Methods and Resources in Germanic Variationist Linguistics

John Nerbonne, Groningen, Freiburg & Tübingen

Verena Blaschke, Hinrich Schütze, and Barbara Plank, Munich

Summary

Variationist linguistics, encompassing dialectology and sociolinguistics, studies how linguistic variation is distributed and the dynamics behind the distribution. This article aims to present the most important current resources – methods and data and software archives – for research in Germanic variationist linguists. It is novel to include a chapter on resources in a collection such as this Encyclopedia, so we begin by motivating its inclusion, justifying why contemporary scholars are expected to make resources available to the discipline. With respect to methods, the emphasis is on analytical methods as opposed to methods for field work, site selection, or interviews, and the focus is on software for data analysis. With respect to archives, we emphasize digital repositories. We report on resources important in the variationist research community, i.e., dialectology and sociolinguistics, but also on resources used in the growing community of computational linguists interested in variation.

Keywords

Germanic languages, English, German, Dutch, Scandinavian languages, Luxembourgish, variationist linguistics, dialectology, sociolinguistics, computational linguistics.

Introduction

There is a growing consensus that research resources ought to be openly available to researchers,[i] meaning that data and processing resources should not be the possession of single researchers, groups or even institutes, instead making open data and science a priority. This might sound as if it could conflict with the need for confidentiality for some kinds of data (e.g., interviews in counseling), but there is also agreement that some sorts of linguistic data require confidentiality. Still, any deviation from openness ought to be justified. Before explaining the concept in more detail, let's review its motivation, focusing first on data.

Research data is normally expensive to create. In variationist linguistics obtaining data often involves field work, which in turn requires travel, organization and record keeping (e.g., for permission forms), but minimally requires surveys that need to be designed, and it may just as well involve experimentation, which again calls for design, subject recruitment, experiment administration, data recording and preliminary organization. The biggest expense may often be the time required of researchers and assistants, but this is genuine, and often runs to twenty to fifty percent of grant budgets. And it's a terrible waste when work needs to be duplicated only because researchers are unwilling to part with it. Grant agencies, at least in North America and Europe, agree on the added efficiency of open data and now commonly require that a data management plan, including plans for open archiving, be submitted with proposals. Many journals require that a statement of data availability accompany submission.

In addition to the inefficiency which unshared data represents, there is an added disciplinary advantage to openness, namely that it facilitates reanalysis and replication, two activities that increase the reliability of scholarly work, and thereby discourage fraud in research.

The acronym 'FAIR' has been coined to summarize the properties we look for in open data (Wilkinson et al. 2016), where the letters stand for *findable*, *accessible, interoperable*, and *reusable*. To be *findable*, data must be accompanied by appropriate metadata and indexed by web search engines. Discipline-based collections of data and metadata, one of which will be discussed prominently below, can be particularly useful. Data is *accessible* when researchers can retrieve it, perhaps after identifying themselves, and it is *interoperable* when it is stored in a format that is shared with others and can be integrated in applications of different sorts. Non-proprietary, or at least common formats are preferred. Finally, data is best *reusable* when it is

thoroughly described, and when metadata is rich and clear to colleague researchers (Wikipedia contributors, 2023).

The desire for scientific sharing extends beyond data and there is likewise a movement advocating open standards for research software (Lamprecht et al. 2020). Open-source software makes the source text of its programs public, usually accompanied by a license indicating the sort of use allowed (e.g., use in research, but not in products sold for profit). GitHub and Zenodo serve *de facto* as archives for research software, ensuring openness but not solving making work easily findable. The European CLARIN project maintains an inventory of language resources including both data and analytical resources, which is complemented by a directory (site map). Research software is often reused successfully, but ensuring replicability (one of the aims of reuse) entails the control of factors such as the version of the software, the operating system, as well as the settings those allow. The software we'll review below is in general open-source and freely available.

We proceed from the same set of languages in focus throughout this section of the encyclopedia, i.e., English, German, Dutch and the Scandinavian languages, but in addition all of their dialectal, ethnic and social variants.

Two Communities: Variationist Linguistics and Natural Language Processing

Variationist linguistics is a substantial, traditional subfield of linguistics called into existence by Chambers and Trudgill ([1]1980), which championed the view that dialectology and sociolinguistics ought to be viewed (and practiced) as branches of the same subfield of linguistics. In general, the proposal has met with approval, and leading sociolinguistics journals such as *Language Variation and Change* now also publish articles on dialectology, and leading

dialectological journals such as *Zeitschrift für Dialektologie und Linguistik* publish sociolinguistic articles as well.

Computational Linguistics (CL) or Natural Language Processing (NLP) is a relative newcomer in the study of variation. While the field as such has overall become more insular, with the ratio of linguistics works cited in NLP publications being on a decline (Wahle et al. 2023), there is also an increased interest in studying variation. For example, the specialized Workshop on "NLP for Similar Languages, Varieties and Dialects" (VarDial) has established itself as a long-standing workshop series interested in processing language variation, and has co-located with meetings of the Association for Computational Linguistics (*ACL)
venues for over ten years. In its latest edition, twenty and more papers on language variation are presented (e.g., Scherrer et al. 2023). At the same time, more papers on NLP for dialects appear at premier *ACL conferences. Finally, the CL emphasis on resource sharing augments its importance. We will discuss this further in the NLP section below.

There are several good reasons to present methods and data resources from variationist studies together with those from NLP in spite of the fact that the two research communities are still quite disjoint. Both communities stand to benefit from closer cooperation. Variationist studies can clearly benefit from the tools that have been developed in NLP such as tools for processing corpora (removing formatting, tokenizing words, etc.), and especially tools which can detect latent structure such as parts of speech, morphological boundaries, or phrasal and/or dependency structure, which is particularly difficult. But NLP also stands to gain. First, it is inherent to the goal of general NLP that one be able to process language samples in all languages, and it only stands to reason that social and geographical varieties be included. Ziems et al (2023) argue for the value of attending to varietal differences in developing practical applications in non-American English. Their goal is to avoid performance differences when different varieties are

used (robustness), and to provide tools to modify (partially translate) existing applications to improve their robustness with respect to differences. They used the eWAVE collection (see below) as a data source. Second, variationists have developed theoretical ideas about the distribution of variation, and NLP experts will be in a better position to contribute to the theoretical ideas if they are aware of them. Similarly, variationists know where varieties are found that might benefit from or challenge NLP work. Third, adapting NLP tools is often more efficient when the tools have already been developed for related varieties, as mentioned in the survey by Zampieri et al. (2020).

The departure point of CL means that methodologically sophisticated work is highlighted, and it turns out that work in syntax, semantics and reference are also relatively much more frequent than in the variationist community per se. This suggests to us that a survey of resources for variation in Germanic ought to *include both strands of research*, and we do not attempt to conceal our hope that including work from both research lines might lead to fruitful collaboration.

Methods

In general, the analysis techniques are agnostic about the languages they apply to, so that we needn't restrict ourselves to techniques that have been proven useful on Germanic varieties. We further emphasize that this article will not attempt to *explain* the methods currently in use, although we'll attempt to cite works that do. Instead, we set ourselves the task of identifying resources, i.e., existing software packages, that implement the methods. We judge that a focus on software packages best fits the needs of most variationist linguists. We deliberately omit presentation of methods that rely on apparatus beyond the personal computer, e.g., on laryngoscopy or electromagnetic articulography, even though both have seen some use in

variationist linguistics (Moisik et al. 2014; Wieling et al. 2016). This restriction is not severe with respect to the volume of work, it enables a sharper focus, and we hope that it emphasizes methods that one can more easily adopt.

The restriction to methods available in software packages implies that we must also omit discussion of data collection (the activity, not the results). Four chapters of the methods section of the recent Wiley *Handbook of Dialectology* (Boberg et al. 2018) covered the methods of data collection thoroughly, in particular, sampling methods (Macauley 2018), the use of questionnaires (Llamas 2018), written surveys (Chambers 2018), and field interviews (Bailey 2018). We know of no software packages for planning and conducting field work, even if methods in that line are still topics of research (Wissner 2023). For this reason, we'll structure this section around the linguistic sorts of data under analysis: acoustic phonetics, phonetic transcription, orthographic transcriptions of speech, (tabular collections of) nominal data, and written (and sometimes, edited) language. This choice was made in order to cover the sorts of data popularly studied in variationist linguistics, i.e., phonetic, lexical and syntactic data. After discussing the resources available for analyzing material at the different linguistic levels, we turn to a discussion of the statistical tools used to analyze sets of variationist data.

Acoustic Phonetics

Praat (aka: PRAAT) was developed in the late nineties and the early part of the 2000s by Paul Boersma and David Weeninck at the University of Amsterdam (Boersma & Weenink 1992-2011). It is focused on the analysis of digital acoustic files, operates under all popular operating systems, and is used throughout the phonetic sciences. In particular, it has changed the work of variationist researchers, where it has been immensely popular. Boersma maintains an informative web site on Praat (https://www.fon.hum.uva.nl/praat/)

Labov (1972) championed the use of formant analysis for vowels in sociolinguistics, and it has been a staple among the methods used there ever since. Thomas (2018) notes that dialectologists have been more hesitant in adopting formant analysis, noting honorable exceptions, however (*inter alia* Labov et al. 2006; Eriksson 2004; and Leinonen 2010). It comes as no surprise therefore that Praat has frequently been used for formant extractions and comparisons, and in fact there is work reporting on automating this (Bareda 2021). Thomas (2018) discusses various prerequisite techniques needed for effective use, including vowel extraction, formant normalization, and common graphs used in reporting.

Vowel duration also interests variationist linguistics (Jacewicz et al. 2007), and Praat is a tool of choice for this as well. Voice onset time (VOT) is also regularly studied in production as a subphonemic indicator of social or geographic status (Docherty 2011), but also in perception experiments, for which Winn (2020) offers a Praat script to manipulate VOT. Stress and intonation are likewise examined regularly by variationists (Thomas 2018), and here again, Praat offers useful analysis software.

It is impossible to do justice to all of Praat's capabilities in an article such as this, but one should add that it also comes with script facilities for automating workflows, speech generation facilities, and excellent visualization tools (Boersma & van Heuven 2004), and that it is effective in research in corpus phonology (Boersma 2014). Finally, there is an active community of Praat users, regularly reporting (and fixing) problems as well as contributing useful code (https://groups.io/g/Praat-Users-List).

Yuan & Liberman (2008) introduced a "forced alignment and vowel extraction" technique, which can automatically identify vowels in speech and compare their formant curves, and which has already been used a good deal by sociolinguists, who often focus on vowel comparison.

Bartelds et al. (2023) use an NLP method to improve orthographic transcriptions of speech material, and Bartelds et al. (2022) introduce a technique for comparing acoustic speech samples based on work on large language models (LLMs), which is discussed below.

Phonetic Transcription

Acoustic recordings have the pride of place in the analysis of variationist phonetics, but large collections of dialect pronunciations have been collected since the late nineteenth century, when acoustic recording was just being developed. Even in the twenty-first century, data collections are often transcribed, e.g., to aid in search. Starting with Kessler (1995) there have been numerous attempts to enlist the edit-distance algorithm (Sankoff & Kruskal 1999, [1]1983) in the task of measuring the difference between two phonetic transcriptions of the same word. The algorithm finds the least costly set of operations mapping one sequence to another and its application induces an alignment that is illustrated in Table 1.

Table 1. The alignment induced by the application of the edit-distance algorithm to two pronunciations of the word for 'milk' in the Netherlands (Heeringa 2004). A roughest assignment of operation costs is shown.

Grouw (Frisian)	m	ɔ	l		k:	ə		
Haarlem	m	ɛ	l	ə	k:			
		1		1		1	$\sum$	$= 3$

Heeringa (2004) initiated a series of attempts to refine the operations costs based on the phonetic similarity of the segments involved. Heeringa showed the value of a syllabicity constraint, effectively making consonant-vowel correspondences impossible, and experimented with segment correspondence costs based either on distance in canonical spectrograms or based on

various feature systems. Meta-studies on consistency and validity have focused on aggregate distance (see below). Heeringa et al. (2002) introduced Cronbach's α from psychometrics as a measure of reliability, and showed that sample sizes of 35 or more comparable word pronunciations were reliable ($\alpha \geq 0.8$). Gooskens & Heeringa (2004) showed that aggregate pronunciation distance correlated well with lay judgments of dissimilarity, and Wieling et al. (2012) were able to show that a data-driven approach, inducing segment distances from alignments improved results, and Wieling et al. (2014) validated the measure by showing that it correlated strongly ($r \approx 0.8$) with human judgments of "native-like" pronunciation.

While the procedure was developed on Dutch (Heering 2004), it has also been applied to German (Nerbonne & Siedle 2005), Norwegian (Gooskens & Heeringa 2004), Swedish (Leinonen 2011), and American English (Nerbonne 2006). The citations are merely examples since there have been many applications.

An important competitor to edit distance is List's (2012) Sound Class algorithm, which is focused on application in historical linguistics, and which is similar to edit distance while discounting common historical changes. For example, the substitution of one back unrounded vowel for another ([a] for [ɑ]) would incur no costs in List's system. Another competitor is n-gram analysis (Kondrak 2005), which will not be discussed at length here, but which is sensitive to phonetic context. It measures sequence distance well but has primarily been applied to other areas, e.g., the detection of confusable drug names (Kondrak & Dorr 2006). Heeringa et al. (2006) introduces a context sensitivity by applying edit distance to bigrams and trigrams (pairs and triples, respectively, of segments), and showing that it indeed improves simpler versions.

Bartelds et al. (2020) is an important attempt to apply a continuous version of edit distance, so-called dynamic time warping, to acoustic material directly. Extending this line of work, they

found that neural representations based on pre-trained transformer architectures outperform the previously proposed acoustic features (Bartelds et al., 2022). This line of research holds the promise of wedding the acoustic analysis (above) with the sequence analysis presented in this section. If successful, the technique could obviate the need for phonetic transcription, a difficult and time-consuming step in analysis.

Categorical Data

Data from all linguistic levels can be compared at a categorical level (same vs. different) if it is suitably prepared. Lexical data is almost always analyzed this way. One construes the variable as a concept for which lexicalizations are elicited and recorded. The result may be, e.g., that ten of fifty villages polled use the word *preacher* for the local clergyman, while the other forty use *minister*. These are collected in a data table such as Table 2.

Table 2. Data from two sites on the lexical realization of five concepts, where data is missing from one cell. Normally the fractional overlap or difference among the instantiated concepts is used. The sites agree on one of the four instantiated concepts (25%), and disagree on three (75%).

Site	Lexical Realizations				
	dog	hat	clergyman	toilet	smallest finger
Brownsville	‘dog’	‘hat’	‘preacher’	‘WC’	‘pinkie’
Whiteplain	‘dog’	‘cap’	‘minister’	‘latrine’	Ø

One hundred or more such responses are collected, then one adds the number of like (or unlike) responses (see Table 2), so that the result can be summarized in a site × site (or variety × variety) table – such as Table 3.

Table 3. An excerpt of a table showing mean distances from one site to another. Note that each site is 0.0 units from itself, so that all the cells on the diagonal are zero. Further, distances are symmetric, so the cell (Bucks, Berks) has the same value as (Berks, Buck).

	Philadelphia	Bucks	Montgomery	Berks	Lancaster
Philadelphia	0.000	0.082	0.080	0.168	0.103
Bucks	0.082	0.000	0.075	0.173	0.111
Montgomery	0.080	0.075	0.000	0.170	0.101
Berks	0.168	0.173	0.170	0.000	0.146
Lancaster	0.103	0.111	0.101	0.146	0.000

Séguy (1971) was the first to suggest this sort of analysis, and Goebl (1982) soon followed, and both focused on French and other Romance varieties. Neither Séguy nor Goebl restricted themselves to lexical data, however; instead, they also included phonetic, morphological, and even syntactic data, but always construed categorically. Thus, they might encode whether a voiceless stop is realized as + or −aspirated, or whether a morphological plural is realized as *kine* vs. *cows*, or what order verbs in a cluster take, governed before ungoverned or vice versa. Notice that such features normally require human intervention, both to find relevant examples and then to judge which category they fall into. This makes them less replicable and less rigorous.

Syntactic Analysis

Spruit (2008) analyzed the *Syntactic Atlas of Netherlandic Dutch* (SAND), using Hamming distance (percentage of disgreeing features), in other words, treating the data as categorical. The SAND is a very large collection of judgments of dialect speakers on nearly 1,200 syntactic variables in a wide range of syntactic contexts in 267 locations in the Netherlands and Flanders. Szmrecsanyi & Anderwald (2018), focused on a specialized sort of variationist linguistics, namely World Englishes (Kortmann & Lunkenheimer 2012), emphasize the use of dialect corpora for their closer proximity to unselfconscious (usage-based) speech and suggest that researchers may define a catalog of interesting features (including perhaps probabilistic features) that can be detected automatically in the corpora. Dunn (2019) demonstrates syntactic dialectometry based on the frequency of constructions in the common crawl and Twitter (now X). While his focus is on national varieties, he is able to distinguish dialectal and social varieties well using constructions detected by a parser. His work is also interesting for taking a first step toward evaluating syntactic dialectology.

Natural Language Processing offers various methods for extracting features automatically, facilitating large-scale comparisons of varieties. We discuss two promising tools, part-of-speech tagging (POS tagging) and parsing. A POS tagger assigns one minimal syntactic category to each word in a sentence. In the case of ambiguity, the tagger aims to provide the category of the word as it's understood in the text, and the set of tags may of course vary. We provide a Bavarian example in Table 4.

Table 4. 'The Lammer (river) has fairly clean water' (sentence originally via https://bar.wikipedia.org/wiki/L%C3%A5mma; CC BY-SA 4.0) with Universal Dependencies-

style part-of-speech tags (DET=determiner, PROPN=proper noun, ADV=adverb, ADJ=adjective).

D'	*Lomma*	*hod*	*a*	*rechd*	*a*	*sauwas*	*Wossa*
The	Lammer	has	a	fairly	a	clean	water
DET	PROPN	VERB	DET	ADV	DET	ADJ	NOUN

Nerbonne et al. (2010) use trigrams (three-element sequences) of POS tags to detect shift effects in the English of Finnish immigrants to Australia. The approach is general and might profitably be applied to the comparison of varietal syntax.

Parsing assigns a syntactic structure to a sentence, which naturally also depends on the syntactic theory behind the work, and on the quality of the parser. A popular choice in today's NLP is dependency grammar, in which dependent phrases and words are linked to their grammatical heads, and where the link is decorated with the sort of dependency involved. The Universal Dependencies (UD) project provides dependency analyses for corpora in nearly 150 languages (as of 01.12.2023), where the analyses are linguistically informed and validated (de Marneffe et al. 2021), even if the corpora vary a lot in size. The analysis graphs are dependency parse trees, as Figure 1 illustrates. Levshina (2019) has used the UD analyses for a quantitative study of typology, suggesting that similar efforts would be promising in varietal syntax.

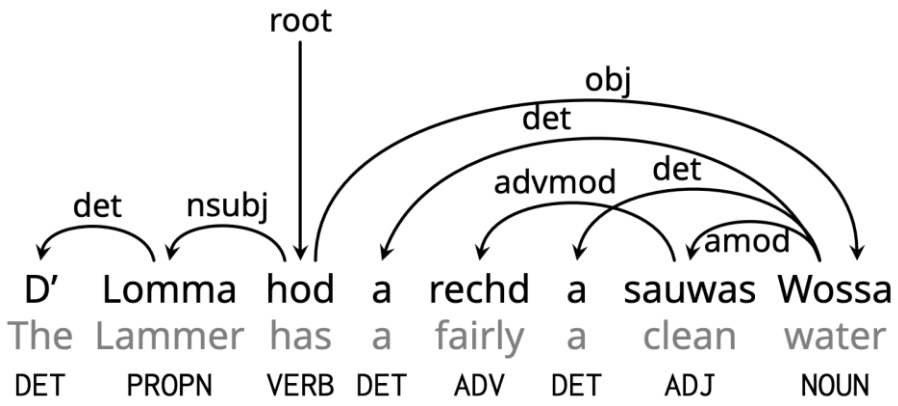


Figure 1. The same sentence as in Table 4, with Universal Dependency annotations (det=determiner, nsubj=nominal subject, obj=object, advmod=adverbial modifier, amod=adjectival modifier). The (unannotated) sentence is originally from Wikipedia (via https://bar.wikipedia.org/wiki/L%C3%A5mma; CC BY-SA 4.0) and the annotations were published by Blaschke et al. (2023b).

The UD project offers tools for querying the corpora and counting instances of a specified configuration.

Natural Language Processing

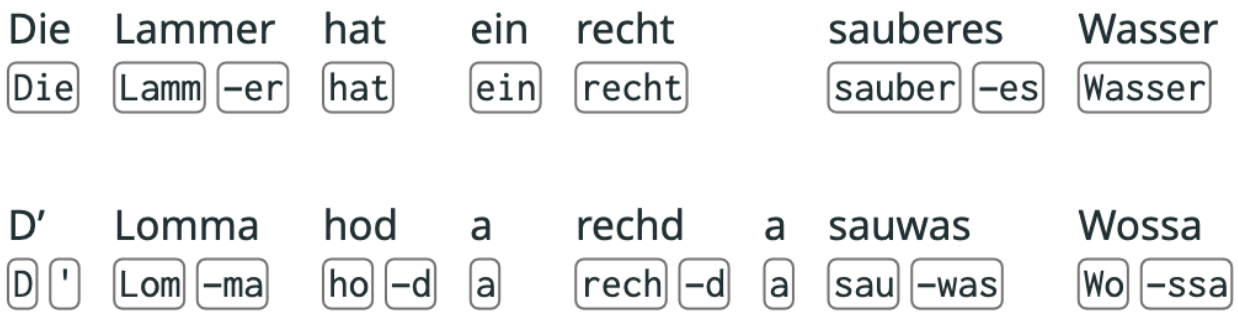
NLP has largely focused on modeling and annotating data from standard languages with many speakers. More recent work has expanded this focus to exploring language variation and learning from small amounts of (often non-standard) data in the context of NLP. We here present a brief overview of such efforts; for more information see Zampieri et al. (2020).

Language models are currently omnipresent in NLP. They are used to encode text for applications such as the ones mentioned above (POS tagging, parsing), but can also be used for language generation (e.g., machine translation, chatbots). Their first step to processing text input is segmenting it into small units ("subwords"). These subwords are generally shorter than words.

Common character sequences form units, and less common sequences tend to be split up into smaller subwords. This segmentation typically purely depends on frequency statistics in a large corpus (rather than linguistically-informed morphological splits). The top of Table 5 shows the subword segmentation of a German sentence produced by a German language model. Mielke et al. (2021) present an overview of segmentation techniques. Each subword is associated with a vector (similar to vectors in distributional semantics; Sinha et al., 2021), and a sequence of such vectors serves as the input for all subsequent calculations. A statistically meaningful division into subwords can easily be constructed for corpora with vast amounts of data written in a predictable orthography, in which case the subword segmentation remains useful when applied to new data from the same language.

Representing texts written in a language for which little data is present, and/or where there is substantial orthographic variation (as a result of phonetic variation and idiosyncratic choices) thus poses a challenge. A model that works very well on German data might perform very poorly when evaluated on Bavarian data (Table 5). Training a new model on Bavarian data might not be possible due to a lack of available data. Several approaches have been suggested that might mitigate the issue: keeping the subword-based models but changing the way that, e.g., German is split into subwords so that the model's subword representations are more favorable for transfer to a related, non-standardized variety (Aepli & Sennrich, 2022), or doing away with the popular subword representations in favor of purely character-based models (El Boukkouri et al., 2020) or visual representations (Salesky et al., 2021).

Table 5. 'The Lammer (river) has fairly clean water', in German (top) and Bavarian (bottom), as split into subword tokens by the German language model GBERT (Chan et al., 2020).



Besides orthographic variation providing a challenge to NLP methods, recent work has also investigated the (lack of) robustness towards syntactic variation, for instance in the context of syntactic differences between English dialects (Ziems et al. 2023). In studies like this one or the one by Kantharuban et al. (2023), the authors make use of insights from dialectology to analyze shortcomings of language models.

Other NLP work focuses on applications for laypeople that should be robust to, or embrace, linguistic variation. This includes nation-level variation, for instance in the context of text retrieval models that ought to not be affected by English spelling variation (Chari et al., 2023) or machine translation models that should translate into the appropriate national variety (Riley et al., 2023). Other lines of research have focused on extending virtual assistants to queries in Bavarian and Swiss German (van der Goot et al., 2021; Aepli et al., 2023) and on automatically creating German transcriptions for Swiss German audio data (Plüss et al., 2020, Gerlach et al., 2022). The latter task not only involves transcribing speech, but also adjusting the syntax to account for structural differences between Swiss and standard German.

For variationist studies, NLP can be a useful tool in several ways: data can be (pre-)annotated automatically with morphosyntactic (or other) information and audio data can be transcribed. We can also use language/dialect identification techniques (Zampieri et al., 2020) to create web

corpora in relevant language varieties. Additionally, neural representations can be used to analyze distances between linguistic features or varieties (Demszky et al. 2020; Kuparinen & Scherrer 2023; Hovy & Purschke, 2018; Nguyen & Grieve, 2020).

Aggregating Individual Differences

Although phonetics differences are often studied without summing over multiple variables (the formant differences of multiple vowels, perhaps), dialectometry normally proceeds by aggregating over many variables (Séguy 1971; Goebl 1982). Nerbonne (2009) reminds dialectologists that moving to an aggregate level is justified once one wishes to characterize the variety as a whole, so that exceptions become less vexing, and that aggregation obviates the need to select a small number of variables for closer study. The major advantage is the opportunity it supports to characterize general tendencies in the distribution of linguistic variation. He further argues that abstracting from details in fieldwork recordings is also tantamount to an aggregating step, albeit one restricted to a single variable.

In an article on methods in variationist linguistics, it is imperative to note that, while dialectometry has adopted the aggregating perspective enthusiastically, sociolinguistics has not. Nerbonne et al. (2013) speculate that many changes triggered by social factors, such as standardization efforts, school reforms, or migrations, deserve the sort of more comprehensive examination made popular in dialectometry, but a key desideratum in sociolinguistics, that of studying (individual) ongoing sound changes, is not in focus (but see below).

We add here a point often glossed over in introductions, namely that many of the further analytical steps presently require that the aggregating step result in distances in the mathematical sense, i.e., a characterization which is symmetric $d(v_1, v_2) = d(v_2, v_1)$, where the distance of any variety to itself is zero $d(v_i, v_i) = 0$, and where the closest distance of any

element to any other is always direct, never via a third $d(v_1, v_2) \leq d(v_1, v_i) + d(v_i, v_2) \, \forall v_i$. This requires care in situations where, e.g., there may be multiple responses (Aurrekoetxea 2020). Principal component analysis and factor analysis do not require distance tables as inputs, however, and have also been used, albeit less often (Pröll et al. 2021; Nerbonne 2006).

We turn to the analyses of distance matrices. Although there are specialized packages for analyzing variationist data, virtually all of them use or are indebted to the open-source R-project for statistical computing (https://www.r-project.org/). Anyone interested in novel statistics for variationist analysis would do well to consult it.

Clustering

Traditional dialectology often concluded that dialect varieties were geographically distributed into discrete regions, which amount to a partition of the data collection sites. Goebl (1982) introduced clustering, which detects the most similar groups in a distance table. Clustering outputs a dendrogram, a tree of dialect similarity. The quality of clustering is measured by the cophenetic correlation, i.e., the degree to which the distances in the input table correlate with the distances in the output dendrogram. The groups detected by clustering normally project geographically to regions, enabling a comparison to older work. Simple clustering is unstable, meaning that small differences in the input can lead to very different results, leading Mucha & Haimerl (2005) and others to suggest so-called bootstrap methods. Wieling & Nerbonne (2011) introduced a further refinement, namely bipartite spectral graph partitioning, where varieties together with their characteristic features are partitioned into groups.

Multidimensional Scaling

A further opportunity to explore distance tables is offered by multidimensional scaling (MDS), first introduced to dialectometry by Embleton (1993). Given a set of data collection sites and the

distance between all the pairs of sites, MDS provides as good a representation as possible of the differences among them, normally in a small number of dimensions. The quality of the dimension reduction is measured in stress, a degree of distortion, or correlation with input distances. One of the measures should always be reported. Dialectometrists using MDS can often represent large sets of sites faithfully (i.e., with $r \geq 0.87, r^2 \geq 0.8$) representing 80% and more of the variance) in only three dimensions, leading to insightful three-color maps (introduced in Nerbonne et al. 1999). The visualization of MDS illustrates how continuous the geographic distribution of variation is, thus supplementing the partitioning view induced by clustering. Some software packages allow researchers to compare clustering and MDS results e.g., Gabmap (Nerbonne et al. 2011), enabling a check on clustering results.

Analyzing Distance Matrices

The aggregate view on variation has spawned new works analyzing the influence of social, geographic, and linguistic features on linguistic differences.

Geographic Influence and Generalized Additive Modeling

Dialectometry organizes the aggregate sums of differences in distance tables (between all pairs of sites such as Table 3 above), and analyzes their dependence on geographic distance using regression (Séguy 1973), and reporting the correlation coefficient as a measure of quality. Because distances are not independent measures, it is important to check significance using a Mantel test (Mantel 1967).

No one has ever postulated that space directly influences language, but distance can serve as a proxy for the chance of contact. Gooskens (2005) showed that travel time was a better predictor than simple distance, confirming the fundamental idea. Nerbonne (2010) examined six language

areas, including four Germanic areas, showing that aggregate linguistic differences were predictable based on the logarithm of geographic distance, where $0.16 \leq r^2 \leq 0.37$. Distance is a simplified, one-dimensional reduction of geography, which prompted Wieling (2012) to apply generalized additive modeling (GAM) to Dutch dialect differences (see Figure 2). GAMs enable the analysis of explanatory variables in potentially non-linear combinations. Technically, functions representing the interaction of the individual variables are added and optimized, in our case modeling the interaction between longitude and latitude. It is worth noting that the regression step estimates the distances from individual sites to a single alternative, namely the standard language. The result therefore depicts not the dialect landscape, but only the degree of difference to the standard.

A recent addition to the regression family of techniques is multiple regression on linguistic distance matrices (MRM), which Huisman et al. (2021) apply to the Dutch-Belgian Limburgish dialect continuum. The authors use MRM to analyze the multiple effects of geographic distance, population size, separation by water, national border, dialect area, semantic density, and concept salience to provide a more comprehensive analysis of the influences on dialect differences.

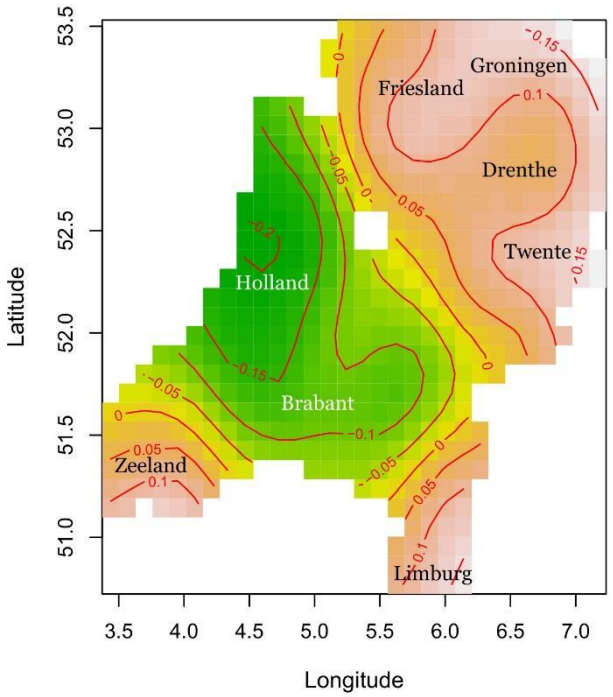


Figure 2. The result of applying Generalized Additive Modeling (GAM) to pronunciation in the Netherlands. Here the influence of geography can be measured outside the strictures imposed by linear (or logarithmic) distance. Dark green represents pronunciations close to standard Dutch, light beige very different pronunciations. From Wieling (2012:90).

Aggregates and Individual Linguistic Items

The aggregating step described above highlights the properties of varieties but comes with a clear disadvantage, namely that differences in individual variables risk being obscured among the dozens or even hundreds of variables they are combined with. For this reason, there has been continuous interest in techniques that combine aggregate and individual perspectives. Heeringa (2004: 267) calculates the correlations between sample words and the most important MDS dimensions, and Prokić et al. (2012) introduce a technique for determining the most characteristic elements in dialect regions. Rubehn et al. (2024) extend this line of work to detect characteristic sound correspondences.

Mixed-Effects Modeling

Johnson (2009) had introduced mixed-effects regression modeling to sociolinguistics, demonstrating several advantages over the logistic regression model which had seen most use in sociolinguistics. A mixed-effects model can use all the independent variables ("fixed effects") used in a standard regression model, to which so-called random effects are then added. Fixed effects distinguish only a few classes, such as gender or educational level, while random effects are used to model individual elements, for example speakers, data collection sites or individual linguistic items such as words. Wieling (2012) introduced mixed-effects modeling into dialectology, focusing on Dutch pronunciation.

This section demonstrates that contemporary variationist linguistics is no longer subject to the criticism leveled by Woolhiser (2005) and Loporcarno (2009). The relation between aggregate analyses and their linguistic foundations can be adduced using more modern techniques.

We have not exhausted the sorts of analyses often used in variationist linguistics, and would like to briefly mention worthwhile work that hasn't been discussed above. Geographic information systems (GIS) are popular throughout natural and cultural studies where the relation to geography is central, e.g., ecology, engineering, and demography. GIS has developed measures of the strength of (spatial) autocorrelation, the strength of spatial dependence and continuity, and interpolation techniques. Grieve (2018) has championed their use in variationist linguistics.

Before concluding this Methods section, we refer the reader to the extensive list of software packages available for analyzing variationist data (Links to digital material).

Data

When we turn to finding relevant data, we note that there is an international Registry of Research Data Repositories (https://re3data.org), which aims to maximize data discovery, for data archived

using the FAIR principles (see section above on Motivation), but the registry still provides rather few links to linguistic data. We mention it here because we suspect that, if the impetus toward data sharing is to be successful, higher-level registries such as r3data.org will be as necessary as will the data repositories themselves, perhaps organized politically, or perhaps by discipline.

Berez-Kroeker et al. (2022) is a comprehensive survey with recommendations for good linguistic data management. There are six chapters of "use cases" dealing with variationist data, e.g., Kendall and Farrington's (2022) exposition of the ideas behind CORAAL, the *Corpus of Regional African American Language*. This would be a good source to consult for advice on preparing data for archiving, including acoustic and audiovisual data.

We focus here on data available in a form that allows one to download entire datasets. We regret needing therefore to ignore data that has served variationist linguistics admirably, such as the Survey of English Dialects (SED) (Orton 1962),[ii] which the British library does make available in a site focused on popular education, but which limits downloads to a small sample of recordings. This is undoubtedly useful, e.g., for checking on data, but most contemporary research will wish to proceed from more comprehensive data samples.

We turn to data resources for studying variation in Germanic languages. Blaschke et al. (2023b) provide a rigorous survey of available data, albeit with a focus on data interesting for experimentation in NLP. We summarize this survey here and the collection of resources that has continued growing after its publication, by now including over 100 accessible and downloadable datasets focusing on variation in Germanic languages and/or more broadly on Germanic languages with few speakers. These differ in a variety of aspects that we explain below.

First, datasets differ regarding the *research purpose* they were created for and the *sources* of the underlying data. For instance, the LIA project is based on a large collection of audio recordings

made for language documentation purposes (Hagen et al., 2021a) and has been used to research linguistic variation across Norwegian dialects (Hagen et al., 2021b, *inter alia*). The recordings were transcribed and a selection was later annotated with syntactic information (Øvrelid et al., 2018), which has been used for NLP research on non-standard varieties (Blaschke et al., 2023a). Corpora like the Swiss German ArchiMob (Samardžić et al., 2016) and Swiss Parliaments Corpus (Plüss et al., 2021) are also based on transcribed and (automatically) annotated audio data, but use audio data collected for non-linguistic purposes as their basis. Atlas-like datasets like the Sound Comparisons project (Paschen et al., 2019, *inter alia*) document phonetic variation across sets of cognates and related language varieties. Other corpora are created directly with specific NLP tasks in mind, for instance the SwissDial dataset (Dogan-Schönberger et al., 2021) for Swiss German speech recognition and synthesis as well as topic classification. Yet others, like OSCAR (Abadji et al., 2022), are the product of applying language identification tools to vast amounts of web-crawled data, resulting in a large corpus containing subcorpora of documents that are likely to be expressed in, e.g., West Frisian, Low Saxon, Luxembourgish, and Swiss German. Some other web-based corpora only include social media data, like the African-American Vernacular English TwitterAAE dataset (Blodgett et al., 2018), or data from contributors to collaborative, language-specific projects, such as the example sentence collection on Tatoeba or the array of Wikipedias in various Germanic languages. With uncurated or loosely curated web corpora, it is important to be wary of data quality issues, such as the potentially low accuracy of language identification tools and the possible inclusion of non-linguistic material like random character sequences or HTML remnants (Kreutzer et al., 2022), as well as the possibility of texts being written by writers not proficient in the language variety (like much of

the Scots Wikipedia being written by a non-speaker whose writing merely mimicked Scots; Brooks & Hern, 2020).

Second, datasets differ in their *modality* – are the data written or in audio form or both? If a dataset is written, the *written representation* can also differ greatly from that of other datasets. This variability in written representations has also been discussed by Tagliamonte (2007) and Gaeta et al. (2022). We can broadly distinguish between four written styles, for which we give examples in Table 6 (for more examples, see Blaschke et al., 2023b). Data can be written in a *widespread normalized orthography,* for instance a Norwegian Bokmål transcription of Norwegian audio data (1a), but there are also cases where a language is transcribed in the orthography of a closely related standard language, like Elfdalian transcribed in Swedish (2b) or Swiss German in standard German (4a). Some low-resource languages also have their *own orthographies,* e.g., the Elfdalian standard orthography that was introduced in 2005 (2a) or the Nysassiske Skryvwyse (one of multiple proposed orthographies for Low Saxon; 3a). Yet other text samples are *phonetic or phonemic transcriptions,* which vary from close transcriptions like the modified X-SAMPA used in the NB Tale corpus (1b), which can easily be converted into IPA (1c), to broader, orthographically inspired transcriptions like the one used in ArchiMob (4b). Other texts are written in *ad-hoc idiosyncratic pronunciation spellings,* like sentence 3a in the table.

Table 6. Examples of different written representation types, based on Blaschke et al. (2023b). The superscript 1 and 2 in example 1c are used to express Norwegian pitch accent tones.

1a	Har	du	noen	gang	sett	stokkmaur	[...]
1b	h"A:	d`"}:	n""u:@N	g"AN	s"et	st""Okm%A}4	[...]

1c [1]hɒː [1]ɖʉː [2]nuəŋ [1]gɒŋ [1]set [2]stɔkmˌɒʁ [...]

‘Have you ever seen carpenter ants [...]?’

From the Norwegian NB Tale corpus (Språkbanken).

2a wen wa wen war eð før ien månað ? juni ?

2b vad va- vad var det för en månad ? juni ?

‘What, wa-, what month was it? June?’

From the Elfdalian part of the Nordic Dialect Corpus (Johannessen et al., 2009).

3a He hadde ene Frau mit fiev Kinder [...]

3b Hee hadde eyne vrouw mid vyv kinder [...]

‘He had a wife and five children [...]’

From UD Low Saxon LSDC (Siewert et al., 2021).

4a können sie ihre jugendzeit beschreiben

4b chönd sii iri jugendziit beschriibe

‘Can you describe your youth?’

From the Swiss German ArchiMob corpus (Samardžić et al., 2016).

Third, some datasets are *annotated* while others are not. Some datasets have automatically generated annotations, like part-of-speech tags in the South Tyrolean DiDi corpus (Frey et al., 2016), but we here focus on annotations that were either done manually or that were produced

automatically but manually corrected. Most of the annotated datasets we are aware of come with *morphosyntactic* annotations: most commonly, part-of-speech tags as in NorDial (Mæhlum et al., 2022) , syntactic annotations like dependencies as in TwitterAAE (Blodgett et al., 2018) or phrase structure annotations like for the Swiss German corpus by Schönenberger and Haeberli (2019), and morphological details (Siewert et al., 2021), and there are also corpora that include several of these dimensions. A few datasets also contain information about specific syntactic phenomena, like the subset of *Stemmen uit het verleden* (Van Keymeulen et al., 2019) annotated by Lybaert et al. (2019) or the Nordic Word Order Database (Lundquist et al., 2019). A smaller number of datasets comes with *content-related* annotations, such as the subset of SB-CH annotated with sentiment judgments (Grubenmann et al. 2018), the L-WNLI dataset (Lothritz et al., 2022) with textual entailment annotations for pairs of statements, and the xSID/SID4LR dataset (van der Goot et al., 2021; Aepli et al., 2023), which provides manual annotations related to conversational intent. Additionally, some datasets provide *parallel data*: xSID also provides sentence-level translations into other languages, TaPaCo (Scherrer, 2020) contains paraphrases of Low Saxon and Gronings sentences, and the Wenkersätze collection (Wenker, 1889–1923; Schmidt et al., 2020–) provides translations of German sentences into many different German dialects and regional languages (more on this below). Datasets providing both audio and written data, e.g., the Upper Saxon SXUCorpus (Herms et al., 2016), and/or different kinds of written representations (as in Table 6) could arguably also be included in this "parallel data" category. As mentioned in the previous paragraph, some datasets compare *many varieties in parallel* whereas others focus on a *single variety*, for instance Texas German (Blevins 2022) or Walser German (Garner et al. 2014). Those including a number of varieties either cover a large number of varieties from a larger geographic region (e.g., Paschen et al., 2019, or Rabanus et al., 2023)

or focus on a more concentrated selection, e.g., the different Faroese dialects in the dataset by Simonsen et al. (2023).

The overall trends we can make out are that there are comparatively few (manually) annotated resources, and datasets are distributed unevenly across language varieties (with Swiss German being especially well-represented among Germanic low-resource varieties). The website https://github.com/mainlp/germanic-lrl-corpora contains currently the best overview of the datasets for Germanic low-resource varieties that we are aware of and that are downloadable (at least in part) for academic research. We invite all researchers to contribute suggestions to this repository for any relevant datasets we may have missed.[iii]

Conclusions and Prospects

Our first goal was to present an overview of the most important resources for studying German variationist linguistics, i.e., dialectology and sociology. We've paid attention to both methods and data. It has become apparent that work in NLP is taking a variationist turn, as witnessed by the ten-year series of VarDial workshops and the extensive registry of data from low-resource Germanic languages (Blaschke et al. 2023b), most of which had been culled from reports on computational research. The time seems ripe therefore to address the variationist and the NLP communities jointly, perhaps stimulating a closer collaboration between them. This was a secondary goal of the overview.

We argued above that both groups stand to benefit from collaboration, the variationists from the computational tools in NLP and the NLP-ers from the variationist' awareness of where dialectal and social differences are found. But there may be further opportunity for collaboration in applied work, i.e., work aimed at improving products or processes in business, government or elsewhere, e.g., work in assisting speakers with standard language, work in instructing

newcomers in areas where dialects are prevalent, or work in detecting speakers' profiles for marketing purposes. There is an enormous popular interest in language variation that any collaboration would benefit from.

We have not included work in social media in this article as it falls outside the usual framework of variationist linguistics. But it is clear that the language of social media is pervaded by dialectal and social features. Eisenstein et al. (2014) is a computationally sophisticated study of Twitter that examines lexical diffusion in Twitter and detects non-standard geographical and social features, and it has spurred a good deal of work among computational linguists. Nguyen et al. (2020) surveys more of this work.

Prospects

We noted above that syntactic variation is an area where the variationist-NLP collaboration seems poised to blossom. We have barely mentioned morphology because the amount of work is smaller, but computational tools are well developed for morphological analysis, which might serve as a tool to detect latent structure for the purpose of comparison.

Validation of methods remains underdeveloped. Reflecting the view that variation serves as an indicator of provenance, Gooskens & Heeringa (2004) introduced a perception-based validation of edit distance, effectively seeking correlations of edit distances with lay judgments of Norwegian dialect similarity. Wieling et al. (2014) used a similar scheme to demonstrate the superiority of PMI-based edit distance (see above in Section "Edit Distance"). But so far, validation has been applied only to the aggregate levels of human judgments and edit-distance measures. Validation based on single-utterance comparisons would enable more sensitive comparison, and the validation of lexical and syntactic measures has just begun.

Scholarship and Further Reading

Wieling & Nerbonne (2015) is a good overview of developments in dialectometry until 2014. Several articles in the OUP Research Encyclopedia of Linguistics are useful background, including "The History of Variationist Germanic Linguistics," Social Variation in Germanic," "Usage-based Approaches to Germanic Languages," and "Corpus and Computational Linguistic Approaches to Germanic Languages."

Links to Digital Material on Methods and Data

Methods

The R-project, the open-source project for statistical computing. Within R, ShinyDialect is a web application emphasizing tools for drawing isoglosses, and LED-A is a web application that offers many of the functionalities of Gabmap (see below) while adding functionality to compute not only aggregate differences but also differences between individual words. Its user interface is also novel. DialectR aims to facilitate the integration of dialectometric tools within novel R workflows.

Praat: Doing phonetics by computer.

The Bavarian Archive for Speech Signals (BAS). Munich, offers tools for segmentation and labeling speech, including dialectal material. See MAUS, in particular.

Gabmap, a web application for conducting dialectometric analyses.

Diatech, under development since 2013, which aims to create a web application emulating Goebl and Haimerl's downloadable Visual Dialectometry, which is no longer supported.

Geoling focuses on the analysis of individual linguistic variables.

Although LingPy (List & Forkel, 2021) focuses on tools for historical linguistics, it has also been used for sequence comparison and aggregate analyses.

The REDE SprachGIS (Schmidt et al., 2020–) provides tools for mapping and combining data from a range of German dialect atlases.

Data

Blaschke et al. (2023b) is a thorough summary of data available for research data on variation in Germanic, including data from the Institut für deutsche Sprache and Deutsch heute such as *Archiv für gesprochenes Deutsch* and *Deutsch heute*. An updated online version is available at https://github.com/mainlp/germanic-lrl-corpora.

Fischer & Limper (2019) provide an overview of online platforms and applications for German dialects, with an updated list of resources available at https://regionalsprache.de/regionalsprachenforschung-online.aspx.

The SPeech Across Dialects of English (SPADE) project shares a collection of speech corpora for studying variation in British and North American English dialects: https://osf.io/4jfrm/.

Acknowledgments

For providing information on various sources we are grateful to Alexandra d'Arcy, Sheila Embleton, Marina Frank, Peter Gilles, Elvira Glaser, Katharina Korecky-Kröll, William Kretzschmar, Jr., Alfred Lameli, Alexandra Lenz, Warren Maguire, Jonnie Robinson, Yves Scherrer, Philipp Stöckle, Matthew Sung, Clive Upton, Eric Wheeler, Martijn Wieling, and the members of Nerbonne's 2023 class in dialectometry in Tübingen.

References

Abadji, J., Ortiz Suarez, P., Romary, L., & Sagot, B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4344–4355. European Language Resources Association.

Aepli, N., Çöltekin, Ç., van der Goot, R., Jauhiainen, T., Kazzaz, M., Ljubešić, N., North, K., Plank, B., Scherrer, Y., & Zampieri, M. (2023). Findings of the VarDial Evaluation Campaign 2023. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 251–261. Dubrovnik, Croatia: Association for Computational Linguistics.

Aepli, N., & Sennrich, R. (2022). Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. *Findings of the Association for Computational Linguistics: ACL 2022*, 4074–4083. Dublin, Ireland: Association for Computational Linguistics.

Aurrekoetxea, G., Nerbonne, J., & Rubio, J. (2020). Unifying analyses of multiple responses. *Dialectologia: revista electrònica*, 59-86.

Bailey, G. (2018). Field interviews in dialectology. In Boberg et al. (2018), 284-299.

Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, *7*(1), 20200051. DOI:10.1515/lingvan-2020-0051

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020). A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence* 3.

Bartelds, M., de Vries, W., Sanal, F., Richter, C., Liberman, M., & Wieling, M. (2022). Neural representations for modeling variation in speech. *Journal of Phonetics* 92.

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. Proc. 61th Annual Meeting of the ACL. Shroudsburg, PA: ACL. 715-729. 10.18653/v1/2023.acl-long.42

Berez-Kroeker, A. L., McDonnell, D., Koller, E., Collister L. B. (2022). The Open Handbook of Linguistic Data Management. Cambridge, MA.: MIT Press.

Blaschke, V., Schütze, H., & Plank, B. (2023a). Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 40–54. Association for Computational Linguistics.

Blaschke, V., Schütze, H., & Plank, B. (2023b). A Survey of Corpora for Germanic Low-Resource Languages and Dialects. *Proc. 24th Nordic Conf. Computational Linguistics (NoDaLiDa)*. 392-419.

Blevins, M. (2022). *Texas German Sample Corpus* [Data set]. Texas Data Repository. https://doi.org/10.18738/T8/IOX9ZA

Blodgett, S. L., Wei, J., & O'Connor, B. (2018). Twitter Universal Dependency parsing for African-American and mainstream American English. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1415–1425. Association for Computational Linguistics.

Boberg, C., Nerbonne, J., & Watt, D. (Eds.). (2018). *The handbook of dialectology*. Boston: John Wiley & Sons.

Boersma, P. (2014). The use of Praat in corpus research. In: Durand, J. et al. (Eds.). (2014). *The Oxford handbook of corpus phonology*. 342–360. Oxford: Oxford University Press.

Boersma, P., & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, 5(9/10), 341-347.

Boersma, P. & Weenink, D. (1992–2022): Praat: doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

Brooks, L., & Hern, A. (2020, August 26). Shock an aw: US teenager wrote huge slice of Scots Wikipedia. *The Guardian.*

Chambers, J. K. (2018). Written dialect surveys. In: Boberg et al. (2018), 268-283.

Chambers, J. K., & Trudgill, P. ([1]1980, 1998). *Dialectology*. Cambridge: Cambridge University Press.

Chan, B., Schweter, S., & Möller, T. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796International Committee on Computational Linguistics.

Chari, A., MacAvaney, S., & Ounis, I. (2023). On the effects of regional spelling conventions in retrieval models. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2220–2224. New York, USA: Association for Computing Machinery.

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255-308.

Docherty, G., Watt, D., Llamas, C., Hall, D., & Nycz, J. (2011). Variation in Voice Onset Time along the Scottish-English Border. In *ICPhS* (Vol. 4), 591-594.

Dogan-Schönberger, P., Mäder, J., & Hofmann, T. (2021). *SwissDial: Parallel multidialectal corpus of spoken Swiss German.* ArXiv.

Dunn, J. (2019). Global syntactic variation in seven languages: Toward a computational dialectology. *Frontiers in Artificial Intelligence*, *2*, 15.

El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., & Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. *Proceedings of the 28th International Conference on Computational Linguistics*, 6903–6915. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Embleton, S. (1993). Multidimensional scaling as a dialectometrical technique: Outline of a research project. In *Contributions to Quantitative Linguistics: Proc. 1st International Conf. Quantitative Linguistics, QUALICO.* 267-276. Springer: Berlin & New York.

Eriksson, A. (2004). SweDia 2000: A Swedish dialect database. In: *Babylonian Confusion Resolved. Proceedings of the Nordic Symposium on the Comparison of Languages*. Copenhagen Business School, CBS. Copenhagen Working Papers in LSP, No. 2004-01 33-48. https://hdl.handle.net/10398/5761c2c0-c021-11db-9769-000ea68e967b.

Fischer, H., & Limper, J. (2019). Regionalsprachliche Forschungsergebnisse online. In: Herrgen, J., & Schmidt, J. E. (Hrsg.): *Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Band 4: Deutsch.* Unter Mitarbeit von Hanna Fischer und Brigitte Ganswindt. Berlin/Boston: De Gruyter Mouton. (Handbücher zur Sprach- und Kommunikationswissenschaft. 30.4), 879-897.

Frey, J.-C., Glaznieks, A., & Stemle, E. W. (2016). The DiDi corpus of South Tyrolean CMC data: A multilingual corpus of Facebook texts. *Proc. Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016).*

Gaeta, L., Angster, M., Cioffi, R., & Bellante, M. (2022). Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations. *Corpus* 23.

Garner, P. N., Imseng, D., & Meyer, T. (2014). Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch. *Proc. Interspeech 2014*, 2118-2122.

Gerlach, J., Mutal, J., & Bouillon, P. 2022. Producing Standard German subtitles for Swiss German TV content. *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, 37–43. Dublin, Ireland: Association for Computational Linguistics.

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

Gooskens, C., & Heeringa, W. (2004) Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, *16*(3):189-207.

Grieve, J. (2018). Spatial statistics for dialectology. In Boberg et al. (2018), 415-433.

Grubenmann, R., Tuggener, D., von Däniken, P., Deriu, J., & Cieliebak, M. (2018). SB-CH: A Swiss German Corpus with Sentiment Annotations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). European Language Resources Association (ELRA).

Hagen, K., Kristoffersen, G., Vangsnes, Ø. A., Åfarli, T. A. (Eds.). (2021a). *Språk i arkiva: Ny forsking om eldre talemål frå LIA-prosjektet* (pp. 7–17). Novus forlag.

Hagen, K., Vangsnes, Ø. A., Åfarli, T. A. (2021b). Om LIA-prosjektet og artiklane i denne boka. In Hagen et al. (2021a).

Heeringa, W. J. (2004). Measuring dialect pronunciation differences using Levenshtein distance. PhD diss., University of Groningen.

Heeringa, W., Nerbonne, J., & Kleiweg, P. (2002). Validating dialect comparison methods. In *Proc. 24th Conf. Gesellschaft Klassifikation, Passau 2000*, 445-452. Berlin: Springer.

Herms, R., Seelig, L., Münch, S., & Eibl, M. (2016). A corpus of read and spontaneous Upper Saxon German speech for ASR evaluation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4648–4651. European Language Resources Association (ELRA).

Hovy, D., & Purschke, C. (2018). Capturing regional variation with distributed place representations and geographic retrofitting. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4383–4394. Association for Computational Linguistics.

Huisman, J. L., Franco, K., & van Hout, R. (2021). Linking linguistic and geographic distance in four semantic domains: computational geo-analyses of internal and external factors in a dialect continuum. *Frontiers in artificial intelligence*, *4*, 668035.

Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, *112*(41), 12752-12757.

Johannessen, J. B., Priestley, J. J., Hagen, K., Åfarli, T. A., & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus–an advanced research tool. *Proceedings of the 17th Nordic*

Conference of Computational Linguistics (NODALIDA 2009), 73–80. Northern European Association for Language Technology (NEALT).

Johnson, D. E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and linguistics compass*, *3*(1), 359-383.

Kantharuban, A., Vulić, I., & Korhonen, A. (2023). Quantifying the dialect gap and its correlates across languages. *The 2023 Conference on Empirical Methods in Natural Language Processing.* Singapore: Association for Computational Linguistics.

Kendall, T., & Farrington, C. (2022). Managing Sociolinguistic Data with the Corpus of Regional African American Language (CORAAL). In: Berez-Kroeker et al. (Eds.) (2022), 185-193. https://doi.org/10.7551/mitpress/12200.003.0019

Kessler, B. (1995). Computational dialectology in Irish Gaelic. *7th Conf. European Chapter ACL*, 60-66. Dublin, Ireland. Association for Computational Linguistics.

Kondrak, G. (2005) N-gram similarity and distance. In: *International symposium on string processing and information retrieval*, 115-126. Berlin: Springer.

Kondrak, G., & Dorr, B. (2006). Automatic identification of confusable drug names. *Artificial intelligence in medicine*, *36*(1), 29-42.

Kortmann, B., & Lunkenheimer, K. (Eds.). (2012). *The Mouton world atlas of variation in English*. Berlin: Mouton de Gruyter.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Ortiz Suarez, P., , … Adeyemi, M.

(2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics,* 10, 50–72.

Kuparinen, O., & Scherrer, Y. (2023). Dialect representation learning with neural dialect-to-standard normalization. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023),* 200–212. Association for Computational Linguistics.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.

Lamprecht, A. L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., ... & Capella-Gutierrez, S. (2020). Towards FAIR principles for research software. *Data Science*, *3*(1), 37-59.

Leinonen, T. (2010). An acoustic analysis of vowel pronunciation in Swedish dialects. PhD diss., University of Groningen.

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, *23*(3), 533-572.

List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117-125.

List, J.-M. and Forkel, R. (2021): *LingPy. A Python library for historical linguistics.* Version 2.6.9. URL: https://lingpy.org. With contributions by Greenhill, S., Tresoldi, T., Rzymski, C., Kaiping, G., Moran, S., Bouda, P., Dellert, J., Rama, T., & Nagel, F. Leizpig: Max Planck Institute for Evolutionary Anthropology.

Llamas, C. (2018). The dialect questionnaire. In: Boberg et al. (2018), 253-267.

Lothritz, C., Lebichot, B., Allix, K., Veiber, L., Bissyande, T., Klein, J., Boytsov, A., Lefebvre, C., & Goujon, A. (2022). LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5080–5089. European Language Resources Association.

Lundquist, B., Larsson, I., Westendorp, M., Tengesdal, E., Nøklestad, A. (2019). Nordic Word Order Database: Motivations, methods, material and infrastructure. *The Nordic Atlas of Language Structures* (NALS) 4(1).

Lybaert, C., De Clerck, B., Saelens, J., & De Cuypere, L. (2019). A corpus-based analysis of V2 variation in West Flemish and French Flemish dialects. *Journal of Germanic Linguistics*, 31(1), 43–100. Cambridge University Press.

Loporcaro, M. (2009) *Profilo linguistico dei dialetti italiani*. Roma/Bari: Laterza.

Macaulay, R. (2018). Dialect sampling methods. In*:* In Boberg et al. (2018), 241-252.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209-220.

Mæhlum, P., Kåsen, A., Touileb, S., & Barnes, J. (2022). Annotating Norwegian language varieties on Twitter for Part-of-speech. *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, 64–69. Association for Computational Linguistics.

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). *Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP*. ArXiv.

Moisik, S. R., Lin, H., & Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *Journal of the International Phonetic Association*, *44*(1), 21-58.

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, *21*(4), 463-475.

Nerbonne, J. (2009), Data-Driven Dialectology. *Language and Linguistics Compass* 3(1), 175-198. DOI: 10.1111/j.1749-818x2008.00114.x

Nerbonne, J., Heeringa, W., & Kleiweg, P. (1999). Edit distance and dialect proximity. Sankoff & Kruskal (Eds.) *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison*. v-xv.

Nerbonne, J., Lauttamus, T., Wiersma, W., & Opas-Hänninen, L. L. (2010). Applying language technology to detect shift effects. *Language Contact: New Perspectives. Amsterdam: Benjamins. Series IMPACT: Studies in Language and Society*, 27-44.

Nerbonne, J., & Siedle, C. (2005). Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik*, 129-147.

Nerbonne, J., van Ommen, S., Gooskens, C., & Wieling, M. (2013). Measuring socially motivated pronunciation differences. Borin, L., & Saxena, A. (Eds.). *Approaches to measuring linguistic differences*, 107-140. Berlin: De Gruyter.

Nguyen, D., & Grieve, J. (2020). Do word embeddings capture spelling variation? *Proceedings of the 28th International Conference on Computational Linguistics*, 870–881. International Committee on Computational Linguistics.

Orton, H. (1962). *Survey of English dialects: Introduction*. Arnold: Leeds.

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., & Johannessen, J. B. (2018). The LIA treebank of spoken Norwegian dialects. *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (LREC 2018), 4482–4488. European Language Resources Association (ELRA).

Paschen, L., Heggarty, P., Maguire, W., Michalsky, J., Dërmaku-Appelganz, D., & Boutilier, M. (2019). Sound comparisons: Germanic. https://soundcomparisons.com/Germanic

Plüss, M., Neukom, L., Scheller, C., & Vogel, M. (2021). Swiss Parliaments Corpus, an automatically aligned Swiss German speech to Standard German text corpus. *Proceedings of the Swiss Text Analytics Conference 2021*.

Plüss, M., Neukom, L., & Vogel, M. (2020). GermEval 2020 Task 4: Low-Resource Speech-to-Text. *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS).*

Pröll, S., Elspaß, S., & Pickl, S. (2021). Areal microvariation in German-speaking urban areas (Ruhr Area, Berlin, and Vienna). In *Urban Language Research. Variation-Contact-Perception*. 227-251. Amsterdam: John Benjamins.

Prokić, J., Çöltekin, Ç., & Nerbonne, J. (2012, April). Detecting shibboleths. In *Proc. EACL 2012 Joint Workshop of LINGVIS & UNCLH*. 72-80.

Rabanus, S., Kruijt, A., Alber, B., Bidese, E., Gaeta, L., & Raimondi, G. (2023). *AlpiLinK Corpus 1.0.2.* In collaboration with Mas, P. M., Bertollo, S., Casalicchio, J., Cioffi, R., Cordin, P., Cosentino, M., Dal Negro, S., Glück, A., Kokkelmans, J., Murelli, A., Padovan, A., Pons, A., Rivoira, M., Tagliani, M., Saracco, C., Siviero, E., Tomaselli, A., Videsott, R., Vietti, A., & Vogt, B. DOI:10.5281/zenodo.10224351.

Riley, R., Dozat, T., Botha, J. A., Garcia, X., Garrette, D., Riesa, J., Firat, O., & Constant, N. (2023). FRMT: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics* 11, 671–685.

Rubehn, A., Montemagni, S., & Nerbonne, J. (2024). Extracting Tuscan phonetic correspondences from dialect pronunciations automatically. *Language Dynamics and Change*, *14*(1), 1-33.

Salesky, E., Etter, D., & Post, M. (2021). Robust open-vocabulary translation from visual text representations. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7235–7252, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Samardžić, T., Scherrer, Y., & Glaser, E.. (2016). ArchiMob - A Corpus of Spoken Swiss German. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4061–4066. European Language Resources Association (ELRA).

Sankoff, D., & Kruskal, J. (1999, [1]1983). *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*. Intro. J. Nerbonne. Stanford: CSLI Press.

Scherrer, Y. 2020. TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6868–6873.

Scherrer, Y., Jauhiainen, T., Ljubešić, N., Nakov, P., Tiedemann, J., & Zampieri, M. (2023). *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects* (VarDial 2023). Shroudburg, PA: The Association for Computational Linguistics.

Schmidt, J. E., Herrgen, J., Kehrein, R., Lameli, A., & Fischer, H. (2020–). *Regionalsprache.de (REDE III). Forschungsplattform zu den modernen Regionalsprachen des Deutschen.*

(Engsterhold, R., Girnth, H., Kasper, S., Limper, J., Oberdorfer, G., Pistor, T., & Wolańska, A., Eds.).

Schönenberger, M., & Haeberli, E. (2019). Ein geparstes und grammatisch annotiertes Korpus schweizerdeutscher Spontansprachdaten. *Germanistische Linguistik*, 241–243, 79–104.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Rev. Ling. Romane*, *35*(139-140), 335-357.

Siewert, J., Scherrer, Y., & Tiedemann, J. (2021). Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation. *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021),* 242–246.

Simonsen, A., Lamhauge, S. S., Debess, I. N., & Henrichsen, P. J. (2022). Creating a Basic Language Resource Kit for Faroese. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4637–4643. European Language Resources Association.

Sinha, K., Jia, R., Hupkes, R., Pineau, J., Williams, A., & Kiela, D. (2021). Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2888–2913. Association for Computational Linguistics.

SPeech Across Dialects of English (SPADE): Large-scale digital analysis of a spoken language across space and time (2017-2020). ESRC Grant ES/R003963/1, NSERC/CRSNG Grant RGPDD 501771-16, SSHRC/CRSH Grant 869-2016-0006, NSF Grant SMA-1730479. (Digging into Data/Trans-Atlantic Platform).

Språkbanken. *NB Tale – speech database for Norwegian.* National Library of Norway. Last updated 2015-12-22.

Spruit, M. R. (2008). *Quantitative perspectives on syntactic variation in Dutch dialects*. Netherlands Graduate School of Linguistics.

Szmrecsanyi, B., & Anderwald, L. (2018). Corpus-based approaches to dialect study. In: Boberg et al. (2018), 300-313.

Tagliamonte, S. A. (2007). Representing real language: Consistency, trade-offs and thinking ahead! In Beal, J. C., Corrigan, K. P., & Moisl, H. L. (eds): *Creating and Digitizing Language Corpora, vol. 1: Synchronic Databases*, 205–240. London, UK: Palgrave Macmillan.

Thomas, E. R. (2018). Acoustic phonetic dialectology. In Boberg et al. (2018), 314-329.

van der Goot, R., Sharaf, I., Imankulova, A., Üstün, A., Stepanović, M., Ramponi, A., Khairunnisa, S. O., Komachi, M. & Plank, B. (2021). From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2479–2497. Association for Computational Linguistics.

Van Keymeulen, J., De Tier, V., Breitbarth, A., Ghyselen, A.-S., & Farasyn, M. (2019). Het dialectologische corpus “Stemmen uit het verleden” van de Universiteit Gent. *Volkskunde*, 120(2), 193–204.

Wahle, J. P., Ruas, T., Abdalla, M., Gipp, B., Mohammad, S. M. (2023). We are who we cite: Bridges of influence between Natural Language Processing and other academic fields. *The 2023 Conference on Empirical Methods in Natural Language Processing.* Singapore: Association for Computational Linguistics.

Wenker, G. (1889–1923): *Sprachatlas des Deutschen Reichs.* Marburg. [Published as *Digitaler Wenker-Atlas (DiWA)*; www.regionalsprache.de].

Wieling, M. B. (2012). A quantitative approach to social and geographical dialect variation. PhD diss., University of Groningen.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., & Nerbonne, J. (2014). Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change*, *4*(2), 253-269.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307-314.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., ... & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122-143.

Wikipedia contributors. FAIR data. *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 27 Sep. 2023. Web. 7 Oct. 2023.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *The Journal of the Acoustical Society of America*, *147*(2), 852-866.

Wissner, Inka. (2023) Investigating Pan-Romance prepositional Adverbials: A methodology for field research in Romance. *Revue Romane* 22: 1-35. DOI: 10.1075/rro.20019.wis

Woolhiser, C. (2005) Political borders and dialect divergence/convergence in Europe. In Auer, P. et al. (eds.) *Dialect Change. Convergence and divergence in European languages*, CUP: New York, 236-262.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878.

Zampieri, M., Nakov, P., & Scherrer, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, *26*(6), 595-612. doi:10.1017/S1351324920000492

Ziems, C., Held, W., Yang, J., Dhamala, J., Gupta, R., & Yang, D. (2023). Multi-VALUE: A framework for cross-dialectal English NLP. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 744–768. Toronto, Canada: Association for Computational Linguistics.

[i] See e.g. the European Union's open science policy https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

[ii] See Chambers & Trudgill (1998:19) for a more complete presentation of works based on the SED.

[iii] We keep contact options for contributions up to date on the website itself.