Language Change in Dialect Continua: A Survey on Diachronic and Diatopic Variation in NLP

Melis Çelikkol **Lydia Körber**
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
(celikkol|koerber)@cl.uni-heidelberg.de

Abstract

Language changes and varies constantly on all linguistic levels. A lot of communities in the world who speak non-standard language are under-represented, if not excluded from machine learning approaches. Although there has been research in diatopic and diachronic change separately, so far there is no study that actively applies both with machine learning; and most of the studies covering this may already be outdated. There are a lot of details to pay attention to when one applies NLP methods with this knowledge due to its interdisciplinary nature. Our aim is to provide a collection of literature focusing on diatopic and diachronic change. Various language and dialect work is presented in order to analyze language change in dialect continua. We hope that it will prove itself a first step to trigger research in this area.

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).

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 computer-mediated communication and social media data tend to exhibit faster

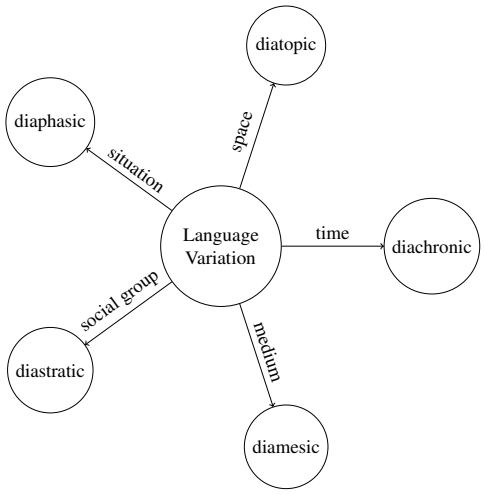

Figure 1: Language variation and the diasystem.[1]

language change patterns, as they are more prone to current events and also show a faster diffusion of new linguistic variants (Eisenstein et al., 2014). This also becomes visible in the temporal degradation of NLP applications, e.g. headline generation models decrease in performance after a few years, emoji prediction models even within a month (Søgaard et al., 2021). As insights from (socio-)linguistics show, diachronic and synchronic variation are closely linked (Beeching, 2006); and often language change manifests first in synchronic variation before entering a diachronic level. There is a strong spatial component in language change, as the propagation and diffusion of a variant is caused by contact between people and speech communities (Jeszenszky et al., 2018). While *isoglosses* assume separate variants that are divided by exact boundaries, the consensus among dialectologists and sociolinguists today is to speak of *dialect continua*, which assume gradual transitions between core areas of dialects (Jeszenszky et al., 2018). In these continua, according to the *wave model*, language change is propagated from a certain locus at

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

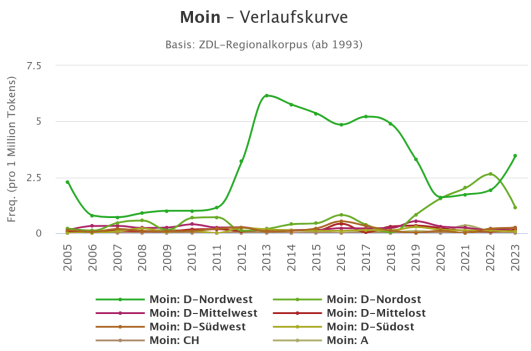

Figure 2: Diachronic usage of *Moin* in the years 2005-2023 in a regional newspaper corpus of German across dialect areas in frequency per 1M tokens.[2]

a certain point in time and spread layer-wise, radiating from the central point of contact (Wolfram and Schilling-Estes, 2017). Contact and isolation are assumed to be the most important drivers of language change, as denoted by the contact effect (Jeszenszky et al., 2019). Furthermore, languages and varieties do not change at the same pace, the speed of language change differs across dialects (Trudgill, 2020). This intersection between diatopic and diachronic variation is also subject to research in Natural Language Processing and is examined in the present survey.

An example of diatopic variation over time can be seen in figure 2 in the usage of the German dialect word *Moin* (Good morning), which is mainly used in Northern dialectal areas of Germany. A query in the ZDL-Regionalkorpus (Nolda et al., 2021, 2023), a collection of regional newspaper texts from Germany, Austria, and Switzerland, reveals its dominance in two Northern German dialect areas (D-Nordwest, D-Nordost), while it also rose in usage in other dialectal areas at specific time points, e.g. in south-western (D-Südwest) in 2016, and in middle-western dialect areas (D-Mittelwest) in 2019.

This work aims at exploring the intersection of diachronic and diatopic variation in NLP research. Research questions on this topic include how to detect and quantify language change in dialect continua on historical and contemporary data, as well as how to build and process diachronic-diatopic datasets. Previous approaches have used statistical and machine learning-based methods to compute distance and similarity of different varieties on various linguistic levels (graphemics, syntax, semantics), and built diachronic-diatopic datasets of written and spoken language data. The scope of this survey is to cover key methodologies, datasets, strengths, limitations, and gaps in the existing literature. It encompasses seminal works with Indo-European languages and their varieties as well as recent developments in the field. 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 in 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).

A comprehensive survey on diatopic language modelling is Zampieri et al. (2020), who evaluate computational methods for processing similar languages, language varieties, and dialects, with a focus on diatopic language variation and integration in NLP applications. The authors identify 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. Very recently, a benchmark for the evaluation of 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. There exists a designated series of workshops on NLP for Similar Languages, Language Vari-

[2]Usage graph for *Moin*, created with Digitales Wörterbuch der deutschen Sprache (DWDS, Digital Dictionary of the German Language), https://www.dwds.de/r/plot/?view=1&corpus=regional&norm=date%2Bclass&smooth=spline&genres=1&grand=0&slice=1&prune=0&window=0&wbase=0&logavg=0&logscale=0&xrange=2005%3A2023&q1=Moin, accessed 12.03.2024.

eties, and Dialects (VarDial)[3], which also proposes shared tasks on different NLP tasks in dialects and other varieties, as well as on 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 Methods

The presented methods with tasks and datasets are listed in table 1.

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 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 without anyone in the group's knowledge apart from the agent who recorded them. 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 a few things missing 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 longest, childhood and current) and speech defects. After collecting data with these factors in the equation, the corpus is balanced according to the speaker's sex, education (binary as tertiary/non-tertiary), age (binary as >35 or <35), and childhood region of residence.

DIALEKT on the other hand presents a collective of traditional 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 say that DIALEKT also considers generational difference considering the birth years of speakers were between 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, going through annotation in two tiers. For ORTOFON, the first one is close to Czech orthography while the second one 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 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.

Ramponi and Casula (2023) present DIATOPIT, a corpus built by analyzing Twitter posts of non-Standard Italian use. The study is one of the newer appraoches we mention, and also includes experiments on the representativeness of the resulting corpus with regards to actual language use across different regions of Italy. Ramponi and Casula (2023) use Twitter APIs to locate non-standard use of language across Italian borders. They collect data that comes from accurate coordinates throughout two years to ensure no occasional visitors will disturb the data. They 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 don't 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.

The research also contains a number of experiments to test DIATOPIT's representativeness of real varieties of Italian, which is shown to yield satisfying results in their metrics. Ramponi and Casula (2023) list a variety of goals for their cor-

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

Lang.	Authors	Task	Dataset
cz	Kopřivová et al. (2014); Komrskova et al. (2017)	corpus construction	ORTOFON, DIALEKT
it	Ramponi and Casula (2023)	geolocation identification	DIATOPIT
pt	Pichel Campos et al. (2018)	language distance	DiaPT
pt	Zampieri et al. (2016)	time span prediction	Colonia
gsw	Jeszenszky et al. (2018)	variant transition modelling	SADS
gsw	Jeszenszky et al. (2019)	variant usage prediction	SADS
de	Dipper and Waldenberger (2017)	corpus analysis	Anselm
de	Waldenberger et al. (2021)	corpus analysis	ReM
en-fr	Montariol and Allauzen (2021)	semantic change detection	Le Monde, New York Times[4]

Table 1: An overview of the presented methodological papers. Language codes of each study are presented alongside the citation, NLP-task and dataset. Detailed information on the dataset statistics is provided in table 2.

pus, but what we can say truthfully is that DIATOPIT's 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.

A different approach works with historical Portuguese to identify diachronic periods within the historical evaluation 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 quality of a system, Pichel Campos et al. (2018) note that this is the first attempt utilizing perplexity this way. They use the PLD measure to calculate diachronic language distance between periods of historical Portuguese. For their corpus, they look at 6 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 where possible. Pichel Campos et al. (2018)'s 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: Differently to the other Portuguese approach mentioned above, Zampieri et al. (2016) don't build a corpus but utilizes the Colonia corpus which is an already existing historical Portuguese corpus with texts from the 16th century to the early 20th century.

They utilize POS tags (morphosyntatic information) and lexicon (either as bag-of-words or word n-grams). Their method is supposedly applicable with any diachronic corpus across various languages as well, as long as POS annotation is available on it.

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, where one dialectal variant transitions into another, i.e. where language change is taking place. They base their analyses on the SADS dataset, 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 in mind, these studies model not only two, but three dimensions: diachronic, diatopic, and diastratic by taking the social variable age into account.

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 variants related to different linguistic levels (morphological, phonological, and graphemic) and align with

[4]These corpora are not listed in the datasets table 2, as they are not described in detail.

Lang.	Dataset	Tokens	Source/Register	Time Range	Modality	Related Work
cz	ORTOFON	1.24 M	dialogue[5]	2012-2017	spoken	(Kopřivová et al., 2014)
cz	DIALEKT	126,131	monologue	1960s-1980s	spoken	(Kopřivová et al., 2014)
it	DIATOPIT	388,069	Twitter	2020-2022	written	(Ramponi and Casula, 2023)
pt	DiaPT	-	historical text	1100-2000	written	(Pichel Campos et al., 2018)
pt	Colonia[6]	5.1 M	media, historical text	1500-2000	written	(Zampieri et al., 2016)
gsw	SADS[7]	-[8]	linguistic survey	2000-2002	written	(Jeszenszky et al., 2018, 2019)
de	Anselm[9]	30,000	religious text	1350–1600	written	(Dipper and Waldenberger, 2017)
de	ReM[10]	2.5 M	historical text	1050-1350	written	(Waldenberger et al., 2021)
de	ZDL-Reg.[11]	11.78 B[12]	regional newspaper	1993-2024	written	-

Table 2: An overview of all the research and the data they use. Language codes of each study are presented alongside the tokens, type of source the data was borrowed from, range of time the study takes into account to capture change, as well as modality and finally related work. If the corpus collection and curation is described in a different paper, this is noted in a footnote along with the dataset name.

findings from historical linguistics on specific phenomena, such as the High German consonant shift. The authors extend their work in a subsequent study (Waldenberger et al., 2021). They use a different dataset, Reference Corpus of Middle High German (Referenzkorpus Mittelhochdeutsch, ReM), and generate difference profiles based on weighted Levenshtein distance, this time including word boundaries as well which allows 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.

An example of using diachronic word embeddings to model semantic change in two languages, English and French, is the work by Montariol and Allauzen (2021).[13] They propose learning word embeddings from a synthetic corpus with a CBOW (continuous bag-of-words) approach and M-BERT and experiment 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. The approach is very interesting and could be applicable to dialect data as well, given a sufficient amount of training data for the embeddings and a sense-annotated corpus for the generation of synthetic evaluation data. 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.

4 Data

We present an overview of diachronic and diatopic datasets in different languages in table 2. Besides the corpora used in the approaches presented in section 3, we mention another corpus of regional newspaper data in German, the ZDL-Regionalkorpus (Nolda et al., 2021, 2023). It has not been subject to methodological diachronic-diatopic studies yet, but it provides a good basis also for machine learning research due to the large dataset size.

Different text sources have been used for 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 (Jeszen-

[13] This paper does not work with dialectal data, but 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, as well. Because the datasets used are not described in detail, we decided to not include them in the dataset section.

[5] DIALEKT and ORTOFON both use everyday life conversation.

[6] (Zampieri and Becker, 2013)

[7] (Glaser and Bart, 2015)

[8] The dataset does not contain natural language data, but 118 multiple-choice questions about 54 (morpho-)syntactic phenomena.

[9] (Dipper and Schultz-Balluff, 2013)

[10] (Petran et al., 2016)

[11] (Nolda et al., 2021)

[12] The corpus is dynamically enlarged. The number of tokens is taken from https://www.dwds.de/d/korpora/regional, accessed 05.03.2024.

szky 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.

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

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.

Finally, 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 offers data from (i.e. there are 38 documents available to us from the 19th century, where the corpus has only 13 available for the 16th). Due to this, Zampieri et al. (2016) generate artificial texts with +-330 tokens for their train and test sets in order to conduct their main experiments.

5 Experiments

Experimental set-ups and results of the presented studies are difficult to compare, as the tasks and datasets used are very different. Some of the papers focus on corpus construction (Kopřivová et al., 2014) or qualitative analysis (Dipper and Waldenberger, 2017). However, the studies mentioned in section 3 that do present quantitative results, either in measuring language distance, in predicting geolocation or dialect variant usage, will briefly be compared here.

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

Ramponi and Casula (2023) test 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 use a macro-averaged precision (P), recall (R), and F1 score metric for evaluation. Multiple baselines are used: both models are compared on BERT-based models[14]. Additionally, for CG they use a most frequent baseline, Logistic Regression (LR) and SVM Classifier, and for FG a centroid baseline and a regression model based on k-nearest neighbours alongside a decision tree regressor. Results averaged across 5 runs with random, different seeds for shuffling the data and initializing the models are presented. For CG, AlBERTo obtains best results, and LR the worst. SVM proves to be competitive for the task. In FG's case, AlBERTo has the best scores again. Interestingly, the decision tree performs competitively despite being a much more cost-efficient system.

As Pichel Campos et al. (2018)'s goal is to compare 6 periods of historical European Portuguese, they implement PLD with 7-gram models alongside a linear interpolation based smoothing technique. They test 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 of the six 7-gram models. For the latter instance, they adjust DiaPT 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 consists of 34 symbols.

Results of both instances are presented with tables and figures. It is interesting to see that the patterns follow a similar fashion in both experiments. Pichel Campos et al. (2018) find that the distance between periods are correlated with chronology, and there is not a huge divergence within the different periods tested. 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.

[14] Both monolingual and multilingual, Italian-only models include: AlBERTo, UmBERTo. Multilingual models include: mBERT, XLM-R.

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 Multinominal Naive Bayes (MNB) to predict which century a text belongs to with words and POS tags.

Afterwards, they start their main experiments where they use 1500 artificially generated documents. An SVM classifier is used for prediction, and Zampieri et al. (2016) present a performance increase due to the implementation of POS tags or words represented as uni-, bi-, and trigrams. The results are presented across a baseline of 20% picked randomly. POS trigrams yield 90.7% accuracy when tested with diachronic prediction 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.

The final step replicates the main experiments across a smaller time span of 50 years. Their findings show that the similarity in grammatical structure of the times in comparison is parallel to the challenge level the classifier must face. Zampieri et al. (2016) present a confusion matrix alongside figures to depict challenge in differentiation. It is noted that POS tags perform the best with trigrams.

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 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 values (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 variation is at a survey site given the respondent age with k-nearest neighbor survey sites based on Euclidean distance, varying the values of k from 5 to 50. They conclude that the significance of age as predictor variable is correlated with space: When present, the usage of this variant is characteristic of certain age groups at survey sites with high significance of logistic regression, while in areas with low significance values, it is used by different age groups. The authors explain this with the finding from sociolinguistics, that lexicon is more prone to change with respect to speaker age than syntax.

Dipper and Waldenberger (2017) and Waldenberger et al. (2021) combine quantitative with an in-depth qualitative analysis. Both do not experiment with different 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. This shows that the work is highly insightful into nuances of linguistic change across different regions and periods from a historical linguistic perspective. The second study (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.

Montariol and Allauzen (2021) experiment with two kinds of embeddings, continuous bag-of-words (CBOW) and BERT, to detect divergence scenarios in embeddings of the two languages in datasets of the English and French newspapers New York Times and Le Monde. There is a clear trade-off between performance and efficiency, as the CBOW model with incremental training is computationally the most efficient, while BERT with k-means clustering yields the highest results, reported in accuracy.

Montariol and Allauzen (2021)'s findings are as follows: The scenario of stable words in both languages mainly applies to everyday-life words. Drift in the same direction mainly occurs with concepts related to technology and society. Divergent word meanings that stay stable in one language, but drift

in the other, are mostly 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 see which of the findings are confirmed in closer related language varieties as well.

6 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 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 NLP. Although some contributive approaches that experiment with diachronic word embeddings on a multilingual level exist (Montariol and Allauzen, 2021), there is a lack of state-of-the-art approaches for machine learning currently.

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

6.1 Open Challenges

An interesting comparison in the outcome of the Portuguese papers can be made: While Pichel Campos et al. (2018) deduct that there's not a large divergence between the historical periods of European Portuguese, Zampieri et al. (2016)'s findings suggest that there are properties that differ between them at an important level. This may be due to the latter paper's approach relying on AI generated texts for training rather than relying on original documents purely. That might also be the driving force behind including a confusion matrix in their study, alongside a set of patterns of language change, in an attempt to back the reliability of their method.

The reliability of Ramponi and Casula (2023) is also worth mentioning: they rely on the belief that the locals may write things in a way that doesn't fit within 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 a lot of data reliability. If their assumption is right, in extreme cases some societies may remain completely under-represented. However, if it's wrong, there may be some posts that disrupt the truthfulness of data. Considering their processes, and the knowledge that they had 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. 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. So Twitter APIs may not provide 100% truthful data either, though this may be minimal to consider in most cases.

Additionally, although Kopřivová et al. (2014) mentions tracking the number of generations present in a conversation to be beneficial for building speaker-characters, Jeszenszky et al. (2019) findings suggest that age is not a definitive for prediction. This may be dependent on the goals and scope of one's research, considering DIALEKT & ORTOFON take into consideration other non-traditional variables for the task, i.e. gender.

Some diatopic datasets which also span a wide time range are yet unexplored: 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 which could enable use for data-intensive machine learning or word embedding approaches.

7 Conclusion

While there is a rising interest in modeling diachronic and diatopic variation in the NLP research community, the intersection of both, i.e. language change in dialect continua, remains an under-studied field. 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. This survey has been a first step in closing this research gap, and we hope to give inspiration to future research.

References

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.

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.

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.

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

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.

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.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection.

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.

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.

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.

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.

Peter Trudgill. 2020. Sociolinguistic typology and the speed of linguistic change. *Journal of Historical Sociolinguistics*, 6(2):20190015.

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.