Data-Driven Dialectology and Regional Variation

Jonathan Dunn and Mai Mohamed Eida

Abstract. We review data-driven approaches to the study of regional variation within languages with a special focus on English and Arabic. This kind of data-driven research uses unelicited corpora to represent the linguistic behavior of different populations. The first step in such analysis is to annotate corpora for specific linguistic features or variants. The second step is to develop large-scale quantitative models of variation. This chapter reviews each of these main methodological components and then surveys the findings that result from these data-driven approaches.

Keywords: Dialectology, Dialectometry, Regional variation, Corpus-based sociolinguistics, Computational sociolinguistics, Register

1. Introduction

This chapter surveys data-driven approaches to social and geographic dialects. Data-driven dialectology differs from more traditional participant-based approaches by starting with corpus data. This data is often *unelicited:* while each sample contains language use from a particular population, its context of production is neither controlled nor influenced by the researcher. In other words, a sociolinguistic interview takes place in a controlled environment while a tweet is produced without the researcher's involvement. This approach is also called corpus-based dialectology, because it is centered on corpus data. And sometimes it is called computational dialectology, because large corpora require automated annotations and more involved quantitative models. And, finally, it is sometimes called dialectometry, because it aggregates a large number of variants or features into a single picture of variation. Taken together, this range of methodologies provides a radical reassessment of the nature and extent of linguistic variation (Nerbonne & Kretzschmar, 2013).

The next section considers sources of corpus data that have been used for the study of dialectal variation, including a comparison of digital and non-digital sources. The third section considers how corpora can be annotated for specific linguistic features and how a large number of features can be aggregated together. The fourth section reviews the two major families of quantitative models that have been used to analyze variation, similarity-based and classification-based models. The fifth and sixth sections examine two important confounds facing corpus-based work: register (i.e., variations in the context) and topic (i.e.,

as an irrelevant source of variation). The seventh section concludes by considering what new findings these data-driven approaches to dialect have made possible.

2. Sources of Corpus Data

The defining principle of corpus-based and computational dialectology is to work with unelicited corpus data rather than with interviews or survey instruments, with a few notable exceptions. Some corpora derive from interviews that were collected for non-linguistic purposes (e.g., Hay, et al. 2015) but mostly corpora are derived from digital sources that are not interviews at all, collected but not created (e.g., Coats, 2023). While these two examples are spoken corpora, most data sets are in fact textual: from the web (Davies and Fuchs, 2015; Cook and Brinton, 2017; Dunn, 2020; Ali et al., 2023) or from social media (Mocanu, et al., 2013; Gonçalves and Sánchez, 2014; Ahmed et al., 2022); another common source is published texts like letters to the editors in local newspapers (Graff, 2003; Grieve, 2011, 2012). This line of work thus represents a shift from elicited to unelicited language use, from face-to-face communication to digital sources, and from spoken to written modalities. What is the combined impact of these three shifts?

The first question is whether digital communication carries the same linguistic patterns as face-to-face communication. One approach to this question is to compare dialect surveys with corpus-based models (Grieve, et al., 2019): this type of work shows mixed results depending on the target language and dialect. There is generally close correspondence between the two for English, at least in the UK, thus supporting the idea that digital communication is a reliable source of dialect data. For example, the three-way alternation of *couch* vs *sofa* vs *settee* has the same geographic distribution in both the BBC Voices dialect survey and in tweet-based corpora.

However, work on Egyptian Arabic shows a gap between digital data and dialectal surveys (Eida, et al., 2024). Other work has suggested that the artificial context of dialect surveys tends to over-estimate geographic factors (Szmrecsanyi, 2014), so that we might expect less geographic variation in more natural contexts of production.

A second approach is to compare the variation present in digital and non-digital corpora regardless of dialect (Cvrček, et al., 2020): this type of work shows that digital corpora are, in general, less diverse than traditional corpora. Thus, a very large number of usage tokens does not necessarily provide a large number of usage types.

A third approach is to triangulate dialectal corpora from multiple sources in order to test the reliability of digital sources (Dunn, 2021) or to test the reliability of models of dialects (Dunn and Wong, 2022): this type of work shows that digital communication does provide

consistent representations of dialect, building on foundational work on comparing corpora (Kilgarriff, 2001). However, consistency does not necessarily imply external validity.

The next question is whether written corpora capture dialectal variation or whether processes of standardization serve to remove variation from written data. It is clear that written corpora cannot capture sociophonetic variation; but can they capture variation in other parts of the linguistic system? While studies cited above show that corpora capture the same patterns as dialect surveys, there are two complicating factors. One is that corpora must be evaluated as to how well they represent specific populations (Dunn & Adams, 2020); thus, we need confidence that the data itself represents the intended population. For example, even minor choices during corpus creation impact different populations in different ways (Dunn, 2024a). The second complicating factor is that writing systems can themselves change spoken language (Dąbrowska, 2021): this means that a strict separation between written and spoken registers is perhaps a false dichotomy and we should perhaps make a distinction between literate and illiterate populations. For Arabic, the distinction between written and spoken registers itself depends on country, dialect, socioeconomic status, and literacy. This is because Modern Standard Arabic influences the dialectal written and spoken registers in different ways (Badawi, 1973).

Much of the focus in Arabic work has been on capturing the complex dialectal diversity of the Arab world (Dahou et al., 2025) and less on validating collected corpora against the target populations. While considerable attention in Arabic has been given to increasing dialectal coverage, comparatively little work has evaluated whether the geographic, socioeconomic, and linguistic distributions of corpora correspond to those of their intended speaker populations.

3. Representing Linguistic Features

The next challenge is to choose linguistic features to represent each dialect-specific sample. Given that most data-driven approaches work with written corpora, these features are typically lexical, morphological, or syntactic. There are two major questions here: (i) What features to represent? and (ii) How to aggregate a large number of features together? These two questions together have formed the core research problem of data-driven dialectology.

First, corpus-based approaches tend to focus on lexical and morphosyntactic features. It remains possible, however, to use morphosyntactic features to study spoken corpora (Szmrecsanyi, 2013), providing a broader view that connects variation in both written and spoken corpora. Much corpus-based work has focused on lexical variation, in part because it is easier to operationalize than syntactic variation. For example, early work showed that

distributional semantics provides a better operationalization of lexical variation than non-distributional approaches (Peirsman, et al., 2010). A primary challenge with lexical features is to aggregate across many lexical items (Grieve, et al., 2011), but a secondary challenge is to deal with the instability of distributional semantics itself (Dunn, 2023b). In Arabic, the combination of non-concatenative morphology and phonological differences not consistently represented in orthography places a greater strain on feature representations. For example, both (1a) and (1b) are written in the same way. However, (1a) is in the active voice and (1b) in the passive. The difference can be inferred from context by native Arabic speakers or it can be marked by diacritics. While the examples below are marked by diacritics, Arabic speakers rarely use diacritics in practice, making feature representation and selection more challenging.

(1a) كَتَبَ

katab-a

write.PFV-3SG.M

'He wrote.'

(1b) كُتِبَ

kutib-a

write.PFV.PASS-3SG.M

'It was written.'

Large-scale work on lexical variation has shown that non-geographic sources of variation are strong (Eisenstein, et al., 2014) and thus should be accounted for in some way (Wieling, et al., 2011). This converges with the morphosyntax-based finding that dialect surveys, focused on geography, tend to underestimate the importance of demographics (Szmrecsanyi, 2014). In English, the main non-geographic factor seems to derive from race/ethnicity (Morin & Grieve, 2024). But, in Arabic, socioeconomic status and especially level of education play the dominant role in shaping linguistic variation (Badawi, 1973).

Much work also draws on morphosyntactic features; these range from grammatical alternations like particle placement (Szmrecsanyi, 2013; Grieve, 2016; Grafmiller & Szmrecsanyi, 2018) to deeper constructional representations (Szmrecsanyi, et al., 2019; Szmrecsanyi & Grafmiller, 2023). Many of these approaches assume a fixed set of grammatical features, based on introspection and on previous studies.

Unlike sociophonetics, however, there is no universal feature space for grammar. Thus, another approach is to learn grammatical features as part of the methodology (Dunn,

2018). But, in practice, this creates a large inventory of features: thousands instead of a dozen or less. This raises the question of how to validate a large number of features that are not based on introspection. As discussed in the next section, one common approach is to use a classifier to predict categorical dialect labels, with prediction accuracy on this task providing a means of validation (Dunn, 2019a, b). Thus, the best model of dialect is able to predict, for unseen samples, what dialect they belong to. This is an approach to aggregation that comes with a ground-truth measure of validity. In English, such models often obtain remarkably high accuracies. In Arabic, however, the accuracy can be much lower, depending on the source of ground-truth labels (Abdelali et al., 2021).

An argument for a feature-agnostic methodology, in which the inventory of grammatical variants is discovered and not assumed, is that the choice of features can lead to very different similarity networks between dialect communities (Dunn, 2023a). Thus, the choice of features has a significant impact on conclusions drawn about variation. An alternate view is that studies of variation should rely on survey data from linguistic atlas projects, thus anchoring them against a known ground-truth (Goebl, 2006). In historical terms, such an anchor was important for early proof-of-concept work in dialectometry but perhaps is no longer necessary.

The ability to aggregate a large number of features or variables is the main advantage of data-driven approaches (Nerbonne, 2009; Wieling & Nerbonne, 2015). As discussed in the next section, many aggregation approaches are unsupervised and thus need to be evaluated against known dialect patterns. For example, work based directly on atlas data has suggested that linguistic distance is more valid than geographic distance (Pickl, et al., 2014) because naive measures of geographic distance do not account for obstacles to human communication like mountains or national borders. This kind of evaluation of linguistic vs geographic distance depends on a ground-truth similarity matrix against which distance measures are compared.

4. Models of Variation

Given corpus data that represents different populations and a set of linguistic features or variables that could be subject to variation, the next challenge is to model variation itself. There are two major families of models: classification-based and similarity-based (Dunn, 2026). Classification models (supervised) treat dialects as discrete categories; for example, New Zealand English is no closer to Australian than to American English. A classifier provides a measure of internal validity (its ability to predict dialect membership of held-out samples) and it excels at problems involving a large number of features.

Similarity-based models (unsupervised) measure the distance between dialects, thus treating dialectal variation as a continuous phenomenon. For example, the linguistic distance between two local dialects might correlate with the geographic distance between them, but it might not. Similarity-based approaches do not assume sets of dialect areas, unlike classification models, which reduces the assumptions they make about speech communities. However, they are more difficult to validate and, because they are unsupervised, a large number of features may lead to more noise.

Both families of models are quantitative and large-scale approaches to variation. Both kinds of models challenge traditional qualitative perceptions of dialect and open up new questions about variation and diffusion (Kretzschmar, 1996; Nerbonne & Kretzschmar, 2013). But these models differ in how they approach variation. Classifiers are supervised approaches: they are given training data which contains dialect labels (as assumptions) and asked to find which set of features best characterizes each dialect. As supervised models, they are validated against data with the same labels as the training set: higher prediction accuracy means a more valid model.

Similarity-based methods are unsupervised approaches: they are given samples and must learn both the dialect areas and the features subject to variation together. Unsupervised models have no training data and thus cannot be validated in the same way as supervised models. A major idea has been that data from a linguistic atlas is the best ground-truth for validating similarity-based models (Nerbonne & Heeringa, 2001; Heeringa, et al., 2002; Nerbonne & Kleiweg, 2007). Simply put, the model should predict the same network of similarities between dialects as the atlas. For example, each community is a node and linguistic distance measures the edge weight between nodes, thus creating a network (c.f., Heeringa, 2004). Thus, a corpus-based network should overlap with an atlas-based network. The challenge is that the number of atlases is limited, especially in non-Western contexts. An example of a linguistic feature for measuring similarity is the genitive alternation: "the country's economic crisis" vs "the economic crisis of the country." As shown in Szmrecsanyi, et al. (2019), all dialects of English share this alternation but differ in their usage and conditioning, with similarity based on the conditioning factors.

Classification-based models are more standardized because they build on standard classifiers like logistic regression, decision trees, or support vector machines. This means that the main design decision is the choice of features. These models predominate in the computational literature and surveys of model design and evaluation are available elsewhere (Zampieri, et al., 2020; Joshi, et al., 2025).

Similarity-based approaches are less standardized and often rely on multiple individual components. At a conceptual level, there are two major families: feature-first methods, like

factor analysis or multi-dimensional analysis, group individual variants together into components or bundles that explain the primary dimensions of variation. These components can then be used to estimate the distance between dialect areas (Nerbonne, 2006). The main outputs are, first, the components of variation and, second, the distances between dialect areas.

The second family, sample-first methods like clustering, group samples together without factor analysis. One argument for a sample-first approach is that dimension reduction methods, while providing fewer dimensions to describe, remove a significant amount of variation from the data. Sample-first methods use measures like cosine or Euclidean distance directly on features, without factor analysis to reduce this to a few feature bundles. For example, a sample-first approach has been used to estimate distances between cities in order to create a network of dialect similarities (Dunn & Wong, 2025). While these approaches differ in their architecture, each results in a network of similarity values that can be analyzed in the same way.

Arabic presents additional challenges for both classification and similarity-based models of variation. For example, political borders are less likely to correspond to linguistic boundaries (Behnstedt & Woidich, 2005). Arabic dialect geography consequently contains both dialect continua and transitional areas (De Jong, 2000). This complicates classification-based approaches because the labels supplied to the model generally must be discrete categories: do we represent dialects as broad regional categories (e.g., Egyptian, Levantine, Gulf) or as finer-grained city-level varieties? The performance of the former averages 60% accuracy (Abdelali et al., 2021) while the latter can reach as low as 6% (Abdul-Mageed et al. 2020).

Similarity-based approaches are particularly useful in this context. Studies of Arabic have found a general relationship between geographic and linguistic similarity, for example with varieties across national borders sometimes being more similar than varieties within the same country (Alsudais et al., 2022). The nation-state is a less salient entity.

5. Register Variation as a Confound

Linguistic variation has multiple sources: dialect (both geographic and social) but also register (the context of production). Corpus-based studies have shown that, of these two sources, register variation is stronger and more pervasive (Biber & Conrad, 2009; Biber, 2012). While there is some amount of overlap in the form of joint register-dialectal variation (Morin, et al., 2026), in many particulars the two sources are independent in English (Kruger & van Rooy, 2018). This would imply, for example, that registers are constant regardless of the dialect of the individual who is producing those registers.

While register and dialect are in many ways independent, register variation is a potential confound in the study of both geographic and social dialects. First, most studies of dialect take place within a single register (e.g., face-to-face conversations or social media). This means that the precise relationship between dialect and register remains unknown because it cannot be triangulated (with the important exception of the *International Corpus of English*: Greenbaum, 1996). Second, corpus-based approaches often assume that all samples are drawn from comparable registers. For example, many studies rely on tweets but ignore the presence of different registers within tweets: personal conversations, business announcements, political arguments, advertisements for events, etc. For example, we might expect that a university city would contain many more academic tweets while a rural area would contain many fewer. The confound is that, by assuming a fixed register for all locations, we cannot eliminate register as the ultimate source of variation.

Arabic further complicates this relationship between dialect and register with Modern Standard Arabic (MSA). Formal and written registers generally encourage greater use of MSA, while informal and spoken contexts allow greater use of dialect, although the boundary between the two is not categorical (c.f., Biber, et al., 2020, in English). Speakers may also avoid highly marked dialectal features in contexts where they are communicating with audiences from different regions. Register can therefore affect the visibility of dialectal variation itself: the same speaker may produce strongly local forms in conversation but use MSA in formal writing. As a result, linguistic similarity between two Arabic samples may reflect similarities in register and MSA use rather than underlying dialect similarity.

For example, speakers of both Cairene Egyptian Arabic and Sa'idi Egyptian Arabic use the spelling "جمل" for the word *Camel*. This is the standardized and codified spelling in Modern Standard Arabic, almost always used across all Arabic dialects. However, this spelling homogenizes the dialectal distinctions: Cairene speakers pronounce this [ɡamal] while Sa'idi speakers pronounce this [damal] or [dʒamal]. These alternations are not present in writing as all of these variations are written the same as MSA: "جمل" . Thus, register interacts significantly with dialect.

6. Topic and Content as a Confound

Topic is a second major confound: how do we separate variation in the topic of discussion, which could lead to preferences for one or another variant, from variation in dialect itself? The problem here is the environment in which communication takes place (Szmrecsanyi, 2016): for example, changes in content might change the number of times the genitive of possession is used, thus changing its token frequency. But this could simply be a result of the number of times that *POSSESSION* as a function needs to be expressed. This can be controlled for at the feature level by quantifying variables using percentages rather than

straight token frequencies. Thus, the share of genitives of possession relative to other means of expressing *POSSESSION* is a more robust measure, assuming that the full set of constructions expressing that function is known.

A second approach to removing the confound of topic is to ensure that the dataset is balanced for topic. This approach is limited to contexts in which data is so abundant that samples can be selected to represent specific topics (Dunn & Wong, 2022): since the observations for each dialect have the same distribution over topics, we no longer need to organize the variables into discrete alternations. This approach can be helpful if we are unable to commit to a discrete set of variables which express a single function. For example, for the function of *POSSESSION*, we have the alternation in (2a) and (2b). These two variants account for much of the functional load. But for the function of *TRANSFER*, the alternation in (3a) and (3b) must also be compared with (3c) and (3d). In other words, there are a wider range of options to express the function of *TRANSFER* and we may be unable to define an exhaustive set possibilities. In this case, balancing samples by topic is an easier approach to eliminating the confound of topic.

(2a) Tom's brother
(2b) the brother of Tom
(3a) mailed the package to Tom
(3b) mailed Tom the package
(3c) Tom was mailed the package
(3d) Tom received the package in the mail

Topic remains a confound in Arabic, where data scarcity has made the second approach especially difficult (Eida & Habash, 2025). For example, one important corpus of Sa'idi Egyptian Arabic is an online forum of romantic Sa'idi novels: but this corpus is necessarily full of topics related to romance, marriage, and family dynamics. On the other hand, limited data sources have encouraged the use of translated or adapted resources, which may introduce culturally or topically irrelevant information. For example, MADAR was constructed by translating English and French sentences from the Basic Traveling Expression Corpus (BTEC) into 25 Arabic city dialects (Bouamor et al., 2018). While this design provides strong parallel corpus in Arabic dialects, the topics themselves originate outside the communities being modeled: grabbing a "beer" before taking the train is not relevant to these communities.

7. What These Methods Tell Us About Dialects

The focus so far has been on methodological advances in dialectology: by using unelicited corpora we can consider more samples and by using larger feature sets we can observe

more types of variation. This allows us, in turn, to study larger populations and a more representative slice of the linguistic system. But what do we learn about dialects, both social and geographic, from these new methods?

The main advantage of data-driven methods is that we can view speakers of a language as a global speech community, thus taking into account the language contact caused by colonialism and migration (Kachru, 1990). Because the contemporary world is organized around these expanded speech communities, approaches which focus on small areas in isolation assume an unrealistic network size. For example, simulation-based models with small networks show a significant difference caused by strong ties vs weak ties between individuals (Fagyal, et al., 2010). However, when networks are expanded to a realistic size, capturing the number of relationships and interactions that actual speakers have in the contemporary world, such a distinction falls away (Laitinen, et al., 2020). This indicates that results based on small-scale case-studies have low external validity. At the same time, cases that involve diffusion within a community require demographic information, which is often missing from such large-scale networks (Laitinen & Fatemi, 2022; c.f., Dunn, 2025).

At the same time, digital communication creates non-geographic speech communities in digital spaces. For example, studies of Reddit forums have shown that communities with more extensive and loyal engagement produce more lexical variants (Lucy & Bamman, 2021): this follows our expectation that a stronger network will have more impact on language use. Thus, the digital world creates competing non-geographic speech communities. For example, some digital communities use *haul* with a sense of a reward for some activity, while others use *haul* with a sense of a period of struggle. While each community likely has access to each of these senses, these different patterns of usage would presumably lead to different changes in the lexicon over time.

At the same time, the rise of LLMs and Generative AI is changing the linguistic behavior of social media users (Møller, et al. 2026). In English machine-assisted writing is creating a pressure towards standardization that removes dialectal variation from written digital sources (Gan, et al., 2026). At the same time, LLMs tend to represent inner-circle populations much better than more peripheral populations (Dunn, et al., 2024), implying that machine-assisted writing will have a greater impact on low prestige dialects.

Data-driven work has also helped us understand what parts of the grammar are more likely to vary across global communities (Dunn, 2026). From a usage-based perspective, the grammar is a network that ranges from highly concrete and item-specific constructions (with more lexical properties) to highly abstract and productive constructions (with more syntactic properties). While previous work has had to rely on an artificial selection of grammatical features, recent work has operationalized the grammar as a network (Dunn,

2024). The results show that dialectal variation is concentrated in more concrete and more peripheral (less commonly used) constructions, while register variation is spread more evenly across the grammar. This view of variation across the entire grammar could not previously have been observed. The main conclusion is that both the grammar and the population are complex systems (c.f., Schneider, et al., 2020), so that arbitrary snapshots of those systems provide arbitrary conclusions about the sources of dialectal variation. Thus, data-driven approaches have forced a new conceptualization of variation in social and regional dialects.

References

Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., & Darwish, K. (2021). 'QADI: Arabic dialect identification in the wild.' In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics. 1-10.

Abdul-Mageed, M., Zhang, C., Bouamor, H., & Habash, N. (2020). 'NADI 2020: The first nuanced Arabic dialect identification shared task.' In *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (pp. 97–110). Association for Computational Linguistics.

Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abd-alrazaq, A. A., & Househ, M. (2022). 'Freely available Arabic corpora: A scoping review.' *Computer Methods and Programs in Biomedicine Update*, 2: 100049.

Ali, A. R., Siddiqui, M. A., Algunaibet, R., & Ali, H. R. (2023). 'A large and diverse Arabic corpus for language modeling.' *Procedia Computer Science, 225*, 12–21.

Alsudais, A., Alotaibi, W., & Alomary, F. (2022). 'Similarities between Arabic dialects: Investigating geographical proximity.' *Information Processing & Management, 59*(1), 102770.

Badawi, E. M. (1973). *Mustawayāt al-ʿArabiyyah al-muʿāṣirah fī Miṣr: Baḥth fī ʿalāqat al-lughah bi-al-ḥaḍārah* [Levels of contemporary Arabic in Egypt: A study of the relationship between language and civilization]. Dār al-Maʿārif.

Behnstedt, P., & Woidich, M. (2005). *Arabische Dialektgeographie: Eine Einführung*. Brill.

De Jong, R. E. (2000). *A grammar of the Bedouin dialects of the Northern Sinai littoral: Bridging the linguistic gap between the Eastern and Western Arab world*. Brill.

Biber, D. (2012), 'Register as a predictor of linguistic variation.', *Corpus Linguistics and Linguistic Theory* **8**(1). 9--37.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*, Cambridge University Press, Cambridge, UK.

Biber, D., Egbert, J. and Keller, D. (2020). 'Reconceptualizing register in a continuous situational space', *Corpus Linguistics and Linguistic Theory* **16**(3). 581--616.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). 'The MADAR Arabic dialect corpus and lexicon.' In *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association. 3387–3396.

Coats, S. (2023). 'Dialect Corpora from YouTube.' In Busse, Beatrix; Dumrukcic, Nina; and Kleiber, Ingo (Eds.) *Language and Linguistics in a Complex World*. De Gruyter: Berlin. 79-102.

Cook, P. and Brinton, J. (2017). 'Building and Evaluating Web Corpora Representing National Varieties of English', *Language Resources and Evaluation* **51**(3), 643--662.

Cvrcek, V., Komrskovб, Z., Lukes, D., Poukarovб, P., Řehořkovб, A., Zasina, A. J. and Benko, V. (2020), 'Comparing web-crawled and traditional corpora', *Language Resources and Evaluation* **54**(3), 713--745.

Dąbrowska, E. (2021), *Language Change: The Impact of English as a Lingua Franca*, Cambridge University Press, chapter How Writing Changes Languages, pp. 75-94.

Dahou, A., Dahou, A. H., Cheragui, M. A., Abdedaiem, A., Al-Qaness, M. A. A., Abd Elaziz, M., Ewees, A. A., & Zheng, Z. (2025). 'A survey on dialect Arabic processing and analysis: Recent advances and future trends.' *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*(8), Article 84, 1–45.

Davies, M. and Fuchs, R. (2015), 'Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE)', *English World-Wide* **36**(1), 1--28.

Dunn, J. (2018), 'Finding Variants for Construction-Based Dialectometry: A Corpus-Based Approach to Regional CxGs', *Cognitive Linguistics* **29**(2), 275--311.

Dunn, J. (2019a), 'Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology', *Frontiers in Artificial Intelligence* **2**, 15.

Dunn, J. (2019b), 'Modeling Global Syntactic Variation in English Using Dialect Classification.', In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, Association for Computational Linguistics. 42--53.

Dunn, J. (2020), 'Mapping languages: the Corpus of Global Language Use', *Language Resources and Evaluation* **54**, 999--1018.

Dunn, J. (2021), 'Representations of Language Varieties Are Reliable Given Corpus Similarity Measures.' In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics. 28--38.

Dunn, J. (2023a), 'Syntactic variation across the grammar: modelling a complex adaptive system', *Frontiers in Complex Systems* **1**.

Dunn, J. (2023b), ‘Variation and Instability in Dialect-Based Embedding Spaces.’ In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics. 67--77.

Dunn, J. (2024a), ‘Validating and Exploring Large Geographic Corpora.’ In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 17348--17358.

Dunn, J. (2024b), *Computational Construction Grammar: A Usage-Based Approach*, Cambridge University Press.

Dunn, J. (2025). “Diffusion Across the Grammar: Complexity in Areal Interactions Between Dialects of English.” In Enrique-Arias, Andrés, Carlota de Benito Moreno and Florencio del Barrio de la Rosa (eds.). *The spatial diffusion of linguistic changes: new methods and theoretical perspectives*. Berlin: De Gruyter.

Dunn, J. (2026), *Syntactic Variation from Individuals to Populations: Language as a Complex System*, Cambridge University Press.

Dunn, J. and Adams, B. (2020), ‘Geographically-Balanced Gigaword Corpora for 50 Language Varieties.’ In *Proceedings of the International Language Resources and Evaluation Conference*. European Language Resources Association. 2528--2536.

Dunn, J., Adams, B. and Tayyar Madabushi, H. (2024), ‘Pre-Trained Language Models Represent Some Geographic Populations Better than Others.’ In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. 12966--12976.

Dunn, J. and Wong, S. (2022), ‘Stability of Syntactic Dialect Classification over Space and Time.’ In *Proceedings of the International Conference on Computational Linguistics*. International Committee on Computational Linguistics. 26--36.

Dunn, J. and Wong, S. (2025), 'Language Contact and Population Contact as Sources of Dialect Similarity', *Languages* **10**(8).

Eida, M., & Habash, N. (2025). ‘Beyond Cairo: Sa’idi Egyptian Arabic literary corpus construction and analysis.’ In *Proceedings of the International Conference on Natural Language Processing for Digital Humanities.* Association for Computational Linguistics. 292–304.

Eida, M., Nassar, M., & Dunn, J. (2024). ‘How well do tweets represent sub-dialects of Egyptian Arabic?’ In *Proceedings of the Workshop on NLP for Similar Languages, Varieties, and Dialects.* Association for Computational Linguistics. 41–55.

Eisenstein, J., O'Connor, B., Smith, N. and Xing, E. (2014), 'Diffusion of lexical change in social media', *PloSOne* **10**, 1371.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L. and Lakkaraju, K. (2010), 'Centers and peripheries: Network roles in language change', *Lingua* **120**(8), 2061-2079.

Gan, Q., Dunn, J., Nini, A. and Adams, B. (2026). “A Multi-Dialectal, Longitudinal Corpus of Human-AI Hybrid Language Production.” In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), Palma, Mallorca, Spain, pp. 11286--11299.

Graff, D. (2003). *Arabic Gigaword* (LDC2003T12) [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/ep1n-de95

Goebl, H. (2006), 'Recent advances in Salzburg dialectometry', *Literary and Linguistic Computing* **21**(4), 411--435.

Gonçalves, B. and Sánchez, D. (2014), 'Crowdsourcing Dialect Characterization through Twitter', *PLOS ONE* **9**(11), 1-6.

Grafmiller, J. and Szmrecsanyi, B. (2018), 'Mapping out particle placement in Englishes around the world A study in comparative sociolinguistic analysis', *Language Variation and Change* **30**(3), 385--412.

Greenbaum, S. (1996), *Comparing English Worldwide: The International Corpus of English*, Clarendon Press, Oxford.

Grieve, J. (2011), 'A regional analysis of contraction rate in written Standard American English', *International Journal of Corpus Linguistics* **16**(4), 514--546.

Grieve, J. (2012), 'A statistical analysis of regional variation in adverb position in a corpus of written Standard American English', *Corpus Linguistics and Linguistic Theory* **8**(1), 39--72.

Grieve, J. (2016), *Regional variation in written American English*, Cambridge University Press, Cambridge, UK.

Grieve, J., Montgomery, C., Nini, A., Murakami, A. and Guo, D. (2019), 'Mapping Lexical Dialect Variation in British English Using Twitter', *Frontiers in Artificial Intelligence* **2**, 11.

Grieve, J., Speelman, D. and Geeraerts, D. (2011), 'A statistical method for the identification and aggregation of regional linguistic variation', *Language Variation & Change* **23**(1), 1--29.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J. and LaShell, P. (2015), 'Tracking word frequency effects through 130 years of sound change', *Cognition* **139**, 83--91.

Heeringa, W. (2004), *Measuring dialect pronunciation differences using Levenshtein distance*, Groningen, Netherlands.

Heeringa, W., Nerbonne, J. and Kleiweg, P. (2002), Validating Dialect Comparison Methods: Classification, Automation, and New Media', Springer, Berlin, pp. 445--452.

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G. and Dippold, D. (2025), 'Natural Language Processing for Dialects of a Language: A Survey', *ACM Comput. Surv.* **57**(6).

Kachru, B. (1990), *The Alchemy of English: The spread, functions, and models of non-native Englishes*, University of Illinois Press, Urbana-Champaign.

Kilgarriff, A. (2001), 'Comparing Corpora', *International Journal of Corpus Linguistics* **6**(1), 97--133.

Kretzschmar, W. A. (1996), 'Quantitative areal analysis of dialect features', *Language Variation & Change* **8**(1), 13--39.

Kruger, H. and van Rooy, B. (2018), 'Register variation in written contact varieties of English: A multidimensional analysis', *English World-Wide* **39**(2), 214--242.

Laitinen, M. and Fatemi, M. (2022). "Big and rich social networks in computational sociolinguistics." In Paula Rautionaho and Hanna Parviainen and Mark Kaunisto and Arja Nurmi, ed., *Social and Regional Variation in World Englishes: Local and Global Perspectives*, Routledge, pp. 1--25.

Laitinen, M., Fatemi, M. and Lundberg, J. (2020), 'Size Matters: Digital Social Networks and Language Change', *Frontiers in Artificial Intelligence* **3**(July), 1--15.

Lucy, L. and Bamman, D. (2021), 'Characterizing English variation across social media communities with BERT, *Transactions of the Association for Computational Linguistics* **9**, 538--556.

Mocanu, D., Baronchelli, A., Perra, N., Goncalves, B., Zhang, Q., Vespignani, A., Gonccalves, B., Zhang, Q. and Vespignani, A. (2013), 'The Twitter of Babel: Mapping World Languages through Microblogging Platforms', *PLOSOne* **10**, 1371.

Møller, A. G., Romero, D. M., Jurgens, D. and Aiello, L. M. (2026), 'The impact of generative AI on social media: an experimental study', *Scientific Reports* **16**(1), 9376.

Morin, C. & Grieve, J. (2024), "The semantics, sociolinguistics, and origins of double modals in American English: New insights from social media." *PLoS ONE* 19(1): e0295799.

Morin, C.; Coats, S.; & Dunn, J. (2026), "How register and region shape the language network: evidence from Computational Construction Grammar." *Constructions*, *18*(1).

Nerbonne, J. (2006), 'Identifying linguistic structure in aggregate comparison', *Literary and Linguistic Computing* **21**(4), 463--476.

Nerbonne, J. (2009), 'Data-driven dialectology', *Language and Linguistics Compass* **3**(1), 175--198.

Nerbonne, J. and Heeringa, W. (2001), 'Computational Comparison and Classification of Dialects', *Dialectologia et Geolinguistica* **2001**(9), 69--84.

Nerbonne, J. and Kleiweg, P. (2007), 'Toward a dialectological yardstick', *Journal of Quantitative Linguistics* **14**(2), 148--166.

Nerbonne, J. and Kretzschmar, W. (2013), 'Dialectometry+ +', *Literary and Linguistic Computing* **28**(1), 2--12.

Peirsman, Y., Geeraerts, D. and Speelman, D. (2010), 'The automatic identification of lexical variation between language varieties', *Natural Language Engineering* **16**(4), 469--491.

Pickl, S., Spettl, A., Pröll, S., König, S. E. W., Schmidt, V., Proll, S., Konig, S. E. W. and Schmidt, V. (2014), 'Linguistic distances in dialectometric intensity estimation', *Journal of Linguistic Geography* **2**(1), 25--40.

Schneider, E. W.; Mauranen, A., and Vetchinnikova, S. (ed.) (2020), *Calling Englishes As Complex Dynamic Systems: Diffusion and Restructuring*, Cambridge University Press, pp. 15-43.

Szmrecsanyi, B. (2013), *Grammatical variation in British English dialects A study in corpus-based dialectometry*, Cambridge University Press, Cambridge.

Szmrecsanyi, B. (2014), 'Forests, trees, corpora, and dialect grammars.' In *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Mouton De Gruyter, Berlin, 89--112.

Szmrecsanyi, B. (2016), 'About text frequencies in historical linguistics: Disentangling environmental and grammatical change', *Corpus Linguistics and Linguistic Theory* **12**(1), 153--171.

Szmrecsanyi, B. and Grafmiller, J. (2023), *Comparative Variation Analysis: Grammatical Alternations in World Englishes*, Cambridge University Press.

Szmrecsanyi, B., Grafmiller, J. and Rosseel, L. (2019), 'Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes', *Frontiers in Artificial Intelligence* **2**, 23.

Wieling, M. and Nerbonne, J. (2015), 'Advances in dialectometry', *Annual Review of Linguistics* **1**, 243--264.

Wieling, M., Nerbonne, J. and Baayen, R. H. (2011), 'Quantitative social dialectology: Explaining linguistic variation geographically and socially', *PloS One* **6**, 9.

Zampieri, M., Nakov, P. and Scherrer, Y. (2020), 'Natural language processing for similar languages, varieties, and dialects: A survey', *Natural Language Engineering* **26**(6), 595--612.