Lörres, Möppes, and the Swiss. (Re)Discovering Regional Patterns in Anonymous Social Media Data

Christoph Purschke and Dirk Hovy

Abstract

We study regional similarities and differences in language use on an anonymous mobile chat application. We use a neural network on 2.3 million online conversations to automatically learn representations of words and cities. These linguistic-use-based representations capture regional distinctions that can be clustered and visualized. We find that the resulting regional patterns are closely linked to the traditional division of the German dialects, even though most of the conversations are written in High German. The resulting maps correspond to traditional dialect divisions and language-external spatial structures, with few notable exceptions that can be explained through external factors.

Our method also facilitates two qualitative analyses, allowing us to discover geographically-pertinent words for various regional levels, as well as creating regional group-specific style profiles based on various linguistic resources. The results of our study strongly suggest the existence of “digital regiolects” in anonymous online communication, which share linguistic features with their offline equivalents, but also show substantial differences regarding their structure and dynamics. As a methodological contribution, we show how linguistic theory can drive the application and direction of neural network-based representation learning, and how their judicious application provides the basis for qualitative analysis of large-scale data collections.

1. Introduction[1]

Language's variability and constant change have always been key elements of the sociopragmatic organization of human cultural practice. People use language in speech acts for various purposes in everyday life. Speech acts therefore have practical consequences in the lifeworld (e.g., a legal judgment), but also carry culturally complex social meanings. This includes regional features often linked to social evaluations, such as a person's origin, social status, or intelligence (see for example Kristiansen 2009, Heblich et al. 2015). Taken together, language variation and the evaluation thereof contribute significantly to the structuring of cultural practice, for example with regard to the negotiation of group membership based on linguistic (dis)similarity (Purschke, forthcoming). In the German-speaking area (GSA, including Germany, Austria, and parts of Switzerland),[2] these distinctions were traditionally bound to local communities, creating small-scale local dialects. Through a combination of increased mobility, the institutionalization of standard languages, and changing sociocultural orientations, these fine-grained distinctions have gradually diminished, leading to regionally-bound intermediate varieties that combine resources from the old local dialects with influences from the standard language. Rather than disappearing, regional distinctions persisted at a regional level, allowing people to still use them as linguistic resource to stylize their language use and, therefore, social positioning in interactions. These distinctions still very much govern people's sociocultural orientation, as Falck et al. (2012) have shown in a recent study: dialect similarity was one of the strongest predictors of people's relocation decisions.

In this paper, we take linguistic variation and its social meaning as starting point to examine how these regionally-bound aspects of language use are realized in online communication. There is ample evidence *that* people use regional forms in digital media, both to define social styles in interactions,

[1] We would like to thank Jannis Androutsopoulos, Alfred Lameli, and Dong Nguyen for reading a draft of this paper.

[2] Technically speaking, Liechtenstein and Luxembourg belong to the GSA as well. However, due to data sparsity, neither are represented in our corpus.

and to create social group membership (see for example Nguyen 2017). However, we go further and ask *how* regional variation helps people structure their online communications. Furthermore, we want to differentiate between linguistic resources that can be linked to traditional regional languages and those that arise specifically in online communication.

We base our analyses on a large sample of anonymous online communications. While it provides ample information of regional variation, it does require the use of quantitative analysis methods. We turn to a very recent method from the field of computational linguistics, namely representation learning with neural networks. These models allow us to learn representations that concisely capture linguistic differences and similarities between entities (here: cities). These city representations allow several analyses: they can be clustered to discover larger geographic areas, which we compare to traditional dialect distinction, but they also enable us to visualize dimensions of variation, and to find representative words for geographic areas, ranging from individual cities to entire dialect regions. The latter allows us to pinpoint the pertinent markers of (digital) regional identity.

We hope to show with this combination of large-scale, quantitative methods and qualitative analysis that "rather than replacing traditional methods […], new techniques complement and augment existing humanities methods and facilitate traditional forms of interpretation and theory-building" (Kitchin 2014, 8). We develop our approach in the following sections: in *chapter 1*, we discuss the variationist and computational linguistic backgrounds of our study. In *chapter 2*, we explain the methodology for our study, including the data set and technical details, before we discuss the resulting spatial linguistic structures in our data set (*chapter 3*). *Chapter 4* comprises a discussion of methodological and linguistic aspects of our study, including benefits and limitations of the data-driven approach.

1.1 Regional Variation in German

Variation and change in German dialects are long-standing research topics. One of the main results of classic dialectology is a number of spatial classifications of these dialects. In *Figure 1*, we contrast two such dialect maps, one based on Wiesinger (1983)[3] and another, more recent one, by Lameli (2013).

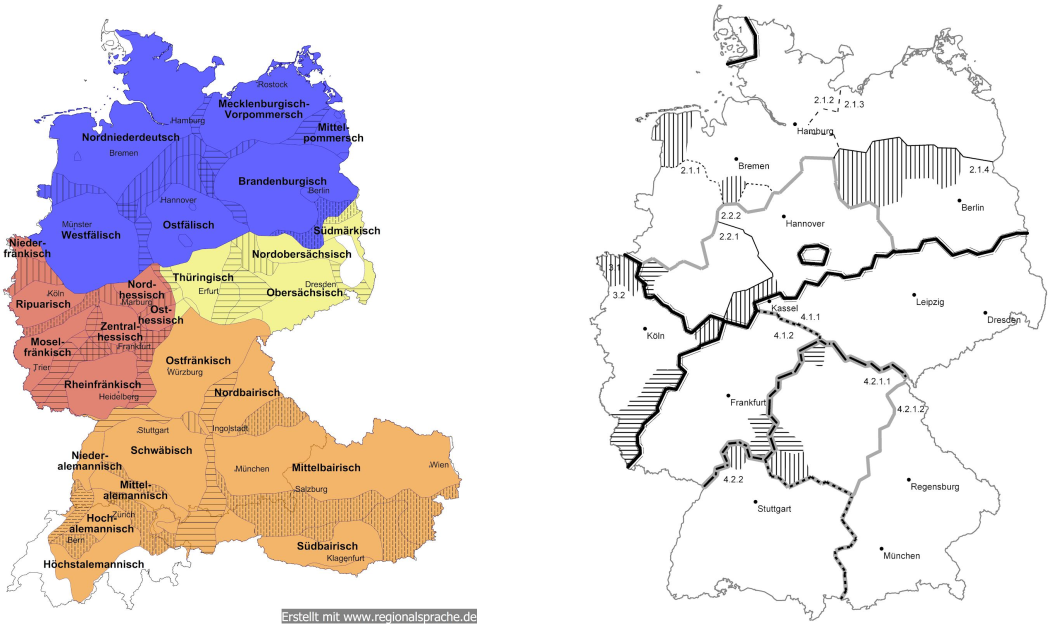

Figure 1: Dialect division of German. Left: according to Wiesinger (1983), right: according to Lameli (2013)

While both maps share many structural division features of regional German varieties, they differ markedly in the ways they have been created. Wiesinger (1983) used the traditional approach of defining and evaluating a comprehensive number of dialect isoglosses, from which he deduced the structural divisions. Lameli (2013), on the other hand, used a sizeable database with (historical) regional variants of a number of words from all administrative districts in Germany to construct a quantitative map with dialectometric methods.

[3] The map has been restricted to the national borders of Germany, Austria, and Switzerland.

Despite these methodological differences, both maps reproduce a similar macro-structure of the German dialects, with Low German (*Niederdeutsch*, blue area) dominating in the north and Upper German (*Oberdeutsch*, brown area) in the south, the latter split further into a Western part that consists of Swabian (*Schwäbisch*) and the Alemannic dialects (*Alemannisch*), an Eastern part that contains the Bavarian dialects (*Bairisch*), and Eastern Franconian (*Ostfränkisch*). In between Low German and Upper German we find a band of varieties, the so-called Middle German dialects (*Mitteldeutsch*), which again are divided into a Western (*Westmitteldeutsch*, red area) and an Eastern group of varieties (*Ostmitteldeutsch*, yellow area). The only substantial macro-difference between the two maps is the westernmost area in the middle of Germany, which is subsumed by Western Middle German in the Wiesinger (1983) map, but shows up as a structurally independent macro-region in the Lameli (2013) map he calls *Westdeutsch* ('Western German'). In the following, we will use a combination of these two maps as basis for our analysis that highlights Western German as an independent macro-area as opposed to the rest of the Middle German varieties.

Over the last 20 years, German variationist linguistics has undergone substantial restructuring, with the advent of new theoretical and methodological approaches. Traditional dialect divisions are no longer of primary interest to variationist linguistics. Instead, many studies focus on the extensive interferences between the old local *base dialects* ('Basisdialekte'), the standard language, and the new intermediate registers, so-called *regiolects* ('Regiolekte'), which lead to the formation of complex modern *regional languages* ('Regionalsprachen'; see Schmidt 2010 or Kehrein 2012). Another important methodological innovation is the consideration of speakers' perception and evaluation of variation for language dynamics and change (Purschke 2011; Stoeckle 2014).

Regarding the dynamics of spoken German, there is a substantial shift of locally-bound use of regional languages towards standard-oriented registers (Purschke forthcoming). At the same time, the advent of digital media produced online language that incorporates many aspects of oral speech (Androutsopoulos 2003). Social media is dominated by informal (conceptually oral) language use (Barton & Lee 2013; Herring 2013) that often makes use of online-specific resources like abbreviations, missing capitalizations, use of acronyms, emoji, logograms, or rebus writing (Dürscheid & Frick 2016; Dürscheid & Starck 2013; Schlobinski 2006). Another distinctive feature in online discourse is the hybridization of elements from different linguistic resources to create specific writing styles and social stances (Thurlow & Mroczek 2011; Androutsopoulos 2007). Nonetheless, regional linguistic resources have proven to be a vital resource for online communication in German (Tophinke & Ziegler 2014), especially in German-speaking Switzerland, where the Swiss-German dialects are the default written variety in digital communication for most language users (Schümann 2011).

Traditionally, regional variation (in German) has been captured by carefully-designed studies that map the distribution of a limited number of pertinent variables. This is true even for recent large-scale studies on regional variation in German or English that make use of a crowdsourcing approach (Leemann et al. 2016; Leemann et al. 2015). However, this approach relies on a number of prerequisites, including the predefinition of variables with high discriminative power for representative sections of the variety in question. These criteria are hard to enforce with online communications. While they provide us with sufficient amounts of data for in-depth studies and often cover a wide range of demographics and regions, they differ thematically and stylistically from offline communication, and therefore often lack any of the traditional variables used to measure and establish variation (see also the stylistic analysis in *Section 4.2*). Working with online communications therefore presents some methodological challenges compared to well-designed interpretation-driven variationist studies, including the composition of the corpus (noisiness, mixed modality, lack of control over social demographics; see Androutsopoulos 2013 for a discussion). However, these data sets provide us with some unique possibilities (beyond simple volume) that allow for a data-driven analysis of language variation and change. Advantages of online communications include their availability as directly machine-readable data, and that they represent unsupervised everyday practice, rather than language use from carefully-designed experiments. These features provide the basis for completely new approaches to variationist analysis. The large amounts of online data allow quantitative modeling and visualization, which we can compare to a hypothesis-driven interpretation. Quantitative modeling thus serves as a complementary step for qualitative analysis.

In this paper, we use a data-driven approach, which does not rely on prior assumptions about the observed variables, but rather aims to uncover patterns contained in the data that we can subsequently interpret. We use a bottom-up approach for spatial structures, by first learning a regionally-sensitive representation based on lexical variation at the city level,[4] and by then clustering and visualizing the learned city-representations to discover larger regional patterns.

Starting from individual interactions that contribute to the overall structure of a speech community, we combine the data-driven analysis – which allows us to find patterns in the large collections of data – with an interpretation-driven approach – which lets us contextualize and explain the patterns we discover. Or, as Kitchin (2014, 8) writes: "It is one thing to identify patterns; it is another to explain them. This requires social theory and deep contextual knowledge. As such, the pattern is not the end-point but rather a starting point for additional analysis."

Consequently, we are trying to merge the individual strengths of both computational linguistics and sociolinguistics. Under the label *computational sociolinguistics*, this combination fosters the idea of a shared perspective between the two disciplines, while at the same time benefiting from the body of knowledge (data-driven and interpretation-driven, respectively) they each provide. This fusion is especially important since – despite recent efforts in computational social science – both disciplines have been mostly unwitting about the respective other's research rationale, methodological potential, and empirical limits.

[4] The granularity is dictated by the availability in the data, see *Section 2.1*.

1.2 Computational Sociolinguistics

The scientific fields concerned with the computational study of language are computational linguistics (CL) and natural language processing (NLP), both of which lie at the intersection of linguistics and computer science. NLP's main focus is the development of engineering solutions to linguistic problems (e.g., Google Translate or Siri), whereas the focus of computational linguistics was the use of computational models to learn about language. Regardless, there is a host of fascinating work in this intersection touching upon sociolinguistic topics, and several recent approaches have shown the power of combining the two fields (Nguyen et al. 2016). These works look at the correlation (and presumed causality) of socio-economic attributes with linguistic features (e.g., Eisenstein et al. 2010, 2011; Eisenstein 2013a, 2013b; Doyle 2014; Bamman et al. 2014b; Eisenstein 2015). Most of this work has focused on lexical differences, and phonological aspects if represented in the data.

Hovy et al. (2015) and Hovy & Johannsen (2016) have explored the use of social media as a source of variation, and showed the prevalence of regional lexical variants reflected in this data, as well as phonotactics in British English and standardization in German. Johannsen et al. (2015) showed a quantitative approach to measure the influence of age and gender on syntactic constructions (see Cheshire 2005). Due to the inherent complexity and scale of the problem, it is hard to evaluate empirically. Their quantitative approach confirmed the hypothesis that syntax changes with age and gender, with differences in the preference for syntactic constructions (women show significantly more verbal conjunctions than men).

Previously, there has already been a number of works on exploring regional variation with statistical methods, from the dialectometric analyses of Dutch (Nerbonne & Heeringa 1997, Prokić & Nerbonne 2008; Szmrecsanyi 2008; Wieling et al. 2011; Pröll et al. 2014; *inter alia*), to work on regional differentiation of African American Vernacular English throughout the US (Jones 2015), based on Twitter data, and works on the regional patterns in the UK (Grieve et al. 2011).

Using similar methodology to us, Bamman et al. (2014a) have shown how regional lexical differences can be learned via distributed word representations (embeddings) to encode meaning differences between US states. Östling & Tiedemann (2016) have shown that distributed representations of national languages, rather than regions, capture typological similarities that can be used to improve machine translation quality. Similarly, Kulkarni et al. (2016), Rahimi et al. (2017a, b) have shown how neural models can be used to exploit regional lexical variation for the task of geolocation, while at the same time enabling dialectological insights.

For a similar approach to ours, but using Twitter as source and extending to the national languages of Europe, see Hovy et al. (2019).

2. Methodological approach

2.1 Data Source

We use data from the social media app "Jodel"[5] in our study, a mobile chat application that lets people anonymously talk to other users within a limited radius around them. The app was first published in 2014, and has seen substantial growth since its beginning. Today, Jodel has several million users in the GSA, but the company is also expanding to new markets in France, Italy, Scandinavia, Spain, and lately the United States.

Jodel users have anonymous accounts and can post anonymously and answer to other users' posts. *Posts* are visible to all users within a radius of about 10km around the user's current location.[6] *Threads* are conversations initiated by an initial post, to which future interlocutors can respond. Within a thread, users can refer to each other through a deictic system referring to the order of posts (i.e., previous user, etc. See *Section 4.1* below). There is also the possibility to up- and downvote posts and answers. The developers constantly implement and test new features, for example the use of

[5] <https://jodel-app.com> (31.10.2017).

[6] In the meantime, the developers have introduced a function called "Home" that lets users stay connected to a second location that they have defined as their "home" location. This of course affects the regional binding of Jodel communities to some extent.

hashtags, a sharing system for threads, changes to the deictic reference system, or a gallery view for images. The main audience of this service is college students, who initially used the app to discuss campus-related topics, but given the anonymity of the Jodel platform, conversations quickly expanded to (and are now dominated by) all kinds of private or even intimate topics.

The nature of Jodel and the ways it can be used by its users therefore influence the structure of our data set and the community practice we can survey in our study in four basic ways:

a) *Anonymity:* Posts and threads in Jodel are anonymous, which means that topics as well as stylizations of speech allow informal registers close to orality. Jodel speech is medially written, but conceptually oral (Koch & Oesterreicher 1985).

b) *Regionality:* Posts and threads in Jodel are regionally bound, due to the limited range of posts based on the user's location. This means that the data fosters the emergence of regional user communities.

c) *Demographics:* Most users of the Jodel app are students, which means that the data mainly covers language use of young adults. It also directly impacts the regional composition of the data: While the majority of students pick a college that is close to their home town, there are still many that move from Hamburg to Munich or even Vienna to study (Statistisches Bundesamt 2016). As a consequence, the data may be affected by inter-areal levelling of region-specific language use, especially in larger or more popular university cities.

d) *Style profile:* The vast majority of posts in Jodel is written in standard German. Regional or even dialectal forms are only common in Switzerland, Austria and and more rural areas in Southern Germany. Still, users actively deploy these forms to mark regionality. Beyond that, we can expect a broad repertoire of linguistic resources that users employ to stylize their online communications.

As a result of these characteristics, the current study differs in many ways from traditional approaches to the study of language variation and change. Normally, such studies survey small groups of carefully chosen participants (in specific situations) which fulfil certain criteria regarding speech competence, age, gender, social status, or mobility. In contrast, our data set combines data from thousands of anonymous users who contribute to conversations by posting. While this allows for a broader population sample, it does relinquish some control over confounding factors. And while this might be seen as a disadvantage compared to classical sociolinguistic study designs, it can also be a decisive advantage, because our data set was not assembled based on predefined criteria (except for list of the locations). Our data sample mirrors true-to-life language use more closely than controlled studies could. In return, we do not have any information about the writers in the sample.

We used the publicly available (albeit unofficial) API to download data from 79 German cities with a population over 100k people, all 17 major cities in Austria ("Mittel- und Oberzentren"), and 27 cities in Switzerland (the 26 cantonal capitals plus Lugano in the very south of the Italian-speaking area).[7] This leads to a list of locations that is relatively evenly spread across the entire GSA, albeit with some gaps in the Northeastern and Eastern Middle parts of Germany, which have a lower population density.

Our collection went on over a period of roughly 2 months between April 11 and June 19 2017, resulting in a total of 2.3 million conversations, or 16.8 million posts. After preprocessing (see 2.2), we end up with 87.8 million tokens. Note that the number of conversations differs widely between the selected cities, ranging from only a couple of dozens to over 40k in cities like Cologne, Vienna, Hamburg, Munich, and Berlin. Using threads as core units of analysis incurs the risks of accidentally including posts from nearby cities, and thereby diffusing regional differences. As we will see, though, this risk is minimal enough to not affect the outcome of the analysis.

2.2 Preprocessing

In order to use the data in algorithms, we need to preprocess it to minimize noise. During this process, however, it is unavoidable to lose some small amount of signal as well. As a first step, we lowercase the entire input, to make it more uniform and easier to process. However, this also

[7] Due to a technical problem, we were not able to collect items from Freiburg im Üechtland.

eliminates the stylistic or grammatical function of capitalization, which in social media is often used for emphasis. In order to find potentially discriminative regional words, we restrict ourselves to content words (nouns, verbs, adjectives, adverbs, and proper names, if they are not contained in a list of common stop words). However, a large class of regionally-distributed content words is that of place names (often the city they are used in, or specific places within theses cities), since people talk about their own region more than about others. We therefore also exclude all words identified as named entities, such as places, people, etc., by using the Python *spacy* package to filter out words based on their part of speech and named entity type, and the *NLTK* package to define stop words and reduce the words to their stem. The last step is necessary to remove the rich inflectional patterns found in German. While there are undoubtedly regional patterns in the inflections, in the current study, we want to focus on variation of lexical items. Note that both part-of-speech tagging and named-entity recognition are stochastic models, so there is a risk of false positives and negatives. One observed effect is that non-standard items are predominantly identified as nouns. While grammatically often incorrect, its only effect is that these items are kept at a higher rate than standard items. Empirically, however, we find that the error rates are low enough that it does not impact the quality of the analysis. One problem with excluding stopwords and place names revolves around the fact that these can be found only in their standard written form not in regional variants thereof, which means that place names and stopwords are only reliably excluded from posts written in High German, but not from posts written in Swiss-German or, say, Bavarian. This may in fact lead to higher coherence for regions with a higher amount of non-standard tokens (as in Switzerland).

2.3 Word and City Representations

Ultimately, we want to represent each city in our data set as a distinct vector, i.e., a list of numbers that capture various linguistic aspects of the city. Ideally, cities with similar language use patterns should end up with similar vectors (as can be measured by, e.g., *cosine similarity*). In order to learn such a representation, researchers have traditionally defined a small set of variables which formed a vector, with each vector position corresponding to one variable. While successful, this approach requires us to pre-define a (preferably) limited set of variables, i.e., to use prior knowledge about dimensions of variation (Lameli 2013). In our approach, however, we would like to abstain from pre-defined variables that could limit our epistemological potential. Rather, we would like to discover inherent patterns of variation in the data. In order to achieve this goal, but still represent each city as a distinct vector, we rely on a neural network method that has recently gained popularity, called *representation learning*.

We use an algorithm, *paragraph2vec* (Le and Mikolov 2014), that learns a vector for cities based on the observed words in said city. It also learns representations for individual words. It achieves all this by learning to predict the vector representing a word by using information about the word's left and right context (as vectors) and the city it was observed in (represented as vector as well). If the algorithm fails to find the correct word, we adjust the representations of both the words and city in such a way as to enable the correct prediction the next time. We repeat this process for all words in all instances (here: threads) in our corpus, until we incrementally reach the optimal prediction for all instances.[8] As a result, we get vector representations of both the words *and* the cities in the same high-dimensional space. This space enables us to compare the vectors of 1) cities to each other (asking: which cities are linguistically most similar to each other), 2) words to each other (asking: which words have a similar meaning, since words that occur in similar contexts receive similar vectors under the method),[9] and 3) words to cities (asking: which words are most similar to/indicative of a city, see *Figure 2*). In our experiments, we use the Python implementation of the algorithm in the *gensim* package.

[8] In practice, the algorithm stops iterating over the data once the prediction accuracy has reached a certain threshold.

[9] The algorithm does unfortunately not distinguish between different senses of a word, and can therefore conflate contexts of ambiguous words. Since we are not conducting a semantic analysis, this property does not affect our results.

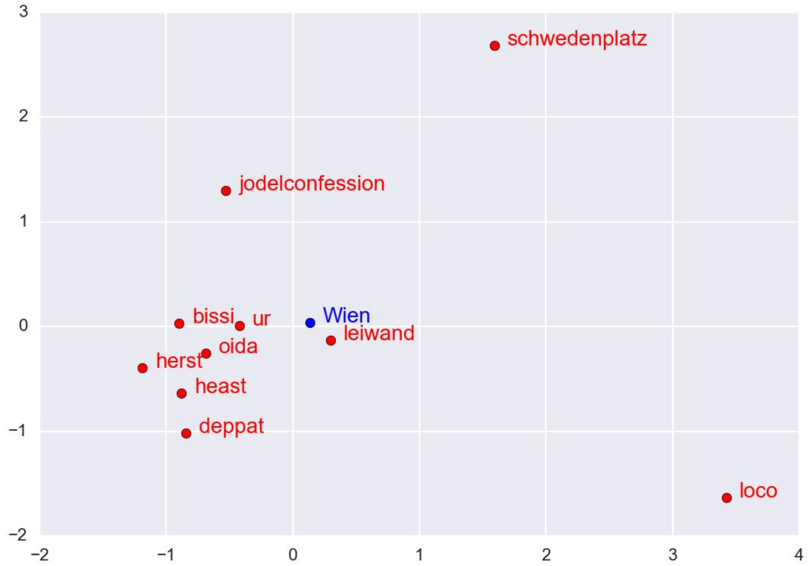

Figure 2: Visualization of learned city representation for Wien (Vienna) and its 10 nearest word neighbors in two dimensions.

Note that the individual dimensions of the vectors do not correspond to any particular feature (e.g., the fifth dimension does *not* correspond to the presence of a particular word), but that they have to be interpreted holistically and in relation: a vector denotes a point in a high-dimensional space. Words occurring in contexts which themselves are similar to each other end up closer together in this space (this can be understood as a soft matching of context).

The number of vector dimensions is a free parameter we have to choose before running the algorithm. More dimensions allow us to capture more fine-grained differences, but also require more data to learn. In our experiments, we use 300 dimensions, based on dimensionality recommendations for distributed representations (Landauer & Dumais 1997; Lau & Baldwin 2016), and on initial empirical tests. Other parameters include the treatment of frequent words: untreated, they would soon come to dominate the representations. Because they occur everywhere, they provide little discriminatory power, leveling out any differences. By down-sampling frequent words, we can shift the discriminatory power to less frequent words, which are typically regional expressions. For all parameters, we follow the settings described in Lau & Baldwin (2016).

As described above, we can use the vectors to compare words to cities. Since vectors are additive (i.e., we can produce a new vector by adding two existing ones), we can also construct artificial centroid vectors made up of several cities, and then find words close to the resulting new vector. This allows us to find words representative of entire clusters (*prototypes*). Note, though, that the new vector is no longer representing a real point in space, but is more akin to the theoretic linguistic center of a dialect region.

2.4 Visualization

Prior research has shown that dimensions of variation can be captured in the principal components of vector representations (Shackleton 2005). We also use a form of *dimensionality reduction* in our study, but not with the intent of controlling the number of dimensions (which we can already control in the *paragraph2vec* algorithm). Rather, we use a method to translate the three first principal components of the inherent variation in our learned word representations into color values. We apply *non-negative matrix factorization* (NMF, a form of dimensionality reduction) to the learned 300-dimensional city representations, and reduce them to 3 dimensions.

We now interpret these three dimensions (the three first principal components) as RGB channels, i.e., we assume that the first principal component signals the amount of red, the second component the amount of green, and the third component the amount of blue for a city (*Figure 3*). This mixture can then be translated into a color value. For example, 0.5 red, 0.5 green, and 0.5 blue would translate into a medium grey. The advantage of this transformation is that it preserves similarities: similar colors signal similar components, which in turn mean that the city representations are based on similar word usage. The approach is somewhat similar to the color assignment used for locations in Pröll et al. 2014.

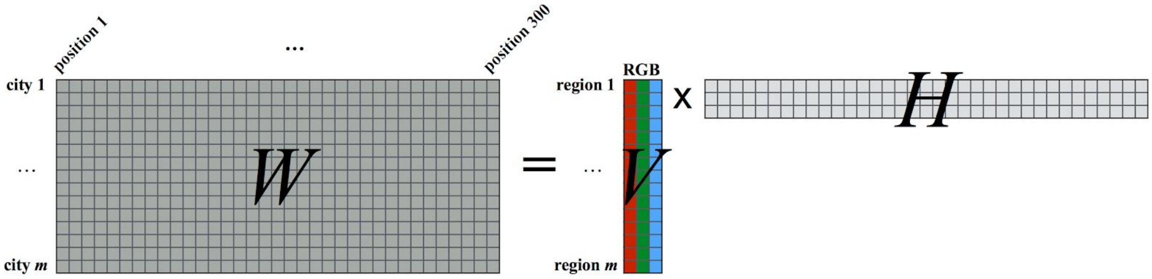

Figure 3: Schematic representation of non-negative matrix factorization. The learned representations of all cities (W) are decomposed into a three-dimensional city view (V), which we use for visualization, and a complementary matrix H.

Applying NMF to the city representations results in a continuous color gradient over the cities in our study (see *Figure 4*). Without adding much linguistic interpretation of the data, we can already see the difference between Switzerland (greenish colors) and the rest of the GSA, namely Germany and Austria. Within Switzerland, we see a distinction between the GSA (lighter green) and the French-speaking area around Lausanne and Geneva (darker tones). On the other hand, we find a continuous transition from red over purple to bluish colors for Germany and Austria. These gradients largely correspond to the dimensions North >> South(East) (i.e., red >> blue) and West >> East (i.e., intense tones >> pale tones). Given that there seems to be a strong connection between the south of Germany (especially Baden-Württemberg and Bavaria) and Austria, a strong connection between most cities in the north of Germany, while the (Western) middle part of Germany defines a transition zone (speaking in colors), it seems likely that these correspondences and differences mirror regional linguistic similarities and differences in German.

Altogether, the map contains 408 locations (333 in Germany, 27 in Austria, 48 in Switzerland), mainly the chosen locations plus smaller cities surrounding the larger ones. The circle size for every location indicates the relative number of conversations per location. Since larger cities tend to have a higher number of college students, in most cases the bigger cities also show the most activity on Jodel. In order to get reliable statistics in our analysis, we restrict ourselves to cities with more than 200 observed conversations. This threshold limits the number of conversations to about 2.1 million conversations (1.82 million in Germany, 173k in Austria, and 146k in Switzerland). Including cities with fewer conversations adds more data points, but often creates noise, as the corresponding representations are based on too little data, resulting in inaccurate vectors.

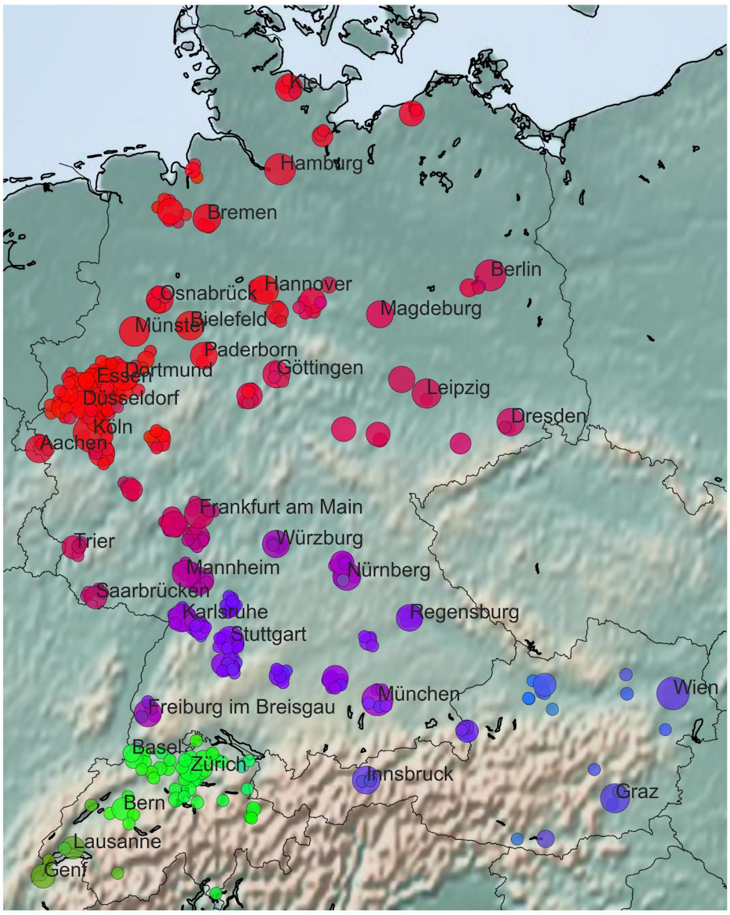

Figure 4: Gradient color map showing the overall similarity between locations based on all data in the sample

2.5 Clustering

The gradient color map of the cities in the last section already suggests the existence of larger areas that can be related to the macro-structure of the German dialects. However, we do not have a selection of pertinent words, but instead a continuous 300-dimensional representation based on word use. We use *agglomerative clustering* over the city representations to discover clusters based on similarities, rather than defining dialect areas based on our knowledge of variation in German. This allows us to assess whether the learned representations capture existing regional distinctions, and the viability of our data-driven approach. If the clusters in our data match existing dialect distinctions (or other sociocultural spatial structures), this provides a compelling argument for the applicability of our methodology. We run hierarchical Ward clustering on the city representations. This algorithm groups cities into a selected number of clusters, starting with the two most similar vectors and working its way through the data by minimizing the variance (sum of squared distances) between clusters, until all cities are assigned to a cluster. Prior work (e.g., Nerbonne & Heeringa 1997; Prokić & Nerbonne 2008; Szmrecsanyi 2008) has used Ward clustering in a linguistic context.

Since the city representations are based solely on word usage in the city, the clustering essentially captures regional patterns of similarity in word usage. Hierarchical clustering allows the introduction of structure through the use of a *connectivity matrix*, i.e., weighted information about the distance between data points, making close neighbors more likely to be merged together. We use the inverse geographic distance of each pair of cities as connectivity weight. I.e., cities that are far from each other (say, Vienna and Hamburg) are less likely to be merged than cities closer together. While this provides the model with a structured component modeling geography, it is important to note that it does *not* predetermine the clustering outcome, as we will see: cities that are close together, but linguistically different still end up in separate clusters. Structured clustering does, however, provide regional stability and more coherent clusters than unstructured clustering.

We can visualize the groupings on a map by assigning each cluster a separate color. Setting different values for the clustering algorithm, we can create more and more fine-grained distinctions. We find that this results in interpretable solutions for up to 15 clusters.

In the following, we discuss several of the clusterings in more detail (due to space constraint, we only show a selection). For every cluster solution, we show the clusters on two maps, the base map, and one that depicts the clustered cities superimposed with a dialect map for the GSA (a combination of the maps by Wiesinger 1983 and Lameli 2013) to check for correspondences and differences with the macro regional linguistic structure of German.[10]

Generally speaking, the juxtaposition reveals that the clusterings closely correspond to the structure of the German dialect areas as defined by traditional dialectology, despite the fact that the vast majority of messages in Jodel is written in High German (except for Switzerland). Still, regional linguistic structure is not the only factor responsible for the clusterings. Therefore, we also check for other potential factors influencing the clustering, like socio-economic migration, student mobility, and the contribution of other than regional linguistic resources. In doing so, we are combining the bottom-up quantitative approach from computational analysis with the top-down interpretative approach that uses sociocultural and linguistic knowledge to analyze the cluster patterns. To get an idea of the linguistic coherence of each cluster, we discuss the prototypical words (based on their averaged similarity vectors) for all cities in a cluster. With each new cluster, we compare the prototypical words for the new cluster in comparison with the one it was separated from. This will give us an idea of what exactly constitutes the linguistic similarity of a cluster. Due to space constraints, we only show the top 25 words in each list.

[10] The cluster maps have been generated with the *sklearn* and *Basemap* packages in Python, while for the dialect maps we have imported the cluster solution to the linguistic GIS *regionalsprache.de* (Schmidt et al. 2008ff.).

3. Tracing regional patterns in online communications

3.1 Three clusters

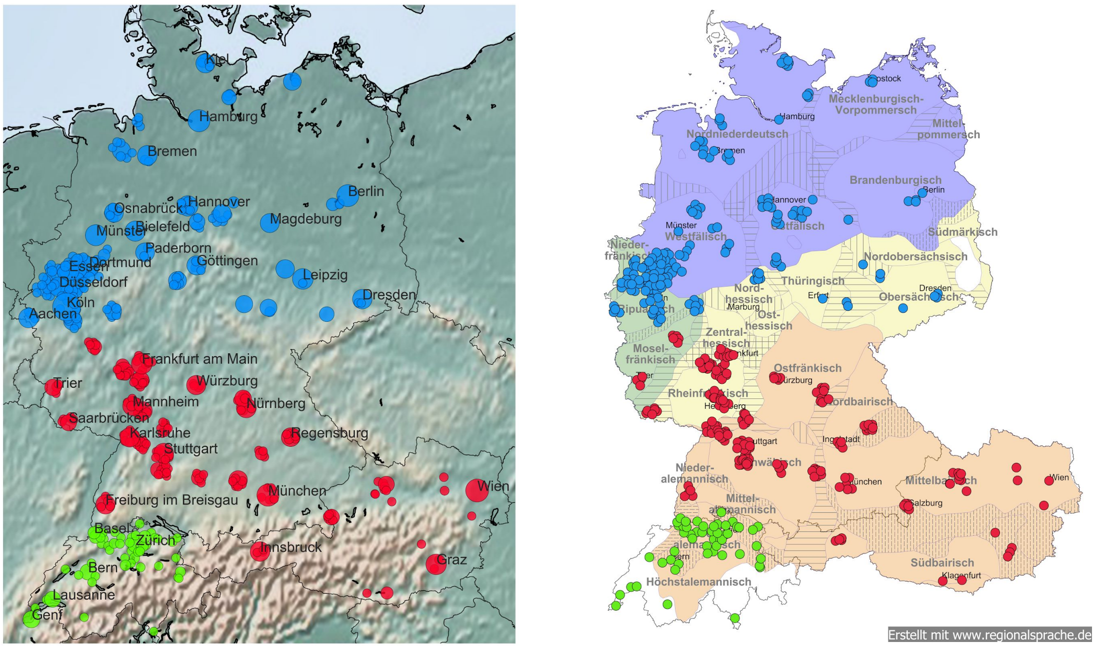

Figure 5: Three-cluster solution

In the three cluster solution (*Figure 5*), we find the basic distinction between Switzerland on the one hand, and Germany and Austria on the other hand, as already expected from the gradient map.[11] Overall, this clear-cut division is contrary to the dialectal continua assumed for the Alemannic dialect area, which combines most of Switzerland and the southwest of Germany. Interestingly, the German and French-speaking areas within Switzerland (even including Lugano in the south, where Italian is the dominant language) appear more similar to each other under the model than they are to the rest of the GSA. Second, we find a distinction of Germany in two clusters: A northern cluster that represents the Low German region (blue area on the dialect map), the Eastern part of the Middle German region

[11] There are two German cities clustered with Switzerland, Rheinfelden (Baden) and Lörrach, both of them located directly on the Swiss border, in close socioeconomic contact to Switzerland.

(yellow area) including Northern Hessian (Kassel), and the northern part of what Lameli (2013) calls "Western German" (green area), i.e., Ripuarian and Low Franconian, and a southern cluster that comprises the Western part of Middle German (yellow area) including Moselle Franconian (green area), and all of the Upper German varieties except Switzerland.

Cluster 1 (Switzerland, 52 locations, 147k threads): *esch, ond, vell, gaht, wüki, nöd, besch, emmer, nor, au nöd, verstahn, muen, wükli, dänn, vode, hett, chan, rechtig, staht, sösch, abig, mached, isch de, lüüt, nanig*
Cluster 2 (Northern Germany, 170 locations, 1.2M threads): *ja gut, erstmal, sieht, drauf, vielleicht, mehr, gut, sehen, schonmal, ahnung, bisschen, gesagt, kommt, allerdings, gucken mal, reicht, achja, bestimmt, garnicht, musst, ansonsten, scheinbar, darauf, schon gut, wahrscheinlich*
Cluster 3 (Southern Germany & Austria, 186 locations, 804k threads): *afoch, voi, nd, i a, oda, möppes, nimma, is a, mei, gscheid, is, ffm, @vj, hnx, vj, lörres, @vvj, bissl, dummwiekarlsruhe, gibt, vermutlich, lässt, gerade, feuerbach, wobei*

If we take a look at the 25 most prototypical words for each clusters, the difference in language use between is striking: The list of prototypical words for cluster 1 only contains words that are written forms of Swiss-German dialect words, including common adverbs (*wüki* 'really', German *wirklich*), verbs (*verstahn* 'understand', German *verstehen*), nouns (*abig* 'evening', German *Abend*), and even conjunctions (*ond* 'and', German *und*). Compared to that, cluster 2 contains only High German words, mostly adverbs (*ansonsten* 'otherwise'), adjectives (*gut* 'good'), verbs (*gucken* 'look'), or nouns (*ahnung* 'suspicion'). Some of them show typical characteristics of digital writing like contraction of collocations (*achja* 'oh well'). In contrast, the items for cluster 3 (the southern part of Germany and Austria) relates to many different linguistic resources: First, we find some High German words (*gerade* 'just', *vermutlich* 'supposedly'). Second, there's a group of highly prototypical words that represent written versions of regional forms that originate from the Upper German dialect area (including Austria), like *gscheid* ('intelligent' or 'properly', Upper German), *mei* ('well' Bavarian interjection), or *voi* ('really', Austrian). The third type of items are Jodel-specific terms that are used for the pragmatic organization of conversations, such as referring to previous messages in a thread (*vj, @vj* for 'vorheriger/s Jodler/Jodel': 'previous author/message'). A fourth type of item seems typical for the Jodel community, but without fulfilling a meta-function like "author reference". This is the case for *möppes* (*Figure 6*), which is commonly used in the Jodel community and refers to ‘a small woman with big breasts’, and lörres (‘penis’). While both words are known in regional Western German (with different meaning in the case of *möpp* ‘mean person’), the words apparently have been re-semanticized in Jodel communications. And lastly, we find discourse labels like *dummwiekarlsruhe* ‘stupid like Karlsruhe’ and examples for references to specific locations like *ffm* (abbreviation for ‘Frankfurt am Main’), *feuerbach* (referring to the quarter in the city of Stuttgart), or *hnx* (‘Heilbronx’, wordplay with Heilbronn and Bronx) which we would expect in the data given the regional binding of the different user communities.

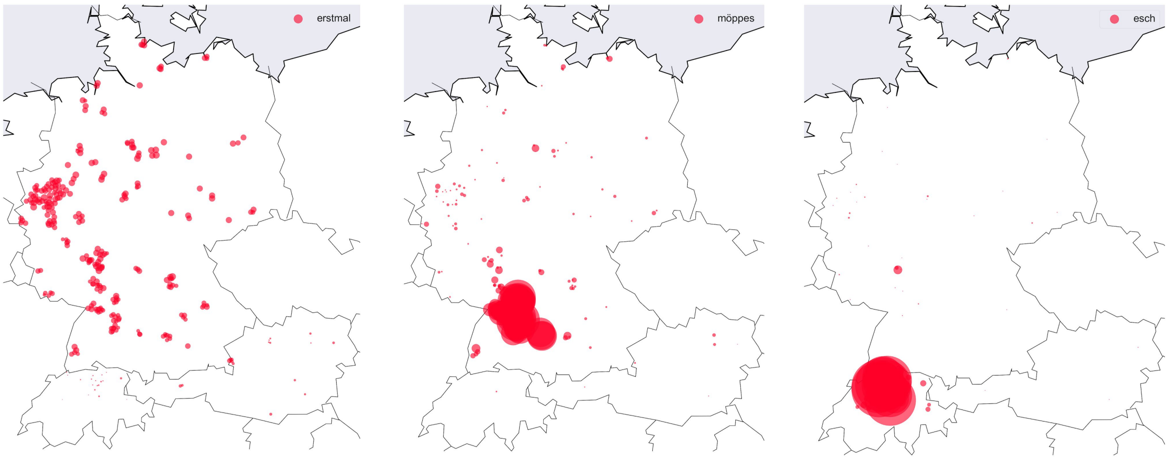

Figure 6: word maps for the items erstmal*,* möppes*, and* esch

The three clusters represent fundamentally different writing styles: the use of standard German forms (cluster 2) versus the transliteration of spoken Swiss-German forms (cluster 1) versus a mixture of different linguistic resources (cluster 3). *Figure 6* shows the regional distribution of three prototypical words in the sample representing different linguistic resources, i.e., *erstmal* (cluster 2, High German), *möppes* (cluster 3, community-specific), and *esch* (cluster 1, Swiss dialect). As we can see, each item shows a different regional spread: while High German items like *erstmal* are evenly spread throughout the entire GSA, other items (and therefore, resources) show regional focuses of distribution (*möppes*) or are exclusive of a specific cluster (*esch*). While the prototypes are calculated for rather large areas, they already demonstrate that our model is able to detect distinctive patterns of similar (or different) language use clearly linked to regional variation. Generally speaking, Swiss-German Jodel users write in their (local) variety of German, whereas Northern German users use standard written German and southern German and Austrian users employ a mixture of different linguistic resources.

3.2 Four clusters

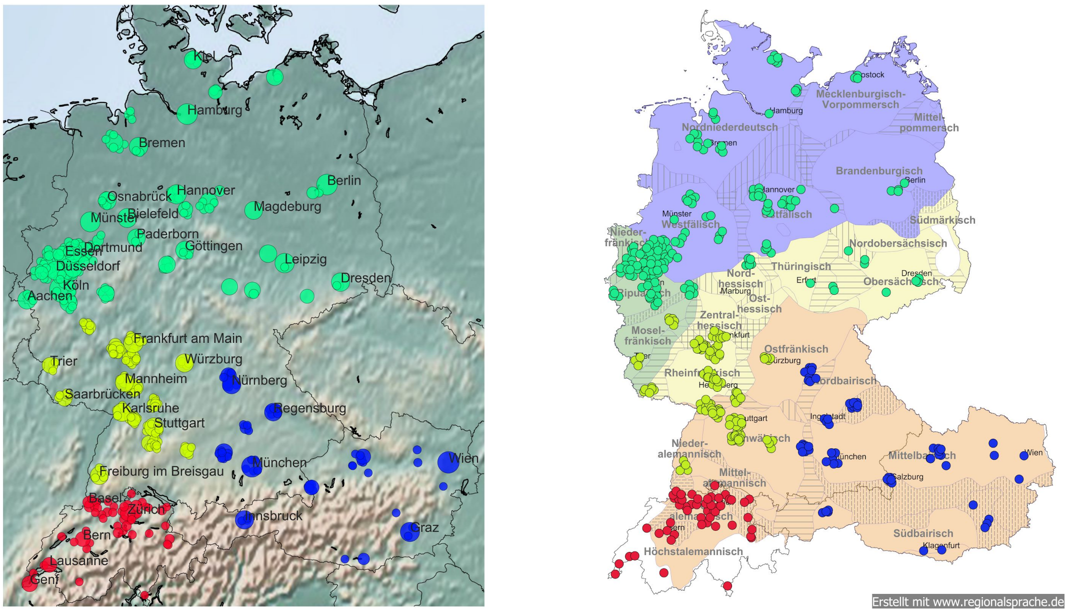

Figure 7: Four-cluster solution

In the four-cluster solution (*Figure 7*), we see the split of cluster 3^{pr12} into a new cluster 3 (blue dots) roughly corresponding to Austria and the German federal states Bavaria (both Bavarian dialects), including the small strip in the east of Baden Württemberg that linguistically belongs to Bavarian ("Bayerisch-Schwaben"), and a cluster 4 (lemon dots) that is constituted by the remaining locations in the Western part of Middle and South Germany. Dialectologically speaking, we see the split between the Western and Eastern Upper German dialects faithfully reproduced in our data, with

[12] The index 'pr' (for 'previous') indicates that the number of the respective cluster is not identical with the respectively labeled cluster in the previous cluster solution. This convention will be used for the following steps as well.

the interesting exception of Würzburg. It is clustered together with the western locations in cluster 4, but in Lamelis (2013) quantitative dialect division it belongs to Eastern Franconian, and is more closely related to Eastern Upper German than to the Western Middle German and Alemannic varieties.

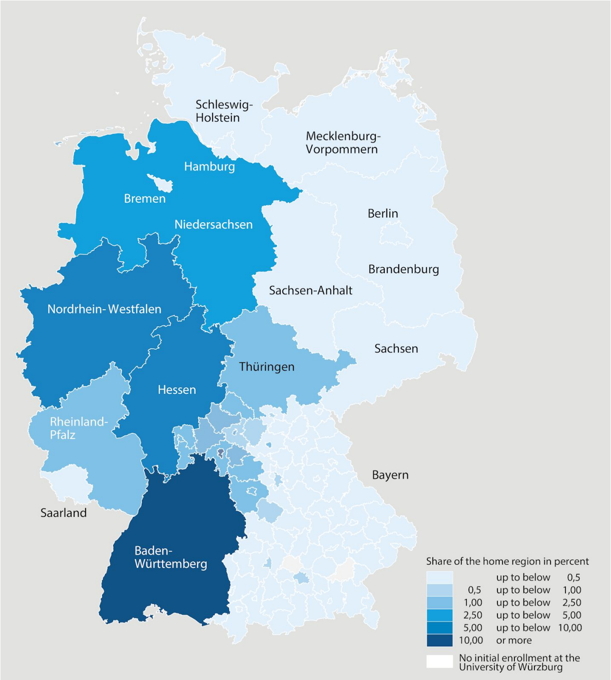

Figure 8: Home regions of students at the University of Würzburg, initial enrollments for winter semester 2006/17 (data source: © Bayerisches Landesamt für Statistik, Munich, 2017)

Geographic proximity does not sufficiently explain this phenomenon, but there is strong evidence that it derives from regional mobility patterns of Jodel users. As mentioned above, most Jodel users can be assumed to be college students. Most of them stay in touch with friends in their home region, or even travel back over the weekends. Some effects of this regional mobility are captured in our data. In the case of Würzburg (see *Figure 8*), there is evidence that its university predominantly attracted students from Baden Württemberg, Hesse, and other Western German federal states over the last years, whereas the percentage of Bavarian students is relatively low.

While this pattern of regional mobility already hints at some of the potential structuring forces for clusters 3 and 4, the prototypical words speak an even clearer language:

Cluster 3 (mostly Bavaria & Austria, 64 locations, 356k threads): *oba, owa, einfoch, hoid, waun, hob, siag, mocht, kan, afoch, ghobt, woa, ana, ois, hobi, voi, najo, obwoi, laung, haum, amoi, fia, faungt, oiso, kema*

Cluster 4 (Western Middle and Southern Germany plus East Franconian, 122 locations, 448k threads): *h7, möppes, sontheim, heilbronner, ffm, @vj, hnx, vj, herrngarten, aurelius, lörres, @vvj, gibt, dummwiekarlsruhe, vermutlich, feuerbach, wobei, pforzheimer, besonders, beispiel, hohenheim, lässt, stuggi, gerade, meinst*

The difference between the locations in the two clusters is quite clear: Among the prototypical words in cluster 3 are dialect words, most of which can be attributed to general Bavarian (*mocht* 'makes', *hoid* 'so, well'). Still, there is some evidence for regional variation within this cluster, for example the forms *laung* 'long' and *waun* 'when' that originate from the Austrian part of Bavarian (see *Figure 9*).

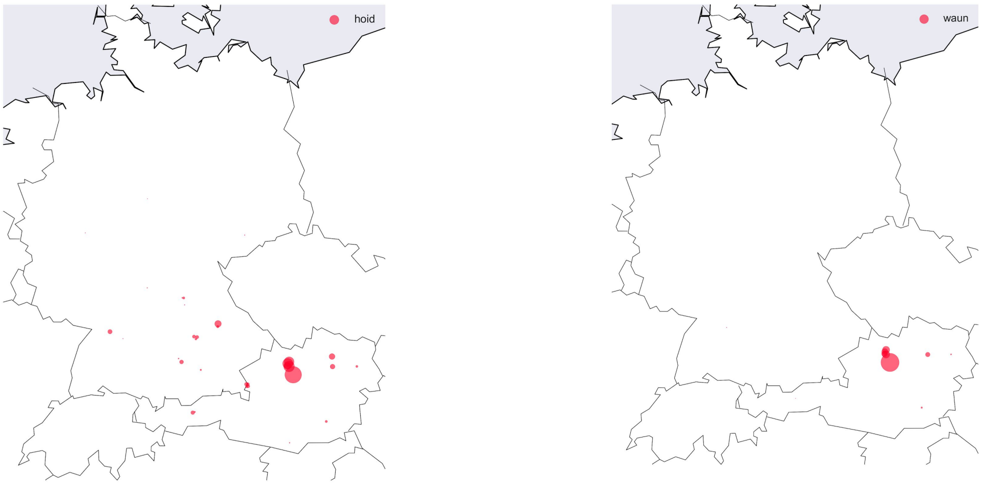

Figure 9: word maps for hoid *and* waun

In contrast, cluster 4 does not contain any dialectal items among the prototypical words. Instead, this cluster shows more location names and community-specific language use (*@vj*, *@vvj*, *lörres*, *möppes*, *dummwiekarlsruhe*), which hints at another structuring factor of the regional clusters in the data. Apart from that, we find some standard German words that do not tell us much about the regional constitution of this cluster.

3.3 Five clusters

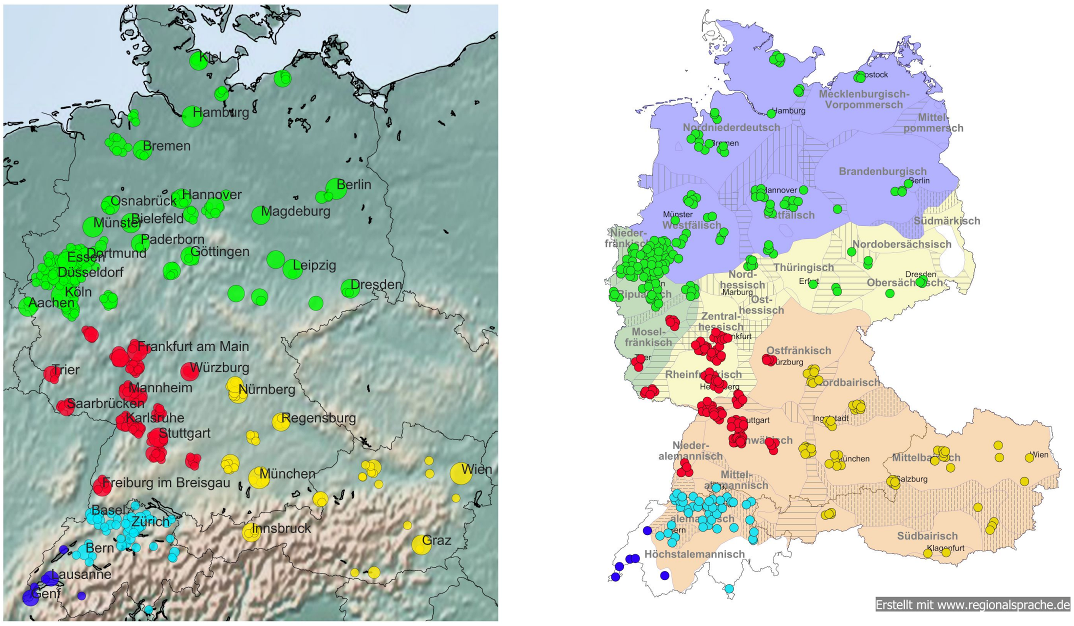

Figure 10: Five-cluster solution

In the five-cluster solution (*Figure 10*), cluster 1^{pr} splits into two new clusters that nicely reflect the border between French and German-speaking Switzerland. Cluster 5 (dark blue dots) represents the French-speaking locations in the West of Switzerland, while cluster 1 (light blue dots) now contains all of German-speaking Switzerland and Lugano in the Italian-speaking part.[13] Note how the border between the two new clusters exactly matches the outer border of the Alemannic dialect continuum. While we would expect the two areas to be quite different in terms of language use, it is interesting to note that under the model, the difference between French and Swiss-German is smaller than the difference between Switzerland and the rest of the GSA (cf. the two-cluster solution). Presumably, French and Swiss-German jointly differ more from standard German than the other clusters in the rest of the GSA do (cf. the prototypical words for clusters 2 and 3 in the three-cluster solution). Another

[13] If we take a look at the nearest word neighbors for Lugano, we can see that all of them do in fact represent standard Italian items. Still, it is unclear why the location gets subsumed with German speaking Switzerland.

reason could be that the Swiss community is relatively small compared to the German and Austrian one: the new cluster 5 only represents only 6 locations.

The prototypical lexical items substantiate this assumption:

Cluster 1 (German-speaking Switzerland & Lugano, 46 locations, 99k threads): *esch, ond, vell, gaht, wüki, nöd, besch, emmer, nor, au nöd, verstahn, muen, wükli, dänn, vode, hett, chan, rechtig, staht, sösch, abig, mached, isch de, lüüt, nanig*
Cluster 5 (French-speaking Switzerland, 6 locations, 42.5k threads): *t'as, je vais, autant, pour le, que ça, peut être, j'ai, en fait, je pense, c'était, une, dans le, trouve, parler, fais, même, sinon, comme ça, je sais pas, que je, pour moi, c'est, à, pour, enfin*

The Swiss-German cluster contains Swiss-German words. Note that the list of prototypical words for cluster 1 contains exactly the same items than it did in the two-cluster solution. The main reason why French appears relatively late (cluster 5), despite the linguistic distance of French to German, might be the small amount of data from Switzerland (42.5k conversations for cluster 5, compared to 99k in German-speaking Switzerland). The list for cluster 5 consists of standard written French, including common collocations (*en fait* 'actually', *dans le* 'in the', *pour moi* 'for me') and examples of informal standard French (*je sais pas* 'i do not know', with the first part of the negation 'ne' omitted). The prototypicality of common French words is due to the fact that we only excluded German stop words during preprocessing.

3.5 Six clusters

We find an interesting division in our data for six clusters (*Figure 11*), which demonstrates both the accuracy of our model, as well as its limits. Linguistically speaking, the new split divides the Western Middle and Upper German areas of cluster 4^{pr} into two clusters: The new cluster 6 contains only locations within the Swabian dialect area, while cluster 4 contains the rest of the locations in the west of Baden Württemberg, Hesse, and Rhineland-Palatinate, plus Würzburg and surroundings. This distinction accurately captures the dialect border between Swabian (i. e., Pforzheim & Heilbronn; Upper German) and Rhine Franconian (i. e., Heidelberg & Karlsruhe; Western Middle German). This result is even more reassuring for our method given that Jodel is highly localized, with a radius of about 10km per post, so we would expect our model to reproduce these local structures.

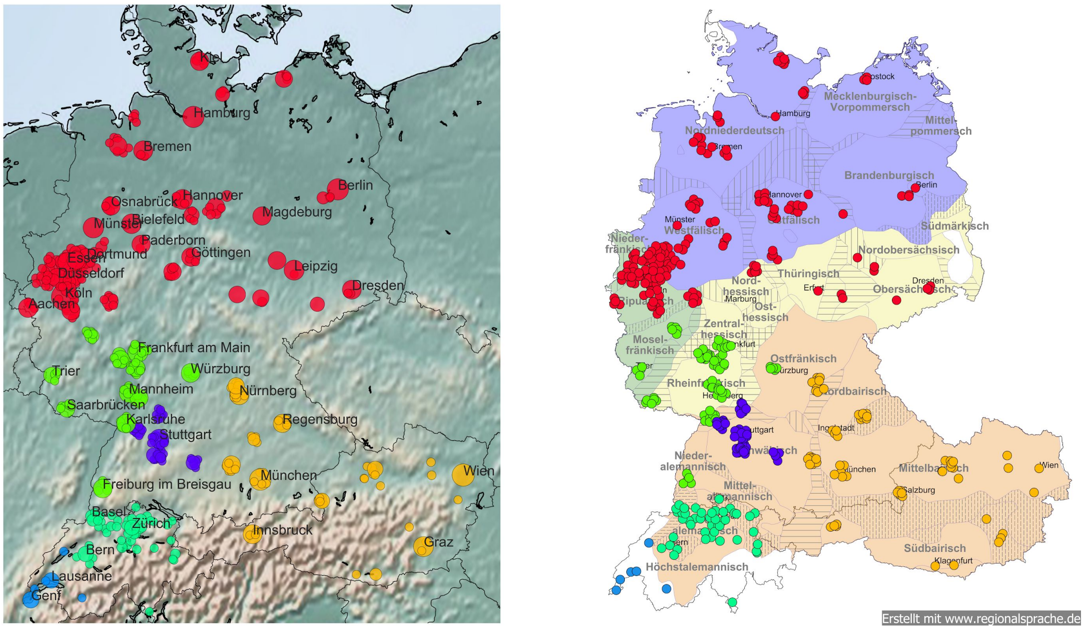

Figure 11: Six-cluster solution

Note that geographic distance can not be the structuring factor in the model, since the distance between some of the locations on opposing sides of the border is closer than 10 km. Also, since our data collection focused on bigger cities in the GSA, the data reflects the effect of socioeconomic mobility and sociocultural orientation, i.e., the influence zone of cities on smaller places around them, which would strengthen proximity effects even more. The fact that these geographic factors do not overwrite linguistic similarities is corroboration of the model.

The prototypical words show strong influence of location and place names for both clusters:

Cluster 4 (Western Middle German, Low Alemannic & East Franconian), 77 locations, 341k threads): *sharks, neckarwiese, city döner, darmstädter, moseleck, bonames, amk, ffm, vj, herrngarten, hda, sgf, lelek, eberstadt, sachsenhausen, bessungen, arheilgen, einzelkampf, aurelius, gibt, 0-6+9, mainzer, @vvj, niederrad, besonders*

Cluster 6 (Swabian, 45 locations, 107k threads): *h7, möppes, tübinger, sontheim, aksaray, heilbronner, mpark, pf, hnx, pforzelona, blaubeurer, cannstatt, newie, bildungscampus, hn., lörres, hhn, greendoor, böckingen, wimpfen, ilsfeld, dummwiekarlsruhe, feuerbach, pforzheimer, blaubeurerstraße*

Both lists predominantly reflect regional locations (*sontheim*, *moseleck* ‘a bar near Frankfurt main station’), facilities (*mpark* ‘abbreviation for Musikpark, a concert location in Heilbronn’, *city döner* ‘a kebab restaurant’), groups of users (*darmstädter*, *heilbronner*), or local elements of culture (*sharks* ‘a strip club in Darmstadt’). However, to some extent the prototypes still reflect regional word use or topics beyond place names, i.e., the notorious *möppes* and *lörres* (cluster 6), but also examples of swearwords, in this case the polish word *lelek* ‘nightjar’ that is common in German ethnolects.

We conjecture that the smaller the clusters become (in terms of the geographic area), the more our data is influenced by local place names and topics that are difficult to interpret for outsiders (e.g., the meaning of *einzelkampf* ‘single combat’). However, note that the prototypical words only represent an average of word usage for all locations in a cluster. Bigger cities and places with more conversations therefore influence these prototypes more than places with a smaller amount of data. E. g., the prototypes for cluster 4 are dominated by location names in the Rhine-Main-Area around Frankfurt (*ffm*, *sachsenhausen* ‘a quarter of Frankfurt’, *0-6+9* ‘area dialing code for Frankfurt’).

With respect to preprocessing (see *Section 2.2*), it becomes apparent that while we can filter out most place names, many creative versions remain, such as abbreviations (*hhn* ‘University of Heilbronn’*, hn* ‘licence plate code for the administrative district of Heilbronn’), neologisms (*pforzelona* ‘humoristic portmanteau of Pforzheim and Barcelona’), facilities (*city döner*), or groups of users (*heilbronner, pforzheimer, tübinger*).

3.6 Seven clusters

In the seven-cluster solution (*Figure 12*), the northern half of the GSA (see cluster 2^{pr} in the cluster three solution) is split, with cluster 7 (purple dots) representing the locations in the Ruhr area and Rhineland while cluster 2 (pale green dots) containing the rest of the Northern German locations. The new cluster 7 is spread across the Western German dialect area as of Ripuarian and Low Alemannic (green area) as well as the Low German dialect area Westphalian (blue area). Again, recent digital language use diverges from the traditional dialect areas. In this case, location density together with the number of users in the Ruhr area and Rhineland (some of the most densely populated in the entire GSA) as well as the high amount of mobility (socioeconomic and sociocultural) in the region lead to a closely interwoven pattern reflected in the clustering.

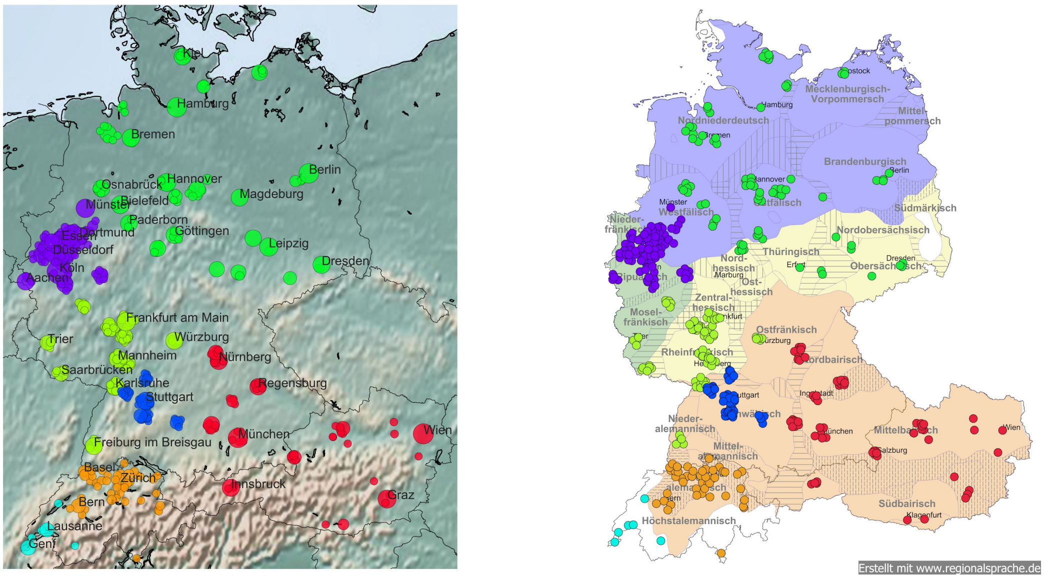

Figure 12: Seven-cluster solution

This effect is also visible in the prototypical words:

Cluster 2 (Northern Germany & East Middle German, 81 locations, 691k threads): *wolfhager, osna, wf, bs, wob, kieler, jedenfalls, gerade, bloherfelde, durchaus, braunschweiger, erstmal, deutlich, schon gut, musst, nochmal, bisschen, dürfte, drauf, schonmal, gehört, vielleicht, hast, achja, außerdem*
Cluster 7 (Ruhr area & Rhineland, 89 locations, 525k threads): *gerade, erstmal, ja gut, mehr, allerdings, garnicht, sieht, gut, scheinbar, kommt, vielleicht, immernoch, ansonsten, bisschen, schonmal, bestimmt, drauf, achja, einfach, gucken mal, zutun, wobei, eben, darauf, ahnung*

Cluster 7 prototypes consist solely of standard German words, as in the two-cluster solution. There is no reference to regional or local place names of any kind. In contrast, cluster 2 contains a mixture of standard German words, references to location names (*osna* 'Osnabrück, *bs* 'Braunschweig', *wb* 'Wolfsburg'), local places (*wolfhager* 'a street name in Kassel with student housing'), and references to local user groups (*kieler, braunschweiger*). This cluster shows the transition from a global structure that reflects macrolinguistic similarities to a regionalized structure that highlights specific regional user communities (and their word use). Given that each cluster represents roughly one quarter of all conversations collected, we can assume an interplay between conversation density and the geographic distribution of place mentions in the cluster: The locations in cluster 2 are spread over entire northern Germany, while the locations in cluster 7 are concentrated in a much smaller area. The prototypical place names we find for cluster 2 belong to local user communities that do not represent the most active communities in term of number of threads in the cluster.[14] This could indicate that in these communities place references are more frequent than in other communities. The locations responsible for the place-name prototypes are among the top 25 locations regarding the number of threads, but do not have the highest total number conversations.

3.7 Eight clusters

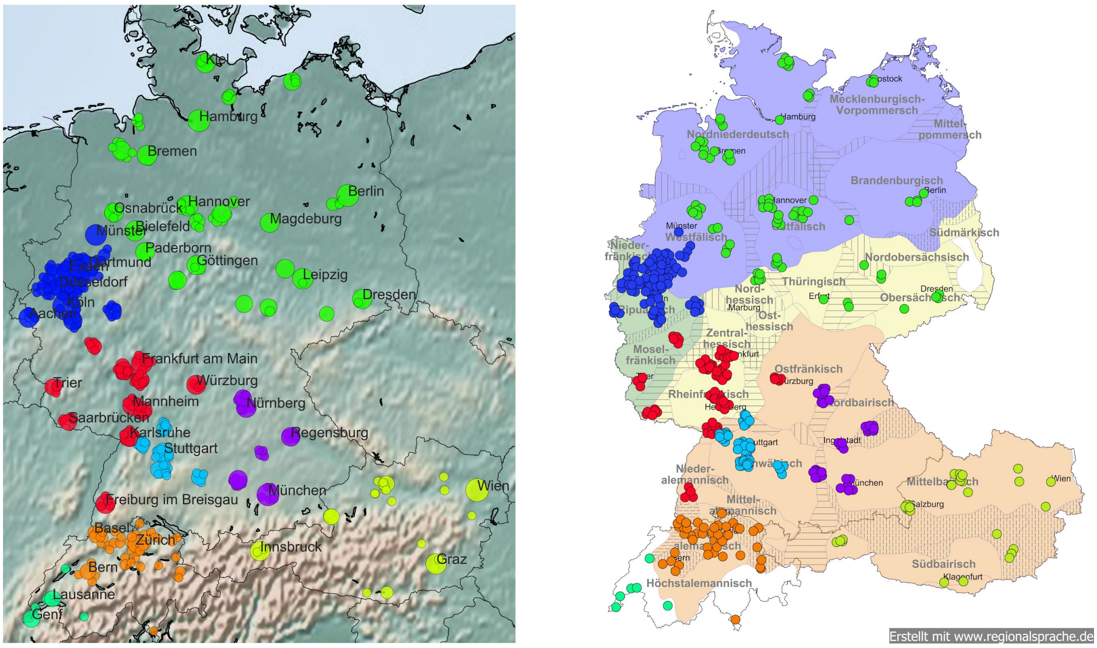

Figure 13: Eight-cluster solution

[14] Number of conversations for the top 25 locations in cluster 2: Hamburg (41809), Berlin (41773), Leipzig (36113), Hannover (35046), Bielefeld (34335), Bremen (32482), Dresden (31205), Göttingen (30210), Magdeburg (29620), **Kiel (29615)**, **Osnabrück (29479)**, Paderborn (29351), Halle (Saale) (27998), Oldenburg (26756), **Braunschweig (26625)**, **Kassel (25932)**, Rostock (25193), Erfurt (23385), Jena (22136), Hildesheim (17987), Chemnitz (15403), Lübeck (14798), Potsdam (9489), **Wolfsburg (6145)**, Bremerhaven (4941)

The eight-cluster solution (*Figure 13*) splits cluster 3^{pr}, which contained cities in Bavaria and Austria. The new cluster 3 consists of the Bavarian locations, while cluster 8 represents all locations in Austria. Given that the national border here is often believed to be a linguistic border,[15] and the long distance between the Bavarian and Austrian locations, this distinction is no surprise. The fact that it occurs so late, however, is. The prototypical words for the two new clusters transform with the split. The prototypes for cluster 3^{pr} combined both Bavarian and Austrian dialect forms. The new list for the Austrian locations (cluster 3), clearly shows the split.

Cluster 3^{pr} (Austrian & German Bavarian), 64 locations, 356k threads): *oba, owa, einfoch, hoid, waun, hob, siag, mocht, kan, afoch, ghobt, woa, ana, ois, hobi, voi, najo, obwoi, laung, haum, amoi, fia, faungt, oiso, kema*

Cluster 3 (Austrian Bavarian, 29 locations, 174k threads): *sowos, oba, owa, nu, einfoch, hoid, waun, hob, oaned, siag, wiakli, mochn, gmocht, kan, fia, afoch, ghobt, woa, freind, ana, aundare, ois, oama, hobi, kinan*

Many of the top words from cluster 3^{pr} (*oba, owa, einfoch, hoid, waun, hob, siag, kan, afoch, ghobt, woa, ana, ois, hobi*) are still prototypical for the Austrian cluster. Therefore, we can classify them as prototypical for Austrian Bavarian (though they might be present in Bavaria as well), whereas the other words (*mocht*, *voi, najo, obwoi, laung, haum, amoi, fia, faungt, oiso, kema*) are replaced by words exclusive to the Austrian community (see *Figure 14*). While this distinction is not selective (see for example *mocht* 'does' in cluster 3^{pr} that is an inflected form of *mochn* 'do' in cluster 3, or the examples of *l*-vocalization in both lists, e.g., *obwoi* 'although' vs. *hoid* 'just, simply'), it still suggests that Austrian and Bavarian words are prototypical in their respective community.

Cluster 8 (German Bavarian, 35 locations, 182k threads): *techfak, schwarz ritter, nbg, göggingen, rgbg, brezn, dult, sax, haunstetten, frauentorgraben, mhwmk, ingolstädter, techfucked, augsburger, nürnberger, dutzendteich, hochzoll, kuhsee, pfersee, lässt, deshalb, jahninsel, plärrer, tennenlohe, gerade*

In contrast, cluster 8 contains only Bavarian locations, but hardly any dialectal forms, except for *brezn* 'pretzel', a well-known Bavarian bakery product, and *dult* 'fair, funfair'. Apart from that, the Bavarian subsample shows references to locations (*göggingen*, *rbgb* 'abbreviation for Regensburg',

[15] In classical dialectology, this distinction does not hold (see *Figure 1*). However, the national border is perceived as a linguistic border by speakers from the region (Kleene 2017).

frauentorgraben ‘a street in Nuremberg’, *plärrer* ‘a place in Nuremberg, also the name of a local fair’, *dutzendteich* ‘lake in Nuremberg, *pfersee* ‘a quarter of Augsburg’), user groups (*ingolstädter, augsburger*), or facilities (*techfak* ‘technical faculty’, *schwarzer ritter, sax* ‘names of a pubs’). Comparing the three prototype lists, we see that for both Bavarian and Austrian Jodel users there is a tendency to use dialectal words in written digital communication. This tendency is stronger in Austria, demonstrated by the fact that all new items for cluster 3 are common Austrian Bavarian dialectal forms, but it gets overruled by the influence of place names, local references, and even standard German items in the Bavarian cluster. This distinction is likely linked to thematic and linguistic routines of the different Jodel communities.

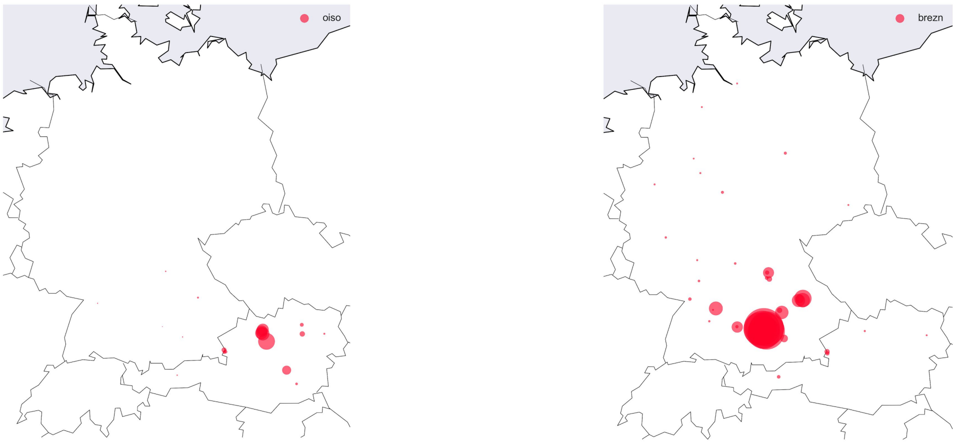

Figure 14: word maps for oiso *and* brezn

3.8 Fifteen clusters

In the 15-cluster solution, we see four important changes (*Figure 15*):

1. Most of the Northern German locations is now split into a Northwestern (purple dots) and a Northeastern part (brown dots). The only Western city still in the Eastern part is Kassel. Unlike for Würzburg, we don’t find a strong influence of students from the Eastern German federal states at the University of Kassel: their share is below 5%; most of the students in

Kassel come from Hesse, Lower Saxony, and North Rhine-Westphalia (Hessisches Statistisches Landesamt 2017). This clustering corresponds to the dialect division in Lameli 2013, though (see *Figure 1*), where Kassel is subsumed under the Eastern Middle varieties Thuringian and Saxonian. However, the prototypical words show no hints for such a regional linguistic similarity.

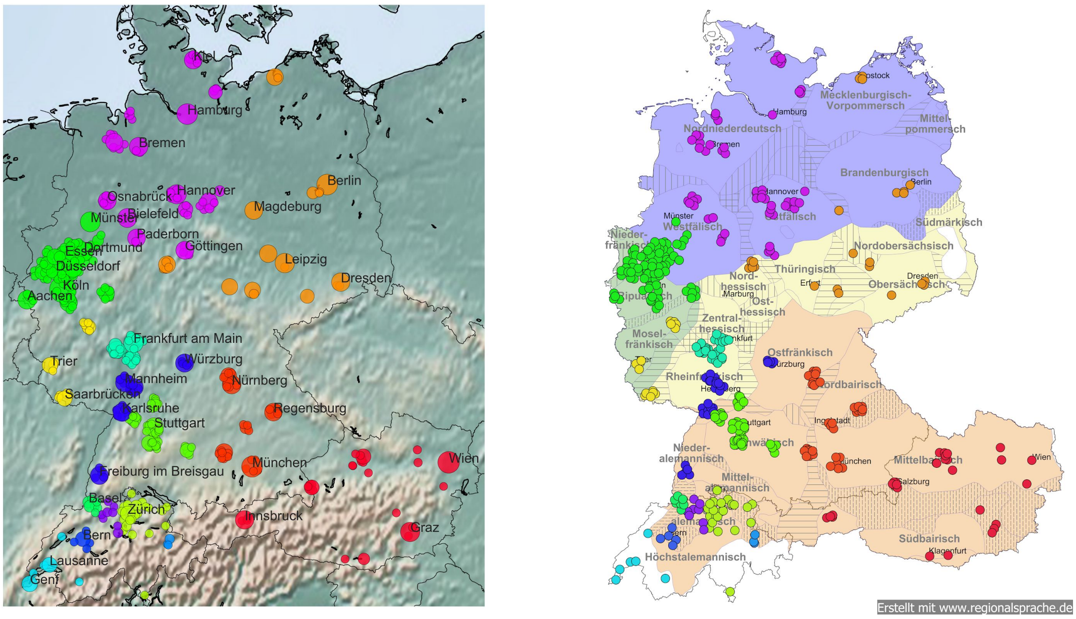

Figure 15: Fifteen-cluster solution

2. The Moselle Franconian locations (yellow dots) form a separate cluster, as opposed to Western Baden Württemberg + Franconia on the one hand and The Rhine-Main area on the other. In this case, we find evidence for an increase of regional lexical items prototypical for the user community (together with an increase in regional and local place references of course). The prototypical words for this small cluster (16 locations and 64k threads) contains many regional variants imitating phonological features: *gudd* 'good', *saarbrigge* 'Saarbrücken', *eijo* 'sure', *awwa* 'but', *bissjen* 'a bit', and *schwenker*, a specific type of barbeque grill used regionally).

3. Cluster 1, containing all of German-speaking Switzerland and Lugano, is split into several local clusters grouped around Switzerland's main cities: Basel (pale green dots), Zurich (lemon dots), and Bern (marine blue dots). Lugano is within the Zurich subcluster. Additionally, we see a subcluster that only contains three locations (skyblue dots), Chur, Lanqquart, and Plessur in the Eastern part of German-speaking Switzerland. In this region, Rhaeto-Romanic is the official language, though German is still relatively common (Lesław 2015). Apart from this, we see another cluster (violet dots) around the cities of Aarau and Luzern between the Basel and the Zurich cluster.

We take a closer look at the most typical lexical items for the Swiss-German subclusters. Traditionally, dialect maps in Switzerland have been mostly restricted to individual maps for words, phonetic features, or morpho-syntactic constructions (Scherrer & Stöckle 2016). While we can not claim any completeness or linguistic validity, our method does afford us the opportunity to investigate larger regional clusters in Switzerland based on entire vocabularies.

Cluster 1 (German-speaking Switzerland & Lugano, 46 locations, 99k threads): *gaht, wüki, nöd, besch, emmer, nor, au nöd, verstahn, muen, wükli, dänn, vode, hett, chan, rechtig, staht, sösch, abig, mached, isch de, lüüt, nanig*

- **Zurich cluster (22 locations, 49k threads):** *gaht, wüki, nöd, nödmal, vo de, au nöd, verstahn, chan, muen, wükli, gahsch, dänn, vode, hett, isch au, demit, chönd, staht, mached, eifach, abig, isch de, isch scho, git, lüüt*
- **Basel cluster (8 locations, 21k threads):** *goht, sehni, drnoch, griegsch, syy, keini, usseht, sunsch, miehsam, mol, iebig, öbbis, miesst, au nid, joor, drugge, kha, unseri, friener, isch e, kei, sälbr, joohr, priefig, bitz*
- **Bern cluster (5 locations, 20k threads):** *geit, viu, gloub, auso, aues, ig, när, ds isch, itz, aube, aui, geng, iz, vilech, ke, ds, nidmau, schnäu, froue, u, ig ha, u när, würklech, angeri, verzeu*
- **Aarau/Luzern cluster (8 locations, 10k threads):** *esch, ond, vell, besch, ech, nor, emmer, au ned, dech, wörkli, wechtig, mech, rechtig, norno, zuekonft, beni, gfonde, brengt, sösch, wössed, drom, esh, dorom, fende, ergendwie*
- **Chur cluster (3 locations, 4k threads):** *miar, diar, dia, leba, werda, aswia, wia, aswo, iar, fraua, akli, liabsta, passiart, könna, niamert, muassi, ihar, kriaga, froga, nögsta, muass, vergessa, eba, glauba, guati*

Comparing the five sublists of prototypical words to the list for cluster 1 shows several interesting aspects of dialectal variation in written Swiss-German. First, despite limited geographic extension and number of conversations, the prototypical words are distinct for all five subclusters. I.e., the Swiss varieties do not share prototypes, a feature we have seen for many of the clusters in the GSA. Since all prototypes contain common words of everyday communication, we have strong reason to believe that the five subclusters represent regional linguistic differences between the different regional user communities.

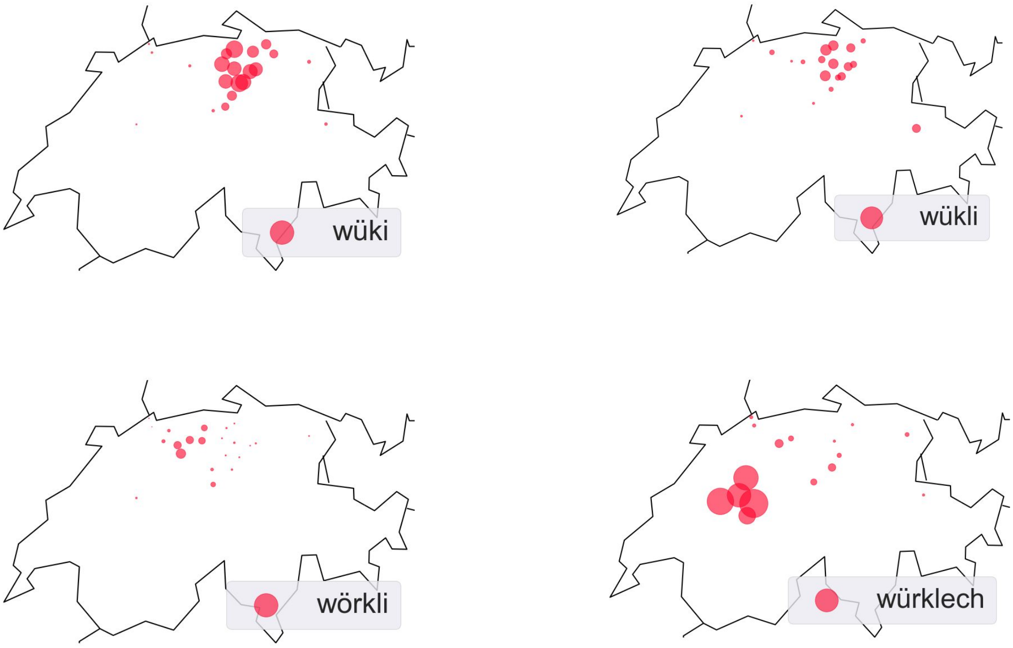

Figure 16: Regional variants of wirklich *in Switzerland:* wüki*,* wükli*,* wörkli *and* würklech

Second, comparing the prototypes for cluster 1 with the subclusters, we find that the Zurich cluster has the biggest impact on cluster 1: several prototypical words from the Zurich cluster are also in cluster 1, which is not the case for the other subclusters. Given that the Zurich cluster makes up half of the Swiss-German data, this dominance is not surprising. Third, comparing the prototypes for the five subclusters, we find dialect equivalents of standard German words that represent the dominant regional variant, e.g., *gaht* (Zurich), *goht* (Basel), *geit* (Bern) for *geht* 'goes' or *wüki* (Zurich), *wükli* (Zurich), *würklech* (Bern), *wörkli* (Aarau/Luzern) for *wirklich* 'really' (see *Figure 16*). These fine-grained distinctions correspond to the prevailing approach of studying Swiss-German variation via individual words. Fourth, we find evidence for regional linguistic features that contribute to the regional linguistic style profile of the different communities. For example, prototypes in the Chur cluster tend to end with either *-a* or *-ar* as suffixes. Similarly, the Bern cluster contains several examples of the idiosyncratic regional *l*-vocalization, as in *viu* 'viel' ('a lot'), *auso* 'also' ('well, so'), *aues* 'alles' ('everything'). These findings suggest that regional variation is the most important linguistic resource for Swiss-German Jodel users, which constitutes the practice of different regional communities.

4. Discussion

4.1 Corpus structure and linguistic resources

The overall aim of this study is to draw the connection between language practice and linguistic structure based on a large corpus of online communications. The previous chapter showed that we can indeed detect (and interpret) many of the structural differences in our data, and link them to linguistic and other aspects of everyday cultural practice. Still, in order to move beyond the bare analysis of the regional clusters found in the data, we need to take a closer look at the different factors structuring our data. In this section, we discuss some aspects of the Jodel environment, and go through some regionally defining characteristics of linguistic resources.

As mentioned previously, Jodel data is characterized by four aspects: its anonymity, user demographics, network regionality, and the (individual and group-specific) style profiles of the users. The anonymity impacts the type of analysis we can perform, but also encourages users to discuss topics we would otherwise not find in (publicly available) online communication and traditional sociolinguistic studies. Discussion often revolve around private or even intimate issues, especially relationship- and sex-related topics. This indicates the kind of interaction that dominates Jodel communications: informal communication that normally happens between peers, with not only thematic, but is also linguistic informality.

The demographics of the Jodel community shape the structure of the data. Strictly speaking, the exact demographics of the community are unknown, because of user anonymity. However, given the rise of Jodel as “campus chat”, and the thematic spectrum in the threads, it is likely correct to assume a young adult audience. This in turn means that the language practice we observe represents the last step in the development of language dynamics in German. I.e., the users’ linguistic repertoires likely contain mainly standard German (and maybe regiolects), but no dialects, except for users from Switzerland, Austria, and some parts of Bavaria.

The regionality of the networks (based on the 10km reach of each post), inevitably structures the data, especially given the sparsity of our location raster. This effect is clearly visible in the geographic distribution for most cities: several smaller locations surrounding one of the larger cities. On the one hand, we can therefore expect our data to be highly pre-structured by the regional binding of Jodel communications: users will resort to existing institutionalized resources. On the other hand, interaction in regionally-bound communities fosters the coining of group-specific words for specific purposes (e.g., *lörres*, *möppes*) as well as technical terms (e.g., *@vj*), and writing conventions (e.g., the regional variants *gaht, goht, geit* ‘goes’ in case of the Swiss-German dialects).

That being said, individual and group-specific style profiles also structure the data, especially with respect to the projection of specific linguistic (and therefore social) identities online. We see the influence of various linguistic resources that contribute to both individual styles (which we did not analyze in this study) as well as community-specific style profiles (Coupland 2007). The analysis showed that there are regional communities which use dialectal or regiolectal forms (Switzerland, Austria), while others employ app-specific pragmatic markers to respond to specific posts (especially users in the middle and southern part of West Germany).

The many individual choices the users make aggregate into group-specific writing styles. Compared to regional varieties, they depend on various linguistic resources, including *lifeworld orientation terms* (i.e, place and location names, recontextualized sociocultural references), *medium-specific organization terms* (i.e., pragmatic markers for author reference), *traditional* *regional items* (i.e., dialectal and regiolectal words), *community-specific regional items* (i.e., coinings for specific thematic aspects, hashtags), typical *items of online communication* (i.e., emoji, logograms, or rebus writing), *foreign-language and borrowed items*, and *standard German items*.

- *Lifeworld orientation terms* include words that refer to places, locations, facilities, or specific sociocultural contexts/items of particular importance for the structuring and interconnectedness of regional communities. Their prevalence is easily explained by the fact that these terms refer to elements in the lifeworld like cities, administrative districts, universities, local user groups, or sociocultural items (like football clubs). These terms thereby create frames of reference for speech acts to contextualize the users' communications and connect them to different aspects of lifeworld practice.

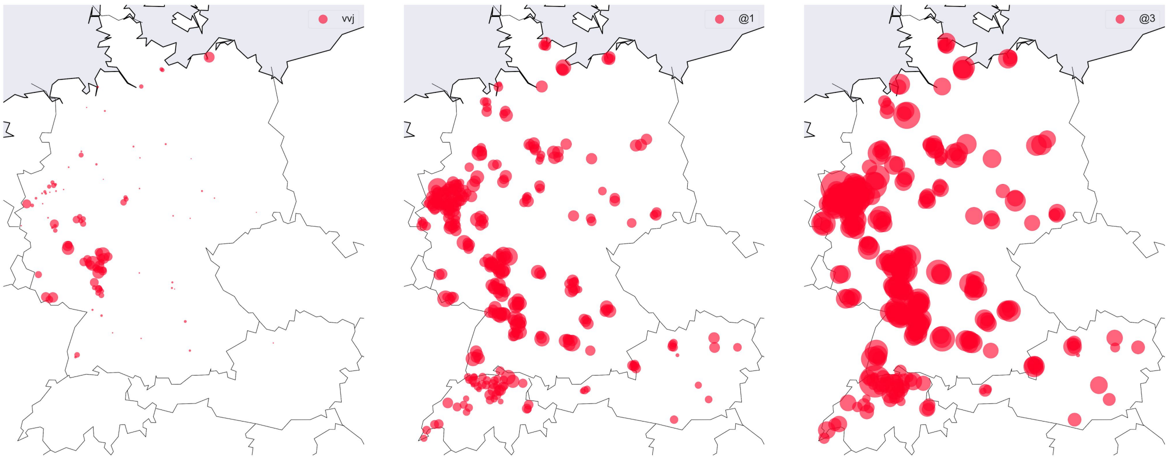

Figure 17: Word maps for vvj *(left),* @1 *(middle) and* @3 *(right)*

- *Medium-specific organization items:* The relatively small group of Jodel-specific communication markers (*vj*, *@vvj*, *oj* 'original Jodler' etc.) contributes significantly to the overall linguistic structure of the community. They serve concrete discourse-pragmatic functions that enable users to answer directly to previous posts and authors, despite their anonymity. As such, they have to be learned by new users in order to efficiently communicate. These markers seem to be bound to a regional community in western middle

Germany, but are also spread across larger areas of the GSA. However, these terms will likely soon change dramatically, since a recent update to the app (May 2017) assigns each unique participant in a conversation an anonymous ID other users can reply to (e.g., *@3*). We therefore expect terms like *vj* and *@vvj* to disappear in a follow-up study, and to be replaced by the ID references. *Figure 17* shows that they are already established in the entire community shortly after their introduction in late May 2017.

- *Traditional regional items:* For some regional communities in our corpus, using dialectal or regiolectal forms is constitutive for the overall structuring of communication practice on Jodel. This is especially true for Switzerland, where the regional dialects are the default writing styles for all interaction, but also for regional forms in the cluster that contains all Austrian locations. Given that our model can not detect stop words in regional varieties (unlike their standard German equivalents), this is an artifact of data preprocessing, but holds for any regional community with regional language use. Indeed, other regions are partly defined by the use of regional words as well (Bavaria, parts of Western Middle and Upper Germany), but mesh there with other linguistic resources to structure the data.
- *Community-specific regional items:* This group of items constitutes the equivalent of traditional regional lexis, as they are thematically motivated and arise out of the interaction with other users in the Jodel community. In some cases (e.g., *lörres* or *möppes*) they originate from regional varieties, and get recontextualized (*lörres*) or re-semanticized (*möppes*) in the wider Jodel community. For example, we find several examples of word formation and word play with *lörres* (and also *möppes*) in our data like *@lörres* 'reference to someone using lörres in a previous post', *singlelörres* 'a single male person', *lörres22* 'male user of 22 years', *justmöppesthings*, or *einsames möppes* 'lonely female user'. Other items arise as discourse labels, such as hashtags for entire conversations, e.g., *dummwiekarlsruhe* 'stupid like Karlsruhe', *dermitdemsonnenbrand* 'the one with the sun burn', *flirtenaufnordhessisch* 'flirting in Northern Hessian', or *aufdemwegzurarbeit* 'on my way to work'. These words also

reflect the thematic scope of many Jodel conversations. They highlight how people creatively use their linguistic resources to innovate precise. regionally-bound, thematically-motivated, and socially-binding words for specific regional communities and communication purposes.

- *Items of online communication:* As with every other social medium, we find online-specific linguistic resources such as abbreviations (*jmd* 'jemand', 'somebody'), acronyms (*uds* 'Universität des Saarlandes'), rebus writing (*3st* 'dreist', 'shameless'), simple and complex emoji constructions (😂, 🆗🆒, 🙈🙋 🤦♂), or logograms (= 'equals'). None of these items are specific to the Jodel community, but might be used, combined, or recontextualized in community-specific ways, especially emoji or abbreviations (e.g., licence plate codes), contributing to regional writing styles of Jodel communities.
- *Foreign-language and borrowed items:* Within the German-based user communities we find hardly any examples of lexical borrowing or loanwords apart from sparse evidence like the use of *lelek* 'nightjar' in the Rhine-Main-area: In the list of 100 prototypical words for this cluster (see *Section 4.2*) we find two more examples for this resource, i.e., *akhis* 'brother' and *sharmut* 'bitch', both of which are common ethnolectal variants in German that originate from Arabic.
- *Standard German items:* For many clusters, standard German forms (including colloquial variants like *gucken* 'look') contribute considerably to the style profile of a community, especially those with many locations or those spread across larger regions of the GSA. Use of standard German generally constitutes the absence of other linguistic resources in the formation of community-specific style profiles. However, the contribution of standard German lexical items defines an important characteristic, at least in some regional communities where regiolects have already replaced the old local base dialects (see *Section 2.1*).

Together, these resources reflect the sociolinguistic structure of language practice in the Jodel community. Regional clusters can therefore be characterized by their use, combination, and mesh-up of these different linguistic resources.

4.2 Linguistic resources and style profiles

Based on the idea of community-specific style profiles, we re-examine the prototypes of the 15 cluster solution. For each cluster, we classify the top 100 prototypical words into one of the seven linguistic resources identified in the last section. This results in style profiles for each regional user community as shown in *Figure 18.*

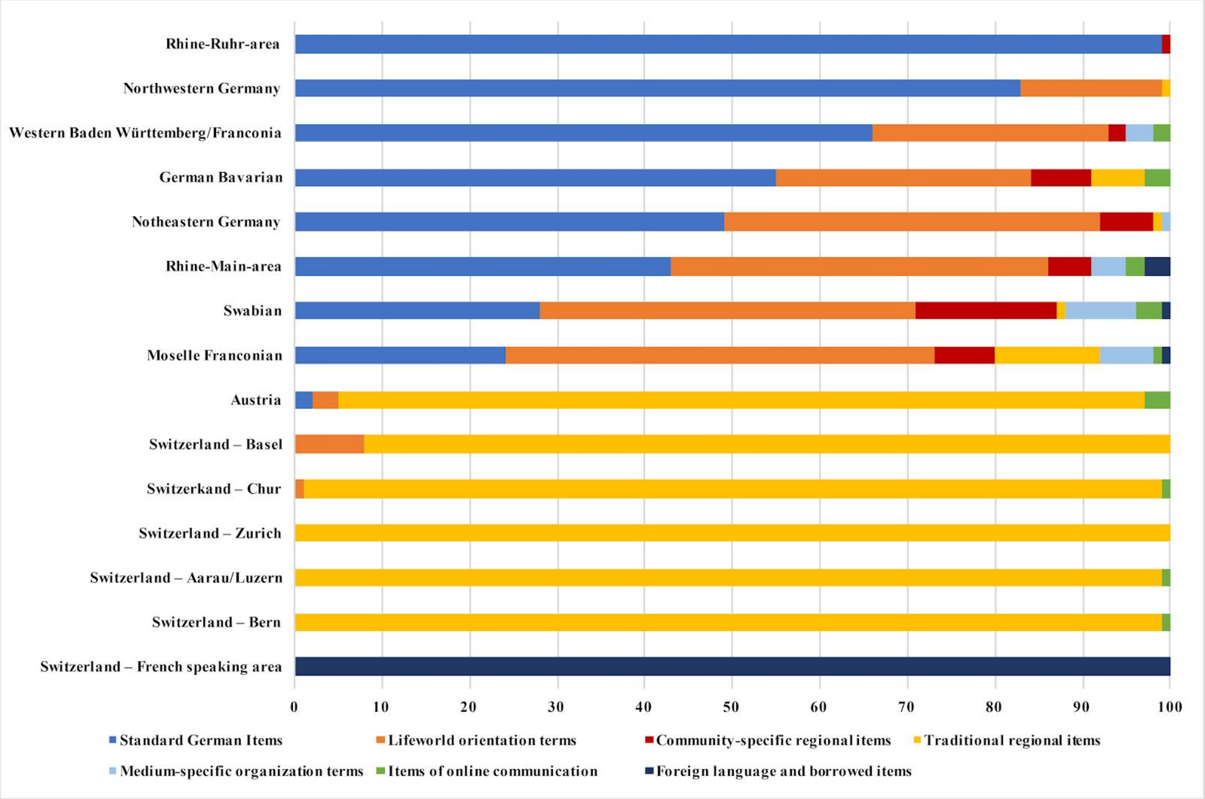

Figure 18: Style profiles for the 15 regional clusters (see Figure 15 above).

Three linguistic resources drive most of the internal structure of the style profiles: Standard German items, lifeworld orientation terms, and traditional regional items. Regional language distinguishes Austrian and Swiss communities from their German equivalents: regional forms are the main (and sometimes only) resource characterizing community prototypes. For the German communities, we see an anti-correlation between the use of Standard German items and lifeworld orientation terms: the more Standard German prototypes we find in a regional community, the fewer lifeworld orientation terms contribute to the community style profile. Apart from these three resources, the use of community-specific regional items is characteristic for some communities.

Especially the western middle- and southern-German communities are characterized by their typical use of medium-specific organization terms. The Rhine-Main-area is the only regional community with at least some influence of foreign language and borrowed items (apart from French-speaking Switzerland of course). Even though the analyzed prototypes are just a fraction of the total vocabulary, we find distinct style profiles that characterize regional communities. Note again that the profiles are influenced by the number of threads and density of locations (e.g., northwestern Germany vs. Rhine-Ruhr-area), the default-variety choice (e.g., French-speaking Switzerland vs. Switzerland/Austria vs. Germany), the regional prototypicality of the vocabularies (e.g., *@3* vs. *@vj*), and the size, scale, and lifeworld anchoring of the different linguistic resources. I.e., compare Standard German, a highly institutionalized resource with extensive vocabulary usable for all lifeworld purposes, to the community-specific organization terms, a newly established resource restricted to few items and purposes usable only in a specific communicative context. And so while the size difference of the associated communities makes them directly incomparable, the style profiles can still be used as an overall diagnostic tool. They provide evidence for (dis)similarities between regional user groups, especially for the smaller and less institutionalized resources. For example, the use of medium-specific organization terms is a shared characteristic between the communities in the south west of Germany. As we have seen in *Figure 6*, terms like *@vj*, *möppes*, and *lörres* seem to be characteristic for these communities, from where they may spread to other regional communities in the entire GSA. These communities foster the establishment and spread of community- and medium-specific vocabulary, and, therefore, substantially influence the structuring of linguistic practice in the entire Jodel community.

4.3 Methodology

While neural networks have become dominant in engineering disciplines in recent years, due to their superior performance, they have also been criticized for their need of large data sets, their sensitivity to parameter settings, and the difficulty to interpret the final models. Within the class of neural networks, however, representation learning is a comparatively safe and simple approach. In our case, the amount of available data and the small number of parameters justify the choice of model. In addition to being relatively simple and well-understood, there exists literature on optimal parameter settings. However, in any quantitative approach, the methodological decision we make influences the results we get. Therefore, we discuss some methodological choices relating to data collection, processing, and visualization.

A main influencing factor of our methodology is its dependence on large data sets: we see diverging results for cities and regions with insufficient coverage. The most prominent cases are Lugano, in the predominantly Italian-speaking southernmost part of Switzerland, and the French-speaking area in Switzerland. Both should theoretically be dissimilar enough to the other (German-speaking) cities to merit its own cluster. However, given the relative uniformity in most of the rest of the GSA, the differences between those two areas and German-speaking Switzerland is still smaller than between all of Switzerland and the rest. To some extent, these distinctions are established during preprocessing: the tools we use are set up for standard German, so we are not able to achieve the same accuracy for dialects or foreign languages. However, since the same holds for regional items in the German user communities, this effect does not reduce the meaningfulness of the distinctions.

A second methodological aspect relates to the geographic raster and data collection. While location activity and sequential access through the API are good reasons to limit data collection to bigger cities, this choice affects the coverage: we do not collect any data for less populated areas, while data from higher population-density regions dominate. The resulting imbalance affects the clustering and style profiles of the regional user communities, for example by changing the ratio of standard German to lifeworld orientation terms within a cluster.

Similar effects hold for several other choices: the removal of named entities and stop words during preprocessing, which affect vocabulary size and the learned representations; the down-sampling of frequent words during representation learning, which impacts prototypes; the assumption that geographically close places are also linguistically more similar, which favors local coherence; and for the limitation to cities with more than 200 threads during clustering, which affects cluster solutions (including more locations leads to fuzzier clusters). For all of these decisions, a grounding in linguistic theory is just as important as an understanding of the algorithmic effects.

Regarding our combination of data-driven and interpretation-driven methods, we conclude that it confers crucial advantages for our type of study, because both methods contribute complementary types of resources. The learned representations can both be visualized as maps and interpreted as prototypes, allowing us to interpret the model output and estimate its quality. The quantitative approach clustering of learned representation allows us to leverage large data sets and forego linguistic presuppositions, while the qualitative analysis of regional style profiles and the structuring of practice relies on well-established knowledge about the sociolinguistic and culture-theoretical structure of language use in practice. The results demonstrate that a judicious application of both can provide unique insights that would be impossible when using either approach in isolation.

5. Conclusion

In this paper, we have investigated the automatic detection of regional patterns in large amounts of young adults' online communications in the German speaking area. We collect a corpus of 2.3 million threads from the anonymous chat app Jodel and fit a representation learning model on the data. The resulting vector representations of words and cities allow us to visualize dimensions of variation, and to use Ward clustering to detect regional patterns.

Our analysis reveals that the regional clusters represent distinct community-specific language practices in Jodel that are fed by different linguistic resources, including community- and medium-specific vocabulary, lifeworld orientation terms, traditional regional items, and standard German. These elements can be used to describe characteristic style profiles of regional user communities. Most notably, we find a difference between user communities that employ regiolects or dialects online (German-speaking Switzerland, Austria, and – to some extent – Bavaria) and communities that are mainly characterized by standard German lexis and the use of region-specific lifeworld orientation terms. The establishment and spread of Jodel-specific linguistic elements can be explained as the result of processes of sociocultural structuration in practice (Purschke forthcoming).

In sum, the distinct regional structures mirror region-specific style profiles of the Jodel community. Remarkably, our quantitative model is able to detect and cluster such regional user groups and style profiles based solely on their written communications in Jodel, without any assumptions about regional language differences in the GSA. Both the clear regional clustering and the close linkage with the German regional languages indicate the viability of our approach. The overall structure of the clusters closely matches the traditional division of German dialects. Diverging cases (Kassel, Würzburg, Switzerland) can be explained through external structures, e.g., national borders, areas of socioeconomic exchange, and sociocultural orientation. Our study demonstrates that the combination of quantitative and qualitative analysis methods can uncover effects in large-scale data, provide the basis for in-depth analyses, hypothesis-generation, and corroboration of existing hypotheses. Our work advances the idea of a complementary use of computational linguistics and sociolinguistics to their mutual benefit. However, we also discuss the need for both disciplines to make informed decisions regarding the processing, analysis, and visualization of the data.

Our results offer a promising starting point for future research in this vein, especially the linguistic dynamics of communities (spread and vanishing of community- and medium-specific items), aspects of discourse organization, the spectrum of topics covered in the threads, and community conventions for regiolectal writing. Future studies should also consider a revalidation of the clustering with a second data set and denser location raster. Another aspect relates to the notion of ‘regional’ as used in the analysis of language variation. As our study reveals, not only dialect vocabulary is bound to regional user communities, but also other linguistic resources like lifeworld orientation terms or community-specific terms are constitutive for regional user groups and style profiles. Thus, our results also invite us to rethink established concepts like “regiolect” in the digital era.

6. Bibliography

Androutsopoulos, Jannis. 2003. Online‑Gemeinschaften und Sprachvariation. Soziolinguistische Perspektiven auf Sprache im Internet. *Zeitschrift für germanistische Linguistik* 31(2): *Deutsche Sprache in Gegenwart und Geschichte*, 173‑197.

Androutsopoulos, Jannis. 2007. Neue Medien. Neue Schriftlichkeit? *Mitteilungen des Germanistenverbandes* 54 (1): *Medialität und Sprache*, 72–97.

Androutsopoulos, Jannis. 2013. Online Data Collection. In Mallinson, Christine, Becky Childs, & Gerard Van Herk (eds.). *Data Collection in Sociolinguistics: Methods and Applications*, 236–249. London: Routledge.

Bamman, David, Chris Dyer & Noah Smith. 2014a. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 828–834.

Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2), 135–160.

Barton, David & Carmen Lee. 2013. *Language Online. Investigating Digital Texts and Practices.* London/New York: Routledge.

Cheshire, Jenny. 2005. Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers. *Journal of Sociolinguistics* 9(4), 479–508.

Coupland, Nikolas. 2007. Style: Language Variation and Identity. Cambridge: CUP.

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In *EACL*.

Dürscheid, Christa & Karina Frick. 2016. *Schreiben digital. Wie das Internet unsere Alltagskommunikation verändert*. Stuttgart: Kröner Verlag.

Dürscheid, Christa & Elisabeth Stark. 2013. Anything goes? SMS, phonographisches Schreiben und Morphemkonstanz. In Neef, Martin & Carmen Scherer (eds.), *Die Schnittstelle von Morphologie und geschriebener Sprache*, 189–210. Berlin: De Gruyter.

Eisenstein, Jacob. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19(2), 161–188.

Eisenstein, Jacob 2013a. Phonological factors in social media writing. In *Workshop on Language Analysis in Social Media, NAACL.*

Eisenstein, Jacob. 2013b. What to do about bad language on the internet. In *Proceedings of NAACL.*

Eisenstein, Jacob, Brendan O’Connor, Noah Smith & Eric Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287.

Eisenstein, Jacob, Noah Smith & Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of ACL.*

Falck, Oliver, Stephan Heblich, Alfred Lameli & Jens Südekum. 2012. Dialects, Cultural Identity, and Economic Exchange. *Journal of Urban Economics* 72, 225–239).

Goldberg, Yoav. 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies* 10(1). San Rafael, California (USA): Morgan & Claypool Publishers.

Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(02), 193–221.

Heblich, Stephan, Alfred Lameli & Gerhard Riener. 2015. The Impact of Regional Accents on Economic Behavior: A Lab Experiment on Linguistic Performance, Cognitive Ratings and Economic Decisions. *PLoS ONE* 10/2: e0113475. DOI: https://doi.org/10.1371/journal.pone.0113475

Herring, Susan. 2013. Discourse in Web 2.0: Familiar, reconfigured, and emergent. In Tannen, Deborah & Anna Trester (eds.), *Discourse 2.0: Language and New Media*, 1–25. Washington: Georgetown University Press.

Hovy, Dirk, & Anders Johannsen. 2016. Exploring language variation across Europe. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association (ELRA).

Hovy, Dirk, Anders Johannsen & Anders Søgaard. 2015. User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW.*

Johannsen, Anders, Dirk Hovy & Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL.*

Hovy, Dirk, Afshin Rahimi, Tim Baldwin, & Julian Brooke. 2019. Visualizing Regional Language Variation Across Europe on Twitter. In: Stanley D. Brunn and Roland Kehrein (Eds.) *Handbook of the Changing World Language Map*. Dordrecht: Springer.

Jones, Tyler. 2015. Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech* 90(4), 403–440.

Kehrein, Roland. 2012. *Regionalsprachliche Spektren im Raum – zur linguistischen Struktur der Vertikale.* (ZDL Beihefte 152). Stuttgart: Franz Steiner Verlag.

Kitchin, Rob. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society,* 1(1).

Kleene, Andrea. 2017. *Attitudinal-perzeptive Variationslinguistik im bairischen Sprachraum. Horizontale und vertikale Grenzen aus der Hörerperspektive*. Dissertation. University of Vienna.

Koch, Peter & Wolf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15–43.

Kristiansen, Tore. 2009. The macro-level social meanings of late-modern Danish accents. *Acta Linguistica Hafniensia* 41, 167–192.

Kulkarni, Vivek, Bryan Perozzi, & Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. *Proceedings of ICWSM*, 615–618.

Lameli, Alfred. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. (Linguistik – Impulse und Tendenzen 54). Berlin/Boston: Walter de Gruyter.

Lameli, Alfred, Volker Nitsch, Jens Südekum & Nikolaus Wolf. 2015. Same Same But Different: Dialects and Trade. *German Economic Review* 16(3), 290–306.

Landauer, Thomas & Susan Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review* 104(2), 211–240.

Lau, Jey & Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *ACL 2016*, 78.

Le, Quoc, & Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*.

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain & Elvira Glaser. 2016. Crowdsourcing language change with smartphone applications. *PLoS ONE* 11(1): e0143060. DOI: https://doi.org/10.1371/journal.pone.0143060

Leemann, Adrian, Marie-José Kolly, Stephan Schmid & Volker Dellwo (eds.). 2015. *Trends in Phonetics and Phonology. Studies from German-speaking Europe.* Frankfurt am Main: Peter Lang.

Lesław, Tobiasz. 2015. Die sprachliche Vielfalt Graubündens – ein Phänomen in der viersprachigen Schweiz. *Linguistica Silesiana* 36, 209–230.

Nerbonne, John & Wilbert Heeringa. 1997. Measuring dialect distance phonetically. *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97).*

Nguyen, Dong. 2017. *Text as Social and Cultural Data. A Computational Perspective on Variation in Text.* Enschede: Universiteit Twente. DOI: 10.3990/1.9789036543002

Nguyen, Dong, Seza Doğruöz, Carolyn Rosé & Franciska de Jong. 2016. Computational sociolinguistics: A survey. Computational Linguistics, 42(3), 537–593.

Östling, Robert & Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 644–649. Association for Computational Linguistics.

Prokić, Jelena & John Nerbonne. 2008. Recognising groups among dialects. *International journal of humanities and arts computing* 2(1/2), 153-172.

Pröll, Simon, Simon Pickl, & Aaron Spettl. 2014. Latente Strukturen in geolinguistischen Korpora. In Elmentaler, Michael, Markus Hundt, Jürgen Erich Schmidt (Hg.): *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder.* (ZDL Beihefte 158), 247–258. Stuttgart: Steiner.

Purschke, Christoph. forthcoming. Language regard and cultural practice. Variation, evaluation, and change in the German regional languages. In Evans, Betsy, Erica Benson & James Stanford (eds.), *Language regard: Methods, variation, and change*, 245–261. Cambridge: Cambridge University Press.

Purschke, Christoph. 2011. *Regionalsprache und Hörerurteil. Grundzüge einer perzeptiven Variationslinguistik.* (ZDL Beihefte. 149). Stuttgart: Franz Steiner Verlag.

Rahimi, Afshin, Timothy Baldwin, & Trevor Cohn. 2017a. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. *Proceedings of Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Rahimi, Afshin, Trevor Cohn, & Timothy Baldwin. 2017b. A neural model for user geolocation and lexical dialectology. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Scherrer, Yves & Philipp Stöckle. 2016. A quantitative approach to Swiss-German − Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica* 24, 92−125.

Schlobinski, Peter (ed.) 2006. *Von *hdl* bis *cul8r*. Sprache und Kommunikation in den Neuen Medien.* Mannheim: Duden-Verlag.

Schmidt, Jürgen Erich. 2010. Language and space: the linguistic dynamics approach. In Peter Auer & Jürgen Erich Schmidt (eds.), *Language and Space. An International Handbook of Linguistic Variation. Vol. 1: Theories and Methods*, 201– 225. (Handbooks of Linguistics and Communication Science. 30.1). Berlin/New York: de Gruyter Mouton.

Schümann, Michael. 2011. Hochdütsch isch en seich – Geschriebenes Schweizerdeutsch bei Twitter. In Ganswindt Brigitte & Christoph Purschke (eds.), *Perspektiven der Variationslinguistik. Beiträge aus dem Forum Sprachvariation*, 239–254. (Germanistische Linguistik. 216–217). Hildesheim: Olms.

Shackleton Jr, Robert G. 2005. English-American speech relationships: A quantitative approach. *Journal of English Linguistics* 33(2): 99-160.

Statistisches Bundesamt. 2016. Studierende an Hochschulen. Fachserie 11 Reihe 4.1. Wintersemester 2015/2016. Wiesbaden: Statistisches Bundesamt. <https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Hochschulen/StudierendeHochschulenEndg.html> (31.07.2017)

Stoeckle, Philipp. 2014. *Subjektive Dialekträume im alemannischen Dreiländereck.* (Deutsche Dialektgeographie. 112). Hildesheim, Zurich & New York: Olms.

Szmrecsanyi, Benedikt. 2008. Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2.1-2 (2008): 279-296.

Thurlow, Crispin & Kristine Mroczek (eds.). 2011. Digital Discourse. Language in the New Media. Oxford: Oxford University Press.

Tophinke, Doris & Evelyn Ziegler. 2014. Spontane Dialektthematisierung in der Weblogkommunikation: Interaktiv-kontextuelle Einbettung, semantische Topoi und sprachliche Konstruktionen. In Cuonz, Christina & Rebekka Studler (eds.), *Sprechen über Sprache. Perspektiven und neue Methoden der Einstellungsforschung,* 205–242. Tübingen: Stauffenburg Verlag.

Wieling, Martijn, John Nerbonne & Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS ONE* 6(9), e23613. DOI: https://doi.org/10.1371/journal.pone.0023613.

Wiesinger, Peter. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert Ernst Wiegand (eds.), *Dialektologie: ein Handbuch zur deutschen und allgemeinen Dialektforschung Vol. 2*, 807–900. (Handbooks of Linguistics and Communication Science. 1.2). Berlin: de Gruyter.