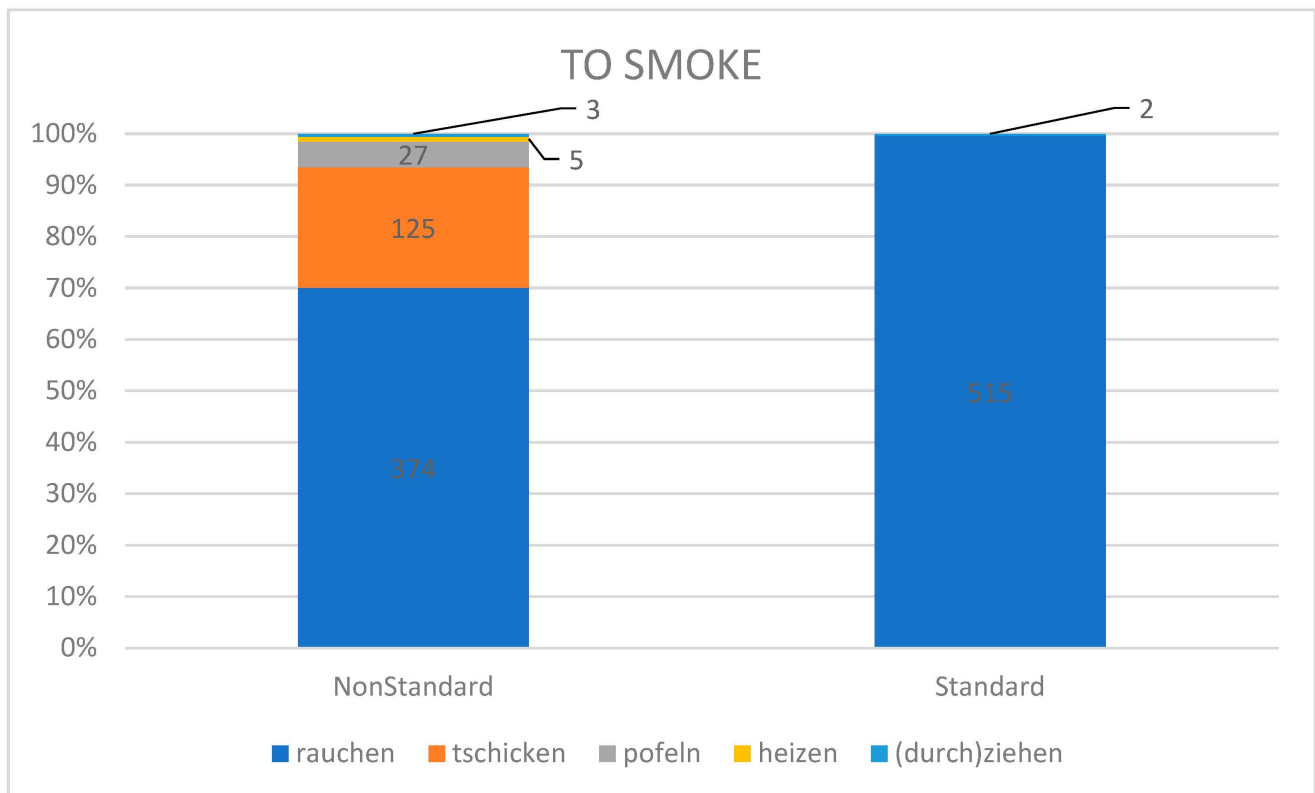


Figure 5. Lexical realizations for TO SMOKE in nonstandard and standard registers (for older speakers).

The lexical concept HORSERADISH illustrates the opposite direction of register-based variation (Figure 6). Although *Kren*, the variant commonly used in Austria, is dominant in both registers, its relative frequency is substantially higher in the nonstandard register (97.2%) than in the standard register (80.2%). The variant *Meerrettich*, which is widespread across much of the German-speaking area—particularly in Germany and Switzerland (cf. Ammon et al., 2016, pp. 419, 468)—is likewise attested in both registers. However, its frequency is nearly ten times higher in the standard register (18.8%) than in the nonstandard register (1.8%). This distribution suggests that the variant predominantly associated with Germany is also more strongly associated with standard register usage in general.

Taken together, these two perspectives reveal that the significant age × register interaction reflects a shift in the linguistic repertoire of Austrian German speakers. Older speakers maintain a clear functional differentiation between registers: in dialectal contexts, they draw on a rich inventory of traditional variants, while in standard contexts, they converge on more uniform forms (or, for some concepts, show uncertainty about standard norms). Younger speakers, by contrast, show reduced variation in the nonstandard register that approaches the levels observed in standard speech. This pattern is consistent with a grad-

ual attenuation of the traditional diaglossic repertoire among younger speakers—one in which the functional separation between dialect and standard varieties appears less sharply maintained. Given the small effect sizes involved, however, this should be understood as a directional tendency rather than evidence of a completed structural shift.

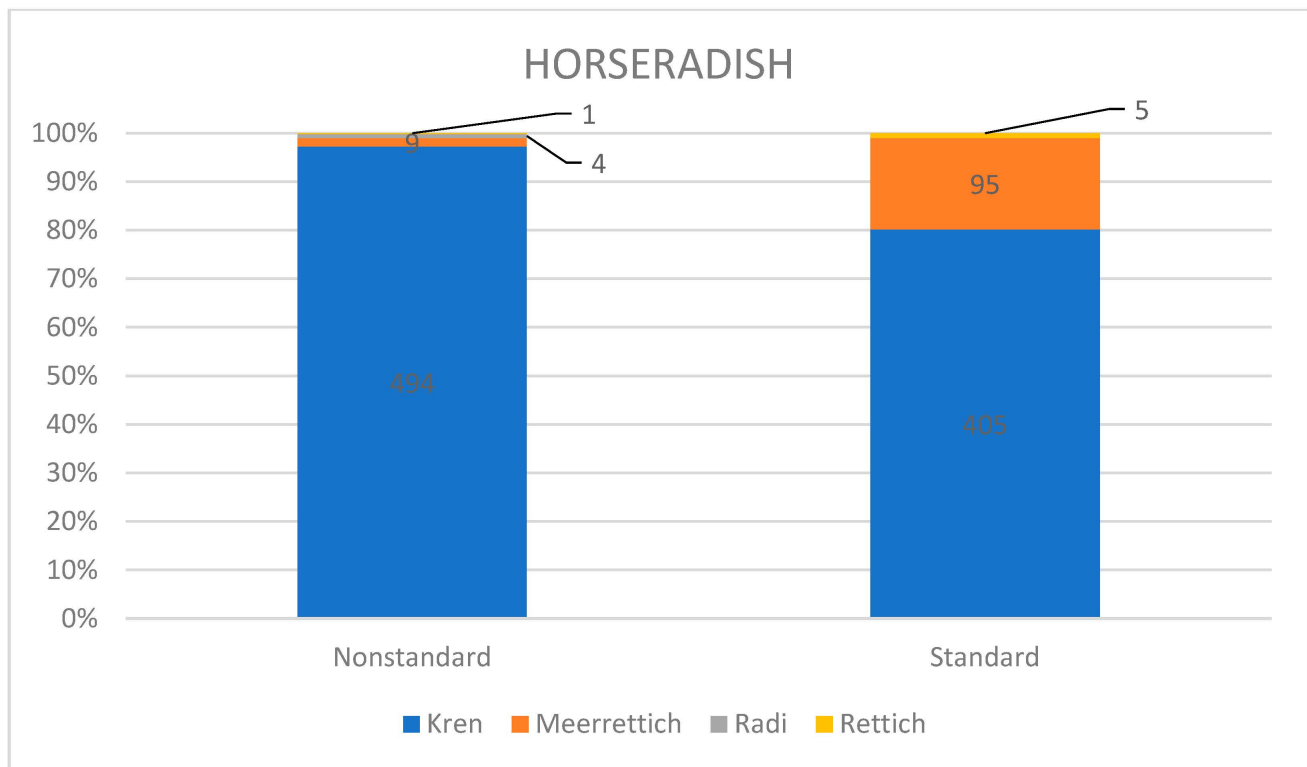


Figure 6. Lexical realizations for HORSERADISH in nonstandard and standard registers (for older speakers).

In sum, the entropy-based analysis reveals a complex picture of lexical variation in Austrian German. At the aggregate level, phenomenon and region emerged as the dominant sources of variation, reflecting both the inherent diversity of different lexical concepts and the dialectal structure of Austria. Age and register, while statistically significant, explained only negligible portions of variance according to conventional effect size benchmarks.

However, statistical significance combined with small effect sizes does not imply that the effects are meaningless. Given that lexical variation is overwhelmingly determined by the phenomenon itself ($\eta^2_p = 0.53$), any additional factors can only explain a small fraction of the remaining variance. The key finding is the consistent *direction* and *pattern* of the effects: the significant interaction between age and register ($p = 0.008$) indicates that lexical change operates differently across stylistic contexts. On average, younger speakers show reduced variation compared to older speakers in dialect and colloquial speech, consistent with a pattern of lexical levelling, while the standard register remains relatively stable across generations.

The phenomenon-specific analyses show that age- and register-related effects are strongly lexeme-dependent. Several everyday lexical concepts such as ROSE HIP, GOOSE-BERRY, CAP, and CARROT exhibit pronounced levelling within the nonstandard register, with younger speakers showing substantially reduced lexical variation compared to older speakers. This pattern points to a loss of dialect-internal alternatives rather than merely a shift in register choice.

At the same time, register-based contrasts among older speakers are not uniform across the lexicon. For many concepts (e.g., ROSE HIP, GOOSEBERRY, TO SMOKE), lexical variation is clearly concentrated in the nonstandard register, while the standard register is highly normed. For other concepts involving competition between supraregional forms and variants restricted to Austria (e.g., HORSERADISH) or to the southern part of the German-speaking area (e.g., PRUNE PLUM)³, higher entropy values in the standard register reflect ongoing competition within the standard lexicon rather than dialectal variability.

Taken together, these findings suggest not a wholesale shift toward a monoglossic repertoire, but a selective reduction of lexical diversity in the nonstandard register among younger speakers, alongside continued register differentiation and lexeme-specific dynamics in the standard register.

Given that region emerged as the second most important factor in explaining lexical variation, the following section employs multidimensional scaling (MDS) to examine the spatial structure of lexical diversity in more detail. This analysis allows us to visualize the linguistic relationships between survey locations and to identify potential dialect areas based on patterns of lexical similarity.

4.2. Multidimensional Scaling (MDS)

To visualize the spatial structure of lexical variation, we employed non-metric multidimensional scaling following the Groningen dialectometry approach (cf. Heeringa & Nerbonne, 2010). For each survey location, we constructed a binary vector across all lexical variants attested in the full dataset. Each dimension of the vector corresponds to one unique variant–concept combination (e.g., *Hagebutte* for ROSE HIP, *Hetscherl* for ROSE HIP, *Stachelbeere* for GOOSEBERRY, etc.), such that a location receives a value of 1 if that variant was attested at least once at that location, and 0 otherwise. Linguistic distances between locations were calculated using the Jaccard coefficient (cf. Jaccard, 1912), which measures the proportion of shared variants relative to the total number of variants present at either location.

Separate three-dimensional MDS solutions were computed for the overall dataset and for each subgroup (age $\times$ register) using the isoMDS algorithm. This approach ensures that the spatial structure within each group is optimally represented, yielding lower stress values than a single combined solution. The three dimensions were mapped onto RGB colour space, such that locations with similar lexical profiles receive similar colours, while linguistically distant locations appear in contrasting colours.

Since each subgroup has its own MDS solution with independent RGB scaling, colours are directly comparable *within* each map but not *across* maps. The relevant comparison between groups is therefore not the absolute colour values, but rather the *patterns* of spatial differentiation: the presence or absence of geographic gradients, the degree of colour homogeneity, and the clustering of locations.

Figure 7 presents the MDS visualization for the complete dataset. The stress value of 19.5% indicates an acceptable fit of the three-dimensional solution to the original distance matrix. The map reveals a clear west–east differentiation, with locations in Vorarlberg and Tyrol (western Austria) appearing in green tones, while eastern Austrian locations (Vienna, Lower Austria, Burgenland) cluster in blue-purple shades. Central Austrian locations show intermediate colours, suggesting a gradual transition rather than sharp boundaries.

To examine whether spatial patterns of lexical variation differ across speaker groups, we computed separate MDS solutions for each combination of age group and register (Figure 8). This approach optimizes the representation of each group’s internal structure.

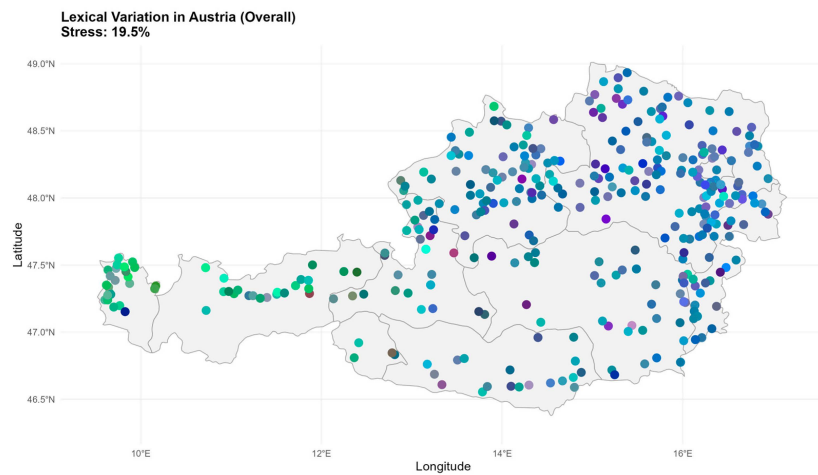


Figure 7. Overall spatial distribution of lexical variation in Austria based on MDS with RGB colour mapping. Similar colours indicate similar lexical profiles. Stress = 19.5%.

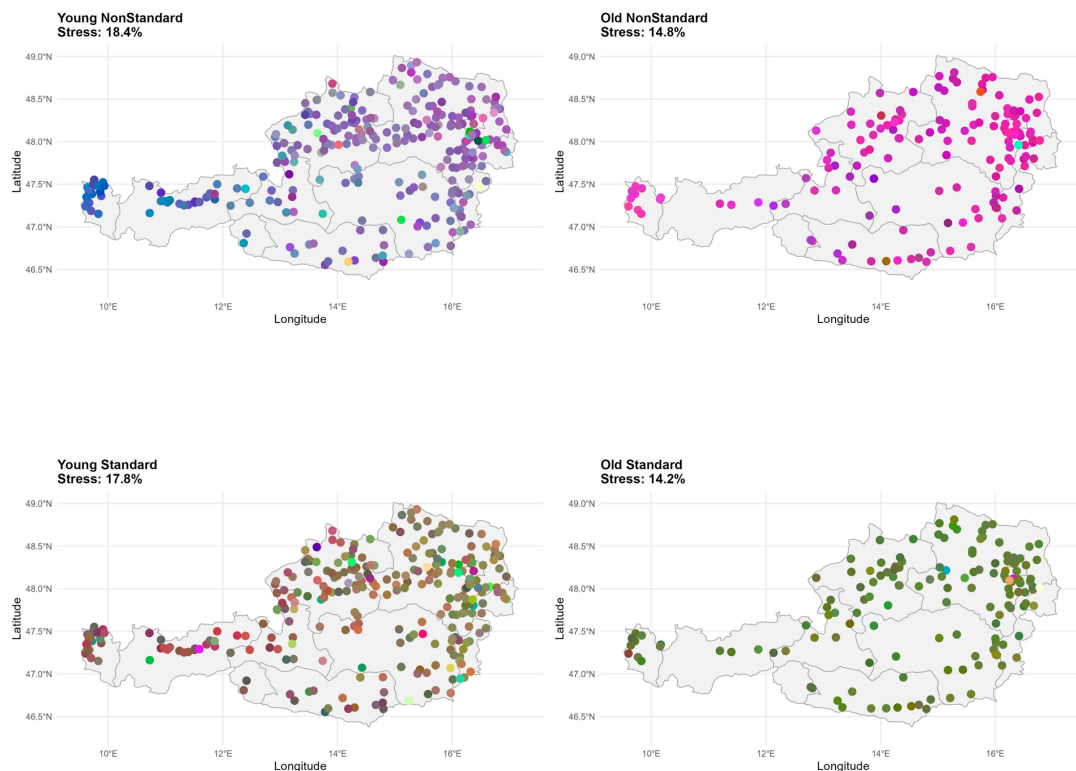


Figure 8. Spatial distribution of lexical variation by age group and register. **Top row:** Nonstandard register (young left, old right); **bottom row:** standard register (young left, old right). Each map represents a separate MDS solution; colours are comparable within maps but not across maps. Stress values range from 14.2% to 18.4%.

Because each subgroup's MDS solution is computed independently, the RGB colour assignments are determined by the internal distance structure of that group alone. Consequently, the same colour in two different maps need not correspond to the same linguistic profile; only the spatial *patterns* of colour differentiation—gradients, clusters, and outliers—are meaningfully comparable across maps.

The four maps reveal notable differences in spatial structure:

Young NonStandard (top left, stress: 18.4%): The map displays a clear west–east gradient, with blue tones dominating in western Austria (Vorarlberg, Tyrol, Salzburg) transitioning to purple–violet shades in central and eastern Austria. This geographic

patterning suggests that younger speakers' dialect and colloquial usage follows regional structuring, with lexically similar locations clustering together.

Old NonStandard (top right, stress: 14.8%): This map shows a more homogeneous pattern dominated by pink-magenta tones across most of Austria, with only limited differentiation. A few outliers appear in different colours (cyan, orange), but overall the spatial structure is less pronounced than among younger speakers.

Young Standard (bottom left, stress: 17.8%): The most heterogeneous pattern among all four groups, with a broad spectrum of colours (red, brown, olive, green, purple) distributed across Austria without a clear geographic gradient. Western Austria shows some red-brown tones, while central and eastern regions display mixed colouring. The pattern suggests that, at the standard level, certain lexical variants are firmly rooted in the west (especially Vorarlberg), whereas the remainder of Austria displays substantial internal variability with no single variant dominating.

Old Standard (bottom right, stress: 14.2%): The most homogeneous pattern, characterized by predominantly olive-green tones with limited variation and nearly all locations sharing similar colouring. The uniform colouring indicates that older speakers, irrespective of region, converge on much the same lexical choices in the standard register, suggesting a markedly more consistent conception of the standard variety than that found among younger speakers.

- The MDS analysis complements the entropy-based findings by revealing the spatial organization of lexical variation. While the variance decomposition identified region as the second most important factor explaining lexical diversity, the subgroup-specific MDS visualizations present a more differentiated picture. Only the overall solution displays a relatively clear spatial pattern, characterized by a west–east division in which Vorarlberg and Tyrol stand out as particularly distinct from the rest of Austria.
- The comparison across subgroups yields an unexpected result: it is the younger speakers who exhibit the clearer spatial structuring, with a west–east gradient that most closely resembles the overall pattern. Older speakers, by contrast, show almost no geographic clustering despite their greater overall lexical diversity. Thus, while older speakers command a larger repertoire of lexical variants, these do not appear to cluster spatially. Younger speakers, conversely, use fewer variants overall, but differentiate more clearly between western and eastern forms.
- This pattern initially appears paradoxical, yet the apparent contradiction dissolves upon recognizing that entropy and spatial differentiation capture fundamentally different dimensions of variation. Entropy quantifies the diversity of competing variants for a given concept, whereas spatial differentiation reflects whether the same or different variants dominate across regions. These dimensions are logically independent: a group may exhibit low entropy while simultaneously showing pronounced spatial differentiation. The results suggest an increasingly sharp west–east contour among younger speakers, pointing toward regional consolidation rather than uniform leveling.

4.3. GAMs

Finally, we employ generalized additive modelling (GAM) to model the possible influence of areality on our measure of diversity (i.e., Shannon entropy) via latitude and longitude. Hereby, we utilize a fine-grained measure of areality, more granular than the major regions we used in the ANOVA in Section 4.1. Our aim thereby is to see if average lexical diversity is possibly more fine-grained than the analysis as major regions in the ANOVA would suggest. We use the *mgcv* package in R (Wood, 2025) for our calculations. Spatial GAMs allow us to utilize coordinates of localities as two-dimensional smooth, allowing for a relatively straightforward representation of geographical reference models

in the form of coordinates in our analysis (see e.g., Ross, 2019). We therefore utilize GAMs to investigate the relevance of areality for lexical diversity in a very fine-grained manner.

For each subgroup (age group $\times$ register, i.e., old nonstandard, young nonstandard, old standard and young standard) we calculate Shannon entropy for each concept in each locality. The mean Shannon entropy for each locality acts as our proxy for lexical diversity. We utilize the mean across all phenomena regardless of whether or not they exhibit entropy within the specific subgroup. We then fit a GAM for each subgroup where we calculate longitude and latitude as one combined predictor for the mean Shannon entropy. We obtained longitude and latitude using the Google Maps API and corrected data manually when needed.

For the subgroup *old_nonstandard* ($n = 151$ localities), the combined smooth of latitude and longitude is not statistically significant ($p = 0.621$), suggesting that mean concept diversity is not influenced by area in our dataset.

The subgroup *old_standard* ($n = 148$ localities) behaves in line with the nonstandard group, where areality also appears to not correlate with Shannon entropy ($p = 0.711$).

For our young speakers, we observe in the nonstandard group ($n = 309$) again no association between latitude and longitude and Shannon diversity ($p = 0.416$). The same goes for the young standard group ($n = 307$), where once again there is no measurable association between lexical diversity and areality ($p = 0.119$).

The results suggest that areality, as modelled in this specific setup as a point in two-dimensional space via coordinates, has no statistically significant influence on average lexical diversity as determined by Shannon entropy in our dataset. This finding contrasts with the ANOVA, where areality modelled by major regions was found to be significant, albeit only with a small effect. This is not necessarily inconsistent, especially when considering the results of the MDS as detailed in Section 4.2. What the result of the GAMs suggest is that overall, lexical diversity can be influenced by areality, but most likely not on such a finely-grained scale that this is detectable on the level of individual localities, especially not with a dataset that is fundamentally not balanced on the level of individual locations. Furthermore, the operationalization of geography via latitude and longitude does not account for factors that might influence the spread of linguistic variants, such as mountains or rivers, which however can be used when classifying major regions. Additionally, as the prior analyses have shown, lexical diversity appears to be dependent on specific phenomena, which is not reflected in such an analysis of average Shannon entropy. While the ANOVA suggests that areality can influence Shannon entropy on the level of major regions, this effect is lost when investigated on the level of locality.

5. Discussion and Conclusions

This study set out to quantify lexical variation in the LexAT21 dataset and to identify the factors that best account for its distribution. Using Shannon entropy as a measure of lexical diversity and combining variance decomposition with multidimensional scaling (MDS) and generalized additive modelling (GAM), we addressed three interrelated research questions.

Our results show that lexical variation is primarily driven by the phenomenon itself, that is, by the lexical concept under investigation. Some concepts exhibit rich inventories of competing variants, whereas others display near-uniform realizations. This phenomenon-level variation far outweighs the effects of speaker age, register, or region. This finding has important methodological implications: aggregate measures of lexical diversity are strongly contingent on the selection of variables, and comparisons across datasets with different concept inventories must therefore be treated with caution.

Although the main effects of age and register are small, a robust interaction between the two emerges. Older speakers maintain a clear functional differentiation between

registers, drawing on a broader range of traditional variants in nonstandard contexts while converging on more uniform forms in standard speech. Younger speakers, by contrast, exhibit markedly reduced variation in the nonstandard register, approaching the levels observed in the standard register. This pattern is consistent with a shift from a diagglossic repertoire—characterized by a functional separation between dialect and standard—towards a more monoglossic configuration in which register distinctions are attenuated. Interpreted in apparent-time terms, these findings are consistent with a process of selective lexical levelling in the nonstandard register: younger speakers show reduced variation in dialectal and colloquial contexts, while the standard register remains relatively stable across generations. The small effect sizes caution against overstating this tendency; the data support a directional interpretation rather than strong claims about the extent or completion of change. This interpretation is supported by real-time evidence from Stöckle et al. (in press), who compared historical WBÖ data with dialectal LexAT21 responses and observed a significant diachronic decrease in lexical entropy.

The spatial analyses yield a seemingly counterintuitive result. While older speakers show significantly higher lexical diversity in nonstandard speech, their MDS maps exhibit relatively homogeneous colour distributions without pronounced geographic structuring. Younger speakers, despite their reduced overall variation, display somewhat sharper west–east contrasts. This apparent paradox can be resolved by recognizing that entropy and spatial differentiation capture distinct dimensions of variation: high entropy reflects the coexistence of many competing variants at individual locations, whereas spatial differentiation depends on whether the same or different variants dominate across regions. Our results indicate that younger speakers are not simply losing variation uniformly; rather, the variants they retain appear to consolidate along broader regional lines, distinguishing “western” from more “general Austrian” forms while reducing the overall size of the variant inventory. This pattern points to regional levelling rather than complete homogenization.

Overall, the spatial structures observed in our data correspond only weakly to traditional dialect boundaries (cf. the dialect classification in Lenz, 2019, p. 319). The only clear division aligns with the Alemannic–Bavarian boundary separating Vorarlberg from the rest of Austria; finer distinctions such as Central, South Central, or South Bavarian are not discernible. This contrasts with dialectometric studies based on traditional dialect atlas data, which typically yield well-defined dialect areas (cf. Pickl, 2013; Scherrer & Stöckle, 2016). However, this difference is not unexpected, given that atlas projects were explicitly designed to document maximally differentiated local dialects, with variables selected for their spatial discriminatory power and informants chosen to represent conservative speech forms.

A more directly comparable point of reference is provided by Pickl et al. (2019), who analysed Austrian data from the *Atlas zur deutschen Alltagssprache* (AdA). Like LexAT21, AdA data were collected via open online surveys with voluntary participation rather than traditional fieldwork. Their factor analysis reveals a tripartite division into western, central, and eastern regions, with boundaries closely aligned with federal state borders. Our MDS results partially converge with this pattern: the west–east distinction is clearly visible, but the central region identified by Pickl et al. (2019) does not emerge as a distinct cluster. Instead, central Austrian locations occupy transitional positions along a gradient.

Several factors may account for this discrepancy. First, Pickl et al. (2019) analysed a substantially larger variable set (191 variables compared to 45 in LexAT21), providing greater statistical power for detecting spatial structure. Second, the AdA dataset includes not only lexical but also—to a lesser extent—phonological, grammatical and even pragmatic variables, which may contribute additional dimensions of geographic differentiation. Third, the analytical approaches differ in important respects: factor analysis assigns locations

to discrete categories, resulting in sharply delimited regions, whereas MDS preserves continuous distance relations and allows for gradual transitions. The absence of a clearly bounded central region in our data may therefore reflect methodological differences rather than a substantive linguistic contrast.

Some limitations of the present study should also be acknowledged. The snowball sampling strategy, while enhancing ecological validity, likely introduces biases towards younger, more educated, and digitally connected participants. Moreover, the analyses presented here are based exclusively on data from the first round of LexAT21; subsequent rounds will enable more robust and fine-grained assessments.

Future research can build on these findings in several ways. The completed second round and the ongoing third round of LexAT21 will substantially increase the available data and allow for testing the stability of the patterns observed here. In addition, more detailed analyses of individual lexical phenomena—such as those conducted for a subset of concepts in this study—may shed further light on the mechanisms underlying lexical change in Austrian German.

In conclusion, this study provides a comprehensive quantitative portrait of lexical variation in contemporary German varieties in Austria. The dominance of phenomenon-level variation, the erosion of register differentiation among younger speakers, and the limited spatial structuring together point to a linguistic landscape in transition—one in which traditional diaglossic repertoires and fine-grained dialect distinctions are increasingly replaced by broader, regionally consolidated patterns of variation.

Author Contributions: Conceptualization, P.S. and M.P.; methodology, P.S. and M.P.; data curation, P.S. and M.P.; writing—original draft preparation, P.S. and M.P.; visualization, P.S. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in whole or in part by the Austrian Science Fund (FWF) [10.55776/F60]. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to national law.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original data presented in the study are openly available via <https://lexat21.lapis-online.at/en> (accessed on 25 June 2026).

Acknowledgments: We thank the informants who contributed to the survey and the reviewers for their valuable comments. Open Access Funding by the University of Vienna.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AdA	<i>Atlas zur deutschen Alltagssprache</i>
ANOVA	Analysis of variance
GAM	Generalized additive modelling
LexAT21	<i>Atlas zur lexikalischen Variation in Österreich im 21. Jahrhundert</i>
MDS	Multi-dimensional scaling

Notes

- ¹ <https://www.dioe.at/en/> (accessed on 25 June 2026).
- ² <https://www.zdl.org/> (accessed on 25 June 2026).
- ³ cf. Atlas zur deutschen Alltagssprache, round 11, questions 1h and 1g (<https://www.atlas-alltagssprache.de/r11-f1hg/?child=runde>) (accessed 28 January 2026).

References

- Ammon, U., Bickel, H., & Lenz, A. N. (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, Der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen* [2., völlig neu beab. u. erw. Aufl.]. De Gruyter.
- Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2023). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (17th ed.). Springer.
- Backhaus, K., Erichson, B., & Weiber, R. (2015). *Fortgeschrittene multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (3rd ed.). Springer.
- Baserow. (2026). Available online: <https://github.com/baserow/baserow> (accessed on 28 January 2026).
- Ben-Shachar, M. S., Lüdtke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5, 2815. [CrossRef]
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons.
- Eichhoff, J. (1977–2000). *Wortatlas der deutschen Umgangssprachen*. A. Francke AG Verlag.
- Eichhoff, J. (1997). Der “Wortatlas der deutschen Umgangssprachen”: Neue Wege, neue Erkenntnisse. In G. Stickel (Ed.), *Varietäten des Deutschen: Regional- und Umgangssprachen* [Institut für deutsche Sprache Jahrbuch 1996] (pp. 183–220). De Gruyter.
- Elsaß, S., & Möller, R. (2003). *Atlas zur deutschen Alltagssprache*. Available online: <https://www.atlas-alltagssprache.de/> (accessed on 28 January 2026).
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE Publications.
- Heeringa, W., & Nerbonne, J. (2010). Measuring dialect differences. In P. Auer, & J. E. Schmidt (Eds.), *Language and space. An international handbook of language variation. Volume 1: Theories and methods* [Handbücher zur Sprach- und Kommunikationswissenschaft 30.1] (pp. 550–567). De Gruyter.
- Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11, 37–50. [CrossRef]
- Kleene, A. (2020). *Attitudinal-perceptive Variationslinguistik im Bairischen Sprachraum: Horizontale und vertikale Grenzen aus der Hörerperspektive*. IDS-Verlag.
- Ko, V., Wieling, M., Wit, E., Nerbonne, J., & Krijnen, W. (2014). Social, geographical, and lexical influences on Dutch dialect pronunciations. *CLIN Journal*, 4, 29–38.
- Koppensteiner, W. (2023). *Standard(s) in Österreich: Attitudinal-perceptive Analysen* [Doctoral dissertation, Universität Wien].
- Langsrud, Ø. (2003). ANOVA for unbalanced data: Use type II instead of type III sums of squares. *Statistics and Computing*, 13, 163–167. [CrossRef]
- Leemann, A., Steiner, C., Studerus, M., Oberholzer, L., Jeszenszky, P., Tomaschek, F., & Kistler, S. (2025). *Dialektatlas: 1950 bis heute*. vdf Hochschulverlag AG.
- Lenz, A. N. (2019). Bairisch und Alemannisch in Österreich. In J. Herrgen, & J. E. Schmidt (Eds.), *Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Band 4: Deutsch* (pp. 318–363). De Gruyter.
- Lenz, A. N. (2025). LexAT21—Ein Lexikatlas und seine Hintergründe. *LexAT21: Atlas zur lexikalischen Variation in Österreich im 21. Jahrhundert. Herausgegeben von Alexandra N. Lenz*. Available online: <https://lexat21.lapis-online.at/de/articles/lexat21-einfuehrung> (accessed on 25 June 2026).
- Lenz, A. N., Dorn, A., & Ziegler, T. (2021). Lexik aus areal-horizontaler und vertikal-sozialer Perspektive—Erhebungsmethoden zur inter- und intraindividuellen Variation. *Sprachwissenschaft*, 46, 387–431.
- Lenz, A. N., Höll, J., & Ziegler, T. (2023). Lexikalische Variation in Österreich: Ausgewählte Austriazismen im Stadt-Land-Vergleich. *Muttersprache*, 133, 82–115. [CrossRef] [PubMed]
- Lenz, A. N., Ziegler, T., Höll, J., Kunzmann, M., & Dorn, A. (in press). Lexikalische Dynamik in Österreichischen Varietäten aus panchronischer Perspektive. In A. Schamberger-Hirt, F. Erhard, M. Schnabel, & V. Schwab (Eds.), *200 Jahre Dialektologie. Forschungsstand und Auftrag. Akten der 15. Bayerisch-Österreichischen Dialektologentagung 2023*. Edition Vulpes.
- LimeSurvey Project Team & Schmitz, C. (2026). *LimeSurvey*. Available online: <https://github.com/LimeSurvey> (accessed on 28 January 2026).
- Mitzka, W., & Schmitt, L. E. (1951–1980). *Deutscher Wortatlas*. Wilhelm Schmitz Verlag.
- Möller, R. (2021). An online atlas of colloquial German: The Atlas zur Deutschen Alltagssprache. In A. Thibault, M. Avanzi, N. Lo Vecchio, & A. Millour (Eds.), *Nouveaux regards sur la variation dialectale—New ways of analyzing dialectal variation* (pp. 197–215). Éditions de Linguistique et de Philologie.

- Nerbonne, J., Heeringa, W., & Kleiweg, P. (1999). Edit distance and dialect proximity. In D. Sankoff, & J. Kruskal (Eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison* (pp. v–xv). Center for the Study of Language and Information.
- Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10, 439–446. [CrossRef]
- Pickl, S. (2013). *Probabilistische Geolinguistik: Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben*. Steiner.
- Pickl, S., Pröll, S., Elspaß, S., & Möller, R. (2019). Räumliche Strukturen Alltagssprachlicher Variation in Österreich anhand von Daten des atlas zur Deutschen Alltagssprache (AdA). In L. Bülow, A. K. Fischer, & K. Herbert (Eds.), *Dimensions of linguistic space: Variation–multilingualism–conceptualisations* (pp. 39–59). Peter Lang.
- Pröll, S. (2015). *Raumvariation zwischen Muster und Zufall: Geostatistische Analysen des Sprachatlas von Bayerisch-Schwaben*. Steiner.
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Ross, N. (2019). *Gams in R*. Available online: <https://noamross.github.io/gams-in-r-course/> (accessed on 25 January 2026).
- Rumpf, J., Pickl, S., Elspaß, S., König, W., & Schmidt, V. (2009). Quantification and statistical analysis of structural similarities in dialectological area-class maps. *Dialectologia et Geolinguistica*, 18, 73–100.
- Scherrer, Y., & Stöckle, P. (2016). A quantitative approach to Swiss German—Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24, 92–125. [CrossRef]
- Schmidt, J. E., Dammel, A., Girth, H., & Lenz, A. N. (2019). Sprache und Raum im Deutschen: Aktuelle Entwicklungen und Forschungsdesiderate. In J. Herrgen, & J. E. Schmidt (Eds.), *Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Band 4: Deutsch* (pp. 28–60). De Gruyter.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423. [CrossRef]
- Stöckle, P. (2014). *Subjektive Dialekträume im alemannischen Dreiländereck*. Olms.
- Stöckle, P. (2016). Horizontal and vertical variation in Swiss German morphosyntax. In M.-H. Côté, R. Knooihuizen, & J. Nerbonne (Eds.), *The future of dialects* (pp. 195–215). Language Science Press.
- Stöckle, P. (2021). Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In A. N. Lenz, & P. Stöckle (Eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (pp. 11–46). Steiner.
- Stöckle, P., Dorn, A., Gellan, A., Kunzmann, M., & Lenz, A. N. (in press). Dialects in Austria: Lexical variation and change. In A. N. Lenz, S. Elspaß, & S. Newerkla (Eds.), *German in Austria. Variation–contact–perception*. Language Science Press.
- Stöckle, P., Hemetsberger, C., & Schütz, M. (2021). Die WBÖ-Belegdatenbank als Quelle für syntaktische Analysen—Möglichkeiten, Grenzen, Perspektiven. *Wiener Linguistische Gazette*, 89, 579–626.
- Stöckle, P., & Vergeiner, P. (2025). Geographical patterns in the Bavarian dialects of Austria and South Tyrol: A real-time comparison using dialectometric methods. *Journal of Language Variation and Sociolinguistics*, 1, 20–41. [CrossRef]
- Venables, B. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer.
- Wiegand, H. E., & Harras, G. (1971). *Zur wissenschaftshistorischen Einordnung und linguistischen Beurteilung des Deutschen Wortatlas* [Germanistische Linguistik] (Vols. 1–2). Olms.
- Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics*, 1, 243–264. [CrossRef]
- Wieling, M., Nerbonne, J., Montemagni, S., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 90, 669–692. [CrossRef]
- Wirtz, M. A., & Pfenninger, S. E. (2025). Identifying and explaining L2 growth spurts: A tutorial on generalized additive models for time-intensive longitudinal data in applied linguistics research. *Research Methods in Applied Linguistics*, 4, 100259. [CrossRef]
- Wirtz, M. A., Pickl, S., Niehaus, K., Elspaß, S., & Möller, R. (2025). Reconciling the social and spatial: An apparent-time analysis of variation intensity in colloquial German. *Journal of Linguistic Geography*, 13, 57–70. [CrossRef]
- Wood, S. (2025). *Mgcv: Mixed GAM computation vehicle with automatic smoothness estimation*. Available online: <https://cran.r-project.org/web/packages/mgcv/index.html> (accessed on 25 June 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.