Natural Language Processing for Dialects of a Language: A Survey

ADITYA JOSHI, University of New South Wales, Sydney, Australia

RAJ DABRE, National Institute of Information and Communications Technology, Koganei, Japan

DIPTESH KANOJIA, University of Surrey, Guildford, United Kingdom of Great Britain and Northern Ireland

ZHUANG LI, Monash University, Clayton, Australia

HAOLAN ZHAN, Monash University, Clayton, Australia

GHOLAMREZA HAFFARI, Information Technology, Monash University, Clayton, Australia

DORIS DIPPOLD, University of Surrey, Guildford, United Kingdom of Great Britain and Northern Ireland

State-of-the-art natural language processing (NLP) models are trained on massive training corpora, and report a superlative performance on evaluation datasets. This survey delves into an important attribute of these datasets: the dialect of a language. Motivated by the performance degradation of NLP models for dialectal datasets and its implications for the equity of language technologies, we survey past research in NLP for dialects in terms of datasets, and approaches. We describe a wide range of NLP tasks in terms of two categories: natural language understanding (NLU) (for tasks such as dialect classification, sentiment analysis, parsing, and NLU benchmarks) and natural language generation (NLG) (for summarisation, machine translation, and dialogue systems). The survey is also broad in its coverage of languages which include English, Arabic, German, among others. We observe that past work in NLP concerning dialects goes deeper than mere dialect classification, and extends to several NLU and NLG tasks. For these tasks, we describe classical machine learning using statistical models, along with the recent deep learning-based approaches based on pre-trained language models. We expect that this survey will be useful to NLP researchers interested in building equitable language technologies by rethinking LLM benchmarks and model architectures.

CCS Concepts: • **Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: NLP, dialects, natural language processing, linguistic diversity, large language models, inclusion

Authors' Contact Information: Aditya Joshi, University of New South Wales, Sydney, Australia; e-mail: aditya.joshi@ unsw.edu.au; Raj Dabre, National Institute of Information and Communications Technology, Koganei, Tokyo, Japan; e-mail: raj.dabre@nict.go.jp; Diptesh Kanojia, University of Surrey, Guildford, United Kingdom of Great Britain and Northern Ireland; e-mail: dipteshkanojia@gmail.com; Zhuang Li, Monash University, Clayton, Victoria, Australia; e-mail: zhuang.li1@monash.edu; Haolan Zhan, Monash University, Clayton, Victoria, Australia; e-mail: haolan.zhan@monash.edu; Gholamreza Haffari, Information Technology, Monash University, Clayton, Victoria, Australia; e-mail: gholamreza.haffari@ monash.edu; Doris Dippold, University of Surrey, Guildford, United Kingdom of Great Britain and Northern Ireland; e-mail: d.dippold@surrey.ac.uk.

1 Introduction

Natural language processing (NLP) is an area of artificial intelligence that deals with processing of human language in its textual form. NLP tasks are broadly viewed as two categories: **natural language understanding (NLU)** and **natural language generation (NLG)**. The former covers language understanding tasks such as dialect identification or sentiment classification, as well as tasks such as morphosyntactic analysis. The latter includes tasks where both the input and the output are textual sequences (for example, summarisation). The state-of-the-art NLP, for both NLU and NLG, is based on Transformer-based models [Naveed et al. 2023; Zhao et al. 2023]. **Large language models (LLMs)** that use decoders in the Transformer architecture have significantly increased attention toward NLP leading to LLM-based applications in several domains such as medicine, business, or law. LLMs released by commercial organisations report an increasingly higher number of parameters and, as a result, improved performances on several NLP tasks. NLP approaches using LLMs are largely viewed as black-box models trained on massive corpora whose composition is not accurately known. This survey dissects one of many attributes in which variations may exist in the training and test corpora: dialects of a language.

Traditionally, a dialect is defined as the regionally or locally based variety of a language [Haugen 1966]. Wikipedia defines a dialect as "a variety of a language that is a characteristic of a particular group of the language's speakers." Zampieri and Nakov [2021] state that dialects are language varieties characterised by systematic patterns of variation. The current notion of dialect has extended to language varieties arising due to factors such as political reasons, country of origin, migration histories, historical factors, register shifts, and so on. In fact, there is an association between perceived social hierarchies and dialects of a language, leading to a term "sociolect" [Kroch 1986]. For the sake of brevity, we use "dialects' as an umbrella term to refer to "dialects/national varieties/cultural variants/sociolects" of a language while acknowledging that the distinction between dialects and language is nuanced [Sandel 2015]. An example of a dialect is the national variety, Australian English, whose phonemes are predominantly derived from Southern British English and other Englishes [Cox 2006; Cox and Palethorpe 2007] but has also developed its own unique vocabulary [Moore 1999]. Overlapping with dialects are Creole languages that develop from the process of different languages simplifying and mixing into a new form (often, a pidgin) and then that form expanding and elaborating into a full-fledged language with native speakers, all within a fairly brief period. Lent et al. [2024] highlight the social and scholarly stigmatisation of Creole languages that has resulted in limited advances in NLP for these languages.

In general, our survey is catalysed by the recent efforts in extending LLMs on NLP tasks for dialects of different languages. As researchers continue to look "under the hood" of LLMs, dialectal differences in training and testing datasets are being increasingly scrutinised, and adaptation techniques to improve their performance on different dialects are being devised. As a result, we hope that this survey will help readers and researchers understand past work in NLP techniques for dialects of a language and contribute to ideas about fair and equitable NLP in the future.

There have been related surveys in the past. Zampieri et al. [2020] describe the available corpora and past approaches to fundamental NLP problems such as POS tagging and parsing, along with applications to NLP. Our survey builds upon theirs in three ways. First, we cover a wider range of downstream tasks such as summarisation and sentiment analysis. Also, this survey contains recent papers, which highlight increasingly growing attention towards NLP for dialects. Finally, the exposition of our survey adopts a deep learning-centric viewby covering deep learning-based approaches, in particular, the recent LLM-based approaches. Another survey by Blodgett et al. [2020] describes biases of different kinds in an analysis of language technologies, including dialectal bias. We derive from their survey to formulate the motivation and trends in NLP for dialects. Similarly, Jauhiainen et al. [2019] present a survey of automatic language identification, which

English: My name is Pushpak

Marathi Transliteration: Maza nav Pushpak ahe   Samvedi Transliteration: maa naav Pushpak haay
Literal Translation: My name Pushpak is   Literal Translation: my name Pushpak is

Fig. 1. An example sentence highlighting the differences between Marathi and its Samvedi dialect.

does not differentiate between dialect or language identification, and mention that dialect identification may be a more challenging task. Finally, extensive surveys focusing on languages from the Middle East have been reported [Darwish et al. 2021; Shoufan and Alameri 2015]. These are surveys of NLP for standard and dialectal Arabic, primarily focusing on dialect identification and synthesis in the form of machine translation. Our survey unifies the efforts in dialects of languages belonging to multiple language families. The contribution of our survey is:

— We present past work in terms of NLU and NLG tasksand include both pre-deep learning and deep learning techniques.
— We highlight trends and future directions and provide summary tables that will help researchers interested in dialectal NLP research.
— The survey covers a broad range of languages from around the world.

The rest of the article is organised as follows: We motivate the need for a discussion on dialects in Section 2. We define the scope of the article and highlight key trends in Section 3. We then cover dialect-specific resources in Section 4. Following that, Section 5 covers several NLU tasks: dialect identification, sentiment analysis, parsing, and NLU benchmarks. Section 6 presents relevant approaches in NLG for machine translation, summarisation, and so on. We conclude the survey and discuss future work in the context of NLP research as well as social/ethical implications in Section 7. The survey contains several summary tables that will be useful for future research.

2 Motivation

2.1 Linguistic Challenges Posed by Dialects

Dialectal differences primarily occur in terms of orthography, syntax, and vocabulary. Some examples of dialectal differences in English are: "I might could help you with that" observed in Southern US, Australian, and New Zealand English [Coats 2022; Morin and Coats 2023] as well as British and Irish English [Coats 2023], "Inside tent can not see leh!" in Singaporean English [Wang et al. 2017], or the uncommon placement of adverbs in native speakers of Asian languages as in "Already, I have done it" [Nagata 2014]. Also, consider the case of the Samvedi dialect of Marathi, one of 42, where we give an example of the Samvedi and Marathi sentences in Figure 1. Samvedi does not exhibit word order differences compared to standard Marathi, but it involves heavy pronunciation relaxation (*ahe* -> *hay* and *maza* -> *maa*) and the usage of older words. Another challenge in handling dialects is that two dialects of the same language can be mutually unintelligible. A classic example of this is the case of the Aomori and Okinawan dialects of Japanese, which has a total of 47 known dialects. Therefore, it is not enough to collect data for one dialect and assume that it will help in NLP for another dialect, which indicates that special attention will need to be paid to each dialect to ensure that it will be well-represented. Dialects assume further importance when people from different cultural backgrounds interact with one another. Wang et al. [2022] show that monophthongal vowels spoken by Australian English speakers may be difficult to be understood by Mandarin English listeners.

Dialects are also associated with pragmatics, with influences derived from macro-social factors such as region, social class, ethnicity, gender, and age [Haugh and Schneider 2012]. For example, Schneider [2012] observes differences in small talk across inner circle varieties of English, i.e., varieties from countries where it is the primary language [Kachru 1992]. They also observed differences between speakers of different ages and genders. This suggests that the notion of "dialect"

can be linked to factors beyond geographical distribution. Merrison et al. [2012] showed that, in student requests to university staff, there were differences in the way obligation was expressed and that these differences were linked to different ways of claiming social standing. Meyer [2014] compares interactions of Australians with people from other cultures in terms of (a) building trust with colleagues, (b) leading teams of a culturally dissimilar background, and so on. An example in the book states that an Australian may invest in shorter small talk than a Mexican with a colleague. Noting the differences in the pragmatic strategies of different dialect speakers provides an important social perspective on dialectal variation. However, these are currently not sufficiently accounted for in NLP.

2.2 Rethinking LLM Benchmarks

There are more English language speakers in countries such as India than the United States, Australia, and England [Dunn 2019]. In addition, an even larger number of speakers have acquired English in a classroom context (e.g., in countries such as China, Germany, or Russia) and use it mainly as a contact language for specific transactional purposes, e.g., business or education. This latter perspective has been described through the notion of English as a lingua franca as "the common language of choice […] among speakers who come from different lingua-cultural backgrounds" [Jenkins 2009]. Despite that, the corpora used to train language models—and, more importantly, the datasets used to evaluate them—do not necessarily reflect dialectal variations within a language. Inoue et al. [2021] examine the performance of BERT-based models for varieties/dialects of Arabic and show that dialect proximity of pre-training and fine-tuning data bears impact on the performance of the downstream task. In the case of GPT-4, the evaluation dataset consists of questions from the MMLU benchmark written in Standard American English. Standard benchmarks used to claim performance of a language model for English primarily contain Standard American English. It has been found that the performance does not extend to NLU tasks for dialects of English [Ziems et al. 2022]. Further, a recent work by Fleisig et al. [2024] analyses the output of ChatGPT for varieties of English and shows that the generated output may be of poorer quality and be prone to stereotyping for non-standard dialects of English. These findings hold for most foundation models that are trained on large amounts of data. The distribution of languages in the training corpora is either not known or difficult to determine.

2.3 Fair and Equitable Technologies

NLP systems that are deployed to serve multicultural communities must be mindful of the variations between different dialects. Evaluation and mitigation of disparity between dialects become an overgrowing need in times when language models claim excellent language performance using datasets from a specific dialect alone. Some examples showing the impact of dialects on the performance of NLP tasks are presented in Table 1. We note that these papers are from the past few years, which have otherwise witnessed a great development in the reported performance of NLP models.

Some implications of dialects in terms of sociological factors are:

(1) **Performance of NLP models and per-capita GDP**: A recent work by Kantharuban et al. [2023] shows the dialectal gap in performance of LLM-based solutions for machine translation and automatic speech recognition for several dialects, similar to Ahia et al. [2023], who show the same for topologically diverse languages. They show a positive correlation between gross domestic product per capita and the efficacy of dialectal machine translation.

(2) **Healthcare monitoring**: Jurgens et al. [2017] show that there exists a disparity between popular dialect speakers and others in the case of healthcare monitoring.[1]

---

[1]They also propose a method to mitigate the disparity.

Table 1. Examples of Adverse Impact on NLP Task Performance Due to Dialectal Variations

NLP Task	Paper	Impact
Language classification	[Blodgett et al. 2016]	Language detection shows lower performance for African-American English.
Sentiment classification	[Okpala et al. 2022]	Text in African-American English may be predicted more commonly as hate speech.
Natural Language Understanding	[Ziems et al. 2022]	Popular models perform worse on GLUE tasks for African-American English text.
Summarisation	[Keswani and Celis 2021]	Generated multi-document summaries may be biased towards majority dialect.
Machine translation	[Kantharuban et al. 2023]	Significant drop in MT from and to dialects of Portuguese/Bengali/etc. to and from English.
Parsing	[Scannell 2020]	Lower performance of parsers on Manx Gaelic as compared to Irish/Scottish Gaelic.
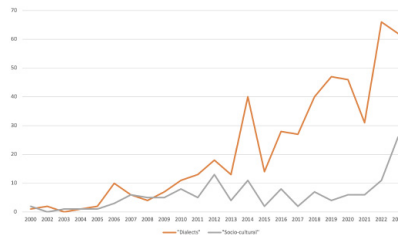
Fig. 2. Number of relevant "papers-per-year" for keywords "dialects" and "socio-cultural" in the *ACL Anthology*.

(3) **Racial biases in hate speech detection**: Okpala et al. [2022] show that hate speech classifiers may lean towards predicting a text as true if it uses African-American English.

(4) **Prejudice in the prediction of employability and criminality**: Hofmann et al. [2024] show that dialects may introduce bias in the output of language models. As a result, a person's output with respect to their employability or criminality may be affected based on the dialects they use.

NLP may not perform as well for dialects of a language, particularly spoken by historically marginalised communities such as the African-American community. This has been shown for language identification where dialects are not predicted as the language, since they differ from the standard version of the language [Blodgett et al. 2016].

An idea closely related to the survey is the "Bender rule" in NLP research. The Bender rule states that the language of datasets used for evaluation must be stated explicitly without assuming English to be the implicit default [Ducel et al. 2022]. We similarly believe that languages are not monoliths, and dialectal differences must be clearly stated. Similarly, Hovy and Yang [2021] show that incorporating dialectal aspects is closely related to social factors of language. As a result, incorporating an understanding of dialects of a dataset is pivoting for fairer NLP tools.

2.4 Recent Work

One observes a renewed interest in using dialects to inform NLP tasks, as shown in Figure 2. The figure was generated using the ACL anthology.[2] For "dialects," we use "dialect," "national variety"

---

[2]https://aclanthology.org/info/development/

(subword for inflections of "variety"), "national variation," and "Creole." For "socio-cultural," we use the words "cultural" and "socio-cultural." We restrict to the year range 2000–2023.

Dialect awareness has been shown to improve the performance of NLP tasks such as machine translation [Sun et al. 2023] and speech recognition [Plüss et al. 2023]. Recent works have also focused on dialect-aware NLP tasks as in the case of machine translation of dialect to standard language translation as in the case of Chinese [Lu et al. 2022] (for Hokkien, a dialect of Chinese).

3 Scope & Trends

The focus of this survey is on NLP approaches that are aware of dialects: either in the form of the choice of the dataset, incorporation in the model, or evaluation along dimensions involving dialect. The survey provides a broad introduction to past NLP research on dialects spoken in different parts of the world. In the forthcoming subsections, we clarify the scope of this article (Section 3.1) and highlight key trends (Section 3.2) that are described in detail in the following sections.

3.1 Scope

We select papers that mention the dialect as an attribute of interest. The focus on dialects is either based on the evaluation datasets or the model innovations to improve performance on dialect-specific datasets.

We keep the following out of scope, primarily to effectively manage the scope of the article:

(1) **Code-mixing**: Code-mixing involves the use of words from two or more languages, often to reduce cognitive load. This survey does not focus on code-mixing.

(2) **Implicit selection biases**: We also acknowledge that selection biases in datasets may introduce dialectal variations. For example, a dataset of tweets downloaded from a specific country is likely to have predominant dialects spoken in the country. However, we cannot locate these papers in particular, or, for social implications, claim that they are based on dialectal variations of a language without the authors mentioning so.

(3) **Accent variations**: Finally, we focus on "text"-based research while acknowledging that the speech processing community has a rich history of using acoustic data centred around accent. To this end, we briefly touch upon speech, since dialects and speech are intertwined to a certain degree. However, our primary focus is on text, since the text aspect has received a lot more attention than the speech aspect. This also sets up a situation where a future survey can expand on the speech aspect of dialectal processing. The focus on the textual form is the typical purview of NLP.

(4) **Systematic review**: This survey is not a systematic review in the sense that we do not exhaustively cover **all** works on dialects due to limited time and article space. Instead, we select key representative papers based on our interpretation of the innovation, which, according to us, cover key progress and innovation in NLP for dialects. We acknowledge that we may have missed out on some important papers in the field. We will incorporate these papers as communicated by readers/reviewers. However, we cover a broad range of approaches in the survey.

(5) **Linguistic studies**: While we acknowledge similar rich linguistic work in terms of understanding dialects, we focus on NLP tasks.[3] For example, dialectometry is a research area that studies variations in dialects of a language [Goebl 1993] but is not included in the survey.

3.2 Trends

Table 2 summarises the papers covered in this survey. We identify three trends in the past work:

---

[3]We cover dialect classification in the section on natural language understanding.

Table 2. State of NLP Research on Dialects

	Languages						Innovation				Problem/Area						
	English	Chinese	Arabic	German	Indic Languages	Other	Dataset	Method/Model	Evaluation/Metric	Benchmark	Dialect Classification	Sentiment Analysis	Machine Translation	Morphology/Parsing	Conversational AI	Summarisation	Speech/Visual
[Nerbonne and Heeringa 1997]	✓								✓		✓						
[Nerbonne and Heeringa 2001]	✓								✓		✓						
[Chiang et al. 2006]		✓					✓								✓		
[Habash and Rambow 2006]		✓						✓					✓				
[Chitturi and Hansen 2008]	✓							✓			✓						
[Paul et al. 2011]	✓			✓	✓	✓		✓					✓				
[Lui and Cook 2013]	✓							✓			✓						
[Abdul-Mageed and Diab 2014]			✓				✓					✓					
[Cotterell and Callison-Burch 2014]			✓				✓				✓						
[Darwish et al. 2014]			✓					✓			✓						
[Doğruöz and Nakov 2014]						✓		✓			✓						
[Estival et al. 2014]	✓						✓										✓
[Jeblee et al. 2014]	✓		✓					✓					✓				
[Zampieri et al. 2014]	✓					✓	✓			✓	✓						
[Jørgensen et al. 2015]	✓								✓						✓		
[Xu et al. 2015]		✓				✓	✓								✓		
[Zampieri et al. 2015]	✓						✓			✓	✓						
[Ali and Habash 2016]			✓					✓							✓		
[Blodgett et al. 2016]	✓							✓			✓						
[Burghardt et al. 2016]				✓			✓				✓						
[Eskander et al. 2016]			✓					✓					✓				
[Goutte et al. 2016]	✓					✓			✓		✓						
[Malmasi et al. 2016]	✓					✓			✓		✓						
[Azouaou and Guellil 2017]						✓	✓				✓						
[Bowers et al. 2017]						✓		✓						✓			
[Criscuolo and Aluisio 2017]	✓					✓			✓		✓						
[Hassan et al. 2017]	✓					✓	✓	✓					✓				
[Jurgens et al. 2017]	✓					✓	✓	✓			✓						
[Mdhaffar et al. 2017]			✓					✓				✓					
[Simaki et al. 2017]	✓							✓			✓						
[Abdul-Mageed et al. 2018]			✓				✓				✓						
[Assiri et al. 2018]			✓					✓				✓					
[Blodgett et al. 2018]	✓		✓					✓							✓		
[Darwish et al. 2018]			✓					✓					✓				
[Erdmann et al. 2018]			✓					✓	✓								
[Elmadany et al. 2018b]			✓				✓	✓					✓				✓
[Elmadany et al. 2018a]			✓					✓							✓		
[Salameh et al. 2018]			✓				✓						✓				
[Baly et al. 2019]			✓				✓					✓					
[Fadhil et al. 2019]			✓					✓	✓						✓		
[Joukhadar et al. 2019]			✓					✓							✓		
[Mulki et al. 2019]			✓					✓				✓					
[Sap et al. 2019]	✓								✓			✓					
[Zampieri et al. 2019]	✓			✓		✓	✓			✓	✓				✓		
[Ahmed and Hussein 2020]						✓		✓							✓		
[Al-Ghadhban and Al-Twairesh 2020]			✓					✓							✓		
[Alshareef and Siddiqui 2020]			✓					✓							✓		
[Demszky et al. 2021]				✓				✓	✓		✓						
[Dunn and Adams 2020]	✓					✓	✓			✓	✓						
[Hanani and Naser 2020]			✓					✓			✓						
[Hou and Huang 2020]		✓				✓		✓			✓						

(Continued)

Table 2. Continued

	Languages						Innovation				Problem/Area						
	English	Chinese	Arabic	German	Indic Languages	Other	Dataset	Method/Model	Evaluation/Metric	Benchmark	Dialect Classification	Sentiment Analysis	Machine Translation	Morphology/Parsing	Conversational AI	Summarisation	Speech/Visual
[Mozafari et al. 2020]	✓							✓				✓					
[Tan et al. 2020]	✓							✓					✓✓				
[Zhao et al. 2020]	✓							✓						✓			
[Ball-Burack et al. 2021]	✓							✓				✓					
[Ben Elhaj Mabrouk et al. 2021]						✓		✓							✓		
[Boujou et al. 2021]			✓				✓				✓	✓					
[El Mekki et al. 2021]			✓					✓				✓					
[Guellil et al. 2021]			✓				✓	✓				✓					
[Keswani and Celis 2021]	✓							✓								✓	
[Kumar et al. 2021]	✓		✓			✓		✓					✓				
[Zhang et al. 2021]	✓							✓					✓				
[Chow and Bond 2022]	✓							✓						✓			
[Coats 2022]	✓						✓										
[Eggleston and O'Connor 2022]	✓							✓						✓			
[Fuad and Al-Yahya 2022]			✓					✓							✓		
[Harris et al. 2022]	✓							✓				✓					
[Husain et al. 2022]						✓		✓				✓					
[Inoue et al. 2022]			✓					✓						✓			
[Kanjirangat et al. 2022]			✓	✓	✓				✓			✓					
[Kaseb and Farouk 2022]			✓					✓				✓					
[Kåsen et al. 2022]						✓		✓						✓			
[Liu et al. 2022]	✓						✓	✓					✓				
[Lu et al. 2022]		✓					✓	✓			✓						
[Okpala et al. 2022]	✓							✓				✓					
[Olabisi et al. 2022]	✓						✓									✓	
[Rajai and Ennasser 2022]	✓		✓						✓				✓				
[Saadany et al. 2022]			✓			✓			✓				✓				
[Artemova and Plank 2023]				✓			✓		✓				✓				
[Held et al. 2023]	✓						✓	✓									
[Kantharuban et al. 2023]				✓		✓			✓								
[Kuparinen et al. 2023]				✓		✓			✓								
[Lameli and Schönberg 2023]				✓			✓				✓						
[Le and Luu 2023]						✓	✓										
[Lent et al. 2024]						✓	✓	✓	✓		✓	✓					
[Maurya et al. 2023]				✓			✓						✓				
[Plüss et al. 2023]				✓		✓	✓										✓
[Ramponi and Casula 2023a]						✓	✓				✓						
[Riley et al. 2023]	✓	✓				✓			✓		✓	✓					
[Zhan et al. 2023]		✓						✓	✓	✓					✓		
[Zhan et al. 2024]		✓						✓	✓	✓					✓		
[Artemova et al. 2024]				✓				✓	✓	✓					✓		
[Ziems et al. 2023]	✓							✓		✓							
[Ahia et al. 2024]						✓	✓	✓		✓			✓				✓
[Talafha et al. 2024]			✓				✓	✓		✓	✓						✓
[Dinh et al. 2024]						✓	✓	✓			✓						✓
[Vidal-Gorène et al. 2024]						✓	✓	✓		✓				✓			
[Dabre et al. 2024]				✓			✓	✓					✓				

Training Datasets	Tasks	Model		Evaluation
Labeled Datasets - Geo-located - Keyword-based - Annotation by native speakers of a dialect - Semi-supervised creation **Treebanks & Lexicons** - Manual annotation - Co-occurrence using statistical methods	- Dialect classification - Sentiment analysis - Parsing - Machine translation - Natural language entailment - ...	**Cross-dialect transfer** - Perturbation rules - Grammar - Lattice path **Dialect invariance** - Character embeddings	**Dialect awareness** - Dialect-specific components/neural layers - Adversarial learning - Multi-task prediction of dialect and task-specific labels - Learning adapters	**Dialectic datasets** **Correlation with downstream tasks** Sociological factors **Applications:** Health monitoring for African-American English

Fig. 3. An overview of approaches in terms of NLP for dialects.

(1) **Tasks in focus**: Older research dealt with dialectal datasets primarily for dialect classification. Past work shows performance degradation when the text contains dialects of a language as compared to the predominant (i.e., standard) form.

(2) **Languages in focus**: The papers reporting work on dialects of Arabic are significantly more than those for dialects of other languages. This has also been accelerated by research forums focusing on Arabic NLP. While the work in English is predominantly for the African-American dialect of English, recent papers examine other dialects such as Indian English, Singaporean English, and so on.

(3) **Mitigation is more than perturbation**: Modifying a sentence or its representation to or from its dialectal variations has been achieved by perturbation techniques of varying complexity. However, recent papers show that dialect mitigation can be integrated into the model architecture itself using adversarial networks [Ball-Burack et al. 2021], hypernetworks [Xiao et al. 2023], and so on.

It may seem that NLP for dialects of a language only pertains to datasets, i.e., it does not need any specialised handling beyond the introduction of a new dataset. However, we observe that the adaptation of NLP techniques for dialects operates at several points in a typical NLP pipeline:

(1) **Training resources**: Labeled datasets (including treebanks) and lexicons in dialects of a language have been reported in the past. This includes datasets with dialect labels along with additional task-specific labels, where the task is an NLP research problem.

(2) **Models**: Models have been enhanced with several techniques, as may be typical of the time of the research. The fact that dialect-aware NLP can benefit from model adaptations and not dataset replacement alone is a key point of the survey.

(3) **Evaluation datasets**: NLP techniques evaluated on datasets in dialects have peculiar observations. Language identification classifiers produce lower performance when the text is in a dialect of a language. The performance of LLMs on dialectal datasets is positively correlated with socio-economic factors.

Figure 3 shows an overview of the approaches in terms of NLP for dialects. There have been different approaches to create labeled datasets, tree-banks, and lexicons. In terms of models, past work varies in terms of NLP tasks and the way dialectal adaptation is handled: dialect transformation (where data is translated between dialects for the purpose of processing), dialect invariance (where models are made invariant to dialects), and dialect awareness (where models include dialect-specific components). Finally, we also describe dialectal datasets and resultant evaluations on downstream tasks including applications such as health monitoring.

4 Resources

Being a data-driven field, NLP techniques rely on resources such as lexicons and textual datasets. In this section, we describe ways in which dialectal datasets have been created.

4.1 Dialectal Lexicons

Dialectal lexicons correspond to word lists or word mappings about a dialect. Although lexicons were popular in early approaches of NLP, a recent paper by Artemova and Plank [2023] highlights the potential of dialectal lexicons and describes an approach to create such lexicons using large language models. Prior to this, research in the creation of dialectal lexicons lies in two categories: the use of online dictionaries and the use of textual corpora.

4.1.1 Online Dictionaries. Azouaou and Guellil [2017] create a lexicon of words mapping French and its Algerian dialect. They use online dictionaries along with a combination of manual and automatic methods to enhance the lexicon. This includes many-to-one mapping of words in the two sets. Similarly, Boujelbane et al. [2013] build bilingual lexicons to create Tunisian dialectal corpora to adapt n-gram models for statistical machine translation.

4.1.2 Textual Corpora. Abdul-Mageed and Diab [2014] present a lexicon of words in dialects of Arabic including Levantine and Egyptian dialects. They present two lexicons: adjectives in news articles and common words in online chat forums. The words are labeled with a combination of manual and automatic techniques, the latter based on statistical techniques such as pointwise mutual information. Similarly, Burghardt et al. [2016] use the web as a corpus to create a lexicon for the Bavarian dialect of German. Starting with a corpus of Facebook comments, they provide a rule-based algorithm to create the lexicon. They first extract unique text forms and then filter non-dialect words based on the Dortmund chat corpus. Harrat et al. [2018]; Younes et al. [2020] discuss various existing resources for the **Maghrebi Arabic dialects (MAD)** including annotated corpora for language identification and morpho-syntactic analysis. MAD include principally Algerian Arabic, Moroccan Arabic, and Tunisian Arabic.

4.1.3 Lexicon Induction Using LLMs. A recent approach by Artemova and Plank [2023] performs German dialect lexicon induction using LLMs.

4.2 Dialectal Datasets

Datasets based on different data sources (such as social media and conversation transcripts) and dialects have been reported. In terms of procuring and labeling these datasets, the following methods have been used:

4.2.1 Recruit Native Speakers of Specific Language Varieties. Estival et al. [2014] create a dataset of audio-visual recordings of 1,000 speakers of Australian English. The dataset is accompanied by a transcript, which was manually created for 100 speakers. Bouamor et al. [2018] present MADAR: a manually curated parallel corpus of sentences in Arabic dialects along with English, French, and Modern Standard Arabic.[4] Similarly, Riley et al. [2023] create a parallel corpus of English sentences and two dialects each of Portuguese and Chinese with the help of native speakers of these dialects. Eisenstein et al. [2023] introduce MD-3, a dataset of conversations between speakers playing the game of Taboo. The dataset consists of speech recordings as well as text transcripts. Seddah et al. [2020] focus on treebank creation for Algerian supplemented with monolingual data obtained from CommonCrawl. They highlight the inherent difficulty of finding annotators and the cost of the same indicating the challenges for dialectal data generation. Riabi et al. [2023] further extend this with additional layers of morpho-syntactic knowledge and correct errors in the same.

4.2.2 Perturbation. Ziems et al. [2022] evaluate natural language understanding for African-American English. They design rules to perturb the dataset from Standard American English to African-American English. They then get them validated by native speakers. Ziems et al. [2023] present Multi-VALUE, a suite of resources to evaluate fairness of LLMs by creating dialectal

---

[4]Obeid et al. [2019] is a demonstration based on the dataset.

variations of a dataset. The suite provides mechanisms to generate 50 dialects of English by applying a set of perturbations. Messner and Lippincott [2024] present a dataset of 19th century American literary orthovariant tokens with a novel layer of human-annotated dialect group tags, to examine language modelling assumptions, and find evidence that choice of tokenisation scheme meaningfully impacts the type of orthographic information in a language model.

4.2.3 Keywords. Wang et al. [2017] create a dataset of Singaporean English sentences by searching for typical Singaporean English terms in online forums. Ramponi and Casula [2023a] take a *complementary* approach to create a dataset of tweets in dialects of Italian along with other languages spoken in Italy (which are not necessarily derived from Italian). They use location-based search to obtain the set of tweets from different regions of interest from within Italy. Following this, they use out-of-vocabulary words to identify words that are indicative of geographical regions and, as a result, dialects. A related dataset is GeoLingIt [Ramponi and Casula 2023b]. In the case of social media, *hashtags* can be used to obtain datasets in certain dialects. Kuparinen [2023] take advantage of dialect awareness week in Finland. They use a hashtag indicating usage of dialects to collect tweets in different dialects of Finnish. In the context of Arabic dialect tweets, Boujou et al. [2021] benchmark is a novel dataset of 50,000 tweets for five dialects of Arabic-Algerian, Lebanon, Morocco, Tunisian, and Egyptian.

4.2.4 Location. Data from particular geographics can be extracted using filters (where the location is known) or inference (where it is not known). Jurgens et al. [2017] use location-based filters available on Twitter at the time. They use language identification classifiers to predict the language and identify dialectal users. Husain et al. [2022] obtain tweets from Kuwait to create a dataset of tweets in the Kuwaiti dialect of Arabic. Coats [2022] creates an unlabeled dataset of YouTube comments. They start with a list of councils in Australia, extract official YouTube channels, and retrieve comments. They manually validate the correctness of the channels. When working with geographically dispersed dialects, *sampling* may also be used. Hovy and Purschke [2018] use Doc2Vec on a large corpus of anonymous online posts to learn document representation of cities and recover dialect areas using geographic information via retrofitting and agglomerative clustering. Dunn and Adams [2020] create a Web-based corpus in different dialects by sampling sentences from different countries. The goal is to build a Web-based corpus where the number of instances is reflective of the population of speakers in a country. The paper states that such a geography-aware corpus can lead to geography-aware representations when language models are trained on them. A criticism to the location-based filtering mechanism is by Goutte et al. [2016].

4.2.5 Dialect-aware Annotation. One such example is by Sap et al. [2019]. They examine racial bias towards African-American English in the case of hate speech detection. They propose race and dialect priming to improve the quality of annotation. To prime the annotators, they propose to ask two questions: (a) is the tweet offensive to them? and (b) is the tweet offensive to anyone? The dialect and race of the speaker are shown to the annotators.

Several datasets exist for varieties of Arabic [Diab et al. 2010], including Palestinian Arabic [Dibas et al. 2022; Jarrar et al. 2017], Gulf Arabic [Khalifa et al. 2016], Egyptian Arabic [Maamouri et al. 2014], and Bahraini Arabic [Abdulrahim et al. 2022]. This is in stark contrast with the lack of availability of datasets for dialects of English or several other languages of the world.

5 Natural Language Understanding (NLU)

This section covers NLU approaches centred around dialects. This includes approaches for NLP tasks such as dialect identification, sentiment analysis, morphosyntactic analysis, and parsing. We also describe approaches reported on NLU benchmarks, which cover multiple tasks.

Table 3. Shared Tasks Related to Dialect Identification

Shared Task	Dialects/Languages
[Zampieri et al. 2014]	Brazilian Portuguese and European Portuguese; American and British English; and Argentinian Spanish and Castilian Spanish
[Zampieri et al. 2015]	American and British English; and Argentinian Spanish and Castilian Spanish
[Malmasi et al. 2016]	Dialects of English, Spanish, French and Arabic
[Zampieri et al. 2019]	Dialects of German, Chinese, Romanian
[Gaman et al. 2020]	Dialects of Romanian, Geolocation-based Varieties
[Aepli et al. 2022]	Dialects of French and Italian
[Aepli et al. 2023]	Dialects of Indo-European and Ural languages (and other tasks)
[Abdul-Mageed et al. 2023]	Dialects of Arabic

5.1 Dialect Identification

The most commonly researched task in the scope of this article is dialect identification. Dialect identification deals with the prediction of the dialect of an input text. Early approaches to dialect identification employed distance-based metrics, namely, Levenshtein, Manhattan, and Euclidean distance with different clustering techniques [Nerbonne and Heeringa 1997, 2002]. They indicate that feature representations are more sensitive and that Manhattan distance and Euclidean distance are good measures of phonetic overlap. Elnagar et al. [2021] is a systematic review of identification of dialects of Arabic. For dialects of Arabic, lexical resources such as lexicons and treebanks and models using SVM or sequential neural layers like BiLSTM have been reported. Jauhiainen et al. [2019] is a survey of automatic language identification. They describe that dialect detection may be more difficult than language detection, since dialects may have lexical or syntactic overlap. In doing so, the survey does not make a distinction between languages and dialects—and treats different dialects as different class labels, while still maintaining a classification approach. However, one sees challenges in this regard. Boujou et al. [2021] present a baseline approach, which utilises classical machine learning. While the majority of past work defines dialect identification as a Boolean/multi-class classification, Baimukan et al. [2022] use a hierarchy of dialect labels based on geographical and linguistic proximity. We now describe details of past work in dialect identification in terms of shared tasks, datasets, and pre-deep learning and deep learning-based approaches.

5.1.1 Shared Tasks. Shared tasks have accelerated past work in dialect identification. These have been primarily led by Workshop on NLP for Similar Languages, Varieties and Dialects, also known as VarDial. The shared tasks are listed in Table 3. Goutte et al. [2016] summarise the findings from the past versions of the shared task from 2014–2016. Similarly, the **Nuanced Arabic Dialect identification (NADI)** shared task is held annually. In the 2024 edition of NADI, a new task has been introduced to estimate the Arabic level of *dialectness* within Arabic sentences.

5.1.2 Datasets. Aji et al. [2022] report a dataset for several languages and dialects spoken in Indonesia. They observe that language identification works well for certain dialects (Ngoko-Central dialect of Javanese, for example). The paper also discusses code-mixing and orthography variations in these languages. Dunn [2019] reports dialect identification on 14 national varieties of English. He shows that cross-domain classification (CommonCrawl versus Twitter) also performs poorly. Cotterell and Callison-Burch [2014] present a multi-dialect, multi-genre corpus of news comments and tweets written in dialects of Arabic. The tweets are manually annotated for

dialect identification on MTurk. Ramponi and Casula [2023a] present a benchmark dataset for dialects of Italian. The benchmark is named as DIATOPIT. There has been recent work on creating a corpus of Norwegian dialect [Barnes et al. 2021]. Also, Alshutayri and Atwell [2018] present a large (200K+ instances) corpus for Arabic dialects and Standard Arabic. The data is sourced largely from tweets but also includes comments from newspapers and Facebook. The data is also being annotated for dialect identification and contains 24K annotated documents. Recently, Talafha et al. [2024] introduced CASABLANCA, a large scale community-driven effort to collect and transcribe a multi-dialectal Arabic dataset, covering eight dialects for Arabic releasing 48 hours of manually transcribed speech data including annotations for transcription, gender, dialect, and code-switching. Le and Luu [2023] present a parallel corpus for dialects of Vietnamese. Further, Dinh et al. [2024] propose a dialect identification and speech recognition dataset and fine-tuned models for 63 provincial dialects of Vietnamese with 102.5 hours of audio, and 19, 000 spoken utterances.

5.1.3   Feature-based Approaches. We now highlight features used for dialect identification.

(1) **Phonological features**: Phonological features are based on markers in the written scripts. Darwish et al. [2014] use lexical along with a lexicon of dialectal Egyptian words, morphological and phonological features in a random forest classifier to detect dialects of Arabic spoken in a geographical region.

(2) **Linguistic features**: Doğruöz and Nakov [2014] present a method to predict dialects of Turkish by using light verb constructions. They use a statistical classifier based on verb-based features (base word, verb order, affixes, etc.) for the task. Xie et al. [2024] discuss an approach to extract distinguishing lexical features of dialects by utilising interpretable dialect classifiers. With focus on varieties of Mandarin, Italian, and Low Saxon, this approach shows promising results on all varieties.

The combinations of the above set of features have also been reported. While Hanani and Naser [2020] work on the detection of dialects from speech, they also use word-level n-gram features. Salameh et al. [2018] perform fine-grained dialect identification for 25 dialects of Arabic, using Naïve Bayes classifier and word and character n-grams as features.

While dialect detection of Arabic has been explored in detail, the pre-deep learning work in the context of dialects of English is comparatively limited, although English is the predominant language for NLP research. Lui and Cook [2013] is an early work in the detection of dialects of English. Specifically, the paper focuses on Australian, British, and Canadian English. Their baseline is the LangID classifier [Lui and Baldwin 2012], where dialects are treated as individual languages. They experiment with classifiers using features such as n-grams and POS-n-grams. This includes a distribution over function words and those in a vocabulary, akin to a clustering algorithm. Simaki et al. [2017] use linguistic, POS-tag-based, and lexicon-based features.

5.1.4   Deep Learning-based Approaches. Deep learning-based approaches for dialect classification span three alternatives: train embeddings to reflect dialectal variations, use end-to-end LLMs, or predict dialect as a result of inference over dialect features.

Embeddings in focus: Abdul-Mageed et al. [2018] label tweets with 10 dialects of Arabic. The city is considered the dialectal granularity. The analysis compares dialectal variants by looking at word embeddings of words across different dialects. They use word2vec representations to show how dialectal words are captured. Goswami et al. [2020] build character-to-sentence embeddings to represent words of different dialects. Unsupervised loss is computed to generate clusters of representations. While they also test on language identification, the dialect identification part is done on Swiss German dialect. Jurgens et al. [2017] use a character-based *seq2seq* model to map dialects. The models used for language identification are RNNs with GRU. Criscuolo and Aluisio

Table 4. Sentiment Analysis Approaches for **Dialectal** Datasets

Paper	Dialects	Modelling Approach
[Mdhaffar et al. 2017]	Dialects of Arabic	SVM/MLP
[Mulki et al. 2019]	Dialects of Arabic	Syntax-ignorant composition to learn word embeddings
[Mozafari et al. 2020]	African-American English	Re-weight instances based on racially biased phrases for hate speech detection
[El Mekki et al. 2021]	Dialects of Arabic	Infer dialect and sentiment label using two channels from BERT encoder
[Guellil et al. 2021]	Dialects of Arabic	Transliteration to map to standard version
[Kaseb and Farouk 2022]	Dialects of Arabic	Infer dialect and sarcasm label; limited backpropagation to maintain label dependency
[Okpala et al. 2022]	African-American English	Adversarial training to debias dialectic variation
[Moghimifar et al. 2023]	English	Socio-cultural norms are inferred using a Markov model variation

[2017] use character n-grams to identify language groups. This is followed by convolutional neural network-based dialect classifiers for each language group.

Fine-tuning LLMs: Ramponi and Casula [2023a] experiment with multiple models including statistical and neural. The fine-tuned AlBERTo model performs the best among umBERTo, mBERT, and XLM-R. Obeid et al. [2020] present CAMeL: a Python toolkit for Arabic language processing. It contains a dialect identifier that gives a distribution over multiple dialects. They use dialectal guidelines provided in Elfardy and Diab [2012].

Detecting dialect features: Demszky et al. [2021] introduce an approach for dialect classification using a novel multi-task approach that employs dialect feature detection. They train two multi-task learning-based approaches using a small number of minimal pairs. They evaluate the output based on 22 dialectal features for Indian English and demonstrate that such models show the capability of learning to identify features with high accuracy. They show the efficacy of this task by applying it to dialect identification and by providing a measure of dialect density.

5.2 Sentiment Analysis

Sentiment analysis is the NLU task of prediction of sentiment polarity of a text. Sentiment analysis encompasses several related tasks, such as sarcasm classification and target-specific sentiment analysis. We discuss past work in sentiment analysis along four directions: experiences from annotation (which highlights the challenge of dialects for sentiment analysis), dialect-aware models, dialect-invariant models, and, finally, de-biasing of sentiment analysis models as a post-processing step. Table 4 summarises approaches for sentiment analysis.

Early guessing for dialects: A recent advancement in dialect identification is early guessing [Kanjirangat et al. 2022]. The approach detects a dialect for an incremental input. Salloum and Habash [2022] also break the input down into its components. Specifically, they present an unsupervised approach that uses unsupervised dialect segmentation for machine translation.

5.2.1 Datasets & Annotation. Several datasets in Arabic sentiment analysis for dialects have been reported, such as Moroccan [Oussous et al. 2020] and Levantine [Baly et al. 2019]. Dialects

can have an impact on annotation itself. Farha and Magdy [2022] show that dialect familiarity helps sarcasm annotation. Mdhaffar et al. [2017] create a dataset of 17,000 Facebook comments labeled with sentiment in Tunisian dialect of Arabic. Assiri et al. [2018] present a sentiment-labeled lexicon of words in the Saudi dialect of Arabic and use simple counting-based sentiment analysis. Husain et al. [2022] use weakly supervised labels for sentiment analysis of tweets in the Kuwaiti dialect of Arabic. The labels are then manually validated and updated.

5.2.2 Dialect-aware Representations. Given the high degree of similarity between dialects, there is a high likelihood for models to make inferences in the same way for different dialects, and thus explicitly modelling dialect awareness into models is important. However, this same degree of similarity makes this dialect-aware modelling challenging. Farha and Magdy [2022] train BERT-based models for sarcasm detection on data annotated by either of the two groups: those familiar with the dialect and those not. They show that familiarity of dialect improves the quality of the models trained on such a dataset. As a result, representations that capture dialects have been used for sentiment analysis. Mdhaffar et al. [2017] present models based on SVM and **multi-layer perceptron (MLP)**. Mulki et al. [2019] use a syntax-ignorant n-gram composition to create embeddings. The classifier model is a dense neural network that works on the addition of word embeddings, with a softmax at the end. Guellil et al. [2021] propose "one" model for sentiment classification in different dialects of Arabic. They use transliteration to map dialects to Standard Arabic. The sentiment analysis model itself uses word2vec features with statistical classifier. Finally, Husain et al. [2022] present statistical models based on SVM along with Transformers-based models like BERT.

5.2.3 Incorporating Dialect Information in Sentiment Prediction. El Mekki et al. [2021] use domain adaptation for sentiment analysis of dialects. Using representations from a BERT encoder, they use two classifiers: sentiment classifier and dialect classifier. The output of the two is later combined for the overall prediction. While this is a two-channel approach, the representation used for the task has also been used to predict dialect of the language. One such example is Okpala et al. [2022], who present an approach for hate speech detection using African-American English. To do so, they re-train BERT with AAE tweets. Finally, adversarial training is needed to regulate the debiasing of the hate speech classifier. Specifically, the adversary takes the final representation learned by the hate-speech classifier and learns to predict the dialect from it. Kaseb and Farouk [2022] present a dialect-aware approach for sarcasm detection called the SAIDS model. SAIDS uses MARBERT to detect dialect and sarcasm. Following that, MARBERT, along with sarcasm and dialect output, are used to detect sentiment. Evaluated on Arabic dialects, SAIDS uses backpropagation only for prediction with respect to the BERT base model. It does not flow through sentiment<->sarcasm or sentiment<->dialect.

5.2.4 De-biasing Sentiment Analysis Models. Making sentiment analysis agnostic to dialects involves removing dialectal biases in the resultant models. A work of this nature is by Ball-Burack et al. [2021], who apply adversarial debiasing to resampled data for harmful tweet detection of tweets written in African-American English. Resampling of the data uses a metric for margin of confidence that selects the set of tweets that are most likely to be mis-classified. Adversarial debiasing involves training an adversary network to debias the classifier by including the adversary network's loss. Similarly, Mozafari et al. [2020] report results on hate speech detection from African-American and Standard American tweets. They re-weight instances based on the presence of phrases that may highlight racial bias. They fine-tune BERT for the task. Finally, Zhang et al. [2021] present an approach to reduce spurious correlation between two attributes: toxicity and African-American Vernacular English. They construct triplets of sentences where the first two have the same toxicity label, and the first and the third have the same dialect label. The objective function of the model consists of a triplet loss over these triplets and a disentanglement loss that

Table 5. Approaches for Morphological Analysis Focusing on Dialects

Paper	Dialects	Highlight
[Habash and Rambow 2006]	Dialects of Arabic	Rewrite rules to adapt morph analysers
[Eskander et al. 2016]	Dialects of Arabic	Dataset proportions for improved performance
[Jørgensen et al. 2015]	African-American English	POS taggers perform worse for the dialect
[Darwish et al. 2018]	Arabic	CRF-based POS tagger; Linguistic features
[Inoue et al. 2022]	Arabic	Fine-tuned LLMs
[Bafna et al. 2023]	Indic Languages	Fine-tuned LLMs

Highlight-> Approach or Key Finding.

ensures the masks for the true attributes are well-separated. Similarly, graphical models have been used to infer socio-cultural norms, since they are closely associated with dialectal variations based on the language and cultural background of the speaker. Moghimifar et al. [2023] present a Markov model to discover socio-cultural norms in emotion classification. Harris et al. [2024] evaluate the zero-shot performance of speech recognition systems across different genders and across four US-based English dialects: SAE, AAVE, Chicano English, and Spanglish. They release a labeled dataset of 13 hours of podcast audio, transcribed by speakers of the represented dialects. While past research has only dealt with African-American English, there may indeed be other dialects, which are considered aggressive and may result in sentiment analysers producing biased output. This is significantly underexplored for dialects of other languages such as the Khariboli (Haryanvi group) dialect of Hindi [Yadav 1974].

5.3 Morphosyntactic Analysis

Morphosyntactic analysis deals with linguistic tasks such as POS tagging and morphological analysis, and has been found to be useful for sense disambiguation, particularly in low-resource settings [Khalifa et al. 2020]. We now describe past work that deals with dialectal variations, as summarised in Table 5.

5.3.1 Classical Approaches. Habash and Rambow [2006] is a seminal morphological analyser for dialects of Arabic called MAGEAD. Using morphological rewrite rules, they show how a morphological analyser can be adapted for dialects of a language. Jørgensen et al. [2015] evaluate on a dataset of African-American Vernacular English and show that the then-prevalent POS taggers perform significantly worse. Darwish et al. [2018] present a CRF-based POS tagger for dialects of Arabic. The POS tagger is trained on a small set of tweets using features derived from the dialects of interest. These features are progressive and negation particles. Eskander et al. [2016] adapt existing morphological analysers to unseen dialects of Arabic by simulating the low-resource dialects.

5.3.2 Deep Learning-based Approaches. Inoue et al. [2022] use CamelBERT trained on Modern Standard Arabic fine-tuned on dialect-specific datasets for morphosyntactic analysis. They observe that training using high-resource dialects helps low-resource dialects as well. In the context of Indic languages, Bafna et al. [2023] explore POS tagging for five Indic dialects by focusing on Hindi-aware LLM adaptation via small dialectal monolingual corpora. Aepli and Sennrich [2022] propose improving cross-lingual transfer between closely related language varieties from the Finnic, West and North Germanic, and Western Romance language branches using character-level noise injection and go on to show consistent improvements for POS tagging. Their approach is further applied to 7 languages from 3 families and a total of 18 dialects [Blaschke et al. 2023] with results showing improvements by varying the level of noise injected during the cross-lingual transfer.

5.4 Parsing

Parsing involves the creation of syntactic parse trees from text. Past work in parsing texts written in dialects of a language lies in three categories. The first category uses an existing parser on a dataset in a dialect of interest. The focus of such work is to create a baseline performance of popular parsers. The second category provides approaches to adapt existing parsers towards texts in the dialect of a language. The third category creates a new parser for the dialect.

5.4.1 Use of Existing Parsers. Eggleston and O'Connor [2022] parse tweets in Standard American English and African-American English and use it to analyse social attributes of an entity, as per sentiment expressed in the tweets. Kåsen et al. [2022] create a tree bank of sentences in the Bokmål variety of Norwegian dialects. They present their results on the UUParser, an existing parser for Norwegian. Roy et al. [2020] present an analysis using Stanford parser and Allen NLP parser on parsing of news headlines in Indian English. Scannell [2020] create a treebank for Manx Gaelic and compare the performance of existing classifiers with Irish Gaelic and Scottish Gaelic.

5.4.2 Adaptation of an Existing Parser. Chiang et al. [2006] show how parsing of Arabic dialects can be done by a sentence transduction approach. This approach parses the standardised version of a dialectal sentence and then links it to the original sentence. The standardisation is achieved using transduction, akin to n-gram decoding. However, Blodgett et al. [2018] use neural networks and present an approach to dependency parsing for African-American English. This approach uses two neural parsers, which are modified with the word embeddings used for initialisation. The word embeddings are trained on the standard and the dialect-specific datasets. Further, Wang et al. [2017] create a dependency parser for Singaporean English. This approach uses a base parser for standard English and stacks it with a series of BiLSTM layers known as the "feature stack" to extract relevant features and an MLP with an output layer to help produce dependency-parsed output. Zhao et al. [2020] use a treebank of learner English sentences labeled with POS tags and dependency information. They propose a factorisation-based parser that first predicts nodes followed by edges in a dependency parse. Dou et al. [2023] evaluate various parsers designed for converting text to SQL, focusing on a multilingual benchmark that covers dialects from seven different languages. This research is significant for its emphasis on semantic parsing, differing from the aforementioned dependency parsing works.

5.4.3 Development of a New Parser. Vaillant [2008] propose a rule-based approach to construct a common syntactic description for a group of Creoles from Haiti, Guadeloupe, Martinique, and French Guiana. Bowers et al. [2017] present a finite-state machine-based parser for the endangered Odawa dialect of Ojibwe spoken in Canada and northeastern United States. This approach uses a phonological module composed of a morphological module where morphological strings are modified by the phonology until they match surface forms of the language.

5.5 NLU Benchmarks

Finally, benchmarks such as GLUE, which provide datasets for NLP tasks such as **semantic textual similarity prediction (STS-B)**, sentiment classification (using the **Stanford sentiment treebank (SST-2)**), **natural language inference (NLI)**, textual entailment, and so on, are an important part of language model evaluation pipeline, as summarised in Table 6. Ziems et al. [2022] show a drop in performance on seven GLUE tasks including SST-2, STS-B, when tested on dialectal English variations of the original Standard American English version. For example, for SST-2, there is a 1.5%–2% drop using fine-tuned RoBERTa.

Dacon et al. [2022] work with African-American English. They first propose CodeSwitch, a rule-based method of perturbing a sentence from **Standard American English (SAE)** to **African-American English (AAE)**. They create perturbed versions of the dataset using

Table 6. Dialect-aware Approaches Evaluated on NLU Benchmarks

Paper	Dialects	Approach
[Ziems et al. 2022]	African-American English	Perturbation to create variants
[Dacon et al. 2022]	African-American English	Adversarial learning
[Held et al. 2023]	Dialects of English	Contrastive loss, Morphosyntactic loss
[Xiao et al. 2023]	Dialects of English	Hypernetworks as LoRA adapters

CodeSwitch and manually evaluate it. They finally evaluate their method on NLI. To do so, they use adversarial learning that ensures that the predicted label is the same if either the SAE or AAE sentences are provided as the input. They refer to this as a disentanglement of language style. Tan et al. [2020] present base-inflection encoding: a mechanism to inject dialectal information into the encoder. They show that their encoding algorithm improves the performance of Vernacular African-American English for SQUAD and MNLI tasks.

Held et al. [2023] model natural language understanding for dialects as a dialect adaptation task. Using Multi-VALUE, they create African-American English variations of the GLUE benchmark (which is primarily written in Standard American English). Following that, they adapt a model pre-trained on Standard American English. To do so, they use: (a) a contrastive loss to ensure the representation of a standard sentence and its dialectal version is as close as possible; (b) a morphosyntactic loss based on word-level alignment between the standard and dialectal sentences. Their results show improved robustness on four dialects based on the GLUE benchmark. Vidal-Gorène et al. [2024] provide a benchmark on lemmatisation, POS-tagging, and morphological analysis for four Armenian varieties—Classical, Modern Eastern, Modern Western, and the under-documented Getashen dialect. They compare traditional RNN models, multilingual encoders, and large language models using supervised, transfer learning, and zero/few-shot learning approaches, and they show how RNNs are strong at POS tagging, but LLMs handle unseen dialectal variations.

A recent work by Xiao et al. [2023] shows how low-rank adapters **Low-Rank Adapters (LoRA)** (a parameter-efficient fine-tuning or PEFT technique that allows fine-tuning LLMs faster by storing weight updates instead of updating all weights) can use linguistic knowledge of dialects to improve zero-shot performance on NLU tasks. They integrate hypernetworks with LoRA adapters for dialect adaptation. Experts encode linguistic information in the form of feature vectors. A hypernetwork then learns to generate adapter weights for LoRA from the feature vectors. They demonstrate the impact of their fine-tuning approach on several GLUE tasks such as MNLI, RTE, and so on. The dataset consists of variants of the GLUE benchmark for five dialects: **African American Vernacular English (AAVE)**, **Indian English (IndE)**, **Nigerian English (NgE)**, **Colloquial Singaporean English (CollSgE)**, and **Chicano English (ChcE)**. Similarly, Liu et al. [2023] use dynamic aggregation of linguistic rules to adapt LLMs to multiple dialects. They first create a synthetic dataset of linguistic transformations using LLM probing. Following that, they train a set of feature adapters to generalise across multiple dialects of interest. They present their evaluation of multiple dialects of English.

DIALECTBENCH [Faisal et al. 2024] is a large-scale benchmark covering 10 NLP tasks focusing on 281 language varieties. Their evaluation shows substantial disparities in performance between the standard and non-standard language varieties, while also identifying language clusters with large performance divergence across tasks. Most recently, the VarDial 2024 evaluation campaign [Chifu et al. 2024] released a dataset on the **choice of plausible alternatives (COPA)** task focusing on three micro-dialects, namely, Cerkno dialect of Slovenian, Chakavian dialect of Croatian, and the Torlak dialect, which is spoken across Serbia, Macedonia, and Bulgaria. This task

Table 7. Representative Examples for Sequence-to-sequence NLP tasks

Paper	Dialect	Approach
Summarisation		
[Olabisi et al. 2022]	African-American /Hispanic English	Representations-based clustering & obtain summaries separately.
[Keswani and Celis 2021]	English Dialects	Diversity-representative sentences & weigh them for summary generation
Machine translation		
[Alam et al. 2024]	891 variations	Contrastive dialect benchmark for MT evaluation
[Riley et al. 2023]	Portuguese and Chinese Dialects & English	Few-shot approach for MT
Dialogue systems		
[Zhan et al. 2023]	Dialectal norms in Chinese culture	Benchmark dataset
[Artemova et al. 2024]	German dialects	Perturbation-based evaluation for dialects
Text generation		
[Sun et al. 2023]	English Dialects	Towards dialect-robust metrics for text generation

requires a computational model to select one of two candidate statements, which is more likely to be the cause or effect of a given premise statement. Collectively, training and test datasets from Var-Dial evaluation campaigns (2014–2024) organised over the years should act as a good benchmark for LLM evaluation of dialects.

5.6 Others

Erdmann et al. [2018] investigate how word embeddings trained on dialect-specific or mixed-dialect corpora perform. In their experiments for text in dialects of Arabic, they show how dialect-specific embeddings can be helpful for dictionary induction. Dictionary induction here refers to alignment tables between dialects of a language. Demszky et al. [2021] report models that predict dialect features using minimal pairs that represent linguistic properties of dialects. They do so for Indian English.

6 Natural Language Generation (NLG)

The previous section showed that NLU for dialects has primarily focused on tasks such as identification of dialects and sentiment analysis. We now present approaches in NLG. NLG deals with **sequence-to-sequence (*seq2seq*)** tasks in NLP, which take a sequence as input and produce a sequence. Challenges in the presence of dialects in a generation task can differ significantly given the task. The data and evaluation methods can be different for tasks, especially where dialectal text is being generated. Some examples of such problems are summarisation, question answering, and machine translation, and they are described in Table 7. While the situation in the case of NLU was already dire, our survey indicates that, for NLG, it is even worse. We will now discuss NLP approaches that deal with dialects of a language in the context of *seq2seq* problems.

Two works reflect advances in the context of *seq2seq* problems:

(1) **Making evaluation metrics dialect-aware**: Sun et al. [2023] state that metrics used to measure text generation may penalise outputs in certain dialects. They propose a metric named NANO, which allows perturbations in the generated output. They show that models pretrained with NANO as the metric can be helpful for dialect-robustness.

(2) **Creating dialectal variants of datasets for benchmarking**: Ziems et al. [2023] present Multi-VALUE, a library that creates dialectal variations of datasets based on a set of manually created rules. They create variants of benchmark datasets and evaluate the variants

for several *seq2seq* tasks including machine translation, question answering, and so on. The models for evaluation are based on modern LLMs such as BERT, ROBERTA, BART, and T5. The library provides a useful resource as well as insights for dialect-aware benchmarking in the future.

6.1 Summarisation

Past work in summarisation, although limited, states that dialect labels may not be explicitly necessary. However, a review of Arabic text summarisation by Elsaid et al. [2022] state the use of "dialect period frameworks" to incorporate semantic information about dialects. In the case of multi-document summarisation, clustering of sentences in the input set is a predominant paradigm. Two such works are noteworthy:

(1) Olabisi et al. [2022] analyse the diversity of dialects in multi-document summarisation of social media posts. They present a dataset that contains summaries of a collection of tweets written in three dialects: African-American English, Hispanic English, and White English. They use extractive summarisation using LONGFORMER-EXT and abstractive summarisation using BART and T5. To bring diversity-awareness in summarisation, they create automatic clusters of input documents based on semantic attributes. They follow a two-stage approach where the summarisers are separately applied, and the resultant outputs are combined again using a summariser.

(2) Keswani and Celis [2021] examine the role of dialect diversity on multi-tweet summarisation. They use a variety of summarisers: typical traditional summarisers such as TF-IDF, TextRank and LexRank, and SummaRunner (a neural summariser that treats summarisation as a sequence classification task). They create a control set: a subset of sentences that represent different dialects in the set of sentences. They introduce a bias mitigation procedure that introduces dialect-awareness in summaries using a parameter that is weighted to increase the score of dialect-diverse sentences in the dialect set.

6.2 Machine Translation

Compared to summarisation, machine translation has been studied a bit more. Recent work is broadly divided into two categories: (i) translation between dialects of the same language and (ii) translation between the dialect of a language and another language. In the rest of this section, we cover the approaches in these categories.

6.2.1 MT between Dialects of the Same Language. The primary goal of inter-dialect translation is the dissemination of information available between a standard dialect and a non-standard one. In this context, the following works are relevant. Mapping from less-used dialects to their most common versions is called **dialect normalisation**. One such work by Kuparinen et al. [2023] provides a dialect normalisation dataset in Swiss German, Slovene, Finnish, and Norwegian. Bouamor et al. [2014] present a multi-dialectal dataset for various dialects of Arabic.
Harnessing pre-trained models: Le and Luu [2023] show that models based on $BART_{pho}$ perform well for dialect normalisation in dialects of Vietnamese. This indicates that denoising-based pre-trained models can be a good source for dialect data generation owing to their infilling capabilities. **Character-level modelling:** Abe et al. [2018] conduct Japanese dialect translation where they use NMT to translate from dialect to standard Japanese using character RNN trained on small datasets collected as a part of their work. Honnet et al. [2018] additionally suggest that normalisation is an important aspect for translating between Swiss German dialects, which is achievable via character-level models. Kuparinen et al. [2023] further show that sliding-window-based approaches are useful, since dialect translation does not need the entire sentence-level context.

Perturbation-based regularisation: Liu et al. [2022] present a *seq2seq* approach for machine translation of Singaporean English to standard English. They use word perturbation and sentence perturbation to prevent overfitting of lexical features. Maurya et al. [2023] used a similar approach for Indian dialects.

Harnessing linguistic features: Erdmann et al. [2017] focus on translation among Arabic dialects in a low-resource setting, where they supplement small parallel corpora with morpho-syntactic information injected into the model for machine translation. In general, incorporating linguistic features into the MT framework is known to significantly boost translation quality in low-resource settings [Chakrabarty et al. 2020, 2022]. Especially, pre-training by leveraging linguistic features, as done by Chakrabarty et al. [2022], should be beneficial for dialectal translation, which is typically a low-resource problem.

Code-mixed training: Lu et al. [2022] use XLM for Translation between Hokkien-Mandarin code-mixed text. They observe that continuous training with code-mixed data enables monolingual language models to provide better performance when applied to code-mixed tasks.

Data creation for MT between dialects: Zbib et al. [2012] and Meftouh et al. [2015] also focus on multi-dialect MT data collection for Arabic, which is, once again, to be noted as one of the most studied languages for dialects. Xu et al. [2015] use a Hidden Markov-based model to create word alignment between dialects of Chinese: Mainland Chinese, Hong Kong Chinese, and Taiwan Chinese. The outcome is a monolingual corpus that contains corresponding words used in the three dialects. Their approach was shown to be effective for three different alignment mapping cases. Rather than use word alignment, Hassani [2017] works on Kurdish dialectal MT using dictionaries and shows that having limited to no parallel corpora is not a significant barrier for inter-dialect translation. YORULECT [Ahia et al. 2024] releases parallel text and speech corpus for four regional variants of Yorùbá by collecting data from within native communities for each variant and providing a benchmark for machine translation, speech recognition, and text-to-speech synthesis. Most recently, Dabre et al. [2024] worked on Kadodi, a dialect of Marathi, for which the local community was involved for parallel corpora creation, further highlighting the need for community involvement.

All these works emphasise that a small amount of parallel data between dialects is always important; however, data synthesis and transfer learning from a high-resource dialect is always impactful, especially in conjunction with character- and word-level perturbation methods.

6.2.2 MT between Dialects and Another Language. The second category, involving the harder challenge of machine translation between a dialect and another language, has received far more attention. We cover notable works below.

Dialect pivoting: An early work in this regard is by Paul et al. [2011]. They present a pivot-based MT approach for the translation of four dialects of Japanese, namely, Kumamoto, Kyoto, Okinawa, and Osaka. To map sentences across dialects, they use a character-based generative graphical model. They then translate the dialects into four Indo-European languages, using standardised Japanese as the pivot language. Jeblee et al. [2014] focus on using modern standard Arabic as a pivot when translating from English to the Egyptian Arabic dialect.

Unsupervised segmentation: Different from Abe et al. [2011] who focus on characters, Al-Mannai et al. [2014]; Salloum and Habash [2022] work on Arabic dialectal translation, which shows that unsupervised word segmentation is just as effective if not better for translation into English.

Evaluating existing translators on dialectal datasets: Kantharuban et al. [2023] show the performance of MT between English and dialects of seven languages. Using state-of-the-art MT systems such as Google NMT and Meta NLLB, they evaluate MT in both directions (to and from English). They report a drastic drop in BLEU for dialects of German, Portuguese, and Bengali.

De Camillis et al. [2023] train NMT systems to evaluate the performance of legal domain translation for Italian ⇔ South Tyrolean German, where their models show better performance compared to Google Translate and DeepL for this niche use case. Similarly, CODET [Alam et al. 2024] is a contrastive dialectal benchmark dataset for evaluating machine translation systems focusing on 891 variations from 12 languages.

Using inferred dialectal labels to guide translation: Sun et al. [2023] add language-dialect information as predicted by a language identifier as an input when training an MT system. They further improve the metrics for robust evaluation of text generation systems for different languages and dialects. They report their results on several dialects of English, Chinese, Portuguese, and so on. They use few-shot prompting to create semantic perturbations to train T5. The results show that dialect-awareness improves the performance of translation. Shapiro and Duh [2019] explore a multidialect system and identify when dialectal identification is useful. Tahssin et al. [2020] focus on dialect identification itself using AraBERT models. Salloum et al. [2014] focus on sentence-level dialect identification for an SMT model selection where the model is optimised for that dialect.

Learning through exemplars: Few-shot learning involves the use of examples in a prompt to guide the generation through a language model. Riley et al. [2023] present a few-shot machine translation approach for translation between English and two variants of Portuguese and Chinese. The parallel corpus is manually created by native speakers. The exemplars from the specific dialect are used to obtain translations of English sentences.

User-generated content: **User-generated content (UGC)** is often mistranslated on social media, especially for low-resource languages like dialectal Arabic. Saadany et al. [2022] train a Transformer to translate from dialectal Arabic to English, where they focus on challenges in translation of UGC and propose a sentiment-aware evaluation metric for translation. They discuss results on multiple test sets, including a hand-crafted test set, and analyse the performance for a semi-supervised approach compared to a baseline NMT system, a pivoting-based system, and Google Translate.

Use of multiple translation models: Translation models that translate between the standard version and a dialect can assist machine translation. Kumar et al. [2021] show an approach for MT from English to Ukrainian, Belarusian, Nynorsk, and Arabic dialects. They use two models: a dialect-to-standard translation model and a standard source-to-target language translation model.

Data creation for MT of dialect to another language: Hassan et al. [2017] explore synthetic data creation using word pairs between dialects based on embeddings. They take seed data, transform it into its dialectal variant, and now have a dialectal parallel corpus. Similarly, Almansor and Al-Ani [2017] focus on using monolingual data and tiny parallel corpora in conjunction with cross-dialectal embeddings to improve MT between dialects. Sajjad et al. [2020] take dialect MT evaluation further and focus on multi-domain coarse-grained analysis of dialects of Arabic via their AraBench benchmark. Hamed et al. [2022] propose an Arabic-English code-switched speech translation dataset, which represents a practical use case, since a vast majority of dialects are often spoken. They state that there is a significant dearth of code-mixed datasets and recommend researchers to focus on the same. Alkheder et al. [2023], recognising the increasing usage of Arabic in several regions of Turkey, expand the MADAR corpus [Obeid et al. 2018] to enable benchmarking of translation between Arabic and Turkish. Contarino [2021] curates LEXB, a parallel corpus between South Tyrolean German and Italian containing nearly 175, 000 parallel segments from the legal domain. To curate parallel data, they use the LexBrowser database[5] and 20 national laws and codes (Civil Code, Criminal Code) translated into German. Igarashi and Miyagawa [2024] curate parallel corpus for Ainu⟺Japanese where text for Ainu was curated from after post-processing

---

[5]http://lexbrowser.provinz.bz.it/

OCR ouput for books and other online resources. Ainu is a critically endangered family of dialects from Northern Japan without any native writing script.

6.2.3 Dialect MT in Shared Tasks. Given that most dialects are spoken and not written, the IWSLT workshop, which focuses on spoken language translation, has been conducting shared tasks on dialects under the banner of low-resource MT. The 2022[6] and 2023[7] workshops featured dialectal speech translation, with resources for text-text as well as speech-text translation. The focus, as is typically the case, is on dialects of Arabic such as Tunisian, Egyptian, and Moroccan. The shared tasks are an excellent source of datasets and benchmarks for dialectal MT. Most recently, the ArabicNLP 2023[8] conference offered a shared task[9] on translation from four Arabic dialects to modern standard Arabic. Over the years, NADI has reported progress on country- and province-level dialect identification, dialectal sentiment analysis, and dialect to MSA machine translation [Abdul-Mageed et al. 2023, 2020, 2021, 2022]. We should also note that the Workshop on Machine Translation (WMT)[10] and the Workshop on Asian Translation (WAT)[11] often feature shared tasks on closely related languages. The 2024[12] edition of IWSLT is expected to focus on North Levantine Arabic.

6.3 Dialogue Systems

Dialogue systems, crucial in facilitating human-computer interaction, are categorised into task-oriented, chit-chat, and hybrid systems. These systems, especially when dialect-aware, face the added challenge of understanding and adapting to linguistic variations.

Task-oriented dialogue system: Task-oriented systems are designed to accomplish specific tasks. They integrate NLU, a dialogue manager, and NLG components. The effectiveness of these systems in handling dialects is pivotal. For instance, Elmadany et al. [2018a]; Joukhadar et al. [2019] study the classification of dialogue acts in Arabic dialect utterances, demonstrating the system's capacity to adapt to dialectal variations. Al-Ghadhban and Al-Twairesh [2020] use the **Artificial Intelligence Markup Language (AIML)** to build a chatbot that assists students with academic enquiries in the Saudi Arabian dialect. The VarDial 2023 campaign [Aepli et al. 2023] reports progress on *slot and intent detection for low-resource language varieties* such as Swiss German, Neapolitan, and South Tyrolean. Artemova et al. [2024] investigate the robustness of task-oriented dialogue systems—specifically, their intent classification and slot detection components—to German dialects by applying perturbations that transform standard German sentences into colloquial variants.

Chit-chat dialogue system: Chit-chat dialogue systems, also known as open-domain systems, primarily focus on daily chat and handle broader interactions. Ali and Habash [2016] employ AIML and rule-based systems to manage dialectal variation in Egyptian Arabic, incorporating features such as short vowels and consonantal doubling. Ahmed and Hussein [2020] also use AIML for Kurdish dialogues. Additionally, Alshareef and Siddiqui [2020] train a Seq2Seq model on a tweet corpus to respond to open-domain Arabic questions.

A specialised subset of chit-chat systems are socially aware dialogue systems, which pay close attention to the influence of social norms and factors. These systems are designed to adhere to the cultural and social norms prevalent in different societies [Hovy and Yang 2021]. In different

---

[6]https://iwslt.org/2022/dialect
[7]https://iwslt.org/2023/low-resource
[8]https://arabicnlp2023.sigarab.org/
[9]https://nadi.dlnlp.ai/
[10]https://www2.statmt.org/wmt24
[11]https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2024/index.html
[12]https://iwslt.org/2024/low-resource

cultures, social norms will no doubt incorporate sociolect, including whatever discourse force it may carry. Ziems et al. [2023] propose a framework to evaluate dialect differences in cross-dialectal English. In addition, Zhan et al. [2024, 2023] propose the socially aware dialogue corpus based on Chinese culture and relevant dialectal norms. Sociolect in a dialogue will dramatically affect human's understanding and behaviours towards speakers. Rajai and Ennasser [2022] summarise existing problems and strategies towards dealing with sociolect in dialogues.

Hybrid system: Hybrid systems combine features of both task-oriented and open-domain systems. An example is the system developed by Ben Elhaj Mabrouk et al. [2021], which answers user queries in various Arabic dialects such as Tunisian, Igbo, Yoruba, and Hausa. This chatbot addresses both official FAQs, especially related to COVID-19, and informal chit-chat, responding to questions in the local dialect.

Awareness of social and societal norms of behaviour is particularly important in dialogue systems that serve specific transactional goals, be it to book a doctor's appointment, to ask questions about income tax, or to make a customer service complaint. Research in interactional sociolinguistics [Gumperz 1982] has—in a rich body of research in different social contexts such as employment interviews [Roberts 2021]—shown that people interpret communicative intent against their own background expectations of what is "normal" or "expected behaviour." This has the potential to exacerbate inequality (for example, by restricting access to employment), in particular, for underrepresented groups such as migrants.

In the case of dialogue systems, a lack of representation of different dialects (e.g., due to the lack of diverse training data) has the potential to cause similar effects: If dialogue systems are not aware of social norms inherent to different dialects, and if what is communicated does not match users' expectations, then communicative intent can be misinterpreted, and underrepresented user groups might become disengaged from the system. The xSID dataset [Van Der Goot et al. 2021] is a multilingual dataset for spoken language understanding and includes a low-resource Austro-Bavarian German dialect and other dialects. Indeed, previous research on dialogue systems has confirmed the importance of alignment of system-style choices with user needs and preferences [Chaves et al. 2022; Følstad and Brandtzaeg 2020; Li and Mao 2015].

7 Conclusion & Future Directions

Dialects are syntactic and lexical variations of a language, often associated with socially or geographically cohesive groups. This article summarises NLP approaches for dialects of several languages. The need for NLP approaches focusing on dialects of a language rest on four motivations: dialects pose linguistic challenges, benchmarks may not have sufficient dialectal representation, dialect-awareness is important for fair NLP technology, and there has been growing recent work in this direction. The survey identified trends in terms of tasks (which shows shifting focus from dialect classification), languages (with more work in Arabic as compared to other languages), and a shifting trend towards mitigation (by either making models dialect-invariant or dialect-aware). Following that, we described different methods to create dialectal lexicons and datasets, ranging from location/keyword-based filtering (of which location-based filtering has been found to be ineffective by Goutte et al. [2016]) to manual (via recruitment of native speakers) and automatic (via automatic perturbation). We then viewed past work in the context of NLU and NLG. For NLU, we covered dialect identification, sentiment analysis, morphosyntactic analysis, parsing, and more recent work in NLU benchmarks. We described how the availability of datasets in multiple languages has fuelled research in dialect identification, which continues to date. Sentiment analysis techniques for dialects included peculiar de-biasing approaches, in addition to dialect invariance and dialect awareness. Approaches to parse dialectal datasets used or adapted existing parsers or developed dialect-specific parsers. Finally, we described how

recent work on NLU benchmarks highlights how adversarial learning and LoRA can be used to reduce the degradation in the performance of dialectal datasets as compared to the standard ones. In the case of NLG, we described work in summarisation, machine translation, and dialogue systems. We described the limited, recent work in multi-document summarisation of dialectal documents. Following that, we discussed approaches for machine translation in the context of dialect normalisation and dialect pivoting, depending on whether the translation is between dialects of a language or between dialects and another language. Finally, we described dialogue systems in the context of task-oriented, chit-chat, and hybrid systems.

Based on our survey, we now identify future directions and social/ethical implications. We hope that the former will be helpful for NLP research for dialects, while the latter will get more researchers interested in this richly investigated yet emergent area of NLP. We believe that NLP researchers should adopt a socio-technical perspective [Johnson and Verdicchio 2017] on their role and consider not only their own possible biases influencing the selection of training data, the design of algorithms, and so on, but also other social arrangements (e.g., users and their behaviours) relevant to specific systems. In their survey of speakers of German dialects, Blaschke et al. [2024] also discuss in detail their needs as users of language technologies.

7.1 Implications to NLP Research

In addition to the trends reported earlier in the article, the following would be potential future directions in the context of NLP:

7.1.1 Focusing on Unexplored Dialects of Languages. NLP for dialects faces problems akin to low-resource languages, in terms of the availability of existing resources and tools. While some dialect families, such as English and Arabic, have seen consistent efforts, dialects for other languages need more focused large-scale efforts for data curation and annotation. While English is arguably the leading language for advances in NLP, efforts remain to be done to fully represent the full diversity of the English language itself through appropriate datasets and models that are curated for specific dialectal tasks. It is not always necessary to create new datasets, given that datasets specific to particular dialects are available. However, caution is advised for dialogue systems, as many existing corpora—with the exception of those focusing on **English as a *lingua franca* (ELF)**—are dominated by written texts, which may not represent the richness of dialectal variations of spoken language.

7.1.2 Rethinking the Pre-training of LLMs. Chow and Bond [2022] present a computational grammar for Singaporean English. Such dialectal representations can be useful to generalise the ability of LLMs. It would be beneficial for LLMs to be able to ingest other kinds of information such as dialect-specific grammatical structures. Ability to pre-train LLMs using data in different formats (not just modality, which is currently a popular paradigm) may improve their performance for diverse datasets such as dialects. Similar impact of pre-training language data distribution is known for cross-lingual meaning transfer within NLP tasks [Qian et al. 2024; Zerva et al. 2024] involving low-resource languages [Nigatu et al. 2024].

7.1.3 Dialect Identification as an Auxiliary Task. Multi-task learning is used to train models for multiple tasks. Dialect identification could be used as one of the tasks to train equitable models. Lent et al. [2024] present a multi-task, multi-lingual dataset of Creole languages. They report the baseline performance of NLU and MT tasks on the dataset using appropriate models. Availability of such large benchmarks will aid the development of new methods and models.

7.1.4 Rethinking LLM Evaluation. Xiao et al. [2023] evaluate LoRA adapters for unseen English dialects, and they say: "a comprehensive examination of PEFT modules for dialects is needed,

which we leave for future work." Similar evaluations can be performed for other NLP approaches. In addition, new evaluation techniques and metrics will be useful to measure dialectal variation and its potential correlation with the performance of NLP tasks. Two recent papers can be of value. Lameli and Schönberg [2023] present a measure for spatial language variation. Using distance between locations as a heuristic for dialectal similarity, they examine variations in dialects of German. Also, Keleg et al. [2023] use a dataset in Arabic labeled with the degree of dialectness to train a BERT-based regression model.

7.2 Ethical & Social Implications

Overall, dialectal NLP presents an excellent avenue for research with huge social implications. We highlight three considerations of relevance.

7.2.1 Social Implications. While everyone speaks a standard dialect, most people tend to feel familiarity with people who speak specific dialects. Furthermore, certain traditions and practices are tied to localities, which are, in turn, tied to dialects. If the goal of NLP research is to make communication seamless, then the correct way to do so is via a strong emphasis on dialects by effectively engaging with speakers of the dialects [Blaschke et al. 2024]. Most dialects around the world are under-represented in modern-day NLP, which can potentially disadvantage them or leave them out of the benefits of LLMs. There is also a growing concern among speakers of specific dialects that their language is dying either due to the pervasiveness of English via the internet, another majority language, or a related dialect, which has higher official support or recognition. We should acknowledge these concerns and make headway into preserving as many dialects as possible, at the risk of losing valuable aspects of the vast tapestry of culture and history.

7.2.2 Dialectal Research by Dialect Speakers. Linguistic research colonisation is the process where researchers who do not speak specific languages nor have connections with them conduct research on said languages. Despite the negative connotation of colonisation, this is not a bad thing, because no one should monopolise working on specific languages. However, it highlights that there are haves and have-nots, where the haves are researchers and organisations with funding who can work on dialects, and the have-nots are the researchers who would like to work on dialects but simply lack funding. Recently, there has been a growing trend where language speakers are reclaiming dominion over research involving their own languages. For example, there has been an explosive growth in the number of researchers and groups such as DeepLearning Indaba, Masakhane from African countries working solely on African languages, and organisations like AI4Bharat in India working on Indian languages. Indeed, they have shown that a dedicated focus on language research by speakers of these languages leads to better NLP systems. We, therefore, propose that the organisations with funding leverage their privilege and support those without funding to ensure that work on dialects is led and owned by groups that are most connected to and impacted by dialects. This will lead to true diversity, equality, and inclusivity in NLP research, which will strongly impact society. Towards this, the emerging sentiment in recent thematic papers in NLP is that communities that speak the dialects must be involved in the development of language technologies for the communities [Bird 2022; Ramponi 2024].

7.2.3 Normalising Working on and Speaking Dialects. One aspect that limits dialectal research is the concept of shame in speaking a certain language or a dialect, an aspect also known as linguistic self-hatred. For example, take the case of Mauritian Creole, whose speakers are dwindling by the day, mainly because the younger generation feels shame in speaking their native language. The same exists for Konkani. While there are no official reports highlighting the same for dialects, it is not far-fetched to consider that linguistic self-hatred will exist here as well. It is time to end this

self-hatred and normalise speaking dialects. By doing so, people speaking dialects will become more enthusiastic about preserving their dialects, and this will inevitably aid research on dialects, thereby positively impacting society. Dialects are closely tied to culture, and such differences have not been captured explicitly beyond the works described in this article.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *LREC*.

Muhammad Abdul-Mageed and Mona T. Diab. 2014. SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *LREC*. 1162–1169.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *ArabicNLP*. Association for Computational Linguistics, 600–613. DOI:https://doi.org/10.18653/v1/2023.arabicnlp-1.62

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *5th Arabic Natural Language Processing Workshop (WANLP'20)*. ACL, 97–110. Retrieved from https://aclanthology.org/2020.wanlp-1.9

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *6th Arabic Natural Language Processing Workshop (WANLP'21)*. ACL, 244–259. Retrieved from https://aclanthology.org/2021.wanlp-1.28

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *7th Arabic Natural Language Processing Workshop (WANLP'22)*. Association for Computational Linguistics, 85–97. DOI:https://doi.org/10.18653/v1/2022.wanlp-1.9

Dana Abdulrahim, Go Inoue, Latifa Shamsan, Salam Khalifa, and Nizar Habash. 2022. The Bahrain corpus: A multi-genre corpus of Bahraini Arabic. In *International Conference on Language Resources and Evaluation (LREC'22)*. 2345–2352.

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Pacific Asia Conference on Language, Information and Computation (PACLIC'18)*. ACL. Retrieved from https://aclanthology.org/Y18-1001

Yusuke Abe, Takafumi Suzuki, Bing Liang, Takehito Utsuro, Mikio Yamamoto, Suguru Matsuyoshi, and Yasuhide Kawada. 2011. Example-based translation of japanese functional expressions utilizing semantic equivalence classes. In *4th Workshop on Patent Translation*. Retrieved from https://aclanthology.org/2011.mtsummit-wpt.10

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *9th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'22)*. Association for Computational Linguistics, 1–13.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'23)*. Association for Computational Linguistics, 251–261.

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of ACL*. 4074–4083. DOI:https://doi.org/10.18653/v1/2022.findings-acl.321

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*. Association for Computational Linguistics, 4392–4409. Retrieved from https://aclanthology.org/2024.emnlp-main.251

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? Tokenization in the era of commercial language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'23)*. 9904–9923.

Hemn Karim Ahmed and Jamal Ali Hussein. 2020. Design and implementation of a chatbot for Kurdish language speakers using Chatfuel platform. *Kurdist. J. Appl. Res.* (2020), 117–135. DOI:https://doi.org/10.24017/science.2020.2.10

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Annual Meeting of the Association for Computational Linguistics (ACL'22)*. 7226–7249. DOI:https://doi.org/10.18653/v1/2022.acl-long.500

Dana Al-Ghadhban and Nora Al-Twairesh. 2020. Nabiha: An Arabic dialect chatbot. *Int. J. Advan. Computer Science and Applications* 11, 3 (2020).

Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov, and Stephan Vogel. 2014. Unsupervised word segmentation improves dialectal Arabic to English machine translation. In *Workshop on Arabic Natural Language Processing (WANLP'14)*. ACL, 207–216. DOI:https://doi.org/10.3115/v1/W14-3628

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of EACL*. 1790–1859. Retrieved from https://aclanthology.org/2024.findings-eacl.125

Dana Abu Ali and Nizar Habash. 2016. Botta: An Arabic dialect chatbot. In *COLING (System Demonstrations)*. 208–212.

Hasan Alkheder, Houda Bouamor, Nizar Habash, and Ahmet Zengin. 2023. Benchmarking dialectal Arabic-Turkish machine translation. In *Machine Translation Summit XIX*. Asia-Pacific Association for Machine Translation, 261–271. Retrieved from https://aclanthology.org/2023.mtsummit-research.22

Ebtesam H. Almansor and Ahmed Al-Ani. 2017. Translating dialectal Arabic as low resource language using word embedding. In *Conference on Recent Advances in Natural Language Processing (RANLP')*. INCOMA Ltd., 52–57. DOI: https://doi.org/10.26615/978-954-452-049-6_008

Tahani Alshareef and Muazzam Ahmed Siddiqui. 2020. A seq2seq neural network based conversational agent for Gulf Arabic dialect. In *21st International Arab Conference on Information Technology (ACIT'20)*. IEEE, 1–7.

Areej Alshutayri and Eric Atwell. 2018. Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers. In *3rd Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT'18)*. 54.

Ekaterina Artemova, Verena Blaschke, and Barbara Plank. 2024. Exploring the robustness of task-oriented dialogue systems for colloquial German varieties. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL'24)*. 445–468.

Katya Artemova and Barbara Plank. 2023. Low-resource bilingual dialect lexicon induction with large language models. In *24th Nordic Conference on Computational Linguistics (NoDaLiDa'23)*. University of Tartu Library, 371–385. Retrieved from https://aclanthology.org/2023.nodalida-1.39

Adel Assiri, Ahmed Emam, and Hmood Al-Dossari. 2018. Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. *J. Inf. Sci.* 44, 2 (2018), 184–202.

Faical Azouaou and Imane Guellil. 2017. ALG/FR: A step by step construction of a lexicon between Algerian dialect and French. In *Pacific Asia Conference on Language, Information and Computation (PACLIC'17)*.

Niyati Bafna, Cristina España-Bonet, Josef Van Genabith, Benoît Sagot, and Rachel Bawden. 2023. Cross-lingual strategies for low-resource language modeling: A study on five Indic dialects. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : Travaux De Recherche Originaux – Articles Longs*. ATALA, Paris, France, 28–42.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *International Conference on Language Resources and Evaluation (LREC'22)*. 4586–4596.

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. 116–128.

Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2019. ArSenTD-LEV: A multi-topic corpus for target-based sentiment analysis in Arabic Levantine tweets. *arXiv preprint arXiv:1906.01830* (2019).

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *23rd Nordic Conference on Computational Linguistics (NoDaLiDa'21)*. Linköping University Electronic Press, 445–451. Retrieved from https://aclanthology.org/2021.nodalida-main.51

Aymen Ben Elhaj Mabrouk, Moez Ben Haj Hmida, Chayma Fourati, Hatem Haddad, and Abir Messaoudi. 2021. A multilingual African embedding for FAQ chatbots. *arXiv e-prints* (2021), arXiv–2103.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Annual Meeting of the Association for Computational Linguistics (ACL'22)*. 7817–7829. DOI: https://doi.org/10.18653/v1/2022.acl-long.539

Verena Blaschke, Christoph Purschke, Hinrich Schütze, and Barbara Plank. 2024. What do dialect speakers want? A survey of attitudes towards language technology for German dialects. *arXiv preprint arXiv:2402.11968* (2024).

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'23)*. 40–54. DOI: https://doi.org/10.18653/v1/2023.vardial-1.5

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050* (2020).

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*. 1119–1130. DOI: https://doi.org/10.18653/v1/D16-1120

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter universal dependency parsing for African-American and mainstream American English. In *Annual Meeting of the Association for Computational Linguistics (ACL'18)*. 1415–1425. DOI: https://doi.org/10.18653/v1/P18-1131

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann et al. 2018. The MADAR Arabic dialect corpus and lexicon. In *International Conference on Language Resources and Evaluation (LREC'18)*.

Rahma Boujelbane, Mariem Ellouze Khemekhem, Siwar BenAyed, and Lamia Hadrich Belguith. 2013. Building bilingual lexicon to create dialect Tunisian corpora and adapt language model. In *2nd Workshop on Hybrid Approaches to Translation*. 88–93. Retrieved from https://aclanthology.org/W13-2813

El Mehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. 2021. An open access NLP dataset for Arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000* (2021).

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for Odawa. In *Workshop on the Use of Computational Methods in the Study of Endangered Languages*. 1–9. DOI: https://doi.org/10.18653/v1/W17-0101

Manuel Burghardt, Daniel Granvogl, and Christian Wolff. 2016. Creating a lexicon of Bavarian dialect by means of Facebook language data and crowdsourcing. In *International Conference on Language Resources and Evaluation (LREC'16)*. 2029–2033. Retrieved from https://aclanthology.org/L16-1321

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Hideki Tanaka, Masao Utiyama, and Eiichiro Sumita. 2022. FeatureBART: Feature based sequence-to-sequence pre-training for low-resource NMT. In *International Conference on Computational Linguistics (COLING'22)*.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2020. Improving low-resource NMT through relevance based linguistic features incorporation. In *International Conference on Computational Linguistics (COLING'20)*. International Committee on Computational Linguistics, 4263–4274. DOI: https://doi.org/10.18653/v1/2020.coling-main.376

Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots. *ACM Trans. Comput.-Hum. Interact.* 29, 2 (2022), 1–38.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*. 369–376.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'24)*. Association for Computational Linguistics, 1–15.

Rahul Chitturi and John Hansen. 2008. Dialect classification for online podcasts fusing acoustic and language based structural and semantic information. In *Annual Meeting of the Association for Computational Linguistics (ACL'08)*. 21–24. Retrieved from https://aclanthology.org/P08-2006

Siew Yeng Chow and Francis Bond. 2022. Singlish where got rules one? Constructing a computational grammar for Singlish. In *International Conference on Language Resources and Evaluation (LREC'22)*. 5243–5250.

Steven Coats. 2022. The corpus of Australian and New Zealand spoken English: A new resource of naturalistic speech transcripts. In *Australasian Language Technology Association Workshop*. 1–5. Retrieved from https://aclanthology.org/2022.alta-1.1

Steven Coats. 2023. Double modals in contemporary British and Irish speech. *Eng. Lang. Ling.* 27, 4 (2023), 693–718.

Antonio Contarino. 2021. Neural machine translation adaptation and automatic terminology evaluation: A case study on Italian and South Tyrolean German legal texts. https://amslaurea.unibo.it/id/eprint/24989/1/Contarino_dissertation.pdf

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *International Conference on Language Resources and Evaluation (LREC'14)*. 241–245.

Felicity Cox. 2006. The acoustic characteristics of/hVd/vowels in the speech of some australian teenagers. *Austral. J. Ling.* 26, 2 (2006), 147–179.

Felicity Cox and Sallyanne Palethorpe. 2007. Australian English. *J. Int. Phonet. Assoc.* 37, 3 (2007), 341–350.

Marcelo Criscuolo and Sandra Aluisio. 2017. Discriminating between similar languages with word-level convolutional neural networks. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'17)*. 124–130.

Raj Dabre, Mary Dabre, and Teresa Pereira. 2024. Machine translation of Marathi dialects: A case study of Kadodi. In *11th Workshop on Asian Translation (WAT'24)*. Association for Computational Linguistics, 36–44. Retrieved from https://aclanthology.org/2024.wat-1.3

Jamell Dacon, Haochen Liu, and Jiliang Tang. 2022. Evaluating and mitigating inherent linguistic bias of African American English through inference. In *International Conference on Computational Linguistics (COLING'22)*. 1442–1454. Retrieved from https://aclanthology.org/2022.coling-1.124

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj et al. 2021. A panoramic survey of natural language processing in the Arab world. *Commun. ACM* 64, 4 (2021), 72–81.

Kareem Darwish, Hamdy Mubarak, Mohamed Eldesouki, Ahmed Abdelali, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA), 93–98.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably effective Arabic dialect identification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1465–1468. DOI : https://doi.org/10.3115/v1/D14-1154

Flavia De Camillis, Egon Waldemar Stemle, Elena Chiocchetti, and Francesco Fernicola. 2023. The MT@ BZ corpus: Machine translation & legal language. In *Annual Conference of the European Association for Machine Translation (EAMT'23)*. 171–180.

Dorottya Demszky, Devyani Sharma, Jonathan H. Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *NAACL*. 2315–2338.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *Workshop on Semitic Language Processing*. 66–74.

Shahd Dibas, Christian Khairallah, Nizar Habash, Omar Fayez Sadi, Tariq Sairafy, Karmel Sarabta, and Abrar Ardah. 2022. Maknuune: A large open Palestinian Arabic lexicon. In *Arabic Natural Language Processing Workshop (WANLP'22)*. Association for Computational Linguistics, 131–141.

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-dialect Vietnamese: Task, dataset, baseline models and challenges. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*. Association for Computational Linguistics, 7476–7498. Retrieved from https://aclanthology.org/2024.emnlp-main.426

A. Seza Doğruöz and Preslav Nakov. 2014. Predicting dialect variation in immigrant contexts using light verb constructions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1391–1395. DOI : https://doi.org/10.3115/v1/D14-1145

Longxu Dou, Yan Gao, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, and Jian-Guang Lou. 2023. Multi-Spider: Towards benchmarking multilingual text-to-SQL semantic parsing. In *AAAI Conference on Artificial Intelligence (AAAI'23)*. 12745–12753.

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Do we name the languages we study? The# benderrule in LREC and ACL articles. In *International Conference on Language Resources and Evaluation (LREC'22)*. 564–573.

Jonathan Dunn. 2019. Modeling global syntactic variation in English using dialect classification. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'19)*. Association for Computational Linguistics, 42–53. DOI : https://doi.org/10.18653/v1/W19-1405

Jonathan Dunn and Ben Adams. 2020. Geographically-balanced gigaword corpora for 50 language varieties. In *International Conference on Language Resources and Evaluation (LREC'20)*. European Language Resources Association, 2528–2536. Retrieved from https://aclanthology.org/2020.lrec-1.308

Chloe Eggleston and Brendan O'Connor. 2022. Cross-dialect social media dependency parsing for social scientific entity attribute analysis. In *Workshop on Noisy User-generated Text (W-NUT'22)*. Association for Computational Linguistics, 38–50. Retrieved from https://aclanthology.org/2022.wnut-1.4

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. MD3: The multi-dialect dataset of dialogues. In *INTERSPEECH Conference*. 4059–4063. DOI : https://doi.org/10.21437/Interspeech.2023-2150

Abdellah El Mekki, Abdelkader El Mahdaouy, Ismail Berrada, and Ahmed Khoumsi. 2021. Domain adaptation for Arabic cross-domain and cross-dialect sentiment analysis from contextualized word embedding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'21)*. 2824–2837. DOI : https://doi.org/10.18653/v1/2021.naacl-main.226

Heba Elfardy and Mona T. Diab. 2012. Simplified guidelines for the creation of large scale dialectal Arabic annotations. In *International Conference on Language Resources and Evaluation (LREC'12)*. 371–378.

AbdelRahim Elmadany, Sherif Abdou, and Mervat Gheith. 2018a. Improving dialogue act classification for spontaneous Arabic speech and instant messages at utterance level. In *International Conference on Language Resources and Evaluation (LREC'18)*.

A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018b. ArSAS: An Arabic speech-act and sentiment corpus of tweets. *Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT'18)*. 20.

Ashraf Elnagar, Sane M. Yagi, Ali Bou Nassif, Ismail Shahin, and Said A. Salloum. 2021. Systematic literature review of dialectal Arabic: Identification and detection. *IEEE Access* 9 (2021), 31010–31042.

Asmaa Elsaid, Ammar Mohammed, Lamiaa Fattouh Ibrahim, and Mohammed M. Sakre. 2022. A comprehensive review of Arabic text summarization. *IEEE Access* 10 (2022), 38012–38030.

Alexander Erdmann, Nizar Habash, Dima Taji, and Houda Bouamor. 2017. Low resourced machine translation via morpho-syntactic modeling: The case of dialectal Arabic. In *Machine Translation Summit XVI: Research Track*. 185–200. Retrieved from https://aclanthology.org/2017.mtsummit-papers.15

Alexander Erdmann, Nasser Zalmout, and Nizar Habash. 2018. Addressing noise in multidialectal word embeddings. In *Annual Meeting of the Association for Computational Linguistics (ACL'18)*. 558–565.

Ramy Eskander, Nizar Habash, Owen Rambow, and Arfath Pasha. 2016. Creating resources for dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In *International Conference on Computational Linguistics (COLING'16)*. 3455–3465.

Dominique Estival, Steve Cassidy, Felicity Cox, and Denis Burnham. 2014. AusTalk: an audio-visual corpus of Australian English. In *9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 3105–3109. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/520_Paper.pdf

Ahmed Fadhil and Ahmed AbuRa'ed. 2019. OlloBot-towards a text-based Arabic health conversational agent: Evaluation and results. In *Conference on Recent Advances in Natural Language Processing (RANLP'19)*. 295–303.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anasta-sopoulos. 2024. DIALECTBENCH: A NLP benchmark for dialects, varieties, and closely-related languages. In *Annual Meeting of the Association for Computational Linguistics (ACL'24)*.

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of Arabic dialect familiarity on data annotation. In *Arabic Natural Language Processing Workshop (WANLP'22)*. 399–408.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in chatGPT: Language models reinforce dialect discrimination. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*. Association for Computational Linguistics, 13541–13564. Retrieved from https://aclanthology.org/2024.emnlp-main.750

Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users' experiences with chatbots: Findings from a questionnaire study. *Qual. User Exper.* 5, 1 (2020), 3.

Ahlam Fuad and Maha Al-Yahya. 2022. AraConv: Developing an Arabic task-oriented dialogue system using multi-lingual transformer model mT5. *Appl. Sci.* 12, 4 (2022), 1881.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, et al. 2020. A report on the VarDial evaluation campaign 2020. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'20)*. International Committee on Computational Linguistics (ICCL), 1–14. Retrieved from https://aclanthology.org/2020.vardial-1.1

Hans Goebl. 1993. Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data. In *Contributions to Quantitative Linguistics: First International Conference on Quantitative Linguistics, QUAL-ICO, Trier, 1991*. Springer, 277–315.

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. Unsu-pervised deep language and dialect identification for short texts. In *International Conference on Computational Linguistics (COLING'20)*. 1606–1617.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *International Conference on Language Resources and Evaluation (LREC'16)*.

Imane Guellil, Faical Azouaou, Fodil Benali, and Hachani Ala-Eddine. 2021. ONE: Toward ONE model, ONE algorithm, ONE corpus dedicated to sentiment analysis of Arabic/Arabizi and its dialects. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 236–249.

John J. Gumperz. 1982. *Discourse Strategies*. Number 1. Cambridge University Press.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Annual Meeting of the Association for Computational Linguistics (ACL'06)*. 681–688.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *7th Arabic Natural Language Processing Workshop (WANLP'22)*. Association for Computational Linguistics, 119–130. DOI : https://doi.org/10.18653/v1/2022.wanlp-1.12

Abualsoud Hanani and Rabee Naser. 2020. Spoken Arabic dialect recognition using x-vectors. *Nat. Lang. Eng.* 26, 6 (2020), 691–700.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2018. Maghrebi Arabic dialect processing: An overview. *J. Int. Sci. Gen. Applic.* 1 (2018).

Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT'22)*. 789–798.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of EMNLP*. Association for Computational Linguistics, 15166–15184. Retrieved from https://aclanthology.org/2024.findings-emnlp.890

Hany Hassan, Mostafa Elaraby, and Ahmed Y. Tawfik. 2017. Synthetic data for neural machine translation of spoken-dialects. In *14th International Conference on Spoken Language Translation*. International Workshop on Spoken Language Translation, 82–89. Retrieved from https://aclanthology.org/2017.iwslt-1.12

Hossein Hassani. 2017. Kurdish interdialect machine translation. In *4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'17)*. ACL, 63–72. DOI: https://doi.org/10.18653/v1/W17-1208

Einar Haugen. 1966. Dialect, language, nation. *Am. Anthrop.* 68, 4 (1966), 922–935.

Michael Haugh and Klaus P. Schneider. 2012. Im/politeness across Englishes. *Journal of Pragmatics.* 1017–1021. https://www.sciencedirect.com/science/article/abs/pii/S0378216612001361

William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task agnostic dialect adapters for English. In *Findings of ACL.* 813–824. DOI: https://doi.org/10.18653/v1/2023.findings-acl.51

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742* (2024).

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1597

Renkui Hou and Chu-Ren Huang. 2020. Classification of regional and genre varieties of Chinese: A correspondence analysis approach based on comparable balanced corpora. *Nat. Lang. Eng.* 26, 6 (2020), 613–640.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. ACL, 4383–4394. DOI: https://doi.org/10.18653/v1/D18-1469

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 588–602.

Fatemah Husain, Hana Al-Ostad, and Halima Omar. 2022. A weak supervised transfer learning approach for sentiment analysis to the Kuwaiti dialect. In *The 7th Arabic Natural Language Processing Workshop (WANLP'22)*. 161–173.

Ryo Igarashi and So Miyagawa. 2024. Enhancing neural machine translation for Ainu-Japanese: A comprehensive study on the impact of domain and dialect integration. In *4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH'24)*. Association for Computational Linguistics, 413–422. Retrieved from https://aclanthology.org/2024.nlp4dh-1.40

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *6th Arabic Natural Language Processing Workshop (WANLP'21)*. Association for Computational Linguistics, 92–104.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of ACL.* 1708–1719. DOI: https://doi.org/10.18653/v1/2022.findings-acl.135

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the Palestinian Arabic dialect. *Lang. Resour. Eval.* 51 (2017), 745–775.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *J. Artif. Intell. Res.* 65 (2019), 675–782.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into Egyptian Arabic. In *Workshop on Arabic Natural Language Processing (WANLP'14)*. Association for Computational Linguistics, 196–206. DOI: https://doi.org/10.3115/v1/W14-3627

Jennifer Jenkins. 2009. English as a lingua franca: Interpretations and attitudes. *World English.* 28, 2 (2009), 200–207.

Deborah G. Johnson and Mario Verdicchio. 2017. Reframing AI discourse. *Minds Mach.* 27 (2017), 575–590.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Workshop on Noisy User-generated Text.* 9–18. DOI: https://doi.org/10.18653/v1/W15-4302

Alaa Joukhadar, Huda Saghergy, Leen Kweider, and Nada Ghneim. 2019. Arabic dialogue act recognition for textual chatbot systems. In *1st International Workshop on NLP Solutions for Under Resourced Languages.* 43–49.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Annual Meeting of the Association for Computational Linguistics (ACL'17).* 51–57. DOI: https://doi.org/10.18653/v1/P17-2009

Braj B. Kachru. 1992. *The Other Tongue: English across Cultures.* Urbana.

Vani Kanjirangat, Tanja Samardzic, Fabio Rinaldi, and Ljiljana Dolamic. 2022. Early guessing for dialect identification. In *Findings of EMNLP.* 6417–6426.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of EMNLP.* 7226–7245. Retrieved from https://aclanthology.org/2023.findings-emnlp.481

Abdelrahman Kaseb and Mona Farouk. 2022. SAIDS: A novel approach for sentiment analysis informed of dialect and sarcasm. In *Arabic Natural Language Processing Workshop (WANLP'22)*.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian dialect corpus treebank. In *International Conference on Language Resources and Evaluation (LREC'22)*. 4827–4832. Retrieved from https://aclanthology.org/2022.lrec-1.516

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'23)*. 10597–10611. Retrieved from https://aclanthology.org/2023.emnlp-main.655

Vijay Keswani and L. Elisa Celis. 2021. Dialect diversity in text summarization on Twitter. In *the Web Conference*. 3802–3814.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 4282–4289.

Salam Khalifa, Nasser Zalmout, and Nizar Habash. 2020. Morphological analysis and disambiguation for Gulf Arabic: The interplay between resources and methods. In *International Conference on Language Resources and Evaluation (LREC'20)*. 3895–3904.

Anthony S. Kroch. 1986. Toward a theory of social dialect variation. In *Dialect and Language Variation*. Elsevier, 344–366.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *International Joint Conference on Natural Language Processing (IJCNLP'21)*. Association for Computational Linguistics, 110–121. DOI: https://doi.org/10.18653/v1/2021.acl-short.16

Olli Kuparinen. 2023. Murreviikko—A dialectologically annotated and normalized dataset of Finnish tweets. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'23)*. Association for Computational Linguistics, 31–39. DOI: https://doi.org/10.18653/v1/2023.vardial-1.3

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of EMNLP*. Association for Computational Linguistics, 13814–13828. Retrieved from https://aclanthology.org/2023.findings-emnlp.923

Alfred Lameli and Andreas Schönberg. 2023. A measure for linguistic coherence in spatial language variation. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'23)*. 133–141.

Thang Le and Anh Luu. 2023. A parallel corpus for Vietnamese central-northern dialect text transfer. In *Findings of EMNLP*. 13839–13855. Retrieved from https://aclanthology.org/2023.findings-emnlp.925

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, et al. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Trans. Assoc. Computat. Ling.* 12 (2024), 950–978. DOI: https://doi.org/10.1162/tacl_a_00682

Manning Li and Jiye Mao. 2015. Hedonic or utilitarian? Exploring the impact of communication style alignment on user's perception of virtual health advisory services. *Int. J. Inf. Manag.* 35, 2 (2015), 229–243.

Yanchen Liu, William Held, and Diyi Yang. 2023. DADA: Dialect adaptation via dynamic aggregation of linguistic rules. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'23)*. 13776–13793.

Zhengyuan Liu, Shikang Ni, Ai Ti Aw, and Nancy F. Chen. 2022. Singlish message paraphrasing: A joint task of Creole translation and text normalization. In *International Conference on Computational Linguistics (COLING'22)*. 3924–3936. Retrieved from https://aclanthology.org/2022.coling-1.345

Sin-En Lu, Bo-Han Lu, Chao-Yi Lu, and Richard Tzong-Han Tsai. 2022. Exploring methods for building dialects—Mandarin code-mixing corpora: A case study in Taiwanese Hokkien. In *Findings of EMNLP*. 6287–6305. DOI: https://doi.org/10.18653/v1/2022.findings-emnlp.469

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *ACL System Demonstrations*. 25–30.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Australasian Language Technology Association Workshop*. 5–15.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *International Conference on Language Resources and Evaluation (LREC'14)*. 2348–2354.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'16)*. 1–14.

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214* (2023).

Salima Mdhaffar, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic resources and experiments. In *Arabic Natural Language Processing Workshop (WANLP'17)*. 55–61.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Pacific Asia Conference on Language, Information and Computation (PACLIC'15)*. 26–34. Retrieved from https://aclanthology.org/Y15-1004

Andrew John Merrison, Jack J. Wilson, Bethan L. Davies, and Michael Haugh. 2012. Getting stuff done: Comparing e-mail requests from students in higher education in Britain and Australia. *J. Pragm.* 44, 9 (2012), 1077–1098.

Craig Messner and Thomas Lippincott. 2024. Examining language modeling assumptions using an annotated literary dialect corpus. In *4th International Conference on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics, 325–330. Retrieved from https://aclanthology.org/2024.nlp4dh-1.32

Erin Meyer. 2014. *The Culture Map: Breaking through the Invisible Boundaries of Global Business*. Public Affairs.

Farhad Moghimifar, Shilin Qu, Tongtong Wu, Yuan-Fang Li, and Gholamreza Haffari. 2023. NormMark: A weakly supervised Markov model for socio-cultural norm discovery. In *Findings of ACL*. 5081–5089.

Bruce Moore. 1999. The vocabulary of Australian English. *Australian National Dictionary Centre, Australian National University*. Retrieved from http://andc.anu.edu.au/sites/default/files/vocab_aussie_eng.pdf

Cameron Morin and Steven Coats. 2023. Double modals in Australian and New Zealand English. *World Englishes* (2023). https://doi.org/10.1111/weng.12639

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS One* 15, 8 (2020), e0237861.

Hala Mulki, Hatem Haddad, Mourad Gridach, and Ismail Babaoğlu. 2019. Syntax-ignorant n-gram embeddings for sentiment analysis of Arabic dialects. In *Arabic Natural Language Processing Workshop (WANLP'19)*. 30–39.

Ryo Nagata. 2014. Language family relationship preserved in non-native English. In *International Conference on Computational Linguistics (COLING'14)*. Dublin City University and Association for Computational Linguistics, 1940–1949. Retrieved from https://aclanthology.org/C14-1183

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* (2023).

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*. ACL.

John Nerbonne and Wilbert Heeringa. 2001. Computational comparison and classification of dialects. (2001).

J. Nerbonne and Wilbert Heeringa. 2002. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica, J. Int. Societ. Dialectol. Geoling.* 9 (2002), 69–84.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's paradox of "low-resource" languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 17753–17774. DOI:https://doi.org/10.18653/v1/2024.emnlp-main.983

Ossama Obeid, Salam Khalifa, Nizar Habash, Houda Bouamor, Wajdi Zaghouani, and Kemal Oflazer. 2018. MADARi: A web interface for joint Arabic morphological annotation and spelling correction. In *International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1415

Ossama Obeid, Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2019. ADIDA: Automatic dialect identification for Arabic. In *Conference of the North American Chapter of the Association for Computational Linguistics (System Demonstrations) (NAACL'19)*. 6–11.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source Python toolkit for Arabic natural language processing. In *International Conference on Language Resources and Evaluation (LREC'20)*. 7022–7032.

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. AAEBERT: Debiasing BERT-based hate speech detection models via adversarial learning. In *International Conference on Machine Learning and Applications (ICMLA'22)*. IEEE, 1606–1612.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *International Conference on Computational Linguistics (COLING'22)*. 6208–6221.

Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. 2020. ASA: A framework for Arabic sentiment analysis. *J. Inf. Sci.* 46, 4 (2020), 544–559.

Michael Paul, Andrew Finch, Paul Dixon, and Eiichiro Sumita. 2011. Dialect translation: Integrating Bayesian co-segmentation models with pivot-based SMT. In *Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, et al. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, 1763–1772. Retrieved from https://aclanthology.org/2023.acl-short.150

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*. Association for Computational Linguistics, 3660–3674.

Al-Khanji Rajai and Narjes Ennasser. 2022. Dealing with dialects in literary translation: Problems and strategies. *Jord. J. Mod. Lang. Lit.* 14, 1 (2022), 145–163.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Trans. Assoc. Computat. Ling.* 12 (2024), 19–38.

Alan Ramponi and Camilla Casula. 2023a. DIATOPIT: A corpus of social media posts for the study of diatopic language variation in Italy. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial’23).* 187–199.

Alan Ramponi and Camilla Casula. 2023b. GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task. In *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.* CEUR.org.

Arij Riabi, Menel Mahamdi, and Djamé Seddah. 2023. Enriching the Narabizi treebank: A multifaceted approach to supporting an under-resourced language. In *17th Linguistic Annotation Workshop (LAW’23).* Association for Computational Linguistics, 266–278. DOI : https://doi.org/10.18653/v1/2023.law-1.26

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Trans. Assoc. Computat. Ling.* 11 (2023), 671–685. DOI : https://doi.org/10.1162/tacl_a_00568

Celia Roberts. 2021. Linguistic penalties and the job interview. University of Toronto Press, 2021. https://utppublishing.com/doi/abs/10.3138/9781845537692

Samapika Roy, Sukhada Sukhada, and Anil Kumar Singh. 2020. Parsing Indian English news headlines. In *International Conference on Natural Language Processing (ICON’20).* 239–242.

Hadeel Saadany, Constantin Orăsan, Emad Mohamed, and Ashraf Tantawy. 2022. A semi-supervised approach for a better translation of sentiment in dialectical Arabic UGT. In *Arabic Natural Language Processing Workshop (WANLP’22).* Association for Computational Linguistics. DOI : https://doi.org/10.18653/v1/2022.wanlp-1.20

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *International Conference on Computational Linguistics (COLING’20).* International Committee on Computational Linguistics, 5094–5107. DOI : https://doi.org/10.18653/v1/2020.coling-main.447

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *International Conference on Computational Linguistics (COLING’18).*

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *52nd Annual Meeting of the Association for Computational Linguistics.* 772–778. DOI : https://doi.org/10.3115/v1/P14-2125

Wael Salloum and Nizar Habash. 2022. Unsupervised Arabic dialect segmentation for machine translation. *Nat. Lang. Eng.* 28, 2 (2022), 223–248.

Todd L. Sandel. 2015. Dialects. In *The International Encyclopedia of Language and Social Interaction.* Wiley, 1–13.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Annual Meeting of the Association for Computational Linguistics (ACL’19).* 1668–1678.

Kevin Scannell. 2020. Universal dependencies for Manx Gaelic. In *Workshop on Universal Dependencies (UDW’20).* 152–157. Retrieved from https://aclanthology.org/2020.udw-1.17

Klaus P. Schneider. 2012. Appropriate behaviour across varieties of English. *J. Pragm.* 44, 9 (2012), 1022–1037.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Annual Meeting of the Association for Computational Linguistics (ACL’20).* ACL, 1139–1150. DOI : https://doi.org/10.18653/v1/2020.acl-main.107

Pamela Shapiro and Kevin Duh. 2019. Comparing pipelined and integrated approaches to dialectal Arabic neural machine translation. In *6th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial’19).* ACL, 214–222. DOI : https://doi.org/10.18653/v1/W19-1424

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Arabic Natural Language Processing Workshop (WANLP’15).* 36–48.

Vasiliki Simaki, Panagiotis Simakis, Carita Paradis, and Andreas Kerren. 2017. Identifying the authors’ national variety of English in social media texts. In *International Conference on Recent Advances in Natural Language Processing (RANLP’17).* INCOMA Ltd., 671–678. DOI : https://doi.org/10.26615/978-954-452-049-6_086

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Annual Meeting of the Association for Computational Linguistics (ACL’23).* 6010–6028. DOI : https://doi.org/10.18653/v1/2023.acl-long.331

Rawan Tahssin, Youssef Kishk, and Marwan Torki. 2020. Identifying nuanced dialect for Arabic tweets with deep learning and reverse translation corpus extension system. In *5th Arabic Natural Language Processing Workshop (WANLP’20).* ACL, 288–294. Retrieved from https://aclanthology.org/2020.wanlp-1.30

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, et al. 2024. CASABLANCA: Data

and models for multidialectal Arabic speech recognition. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*. Association for Computational Linguistics, 21745–21758. Retrieved from https://aclanthology.org/2024.emnlp-main.1211

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with base-inflection encoding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 5647–5663.

Pascal Vaillant. 2008. A layered grammar model: Using tree-adjoining grammars to build a common syntactic kernel for related dialects. In *9th International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG'08)*. 157–164. Retrieved from https://aclanthology.org/W08-2321

Rob Van Der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'21)*. 2479–2497.

Chahan Vidal-Gorène, Nadi Tomeh, and Victoria Khurshudyan. 2024. Cross-dialectal transfer and zero-shot learning for Armenian varieties: A comparative analysis of RNNs, transformers and LLMs. In *4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH'24)*. Association for Computational Linguistics, 438–449.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial Singaporean English. In *Annual Meeting of the Association for Computational Linguistics (ACL'17)*. 1732–1744. DOI : https://doi.org/10.18653/v1/P17-1159

Yizhou Wang, Rikke L. Bundgaard-Nielsen, Brett J. Baker, and Olga Maxwell. 2022. Perceptual overlap in classification of L2 vowels: Australian English vowels perceived by experienced Mandarin listeners. In *Pacific Asia Conference on Language, Information and Computation (PACLIC'22)*. 317–324.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Findings of ACL*.

Roy Xie, Orevaoghene Ahia, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Extracting lexical features from dialects via interpretable dialect classifiers. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'24)*. 54–69. DOI : https://doi.org/10.18653/v1/2024.naacl-short.5

Fan Xu, Xiongfei Xu, Mingwen Wang, and Maoxi Li. 2015. Building monolingual word alignment corpus for the greater China region. In *Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*. Association for Computational Linguistics, 85–94. Retrieved from https://aclanthology.org/W15-5414

R. S. Yadav. 1974. Interactions of written and spoken Hindi. *Ind. Lit.* 17, 3 (1974), 61–66.

Jihene Younes, Emna Souissi, Hadhemi Achour, and Ahmed Ferchichi. 2020. Language resources for Maghrebi Arabic dialects' NLP: A survey. *Lang. Resour. Eval.* 54 (2020), 1079–1142.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third VarDial evaluation campaign. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'19)*. 1–16. DOI : https://doi.org/10.18653/v1/W19-1401

Marcos Zampieri and Preslav Nakov. 2021. *Similar Languages, Varieties, and Dialects: A Computational Perspective*. Cambridge University Press.

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

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*. 1–9.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'12)*. 49–59. Retrieved from https://aclanthology.org/N12-1006

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, et al. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *9th Conference on Machine Translation (WMT'24)*. ACL, 82–109.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds et al. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*. ACL.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma et al. 2023. SocialDial: A benchmark for socially-aware dialogue systems. In *ACM SIGIR Conference*. 2712–2722.

Xiongyi Zhang, Jan-Willem van de Meent, and Byron C. Wallace. 2021. Disentangling representations of text by masking transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 778–791.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

Yuanyuan Zhao, Weiwei Sun, Junjie Cao, and Xiaojun Wan. 2020. Semantic parsing for English as a second language. In *Annual Meeting of the Association for Computational Linguistics (ACL'20)*. Association for Computational Linguistics, 6783–6794. DOI : https://doi.org/10.18653/v1/2020.acl-main.606

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Annual Meeting of the Association for Computational Linguistics (ACL'22)*. 3701–3720. DOI : https://doi.org/10.18653/v1/2022.acl-long.258

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL'23)*. 744–768. DOI : https://doi.org/10.18653/v1/2023.acl-long.44