A resource for lexical variation in a pluri-areal context: Introducing LexAT21, the atlas on lexical variation in Austria in the 21st century

Markus Pluschkovits
Austrian Centre for Digital Humanities
Austrian Academy of Sciences, Austria
markus.pluschkovits@oeaw.ac.at

Daniel Schopper
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
Vienna, Austria
daniel.schopper@oeaw.ac.at

Anja Wittibschlager
Institute for German Studies
University of Vienna, Austria
anja.wittibschlager@univie.ac.at

Jakob Bal
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
Vienna, Austria
jakob.bal@oeaw.ac.at

Kilian Kukelka
Austrian Centre for Digital Humanities
Austrian Academy of Science
Vienna, Austria
kilian.kukelka@oeaw.ac.at

Olivia Reichl
Austrian Centre for Digital Humanities
Austrian Academy of Science
Vienna, Austria
olivia.reichl@oeaw.ac.at

Abstract

The following contribution introduces LexAT21, the atlas on lexical variation in Austria in the 21st century. The linguistic context of this project is briefly covered in section one, suggesting that areal and social/vertical variation is characteristic for the German language in Austria. This is then followed by a brief description of the dataset of the first two survey rounds of LexAT21 and a short sample analysis, showing the potential of the dataset for linguistic research. We then turn to integration into the CLARIN context. The contribution closes by considering the issue of representing different varieties of languages in standards.

1 Introduction and state of research

The field of language variation is a wide and burgeoning one and has evolved considerably during the last hundred years. However, it is far from easy to capture the complex array of variation within the German language empirically. German within Austria is strongly characterized by pluriareal variation (see e.g., Auer 2021 for a discussion) on all linguistic levels, and shows significant variation not just on dialectal levels, but also in its standard varieties (see e.g., Ammon et al. 2016 or Dürscheid et al. 2020). For the longest time, language variation in German has been

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

considered mostly in nonstandard varieties in the tradition of German dialectology.

Especially in German dialectology, the study of variation on the lexical level in dialects has a rich tradition, involving large-scale projects in the past such as the Deutsche Wortatlas (Mitzka & Schmitt1951 ff).[2] However, many of these projects typically consider only parts of the lexical spectrum (such as purely dialectal variation or, less commonly, variation in standard varieties, notably for example in the Variantenwörterbuch des Deutschen by Ammon et al. 2016) and differ in their methodology so much that the comparison of the data between different projects would be academically negligent (see Schmidt et al. 2019: 43–44). Thereby, what is still missing is a systematic, large-scale investigation of lexical variation across varieties in a comparable methodology, ideally also considering variation within individual speakers (cf. Schmidt et al. 2019: 44). LexAT21, the atlas on lexical variation in Austria in the 21st century (see Lenz 2025) aims to fill this research gap for German varieties in Austria.

The aim of this paper is to introduce the project and dataset and showcase the potential of the data for further use. We briefly lay out the project context and elaborate on data collection in Section 2. This is then followed by short sample analysis presented in Section 3. Finally, in Section 4, we close with the integration of the data into CLARIN, with a reflection on glottonyms and their role for this data set.

2 The LexAT21 project and data

LexAT21 is the result of a close collaboration between the Special Research Program "German in Austria: Variation – Contact – Perception", funded by the Austrian Science Fund FWF (Funding Agreement FWF F 0 60, see Budin et al. 2019) and the Austrian Centre for Digital Humanities. The aim of LexAT21 is to investigate variation in lexis in contemporary German in Austria, focusing on both variation between speakers (interindividual variation, controlled by e.g., areal factors such as dialects) and within a speaker (intraindividual variation across repertoires on the dialect-standard axis). To achieve this aim, LexAT21 considers both vertical-social and horizontal-areal variation in the data acquisition and analysis.

The data used for LexAT21 is acquired in semi-regular rounds of online-surveys, designed in the open-source tool LimeSurvey (LimeSurvey 2026). As of writing, the third survey round is being conducted, and data from the second survey round is being analyzed, with the first survey round already published on the LexAT21 platform. All surveys are comprised of roughly 40 linguistic variables[3], which are presented as picture stimuli accompanied by some variation of the written question "Wie nennen Sie den abgebildeten Gegenstand?" ('What do you call this pictured object?'), or, in the case of verbs, "Was macht die Person auf dem Bild? Antworten Sie in einem vollständigen Satz." ('What is the person in the picture doing? Answer in a full sentence.'). Informants then have to type their variants in dedicated boxes, differentiating between their registers. At the beginning of the survey, informants had to declare themselves as either speakers of dialect (presented to them with the term "Dialekt (Mundart)") or speakers of colloquial German ("Umgangssprache oder Alltagssprache"), a salient distinction in the context of German in Austria – from survey round two on, speakers were also able to declare themselves as speaking a standard variety exclusively. Their choice influenced the answer boxes for each given stimulus – if speakers identified with speaking dialect, the label displayed for their nonstandard answer was "Dialekt" (respective "Umgangssprache" for those identifying as speakers of more colloquial varieties), and if speakers stated that they speak neither when speaking German, there was no answer field for answers from the

[2] For a recent overview on lexical variation in other languages, see e.g. Sandow et al. (2025) or Durkin (2012).
[3] Meaning in this context a concept which can be named with more than one lemma (i.e, variant). See Tagliamonte (2025: 65–89) for a recent overview of the term in linguistics.

nonstandard register. The answer box for their standard register, which was present in every case, was labelled with one random glottonym for standard, drawn from a pool of three (“Ihr bestes Hochdeutsch”, “Ihr österreichisches Hochdeutsch”, “Ihr Hochdeutsch”, ‘Your best High German’, ‘Your Austrian High German’, ‘Your High German’, see e.g., Koppensteiner 2023 for a discussion of these terms). The term for the standard variety stayed consistent throughout the survey. Previous analysis of the data shows that the choice of term influences the answers given by the informants, see Lenz et al. (in print). There was an additional field for further terms the informants use but not associate with any of their registers, simply asking “Verwenden Sie noch weitere Bezeichnungen?” (‘Do you use any further terms?’).

Figure 1 illustrates how such a stimulus looks like for the variable HAT. The informant in this example identifies as dialect speaker and had the glottonym “bestes Hochdeutsch” (‘your best High German’) assigned randomly.



Figure 1: Written stimulus on top asking ‘How do you call the garment pictured?’, with answer boxes for ‘dialect’, ‘best High German’, and ‘further answers’.

The mode pictured in Figure 1 (i.e. a picture stimulus in combination with a written task) was used to collect the entire data, with variations on the formulation of the verbal stimuli. All pictures were chosen with a permissive CC-0 license. Overall, 45 different stimuli were tested in survey round one, and 40 stimuli were / are being tested in survey round two and three respectively. The choice of variables for survey round one was made on the basis of a sample of variables from the Deutsche Wortatlas, which allows for diachronic comparison. Variable choice for the following survey rounds was made based on introspection and reference to secondary materials within the LexAT21 team. In addition to the language data, metadata on the informants was collected: brief biographical details such as age, gender and employment with additional, more detailed information on residency history. Furthermore, each speaker was asked to declare themselves as an informant for one specific location (in Austria). Each survey round was considered finished as soon as roughly 1,900–2,000 valid responses were collected (valid here meaning that the reference locality is located in Austria).

At the end of a collection period, the raw answer data is exported from the LimeSurvey instance. We then import the data into an instance of baserow (baserow 2026), an open-source online data platform. In baserow, we are able to view the data in grid views, which we use for the first manual processing the data – among these is the exclusion of informants that are not located in Austria and a brief review of the provided residency history vis-à-vis the locality chosen by the informant to be representative for them (as sometimes, this question is misunderstood and the selected localities are not fully consistent with the informants' residency history). After a basic check of the informant data, we annotate the linguistic data. This is done on a variable-by-variable basis. The data is annotated by

lexical type, usually corresponding to a lemma, but sometimes to a collection of different lemmas (in the case of e.g., compound nouns the analysis might only focus on the lemma of the head, thereby subsuming different modifiers to one lemma). Annotation is done manually instead of programmatically, for two key reasons: first, the effort is – for many variables – quite trivial, and bespoke automatic solutions would simply require more effort than manual annotation. In many cases, filtering in baserow allows to quite easily identify all data points for one specific annotation based on substrings characteristic for a variant (e.g., for TABAK RAUCHEN 'to smoke', all variants subsumed under the annotation *tschicken* will – in contrast to others – include some variation of the grapheme <s>). The second reason for manual annotation is that some of the lemmas simply are only found in nonstandard varieties, and therefore have no normalized spelling version. Often, these variants will also only emerge while the data is being annotated, meaning that dictionary-based solutions would not work. For many variables, the concrete taxonomy of annotations is not known before the data is annotated, which also does not lend itself to an automated solution.

Typically, the complexity of the annotation scheme was highly dependent on the individual variable. We employed some guiding principles to ensure that the different variables remained comparable: variants not designating the variable were annotated as "irrelevant" ('irrelevant') across all variables (e.g., "Zigarette" 'cigarette' as answer for the variable TABAK RAUCHEN 'to smoke' – this was categorized as irrelevant as 'cigarette' is not a variant that refers to the act of smoking). These irrelevant variants were typically rare, suggesting that the choice of stimuli was – in general – appropriate, although some uptick of irrelevant variants can be detected with designations for flora and fauna. A further principle was the subsuming of hapax variants under the label "sonstige" ('other') – some informants reported idiosyncratic variants (e.g., names of family members for the variable OMA 'grandmother'), which were not considered as autonomous variants. We made exceptions for those variants with historic relevance (because they e.g. appear in older atlases) and usually did not subsume them as "sonstige" even if they did not meet the frequency we established for autonomous variants (in most cases five independent appearances).

Once all variables for one survey round are annotated and the annotations are reviewed, the data is transferred from baserow to a postgreSQL relational database. This is done utilizing the baserow API and transforming the data to fit the data model before populating the database with the new round. This workflow is replicable, not just to account for the continuous survey rounds, but also to accommodate changes in the data, such as fixing minor mistakes (e.g., typos in annotation names), whereby – depending on the amount of data that is changed – the data of the survey round may be wiped and imported new.

The database acts as the backend from which the LexAT21 frontend (available online, see Lenz 2025 ff.) pulls the actual data via an API. The data is displayed both on a customizable map and in a table view, offering many options for display and filtering of data. The maps and data tables are easily shareable, as the filter parameters are written directly into the URL. Furthermore, users can switch between the data grid view and the map with the filter parameters staying intact, allowing users to easily visualize data on the map. The map can be customized not just with regard to data but also with regard to visualization parameters, such as base maps and colours of the legend, and can easily be exported in a variety of formats for use in publications and the likes. The datasets and maps are one of the main outputs of LexAT21, and are complimented by the map commentaries.

There are two different types of map commentaries: short comments and full map commentaries. The short commentaries are, as the name implies, simply a brief overview over the dominant variants and typically do not exceed a paragraph in length. Their aim is to concisely inform readers on surface-level variation and typical variants in German as spoken in Austria. The full map commentaries in contrast go into more depth. The full map commentaries, additionally to brief

overviews of the variables in other data sources and atlases, elaborate on the variation in different registers (both standard and nonstandard), and offer curated maps on the individual variable – which are linked back to the mapping tool so readers can explore and customize them. These map commentaries are, just like the data, part of the output of LexAT21 and their integration into the CLARIN infrastructure will be discussed in Section 4. Before turning to this integration, a small sample analysis taken from the LexAT21 dataset aims to lay out the relevance of the data to the community.

3 Sample analysis

This brief sample analysis is based on the variable ZUSAMMENKEHREN ('to sweep') taken from LexAT21 survey round one. For further quantitative analyses see e.g., Stöckle & Pluschkovits (forthcoming) or Ziegler et al. (forthcoming). This present analysis is based on the 1,925 informants of this survey round, out of which 996 identified as speakers of colloquial language ('Umgangssprache') and 929 identified as speakers of dialect. The average informant age of this round was 39, with roughly one third of the informants being below 30, one third between 30 and 50, and one third over 50 years. 1,205 informants identified as female, 703 as male and 17 as diverse. Based on these informants, there were 3,408 relevant individual answers for the variable ZUSAMMENKEHREN (excluding irrelevant and miscellaneous variants, see also Pluschkovits 2025).

The variable ZUSAMMENKEHREN is interesting not just because it was also surveyed in the Deutsche Wortatlas (Mitzka & Schmitt 1953) but also because it has been in the focus of variationist analyses before: as early as Kretschmer (1918: 196), who assured that the variant *kehren* is completely dominant in Austria, but that Viennese speakers consider *fegen* more prestigious (in contrast to Berlin speakers, where *fegen* is dominant but *kehren* is more prestigious). Since then, much has been written about the variation between *kehren* and *fegen*, e.g. in Huesmann (1998) or Muhr (2001), and we find this variable in atlases of colloquial language such as the Atlas deutscher Alltagssprache (AdA 2005) or the Wortatlas der deutschen Umgangssprachen (1978), with more complex variation than what was asserted at the beginning of the 20th century.

In the LexAT21 dataset, we can observe complex variation in all registers. Considering the standard register (Figure 2), we find considerable diversity. It is especially noteworthy that the variant *fegen*, presumed to not be relevant in Austria, actually makes up 12 % of all answers in the standard register[4] with seemingly no areal preference. Furthermore, the simple verb form *kehren* is more common than any individual form with a particle, such as *zusammenkehren* or *aufkehren*.

[4] For brevity, the three different standard registers are considered together.

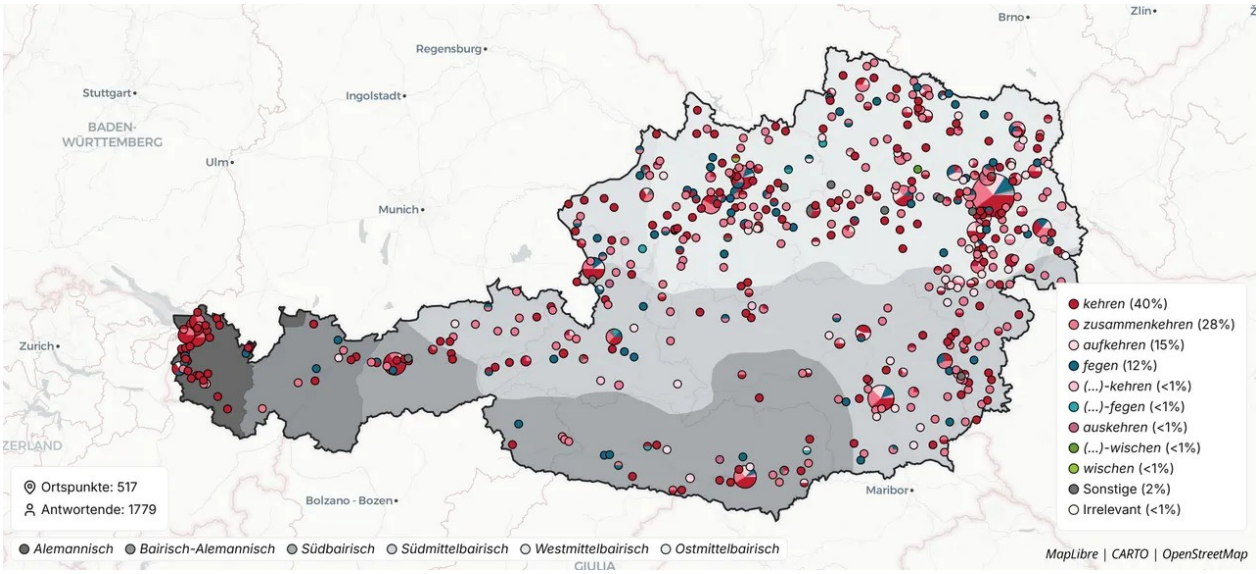


Figure 2: variation in standard registers for ZUSAMMENKEHREN, taken from Pluschkovits 2025

In colloquial language (Figure 3), we can observe this pattern reversed: here, the particle form *zusammenkehren* is clearly more widespread than the base form *kehren*, making up almost half of all data points. Additionally, the variant *fegen* is even more marginal here, with roughly 5 % of all data being *fegen* or a particle verb with *fegen*. We can furthermore observe that colloquial speakers of German tendentially reside more in the eastern than in the western half of Austria (see also Lenz 2025).

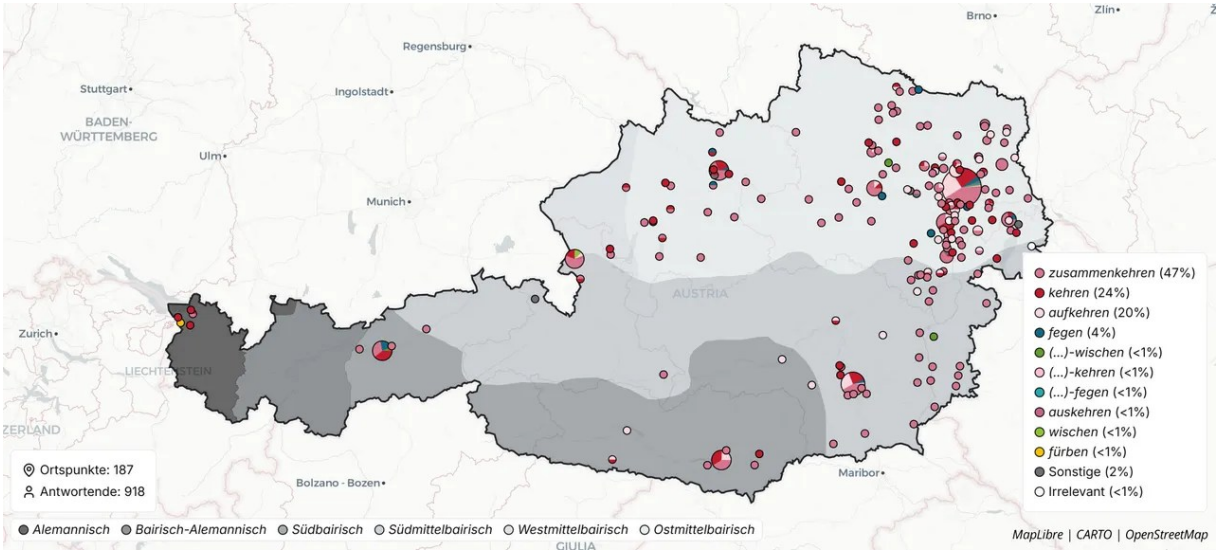


Figure 3: variation in colloquial German for ZUSAMMENKEHREN, taken from Pluschkovits 2025

Figure 4 details the variation of ZUSAMMENKEHREN in dialect. Immediately, one can spot clear differences to the answers of the speakers of colloquial German: *zusammenkehren* is even more dominant than in the colloquial register, and *fegen* even more marginal, accounting for less than 1 % of the data. The variant *fürben*, not present in standard registers and barely present in colloquial German, now accounts for 5 % of all data and is the dominant variant in Alemannic dialects in the west of Austria. We can furthermore observe a different regional spread of data in contrast to colloquial speakers, which are only marginally relevant in the west of Austria – dialect-identified speakers appear all through Austria and constitute the majority of speakers in the west.

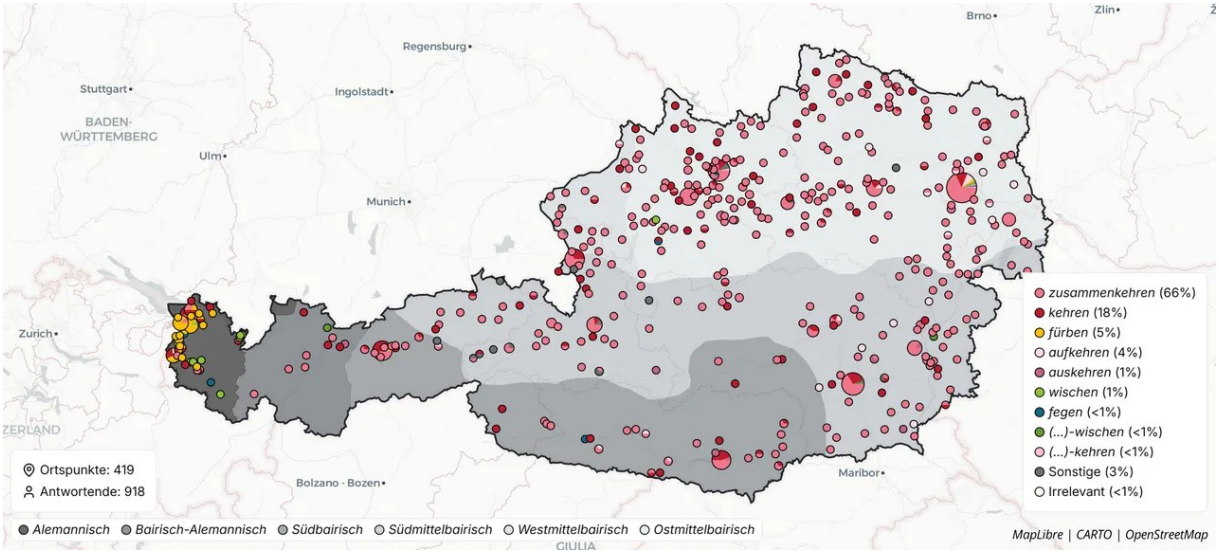


Figure 4: variation in the dialect register for ZUSAMMENKEHREN, taken from Pluschkovits 2025

As can be seen from this brief analysis, language variation in German in Austria is complex and partially dependent on the targeted register(s). We can observe that there is also heterogeneity in standard registers, and that the differentiation in nonstandard registers between colloquial and dialectal speakers can be relevant (see also Lenz et al. 2021 or Lenz et al. 2023). This, of course, has implications for how the data can be reused and shared with the CLARIN community. The final Section 4 is dedicated to circumnavigating this issue and detail how we integrate LexAT21 into the broader CLARIN context.

4. Integration into CLARIN

Both the map commentary and research data are outcomes we consider relevant for the CLARIN context. We will provide both to the community with a permissive license to the Virtual Language Observatory via the repository of the Austrian Centre for Digital Humanities, ARCHE (A Resource Centre for the HumanitiEs, see Żółtak et al. 2021). The data will be provided in two formats, as a csv dump as well as in TEI based XML.

We automated the creation of the TEI versions of both the map commentary as well as the entire corpus. Basis for this creation is a bash-script which pulls the data out of the relational database. We utilize the data of the responses, the corresponding annotations, informant data and metadata as well as the location data and information on the survey round for the construction of the corpus and map commentaries. Each map commentary is considered its own document. We construct the corpus and commentaries via different partials, in an effort to create a replicable export strategy. This is done for two reasons: first, because data is still being re-categorized by researchers by the time it is first imported into the database – the researchers utilize the mapping tool of LexAT21 to uncover possible areal patterns, which might lead to recategorization of variants or renaming of individual lemma (e.g., if a nonstandard lemma is mostly dominant in the Alemannic area, the spelling of the lemma might be changed to reflect this fact). Having an easily replicable workflow means that the production of different versions of the data is trivial. Secondly, as LexAT21 is an ongoing project with multiple survey rounds, the data will expand, which is another argument to make this workflow easily retriggerable. The import step of the data itself is deployed via GitHub Actions.

The first chunk of the export into TEI is the creation of the documents containing the map

comments. In addition to the text of the map commentary itself (i.e., the readable content) with basic markup, the TEI versions of the map commentary includes all necessary information on the provenance of the data – this includes not just the responses of the informants for each map commentary packaged in a `cit` element with an (among others) `ana` attribute but also exact formulation of stimulus as well as a link to the stimulus picture, ensuring that the individual analysis is reproducible and the data generation itself replicable. All possible stimuli (i.e., stimuli for all possible registers that informants could encounter with the exact verbal stimulus) are included as a feature structure for every map commentary.

The map commentaries are accompanied by a feature structure declaration (`fsDecl`), which includes the prose description of how the individual stimuli were constructed (thereby – much like a shorter version of the present paper – contextualizing the dataset).

The corpus document includes the metadata on all speakers as an initial `particDesc` element, including age and sex for every informant. A following `listPlace` details all the localities covered in the corpus. These, just as the informants, will necessarily differ between the individual survey rounds, which is why we will release the data of each survey round as its own (sub)corpus, albeit in the same manner. In an effort to reduce unnecessary redundancy, the linguistic data itself is not included in the corpus element. Instead, we reference the data from the individual map commentaries by utilizing `xi:include` elements with an Xpointer, pointing towards the data in the individual map commentaries.

The corpus object is also an interesting challenge in data modelling. Quite clearly, LexAT21 is sensitive to variation in language, especially across the areal dimension, but also in the social dimension. The project explicitly investigates the difference in registers, i.e., when informants are prompted to give their designations in standard or nonstandard registers. And out of this social variation, the question arises on how these different registers can effectively be identified in the corpus data. There has been a recent uptick in understanding language data from a variationist standpoint and trying to standardize language designations while remaining sensitive to variation and different varieties (see e.g., Romary 2025). However, how can the present possibilities be utilized to account for social variation?

As a first possible solution, the `<langUsage>` element in the header could be used to describe the "languages, sublanguages, registers, dialects" (TEI Consortium: 58). This, however, would only be used to identify the languages (or in this case, registers) of the corpus and has, for the present case, two considerable shortcomings: the first shortcoming is that the concrete tag for language identification should follow the BCP 47 (Philipps & Davis 2009), which in turn necessitates that the individual subtags are registered at IANA. There are currently no subtags that would cover the distinction between *Dialekt* und *Umgangssprache*, especially as those are considered from the perspective of them being glottonyms instead of concrete dialects (e.g., an informant might respond in Alemanic and another one in South-Bavarian, but for this analysis, both of their language use is simply considered *Dialekt*, see also e.g., Kim & Breuer 2017 on subtags for varieties of languages). This could be circumvented by utilizing the private use subtag `x` (i.e., `de-x-`) which would allow to e.g., define a `de-AT-x-dialekt` and `de-AT-x-umgangssprache`. This, however, is not beneficial to keeping the data as interoperable as possible.

The second shortcoming is that the application of such a label has a normative character – if e.g., a subtag for dialect was used, this would imply a certain level of curation on the data that simply is not present (e.g., excluding non-dialectal material). Essentially, we run into the issue of circularity: if we proclaim something to be dialect, we necessarily must have a notion of what constitutes such a

dialect, and by excluding material that does not fit this notion, we have constructed a concept which cannot be disproven. Alternatively, simply all responses elicited for the label *Dialekt* (or *Umgangssprache*) could be declared as such, which does beg the question on what to make of those responses were nonstandard and standard forms correspond perfectly. If nonstandard is defined as being (maximally) distant from standard, the fact that both forms coincide could imply that either the nonstandard or standard form is faulty, caused by either dialectal leveling or someone not being competent in standard language. Furthermore, exclusively labelling nonstandard varieties with a subtag implies that standard varieties of a language are the default, which is a questionable assertion for German, especially as spoken in Austria, at best.

With all this in mind, we consider the current version to represent our data (via a `usg type` attribute of the respective stimulus in the `cit` element) as a workable compromise, that hopefully does not do too much injustice to the data itself. There are, as mentioned above, exciting developments about more-dimensional representation of varieties of languages in TEI (as exemplified by e.g., Romary 2025). This could help make the dataset – which we consider of interest for everyone in the CLARIN context working on German, language variation generally or lexical variation specifically – more easily comprehensible. We hope that our use-case with regard to representing intended varieties based on their glottonyms can be a point of departure for further conversations, specific as this use-case might be.

5. Concluding remarks

The LexAT21 addresses a gap in research and aims to investigate lexical variation throughout Austria along the dialect-standard axis, from dialect through colloquial language to standard, using a consistent survey methodology. The LexAT21 data is collected in a controlled survey-process using multiple rounds of online questionnaires. Exemplary analyses of the linguistic variable "ZUSAMMENKEHREN" ('to sweep') do not only demonstrate a broad spectrum of lexical variants but also highlight significant spatial effects within Austria as well as the effect of the register (dialect, colloquial language, standard).

The technical pipeline begins with the data collection in LimeSurvey, proceeds through the linguistic annotation in Baserow and storage in PostgreSQL, and ultimately results in the presentation of the data in the web frontend. This replicable approach is essential within the project, since it is an ongoing project involving multiple rounds of data collection. Furthermore, integration into CLARIN and ARCHE, as the provision of the raw data in CSV and TEI formats, is intended not only to ensure the long-term preservation but also to enable cross-project reusability and interoperability. Finally, the LexAT21 dataset also shows that labelling language data is not necessarily a trivial act. While all data might be German, the differences in the data depending on the targeted registers demonstrates that no natural language is a monolith and that variation is inherent to natural language use. This, of course, has consequences on the reuse of the data – as the targeted register is essential context of the dataset, it needs to be taken into account for any given reuse of the dataset. This, however, is not necessarily a flaw of the data or limits the reusability – instead, it allows for more control over which part of the data (by e.g. register) is actually reused.

References

Ammon, U., Bickel, H., & Lenz, A. N. (Eds.). (2016). Variantenwörterbuch des Deutschen (2., völlig neu beab. u. erw. Aufl.). De Gruyter. https://doi.org/10.1515/9783110245448 (Original work published 2011)

Auer, P. (2021). Reflections on linguistic pluricentricity. Sociolinguistica, 35(1), 29–47. https://doi.org/10.1515/soci-2021-0003

Baserow. (2026). [Repository]. Github. https://github.com/baserow/baserow

Bülow, L., Wittibschlager, A., & Lenz, A. N. (2023). Variation and change of relativizers in Austria's German varieties. SPRACHWISSENSCHAFT, 48(3), 243–280.

Durkin, P. (2012). Variation in the lexicon: The 'Cinderella' of sociolinguistics?: Why does variation in word forms and word meanings present such challenges for empirical research? English Today, 28(4), 3–9. Cambridge Core. https://doi.org/10.1017/S0266078412000375

Dürscheid, C., Elspaß, S., & Ziegler, A. (2018). Variantengrammatik des Standarddeutschen. Leibniz-Institut für Deutsche Sprache. https://mediawiki.ids-mannheim.de/VarGra/index.php/Start

Eichhoff, J. (1978). Wortatlas der deutschen Umgangssprachen (Vol. 2). A. Francke AG Verlag.

Elspaß, S., & Möller, R. (n.d.). Atlas zur deutschen Alltagssprache. Retrieved https://www.atlas-alltagssprache.de/

Huesmann, A. (1998). Zwischen Dialekt und Standard: Empirische Untersuchung zur Soziolinguistik des Varietätenspektrums im Deutschen. Max Niemeyer Verlag.

Kim, A., & Breuer, L. M. (2017). On the Development of an Interdisciplinary Annotation and Classification System for Language Varieties – Challenges and Solutions. Jazykovedný Časopis, 68, 191–207. https://doi.org/10.1515/jazcas-2017-0029

Koppensteiner, W. (2023). Standard(s) in Österreich: Attitudinal-perzeptive Analysen [University of Vienna]. https://ubdata.univie.ac.at/AC16945911

Kretschmer, P. (1918). Wortgeographie der hochdeutschen Umgangssprache. Vandenhoeck & Ruprecht.

Lenz, A., Höll, J., & Ziegler, T. (2023). Lexikalische Variation in Österreich. Ausgewählte Austriazismen im Stadt-Land-Vergleich. Muttersprache: Vierteljahresschrift Für Deutsche Sprache, 133(1–2), 82–115.

Lenz, A. N. (2025). Atlas on Lexical Variation in Austria in the 21st Century. Austrian Centre for Digital Humanities. https://lexat21.lapis-online.at/en

Lenz, A. N. (2025ff.). LexAT21—Atlas zur lexikalischen Variation in Österreich im 21. Jahrhundert. Konzipiert und entwickelt von Jakob Bal, Kilian Kukelka, Markus Pluschkovits, Marie Rohler, Daniel Schopper und Anja Wittibschlager. Unter Mitarbeit von Amelie Dorn, Jan Höll, Katharina Korecky-Kröll, Wolfgang Koppensteiner, Claudia Mattes, Markus Pluschkovits, Rita Stiglbauer, Florian David Tavernier, Anja Wittibschlager, Theresa Ziegler, Kerstin Lorenz und Eric Schirl. https://lexat21.lapis-online.at/de

Lenz, A. N., Dorn, A., & Ziegler, T. (2021). Lexik aus areal-horizontaler und vertikal-sozialer Perspektive – Erhebungsmethoden zur inter- und intraindividuellen Variation. Sprachwissenschaft, 46(4), 387–431.

Lenz, A. N., Ziegler, T., Höll, J., Kunzmann, M., & Dorn, A. (in print). Lexikalische Dynamik in österreichischen Varietäten aus panchronischer Perspektive. 200 Jahre Dialektologie. Forschungsstand Und Auftrag. Akten Der 15. Bayerisch-Österreichischen Dialektologentagung 2023. Bayerisch-Österreichischen Dialektologentagung 2023.

LimeSurvey Project Team, & Schmitz, C. (2026). LimeSurvey [Repository]. Github. https://github.com/LimeSurvey

Mitzka, W., & Schmitt, L. E. (1951). Deutscher Wortatlas (Vol. 1). Wilhelm Schmitz Verlag.

Mitzka, W., & Schmitt, L. E. (1953). Deutscher Wortatlas (Vol. 2). Wilhelm Schmitz Verlag.

Muhr, R. (2001). Varietäten des Österreichischen Deutsch. Revue Belge de Philologie et d'histoire, 79(3), 779–803.

Phillips, A., & Davis, M. (Eds.). (2009). Tags for Identifying Languages (No. RFC5646; p. RFC5646). RFC Editor. https://doi.org/10.17487/rfc5646

Pluschkovits, M. (2025). ZUSAMMENKEHREN/KEHREN. Atlas Zur Lexikalischen Variation in

Österreich Im 21. Jahrhundert. URL: https://lexat21.lapis-online.at/de/articles/zusammenkehren-kehren

Romary, L. (2025). International standards for the identification and the description of languages and their varieties. In Harmonizing language data Standards for linguistic resources (pp. 35–60). de Gruyter. https://doi.org/10.1515/9783112208212-003

Sandow, R., Grieve, J., Stamp, R., & Schembri, A. (2025). Lexical variation and change: Integrating lexis into variationist sociolinguistics. Language Variation and Change, 37(3), 293–318. Cambridge Core. https://doi.org/10.1017/S0954394525100628

Schmidt, J. E., Dammel, A., Girnth, 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. Unter Mitarbeit von Fischer, Hanna und Ganswindt, Brigitte. (Vols. 1–4, pp. 28–60). https://doi.org/10.1515/9783110261295-002

Stöckle, P., & Pluschkovits, M. (forthcoming). The Structure of Lexical Variation in Austrian Varieties: A Quantitative Analysis. Languages, (Special Issue: Advances in Variationist Linguistics on German—Focus on Lexis and Pragmatics).

Tagliamonte, S. (2025). Analysing sociolinguistic variation (Second edition). Cambridge University Press. https://doi.org/10.1017/9781009403092

TEI Consortium (Ed.). (last updated on 4th september 2025). TEI P5: Guidelines for Electronic Text Encoding and Interchange P5 Version 4.10.2. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Ziegler, T., Koppensteiner, W., Gellan, A., Ebner, J., & Lenz, A. N. (forthcoming). Standard Language in Austria: Lexical Variation. In A. N. Lenz, S. Elspaß, & S. M. Newerkla (Eds.), Handbook on German in Austria. Variation – Contact – Perception. Language Science Press.

Żółtak, M., Trognitz, M., & Ďurčo, M. (2022). ARCHE Suite: A Flexible Approach to Repository Metadata Management. 190–199. https://doi.org/10.3384/ecp18917