Exploring Diachronic and Diatopic Changes in Dialect Continua: Tasks, Datasets and Challenges

Melis Çelikkol[*1] **Lydia Körber**[*1] **Wei Zhao**[2]
[1]University of Heidelberg [2]University of Aberdeen
{firstname.lastname}@stud.uni-heidelberg.de
wei.zhao@abdn.ac.uk

Abstract

Everlasting contact between language communities leads to constant changes in languages over time, and gives rise to language varieties and dialects. However, the communities speaking non-standard language are often overlooked by non-inclusive NLP technologies. Recently, there has been a surge of interest in studying diatopic and diachronic changes in dialect NLP, but there is currently no research exploring the intersection of both. Our work aims to fill this gap by systematically reviewing diachronic and diatopic papers from a unified perspective. In this work, we critically assess nine tasks and datasets across five dialects from three language families (Slavic, Romance, and Germanic) in both spoken and written modalities. The tasks covered are diverse, including corpus construction, dialect distance estimation, and dialect geolocation prediction, among others. Moreover, we outline five open challenges regarding changes in dialect use over time, the reliability of dialect datasets, the importance of speaker characteristics, limited coverage of dialects, and ethical considerations in data collection. We hope that our work sheds light on future research towards inclusive computational methods and datasets for language varieties and dialects.

1 Introduction

Language continuously changes, varies and transforms on all levels of linguistics. Research in sociolinguistics assumes five dimensions of language variation, the so-called diasystem, that are mutually influential: diaphasic (situation), diamesic (medium), diastratic (social group), diachronic (time), and diatopic (space), as shown in Figure 1 (Zampieri et al., 2020).

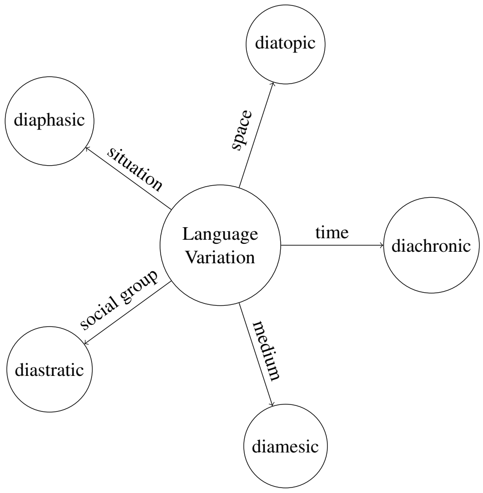

Figure 1: Language variation and the diasystem.[1]

Diaphasic, diamesic, diastratic and diatopic variation can be grouped to synchronic variation, as opposed to diachronic variation which spans several points in time. Diachronic variation is not limited to decades and centuries, but may already be observed within years, months, and even weeks or days. Especially with computer-mediated communication and social media platforms, language change appears to spread at a faster pace (Eisenstein et al., 2014). This exposes a challenge in NLP applications, as models remain static after training and struggle to understand the evolving nature of language[2]. As a result, model performance decreases over time. For instance, headline generation models decrease in performance after a few years, while emoji prediction models do so even within a month (Søgaard et al., 2021). As shown in the (socio-)linguistic work (Beeching, 2006), diachronic and synchronic variation are closely linked in the sense that language change often manifests first in synchronic variation

[*] These authors contributed equally to this work.

[1] Inspired by: http://phylonetworks.blogspot.com/2015/06/the-diasystematic-structure-of.html, accessed on 11.03.2024.

[2] Although there are methods to keep models up-to-date, such as re-training, fine-tuning, and RAG (Retrieval-Augmented Generation) leveraging up-to-date information sources at inference, the process is time-consuming and costly.

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 12–22
August 15, 2024 ©2024 Association for Computational Linguistics

before entering a diachronic level. Additionally, there is a strong spatial component in language change, as language change is caused by contact between people and speech communities (lately by online interactions too), which gives rise to dialects (Jeszenszky et al., 2018). While *isoglosses* separate dialects by drawing the geographic boundaries, the consensus among dialectologists and sociolinguists today is to speak of *dialect continua*, which assume gradual transitions between central areas of different dialects over time (Jeszenszky et al., 2018). In these continua, as proposed by the *wave model* (Wolfram and Schilling-Estes, 2017), language change is propagated from a certain locus at a certain point in time and spread layer-wise, radiating from the central point of contact. This is indeed a result of both spatial (diatopic) and temporal (diachronic) interactions within dialect continua.

An example of diatopic variation over time can be seen in Figure 2: the usage of the German dialect word *bissel* (a bit). A query in the ZDL-Regionalkorpus (Nolda et al., 2021, 2023), a collection of regional newspaper texts from Germany, Austria, and Switzerland, reveals its constant usage in Austria (A), and an increased usage in other, more northern regions over time, first in Bavaria (D-Südost) possibly due to geographic proximity, and a more recent rise in Central Eastern Germany (D-Mittelost).

In this work, we explore the intersection of diachronic and diatopic changes in language variants and dialects within the NLP community. To do so, we investigate nine tasks and datasets across five dialects from three language families to address the following research questions:

- What are the characteristics of dialect datasets across different time periods and geographic areas, and what NLP tasks have been established based on these datasets (§3)?
- What is the current state of computational methods and their results in these dialect-related NLP tasks (§4)?
- What are the challenges in dialect NLP research that have not been addressed in previous works (§5.1)?

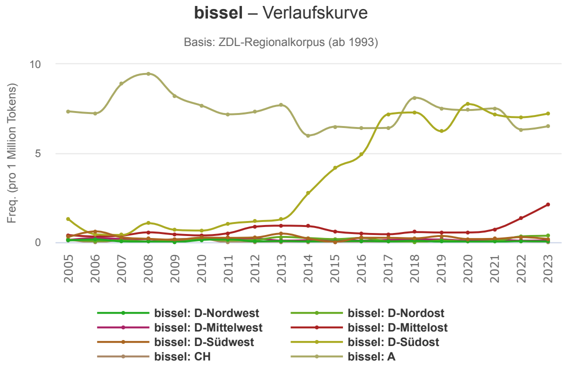

Figure 2: Diachronic usage of *bissel* in the years 2005-2023 in a regional newspaper corpus of German across dialect areas in frequency per 1M tokens.[3]

In this work, we aim at exploring the intersection of diachronic and diatopic variation in dialect NLP research. Research questions on this topic include (a) how to detect and quantify language change in dialect continua over time, and (b) how to build and process diachronic-diatopic datasets. Previous approaches leveraged machine learning methods to compute the distance and similarity of different varieties on various linguistic levels such as graphemics (Waldenberger et al., 2021), syntax (Jeszenszky et al., 2018; Chen et al., 2024), phonetics (Boldsen and Paggio, 2022; He and Zhao, 2024), semantics (Montanelli and Periti, 2023; Ma et al., 2024b,a), and built diachronic-diatopic datasets in both written and spoken modalities (Kopřivová et al., 2014; Komrskova et al., 2017).

Here, we critically review nine tasks and datasets, highlighting their strengths and limitations, as well as identifying challenges that have not been previously addressed. We discuss seminal works in Indo-European languages and their varieties, as well as recent works on this topic. The dialect continua covered here include the Slavic family with the Czech dialect landscape (Kopřivová et al., 2014; Komrskova et al., 2017), the Romance language family with Italian (Ramponi and Casula, 2023) and Portuguese (Pichel Campos et al., 2018; Zampieri et al., 2016), as well as the Germanic language family with Swiss German (Jeszenszky et al., 2018, 2019) and historical German varieties (Dipper and Waldenberger, 2017; Waldenberger et al., 2021).

2 Related work

To our knowledge, there is no survey examining the intersection of diachronic and diatopic variation

[3] Usage graph for *bissel*, created with Digitales Wörterbuch der deutschen Sprache (DWDS, Digital Dictionary of the German Language), `https://shorturl.at/9XVwt`, accessed on 04.07.2024.

in dialect NLP so far. However, there are survey papers examining the diachronic and diatopic components separately, which will be briefly presented here. Diachronic language modeling has been surveyed with regard to embeddings (Kutuzov et al., 2018) and semantic shift detection (Montanelli and Periti, 2023).

The comprehensive survey on diatopic language modelling by Zampieri et al. (2020) evaluates computational methods for processing similar languages, language varieties, and dialects, with a focus on diatopic language variation and integration in NLP applications. The work identifies the availability of suitable data as a key challenge, as the classical NLP data sources like newspaper text and Wikipedia do not cover dialectal data. Instead, social media posts and speech transcripts can be used. More recently, an evaluation benchmark for different NLP tasks in dialects, varieties and closely-related languages, DIALECTBENCH, was published (Faisal et al., 2024), proving that variation is of current interest in the research community. Joshi et al. (2024) survey Natural Language Understanding and Generation in dialects, without taking other axes of the variation diasystem into account. There exists a designated series of workshops on NLP for Similar Languages, Language Varieties, and Dialects (VarDial)[4], which also proposes several NLP shared tasks in dialects and other varieties, such as dialect classification and identification itself. Even though the workshop has featured a number of publications and talks dealing with the intersection of diachronic and diatopic variation over the years (Sukhareva and Chiarcos, 2014; Baldwin, 2018; Vidal-Gorène et al., 2020), this has not been a separate workshop or shared task topic up until now.

3 Tasks and Datasets

In this section, we review the dialect-related tasks and datasets from a unified perspective considering both diachronic and diatopic aspects, and organize them by languages (See Table 1 for tasks and datasets, and Table 2 for data statistics).

3.1 Czech

A very interesting albeit not very recent paper by Kopřivová et al. (2014) explains the building process of their later released ORTOFON and DIALEKT corpora (Komrskova et al., 2017). Although both papers are mention-worthy, we focus on Kopřivová et al. (2014) due to the presentation and depth of explanation for the data collection processes.

The ORTOFON corpus relies on spontaneous conversations recorded between 2012-2017, where nobody was aware that the conversations were recorded except for the person who made the recording. The non-scripted interactions recorded this way are then separated into the closest one of 12 situation categories which were created with the topics of Czech daily-life in mind. What makes this corpus really strong is that Kopřivová et al. (2014) consider several missing elements in other corpora all at once: relationship between speakers is noted alongside the total number of generations present in each conversation, as well as the speaker characteristics, such as education, occupation, region of residence (with subtypes such as longest, childhood and current residence) and speech defects. After factoring these elements into the data collection process, the corpus is balanced according to the speaker's gender, education (binary as tertiary/non-tertiary), age (binary as >35 or <35), and childhood region of residence. As promising as Kopřivová et al. (2014)'s collection methods are to provide natural results, the approach is not discussed in terms of ethics in their presentation.

DIALEKT on the other hand presents a collection of regional dialects from the 1960s-1980s. The DIALEKT corpus includes dialects, some of which are even extinct now. The DIALEKT monologues are all by people who have always lived in rural areas and are all natives to their regions. One can argue that DIALEKT also considers generational difference, as the birth years of speakers range from the end of 19th century to the start of 20th century, although may not be to the extent of ORTOFON in some cases. Another feature of DIALEKT worth mentioning is that it allows users to search for dialect features captured with regards to all levels of linguistic analysis.

Both corpora utilize ELAN linguistic transcription software (Sloetjes and Wittenburg, 2008), going through annotation in two tiers. For ORTOFON, the first layer is close to Czech orthography while the second adapts phonetic transcription. The latter enables collecting features such as stress groups, vowel reductions and cliticization which might have been lost otherwise. For DIALEKT, the first layer is dialectological, and the second

[4] cf. 2024 edition https://sites.google.com/view/vardial-2024/home, accessed on 11.03.2024.

Languages	**Tasks**	**Datasets**
Czech	Corpus Construction (Kopřivová et al., 2014; Komrskova et al., 2017)	ORTOFON, DIALEKT
Italian	Geolocation Prediction (Ramponi and Casula, 2023)	DIATOPIT
Portuguese	Language Distance Estimation (Pichel Campos et al., 2018)	DiaPT
Portuguese	Century Classification (Zampieri et al., 2016)	Colonia
Swiss German	Modeling of Dialectal Variant Transition (Jeszenszky et al., 2018)	SADS
Swiss German	Predicting Which Regions Use Which Dialectal Variants (Jeszenszky et al., 2019)	SADS
German	Investigating Diachronic Changes in Dialects (Dipper and Waldenberger, 2017)	Anselm
German	Investigating Graphemic Variation in Dialects (Waldenberger et al., 2021)	ReM
English, French	Semantic Change Detection (Montariol and Allauzen, 2021)	Le Monde, NY Times[5]

Table 1: An overview of the presented papers in Section 3.

is the ortographic one same as ORTOFON. In this case, the dialectological layer allows distinguishing speech sounds which are special to non-standard varieties of Czech via the use of a set of symbols. These qualities make the corpora later presented by Komrskova et al. (2017) worth of note.

3.2 Italian

Recently, Ramponi and Casula (2023) present DIATOPIT, a corpus built by analyzing Twitter posts of non-Standard Italian use. They use Twitter APIs to locate non-standard use of language across Italian borders. Moreover, they collect data that comes from accurate coordinates throughout two years to ensure no occasional visitors will disturb the data. They also consider a variety of "out of vocabulary" (OOV) tokens that they use to deduct which of the Twitter posts collected may be from a regional language user. OOV tokens contain tokens which may not be special tokens (i.e. hashtag) and also may not exist in the Aspel dictionary for Italian, but do not include common interjections, elongated words, slangs, wrong diacritics or foreign language tokens, as well as named entity tokens. In doing so, the coordinates from Twitter API and the OOV tokens can be matched to create a map of data by the administrative region.

They also include experiments for evaluating the DIATOPIT's representativeness of real varieties of Italian, which is shown to yield satisfying results in their metrics. While they list a variety of goals for their corpus, what we can say truthfully is that the main contribution is to enable a starting point for those interested in applying NLP methods to research varieties of dialects spoken within Italy. It also serves as first example focusing Italian diatopic variation.

3.3 Portuguese

A different approach works with historical Portuguese to identify different time periods within the historical evolution of a language. Pichel Campos et al. (2018) use a perplexity based measure for this task. Perplexity is a metric indicating how well a system fits a text sample, with a lower score being the better score. It is commonly used as a measure to evaluate the quality of a system, Pichel Campos et al. (2018) note that this is the first attempt utilizing perplexity to calculate diachronic language distance between different periods of historical Portuguese. Their corpus includes six time periods of European Portuguese ranging from the 12th century to the 20th century. They collect their data from various open historical text repositories and historical corpora, and keep the original spelling whenever possible. The perplexity-based approach is noted to successfully identify three main periods for European Portuguese, and should be applicable with other languages as well.

There is another study that works with Portuguese: Zampieri et al. (2016) build upon the Colonia corpus that is an already existing historical Portuguese corpus with texts from the 16th century to the early 20th century. Additionally, they include Part-of-Speech tags for the corpus.

3.4 Swiss German

An interesting approach of modeling transition areas between different dialectal variants using logistic functions is proposed by Jeszenszky et al. (2018): The idea is to model geographic areas,

[5] These corpora are not listed in Table 2, as they are not described in detail.

where one dialectal variant transitions into another, i.e. where language change is taking place. They base their analyses on the SADS dataset (Glaser and Bart, 2015), a linguistic survey with questions on different dialectal phenomena in Swiss German which provides detailed geolocations. Even though the method is very elaborate on a geo-linguistic level, a major drawback is that it can only model the transition of two variants, whereas in real-world scenarios, variation patterns are much more complex and numerous variants are assumed to coexist and influence one another. In a subsequent study on the same dataset, the authors focused further on the temporal aspect (Jeszenszky et al., 2019), and also took the age of respondents into account, an approach similar to Kopřivová et al. (2014). With the sociolinguistic diasystem of language variation, these studies model not only two, but three dimensions: diachronic, diatopic, and diastratic by taking the social variable of age into account.

3.5 German

There are two noteworthy diachronic-diatopic studies on historical corpora of German: Dipper and Waldenberger (2017) examine language change across dialects on a graphemic level. They use aligned equivalent word forms (i.e. word forms that have the same normalization to Standard German) from different German regions to derive rewrite rules with insertions, replacements and identity and create mappings based on weighted Levenshtein Distances. The results show differences across linguistic levels including morphology, phonology and graphemics. The results align with findings from historical linguistics on specific phenomena, such as the High German consonant shift. A follow-up work by (Waldenberger et al., 2021) uses a different dataset, Reference Corpus of Middle High German (Referenzkorpus Mittelhochdeutsch, ReM) (Petran et al., 2016), and generate difference profiles based on weighted Levenshtein distance. The work includes word boundaries as well which allows for capturing further linguistic phenomena. The created mappings from one historical and dialectal variety to another are then compared on a graphemic and graphophonemic level. On a broader level, they conduct further statistical analyses by comparing the intersection of shared mappings between texts in a diatopic subcorpus, and find that this measure indeed reflects the similarity of neighboring dialects.

3.6 English and French

An example of using diachronic word embeddings to model semantic change in the English and French languages is the work by Montariol and Allauzen (2021). Although this work does not work with dialectal data, we still decided to include it, as the approach is interesting and could be applied to (non-continuous) dialect data, e.g. Standard German and Swiss German. Since the datasets are not described in detail, we decided to not include them in Table 2. Overall, the work proposes learning word embeddings from a synthetic corpus with the CBOW (continuous bag-of-words) approach and M-BERT (Devlin et al., 2019), and experiments with different training and aggregation techniques. Computing the divergence of word senses in the two languages, they analyze different language change patterns such as stability in both languages, drift in the same direction, and divergence in word senses with culture-specific contexts. Cathcart and Wandl (2020) propose a related approach experimenting with word embeddings to model phonological change in related varieties of historical Slavic languages in a continuous and discrete way. These approaches are quite interesting and could be applicable to dialect data as well, given the availability of a large amount of training data for dialect embeddings and an evaluation corpus that includes sense and phonetic information.

3.7 Data Characteristics

Data Sources. Different text sources have been used for collecting diatopic datasets: While some approaches work with social media data from Twitter (Dunn and Wong, 2022; Ramponi and Casula, 2023), historical corpora mainly contain religious text or official documents (Dipper and Waldenberger, 2017; Waldenberger et al., 2021) and are usually not suited for a geographical analysis on a fine-grained level. The approaches working on Swiss German (Jeszenszky et al., 2018, 2019) do not base their analyses on natural language data, but on a linguistic multiple-choice survey, the Syntactic Atlas of German-speaking Switzerland (SADS). This kind of data can still be very useful, as it provides direct information about specific language phenomena paired with a very fine-grained, reliable geolocation.

Modality. Most of the corpora rely on written language, only Kopřivová et al. (2014) create two spoken language corpora. From a linguistic point

Languages	Datasets	Tokens	Source/Register	Time Span	Modality
Czech	ORTOFON (Komrskova et al., 2017)	1.24 M	dialogue	2012-2017	spoken
Czech	DIALEKT (Komrskova et al., 2017)	126,131	monologue	1960s-1980s	spoken
Italian	DIATOPIT (Ramponi and Casula, 2023)	388,069	Twitter	2020-2022	written
Portuguese	DiaPT (Pichel Campos et al., 2018)	-	historical text	1100-2000	written
Portuguese	Colonia (Zampieri and Becker, 2013)	5.1 M	media, historical text	1500-2000	written
Swiss German	SADS (Glaser and Bart, 2015)	-	linguistic survey	2000-2002	written
German	Anselm (Dipper and Schultz-Balluff, 2013)	30,000	religious text	1350-1600	written
German	ReM (Petran et al., 2016)	2.5 M	historical text	1050-1350	written
German	ZDL-Reg. (Nolda et al., 2021)	11.78 B	regional newspaper	1993-2024	written

Table 2: An overview of the dialect-related datasets discussed in Section 3. ZDL-Reg. is dynamically enlarged; the number of tokens is taken from `https://www.dwds.de/d/korpora/regional`, accessed on 05.03.2024. SADS does not contain natural language data, but 118 multiple-choice questions about 54 (morpho-)syntactic phenomena. Additionally, we include another corpus of regional newspaper data in German, the ZDL-Regionalkorpus (Nolda et al., 2021, 2023)—which has not been explored for diachronic-diatopic studies yet.

of view, this is very effective, since variation usually is much stronger in spoken compared to written language, as most dialects do not deviate markedly from Standard languages in the written modality.

Time Span. The diachronic spans of the datasets also vary strongly: While some historical corpora cover very long periods of time, e.g. the Diachronic Portuguese Corpus (DiaPT) (Pichel Campos et al., 2018) spans almost one millennium, social media-based corpora like DIATOPIT or linguistic survey data like SADS only span two years.

Data Imbalance. It must be noted that the Colonia corpus used by Zampieri et al. (2016) does not contain the same amount of text from each period it covers. For instance, there are 38 documents available from the 19th century, while there are only 13 available from the 16th century. Due to this, Zampieri et al. (2016) generate artificial texts with around 330 tokens for their train and test sets in order to conduct their main experiments.

4 Experiments

Experimental setups and results of the presented studies are difficult to compare, as the tasks and datasets presented in Section 3 are very different. Some of the papers focus on corpus construction (Kopřivová et al., 2014) or qualitative analysis (Dipper and Waldenberger, 2017), while some present quantitative results in the tasks of measuring language distance, predicting geolocation or dialect variant usage, will briefly be compared here.

Czech. Since Kopřivová et al. (2014) aims to build/present corpora, there are no experiments to mention. But one can argue that when ORTOFON and DIALEKT are used interconnectedly, they will present a good outlook on diachronic and diatopic variation in Czech. The work by Kopřivová et al. (2014) is to set apart with their detailed annotation system separated with several parallel layers to accommodate speakers individually. In the follow-up work by Komrskova et al. (2017), the advantages are evident thanks to the use of this multi-tier transcription.

Italian. Ramponi and Casula (2023) evaluate geolocation predictions on two levels: coarse-grained geolocation (CG, i.e. region classification), and fine-grained geolocation (FG, double-regression i.e. for latitude/longitude coordinates). They measure the accuracy of the prediction results in the macro-averaged Precision, Recall, and F1 score. Baseline models are mostly built upon BERT (Devlin et al., 2019). Both monolingual (Italian-only) and multilingual models are investigated, including AlBERTo (Polignano et al., 2019), UmBERTo (Parisi et al., 2020) and mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020). Additionally, for CG they use Logistic Regression (LR) and SVM classifiers, and for FG they use a centroid baseline and a regression model based on k-nearest neighbors alongside a decision tree regressor. Results averaged across five runs with random seeds for shuffling the data and initializing model parameters are presented. For CG, AlBERTo achieves best results, and LR performs the worst. SVM proves to be competitive for the task. In FG's case, AlBERTo achieves the best scores again. Interestingly, the decision tree performs competitively despite being a much more cost-efficient system.

Portuguese. Pichel Campos et al. (2018) aims to compare six time periods of historical European Portuguese. They implement a perplexity-based language distance (PLD) measure with 7-gram models alongside a linear interpolation based smoothing technique. They conduct experiments on two levels: PLD with original spelling, and PLD with transcribed spelling. For the first instance, they compute PLD for each possible train-test pair. For the latter instance, they adjust the Diachronic Portuguese Corpus to have all periods share the same spelling. This is achieved by transliterating all historical periods into Latin scripts and then normalizing it with a generic orthography similar to phonological style. The resulting encoding of spelling normalization consists of 34 symbols, including 10 vowels and 24 consonants.

Overall, the results in both experiments observe a similar pattern. It is shown that language distances between different time periods are correlated with chronology. Moreover, there is not a huge divergence within the different periods investigated. The longest difference between periods scores roughly 6.19 with original spelling and 5.92 with transcribed spelling, which is still lower than the distance between closely related languages, such as Spanish-Portuguese's score of 7.74. The results suggest that, at least for the case of Brasilian-Portuguese, the language has remained similar over time.

For the other study that works with Portuguese, Zampieri et al. (2016) conduct experiments in three steps. They first have a preliminary session where they test a small sample with 87 documents from their corpus. They train SVM alongside Multinomial Naive Bayes (MNB) to predict which century a text belongs to, using both words and Part-of-Speech (POS) tags.

Secondly, they start their main experiments where they use 1500 artificially generated documents, and use the SVM classifier to execute predictions. They observe a performance increase due to the implementation of POS tags or words represented as uni-, bi-, and trigrams. Results show that POS trigrams yield the 90.7% accuracy when tested with century classification of the presented documents. Zampieri et al. (2016) note that this emphasises the existence of difference in structural properties in each time span by an important level; this means changes in structural properties take place at both the word level and beyond, and these changes can be captured through uni-, bi-, and trigrams.

Lastly, they conduct experiments across a smaller time span of 50 years. Their findings show that many time periods exhibit high similarity in grammatical structure. This presents a challenge for century classification of documents. It is noted that POS tags perform the best with trigrams.

Swiss German. Jeszenszky et al. (2018) conceptionalize transitions between dialectal variant areas via logistic regression and intensity maps in an attempt to present spatial distribution of syntactic variants in Swiss. The results show gradual and sharp transitions between variants alongside distinct spatial patterns. Subdivision analyses further elucidated the characteristics of dominance zones and transition areas. Overall, the findings shed light on the spatial distribution and dynamics of linguistic features. A drawback of the methodology is that only 40% of the variables in the SADS dataset (Glaser and Bart, 2015) can be modeled. An important take-away is that the transition of dialectal variants is a highly complex phenomenon, which cannot be fully modeled by only taking the spatial dimension into account.

Jeszenszky et al. (2019) use logistic regression on a global level to model the association of linguistic variation and age with 10-fold cross-validation. The AUC scores (area under the curve) reveal that for more than half of the variants considered, age is not a significant predictor. On a local level, they classify whether a specific linguistic variant is used at a survey site given the respondent age. The survey site is chosen from the k-nearest neighbors based on Euclidean distance, k ranging from 5 to 50. They conclude that the significance of age as predictor variable is correlated with space: When a specific age group within a region is significant, the prediction of which dialect that region speaks is more accurate. However, the prediction becomes less accurate when a region associates with multiple age groups. They attribute this finding to a sociolinguistic fact that lexicon in dialects is more prone to change with respect to speaker age than syntax.

German. Dipper and Waldenberger (2017) and Waldenberger et al. (2021) combine quantitative with an in-depth qualitative analysis. Both do not experiment with complex methods, but conduct a simple frequency-based, statistical analysis. Dipper and Waldenberger (2017) find quantitative proof for morphological, phonological, and graphemic phenomena by deriving replacement rules. They show insightful results into nuances of linguistic change across different regions and periods from a historical linguistic perspective: finding quantitative prove for theories like the High German consonant shift. The second study by Waldenberger et al. (2021) employs slightly more elaborate statistical measures to quantify differences between texts and subcorpora. The results confirm the diatopic and diachronic variation: By analyzing Levenshtein mappings and computing similarity scores, the study demonstrated that texts from closely related dialects exhibited higher similarity scores compared to those from more distant regions. Overall, Upper German texts are found to be more similar to each other than Middle German texts.

English and French. Montariol and Allauzen (2021) experiment with two kinds of embeddings, continuous bag-of-words (CBOW) and BERT (Devlin et al., 2019), to detect whether meaning changes of a word and its translation in English and French are consistent or divergent over time. They show a trade-off between performance and efficiency: While BERT with k-means clustering achieves the best performance, the CBOW model with incremental training is computationally the most efficient and offers very competitive results.

Their findings are summarized as follows: Semantic meanings drifting in the same direction across languages mainly occurs with words related to technology and society. On the other hand, meanings diverging in different directions implies that the meaning of a word might remain unchanged over time in one language, but drift in the other. This is mostly seen for words related to culture-specific concepts or controversial topics. It would be interesting to apply this approach not only to related languages, but to an actual dialect continuum to investigate whether these findings are confirmed in closer related language varieties as well.

5 Discussion

Almost all languages in the world have distinct dialects varying by location that change quickly due to complex factors related to contact. Taking these two dimensions of language variation, diachronic and diatopic, into account can improve the diversity and representativeness of languages covered in this field , and benefit the communities of non-standard language users. Our research shows that the intersection of diachronic and diatopic variation is an under-studied topic in dialect NLP. Although there are some approaches experimenting with diachronic word embeddings on a multilingual level (Montariol and Allauzen, 2021), there is currently a lack of state-of-the-art machine learning and NLP approaches.

This is a challenging topic to work with, considering its interdisciplinary nature combining historical linguistics, dialectology, machine learning and NLP. Perhaps this is a factor contributing to the status of deep learning based NLP methods having not yet been applied to studying language change in dialect continua.

5.1 Open Challenges

Do language variants and dialects change over time? While Pichel Campos et al. (2018) show that the difference in perplexity-based language distances between different time periods of European Portuguese is not substantial, Zampieri et al. (2016) suggest that grammatical structure can be substantially different in some time periods of Portuguese; however, their study was conducted on artificial documents. This means that either perplexity-based language distance fails to capture the differences in grammatical structures of different time periods, or such differences are not present in the real-world historical Portuguese documents investigated. We leave this question to future work.

Is the construction of dialect-related datasets reliable? The reliability of Ramponi and Casula (2023) is also worth mentioning: They rely on the belief that the locals may write things that deviate from Standard Italian just because they speak it so, but they also rely on Twitter language identifiers to deduct whether a tweet is in Italian or not. This, of course, is a double-edged sword and may cut back on data reliability. If their assumption is correct, in extreme cases some societies whose language use deviate from the standard may remain completely under-represented and their twitters might be misclassified as Standard Italian. However, if it is incorrect (i.e., the language use of the locals follows the standard), the tweets written by the locals and those in standard Italian become indistinguishable. Considering their access to speakers of regional Italian varieties (curators), Kopřivová et al. (2014) set a good example they could follow to ensure more varieties are correctly represented. However, one can argue that if someone was to use VPN for any reason, the coordinates would also be set for the entire time of use. Thus, Twitter APIs may not provide completely accurate data either, though this may be minimal to consider in most cases.

Are speaker characteristics important? Additionally, although Kopřivová et al. (2014) show that tracking the number of generations present in a conversation is beneficial for building speaker-characters, Jeszenszky et al. (2019) suggest that age is not a definitive for prediction. This means the usefulness of age information is quite task-dependent. An interesting follow-up work would be to incorporate other speaker characteristics, such as gender and education, into the analysis.

Limited coverage of dialects. There are numerous dialects spoken in the world. For instance, English alone has approximately 160 dialects (Aeni et al., 2021). However, only a small number of dialects have been researched in the NLP and machine learning communities. Future work could establish a data center to manage and update world-existing dialect corpora. Indeed, many corpora are publicly available but are little explored. For instance, the German regional newspaper corpus ZDL-Regionalkorpus (Nolda et al., 2021) has not been used for diachronic analysis so far, despite its size of more than 11 B tokens covering a time span 1993-2024 with regular updates which could enable use for data-intensive machine learning and word embedding approaches.

Ethical considerations in the collection of dialectal data. Although the data collection methods of Kopřivová et al. (2014) promise to provide near authentic results, no ethical issues are mentioned. As the conversations are recorded without the knowledge of the participants to ensure natural quality, it would not have been possible to get individual consent from the participants, although the person recording may have agreed otherwise. This, therefore, shows risk of privacy breach, and may not be an acceptable approach in a lot of data collection cases. Whether this would be acceptable if the speakers are informed after the data is collected may still be questionable to some people's discretion, however, this doesn't change the fact that despite being a breach, Kopřivová et al. (2014)'s approach does provide data as close to real-life situations as possible. This is of value in itself.

6 Conclusion

While there is a rising interest in modeling diachronic and diatopic variation in the NLP community, the intersection of both, i.e. language change in dialect continua, remains an under-studied topic. Even though findings from linguistics and sociolinguistics stress the importance of the diatopic dimension when modeling language change, the topic has not yet received as much attention in computational linguistics and not many methodological advancements have been made. Our work has been a first step in closing this research gap, and we hope to give inspiration to future research.

Acknowledgements

We thank the anonymous reviewers for their thoughtful feedback that greatly improved the texts.

References

Nur Aeni, Like Raskova Octaberlina, Nenni Dwi Aprianti Lubis, et al. 2021. A literature review of english language variation on sociolinguistics.

Timothy Baldwin. 2018. Language and the shifting sands of domain, space and time (invited talk). In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, page 76, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kate Beeching. 2006. Synchronic and diachronic variation: the how and why of sociolinguistic corpora. In *Corpus linguistics around the world*, pages 49–61. Brill.

Sidsel Boldsen and Patrizia Paggio. 2022. Letters from the past: Modeling historical sound change through diachronic character embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6722, Dublin, Ireland. Association for Computational Linguistics.

Chundra Cathcart and Florian Wandl. 2020. In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 233–244, Online. Association for Computational Linguistics.

Yanran Chen, Wei Zhao, Anne Breitbarth, Manuel Stoeckel, Alexander Mehler, and Steffen Eger. 2024. Syntactic language change in english and german: Metrics, parsers, and convergences. *arXiv preprint arXiv:2402.11549*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefanie Dipper and Simone Schultz-Balluff. 2013. The anselm corpus: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the workshop on computational historical linguistics at NODALIDA*, pages 27–42.

Stefanie Dipper and Sandra Waldenberger. 2017. Investigating diatopic variation in a historical corpus. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 36–45, Valencia, Spain. Association for Computational Linguistics.

Jonathan Dunn and Sidney Wong. 2022. Stability of syntactic dialect classification over space and time. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 26–36, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *Preprint*, arXiv:2403.11009.

Elvira Glaser and Gabriela Bart. 2015. *4. Dialektsyntax des Schweizerdeutschen*, pages 81–108. De Gruyter, Berlin, München, Boston.

Siqi He and Wei Zhao. 2024. Exploring sound change over time: A review of computational and human perception. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Péter Jeszenszky, Panote Siriaraya, Philipp Stoeckle, and Adam Jatowt. 2019. Spatio-temporal prediction of dialectal variant usage. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 186–195, Florence, Italy. Association for Computational Linguistics.

Péter Jeszenszky, Philipp Stoeckle, Elvira Glaser, and Robert Weibel. 2018. A gradient perspective on modeling interdialectal transitions. *Journal of Linguistic Geography*, 6(2):78–99.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *Preprint*, arXiv:2401.05632.

Zuzana Komrskova, Marie Kopřivova, David Lukeš, Petra Poukarová, and Hana Goláňová. 2017. New spoken corpora of czech: Ortofon and dialekt. *Journal of Linguistics/Jazykovedný casopis*, 68.

Marie Kopřivová, Hana Goláňová, Petra Klimešová, and David Lukeš. 2014. Mapping diatopic and diachronic variation in spoken Czech: The ORTOFON and DIALEKT corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 376–382, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xianghe Ma, Dominik Schlechtweg, and Wei Zhao. 2024a. Presence or absence: Are unknown word usages in dictionaries? In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Xianghe Ma, Michael Strube, and Wei Zhao. 2024b. Graph-based clustering for detecting semantic change across time and languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1561, St. Julian's, Malta. Association for Computational Linguistics.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *Preprint*, arXiv:2304.01666.

Syrielle Montariol and Alexandre Allauzen. 2021. Measure and evaluation of semantic divergence across two languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258, Online. Association for Computational Linguistics.

Andreas Nolda, Adrien Barbaresi, and Alexander Geyken. 2021. Das ZDL-Regionalkorpus: Ein Korpus für die lexikografische Beschreibung der diatopischen Variation im Standarddeutschen. Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch, pages 317 – 321. de Gruyter, Berlin [u.a.].

Andreas Nolda, Adrien Barbaresi, and Alexander Geyken. 2023. Korpora für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache. Das ZDL-Regionalkorpus und das Webmonitor-Korpus. Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial, pages 29 – 52. de Gruyter, Berlin/Boston.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

Florian Petran, Marcel Bollmann, Stefanie Dipper, and Thomas Klein. 2016. Rem: A reference corpus of middle high german – corpus compilation, annotation, and access. *Journal for Language Technology and Computational Linguistics*, 31(2):1–15.

Jose Ramom Pichel Campos, Pablo Gamallo, and Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to European Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, et al. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR.

Alan Ramponi and Camilla Casula. 2023. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-elan and iso dcr. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Maria Sukhareva and Christian Chiarcos. 2014. Diachronic proximity vs. data sparsity in cross-lingual parser projection. a case study on Germanic. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 11–20, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Chahan Vidal-Gorène, Victoria Khurshudyan, and Anaïd Donabédian-Demopoulos. 2020. Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Sandra Waldenberger, Stefanie Dipper, and Ilka Lemke. 2021. Towards a broad-coverage graphemic analysis of large historical corpora. *Zeitschrift für Sprachwissenschaft*, 40(3):401–420.

Walt Wolfram and Natalie Schilling-Estes. 2017. *Dialectology and Linguistic Diffusion*, chapter 24. John Wiley & Sons, Ltd.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien*, 5:69–76.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: The case of Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4098–4104, Portorož, Slovenia. European Language Resources Association (ELRA).

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.