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

Melis Çelikkol[1], Lydia Körber[1], Wei Zhao[2]

University of Heidelberg[1], University of Aberdeen[2]

Paper in a Nutshell

Motivation: So far no work on the intersection of diachronic and diatopic variation in NLP

Aim: Explore current works on diachronic and diatopic changes in dialect continua → 9 tasks and datasets across 5 dialect continua

Main results: Five open challenges (changes in dialect use over time, reliability of datasets, speaker features, limited coverage of dialects, ethical considerations)

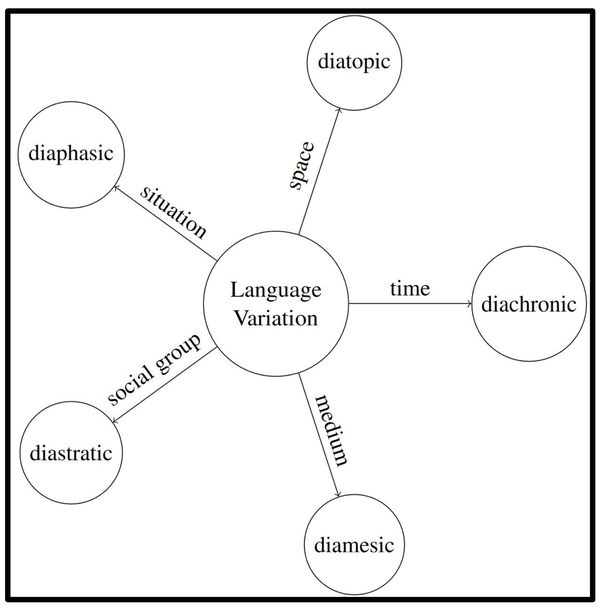
Research Questions

- 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?
- What is the current state of computational **methods** and their results in these dialect-related NLP tasks?
- What are the **challenges** in dialect NLP research that have not been addressed in previous works?

Contributions

- **A starting point** provided for future work on the intersection of diatopic and diachronic variation.
- Highlights the undiscovered **potential** of the **relationship** between diatopic/diachronic change.
- Serves as a **call-to-action** for other researchers.

Illustration



Language variation across 5 dimensions in the diasystem: Our work focuses on the intersection of **diachronic** (time) and **diatopic** (space) variation.

Main results – Tasks

An overview of the presented NLP tasks, languages and datasets:

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]

Main results – Datasets

An overview of the diachronic-diatopic datasets examined:

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

Open Challenges

- contradictory results on the change of dialect variants over time
- reliability of dialect datasets
- importance of speaker characteristics (sociodemographic information)
- limited coverage of dialects
- ethical considerations in data collection

Conclusions and futuristic outlook

- The intersection of diatopic and diachronic change is **understudied** at this time.
- This work serves as a starting point for those interested in **closing the research gap**.

Limitations

- The research mentioned is not directly comparable, as methodologies and approaches are very diverse.
- No non-Indoeuropean languages and dialect continua are included (e.g. Arabic).

Paper link

LChange @ ACL 2024, Bangkok