Diffusion Across the Grammar:

Complexity in Areal Interactions Between Dialects of English

Jonathan Dunn

Abstract. This paper experiments with the diffusion of syntactic variants within a complex system by operationalizing both (i) the network structure of the grammar and (ii) the network structure of the speech community. Drawing on both computational syntax and computational sociolinguistics, synchronic similarity measures between 505 local dialects are used to capture diachronic processes of diffusion. The results show that similarity measures (and thus the processes of diffusion that they reflect) differ according to the neighborhood of the grammar being observed. This means that the diffusion of grammatical constructions operates according to the network structure of the grammar. By implication, this means that studies of individual features in isolation are unable to accurately observe diffusion as a phenomenon in its entirety.

Keywords: computational sociolinguistics, dialectology, construction grammar, complex systems, geographic corpora

1.Introduction[1]

From a constructionist perspective, the grammar is a network that contains many individual nodes representing either (i) a single construction or (ii) a family of constructions (Diessel 2023). This network structure can be induced given observed similarities in the form or usage of constructions (Dunn 2024). If the grammar is stored as a network, then we would expect processes of diffusion to advance through that network structure: for instance, two neighboring constructions should experience the same pressure during diffusion and thus should pattern together. This paper asks whether the diffusion of syntactic structures operates on the grammar as a connected network or on the grammar as a set.

At the same time, from a sociolinguistic perspective, the speech community is also a network in which each individual is a node represented by its degree of contact within the network (Faygal, et al. 2010). In digital settings this network structure can be induced given geo-referenced corpora (Laitinen, et al. 2020). This paper draws on geo-referenced social media corpora to represent usage of English in 505 cities across 111 countries. Because English is now a *lingua franca*, speakers of English form a global speech community with long-distance connections created as a result of immigration, mass media, and digital communication platforms (cf. Dąbrowska 2021). Thus, each local population is a node in a global network of English users. If these new forms of communication create contact and exposure situations which influence diffusion, then we would expect that a global network would describe syntactic similarity differently than a restricted local network. This paper leverages these large geo-referenced corpora to model processes of diffusion that reach beyond local populations. By *diffusion* we mean the spread of constructions from one community to another; synchronic similarity between dialects must result either from previous diffusion processes or from independent drifts

[1] Supplementary material for this chapter is available at: https://doi.org/10.17605/OSF.IO/SUZY9

in the grammar. By *grammar network* we mean to view the grammar as constructions which have relationships with one another: parent and child and sibling, for instance. By *population network* we mean to view dialects as local communities which have different amounts of connection with and exposure to one another.

Even computational models of diffusion have traditionally focused on individual features in relative isolation, thus ignoring the larger network structure of the grammar (c.f., Kodner 2019; Würschinger 2021). This paper instead views language as a complex system (c.f., Beckner, et al. 2009; Schneider 2020). We ask whether processes of diffusion across the speech community network are influenced by relationships within the grammar network. For instance, if the grammar as a whole is subject to a single process of diffusion, then we would expect all nodes (i.e., constructions) within that grammar to change in the same direction, potentially at different rates. On the other hand, if each node within the grammar is subject to its own individual pressures, then one part of the grammar (e.g., transitive verbs) might be changing in the opposite direction from another part (e.g., adpositional phrases).

Hypothesis 1: *If the grammar is stored as a network, then diffusion will operate differently in each neighborhood of this network. We expect to see different similarities between dialects depending on the portion of the grammar under observation.*

For Hypothesis 1 to be considered as true, we need two find two distinct facts: First, variation must be spread across many different constructions, so that similarity relationships between local dialects depend heavily on which constructions are observed. Second, however, variation must not be spread randomly or arbitrarily across the grammar: for instance, this hypothesis does not expect that constructions will be varying uniquely as individual representations, each with its own trajectory. Instead, diffusion must be structured alongside the network structure of the grammar. In particular, this structure of the grammar has two dimensions: level of abstraction and order of emergence. Thus, the hypothesis is not simply that individual constructions will have their own trajectories of diffusion but rather that neighborhoods in the grammar will explain observed similarity networks better than simply an unorganized set of constructions without network structure.

The speech community is also a complex network: close ties within local communities will have the strongest influence, but -- if properly observed -- weaker ties like digital communication and mass media will also have an influence (c.f., Carvalho 2004). For instance, if global influence is not a significant factor in diffusion, then syntactic variation in Christchurch (NZL) should be disconnected from syntactic variation in Chicago (USA). On the other hand, if digital sources do exert an influence, then both of these locations will experience pressure from non-geographic digital sources. This, in turn, means that both will share some of the same diffusion processes. The problem with small-scale work on diffusion is that it assumes that only face-to-face conversations are interactive (Trudgill 2014) and uses networks of an unrealistic size which over-state the importance of strong vs. weak ties (Laitinen, et al. 2020).

Hypothesis 2: *If English forms a global and digital speech community, then a global model will describe variation better than a purely local model. We expect to see diffusion*

processes that create similarity relationships between non-local portions of the speech community network.

This paper uses social media data (tweets) in English to create comparable corpora representing 505 local urban areas in 111 countries. Some of these represent long-standing inner-circle or outer-circle varieties of English, but others represent newer and potentially less-stable expanding circle varieties (c.f., Kachru, 1990). These tweets are aggregated into samples of approximately 3,900 words each and the total corpus contains 82,728 samples or a little over 322 million words. Using Computational Construction Grammar as a feature space (c.f., Dunn 2018a 2019a 2019b; Dunn and Wong 2022), the paper looks at the network of similarity relationships across nodes within the grammar (c.f., Dunn 2023a). This allows us to determine whether all nodes are subject to the same similarity relationships, across 505 local dialects and approximately 2,000 families of constructions.

This work is situated within computational sociolinguistics, in which an unelicited corpus of usage is taken as a representation of dialectal production (Szmrecsanyi 2013; Grieve 2016). This tradition depends on corpora that represent specific dialect communities, whether compiled manually (Greenbaum 1996) or drawn from web pages (Davies 2013; Davies & Fuchs 2015; Cook and Brinton 2017) or pulled from geo-referenced social media posts (Wieling, et al. 2011; Mocanu, et al. 2013; Gonçalves & Sánchez 2014; Eisenstein, et al. 2014; Donoso, et al. 2017). Work in computational sociolinguistics has established that lexical usage on social media mirrors usage as captured by dialect surveys (Grieve et al. 2019) and also that digital communities develop their own non-geographic variants (Lucy and Bamman 2021). Because corpora provide more samples than traditional methods and can be connected with social networks, computational sociolinguistics has also found that categories like gender are more complex when viewed at scale (Bamman et al. 2014). Other work has looked at the influence that inner-circle varieties like American English have in digital spaces (Gonçalves et al. 2018) and at how deeper syntactic variation can be modelled as differences in constraint rankings (Grafmiller and Szmrecsanyi 2018; Szmrecsanyi and Grafmiller 2023). This present paper expands on this corpus-based tradition by viewing local dialects as a network in which each community is exposed to its own usage and to the usage of near-by communities. Only corpus-based evidence could capture these global relationships between local dialects of English.

The idea behind this approach is that increased diffusion between two dialects will result in increased similarity, so that synchronic similarity measures can be used to reconstruct diachronic diffusion processes. The theoretical question is whether the entire grammar is subject to a single, central process of diffusion or whether there is an interaction between the network structure of the speech community and the network structure of the grammar itself. If this study were relying on a small number of manually curated syntactic features, then the answer to this hypothesis would be true but trivial: we already know that individual features can have their own trajectories. But the question here goes deeper in two ways: First, we are viewing the grammar as a complex system and the question is whether there are processes of diffusion which extend beyond individual features and instead apply to large neighborhoods within the grammar. This has never before been tested at scale and thus it remains an open empirical question. Second, construction grammar makes specific claims about the network structure of the grammar; it is possible

that diffusion takes place across large neighborhoods within the grammar but that these pathways of diffusion are not those pathways expected by CxG. Thus, the hypothesis here is not just that diffusion operates as a large-scale pressure on the grammar network but that it does so in a way that is specifically organized around level of abstraction and order of emergence.

At the same time, there is a trade-off in which the geographic breadth provided by social media comes with a shallow time depth. For instance, if different nodes within the grammar were subject to unique rates of diffusion, this would appear within a single time period to be decentralized diffusion when actually it is a single centralized process progressing at different rates. It remains the case, however, that exposure to variants takes place at each period of time. Thus, if diffusion were happening in one direction at different rates this would still mean that some communities are experiencing very different input as learners and this, on its own, would lead to a differentiating grammar. The basic point is that taking seriously the idea that language is a complex system requires grappling with this question of how processes of diffusion operate across the grammar-as-a-network.

Our methodology uses different portions of the grammar (c.f., Section 3) to measure similarity between local populations (defined as individual metro areas). We have two basic measures: (1) *homogeneity* is self-similarity across multiple samples representing the same local population; (2) *pairwise distance* is the similarity between two local populations, again measured across multiple samples. Because there are 505 local populations represented in these experiments, there are 127,260 pairwise relationships. This creates a network that allows us to test which factors influence diffusion, where the result of diffusion is greater similarity between two local populations. The geographic corpora are discussed further in Section 2 and the construction-based syntactic features in Section 3. The methodology for calculating homogeneity and pairwise distance is described in Section 4. The results are then analyzed in Section 5 (focusing on the first hypothesis, that diffusion follows the network structure of the grammar) and Section 6 (focusing on the second hypothesis, that users of English form a global speech community). The supplementary material is available as an open-source repository.[2]

2. Geographic Corpus Data

The corpus data consists of geo-referenced tweets derived from the *Corpus of Global Language Use* (Dunn 2020). The social media portion of this corpus contains publicly-accessible tweets collected from 10k cities around the world. Each city is a point with a 25km collection radius. The location of each tweet is taken from its original co-ordinates, as provided by the Twitter API. Tweets are tagged for language using both the idNet model (Dunn, 2020) and the PacificLID model (Dunn and Nijhof 2022); only tweets which both models predict to be English are included.

The metro areas that represent local populations are centered around international airports. The corpus is assigned to airports by finding the nearest airport to the collection point, with the constraint that the nearest airport must be within the same country and within 100km. The advantage of using airports to represent local populations is that the

[2] https://doi.org/10.17605/OSF.IO/SUZY9

selection of cities is guided by the presence of a major airport, a criterion that can be employed across many countries. The collection points, airport locations, and break-down of corpus size and amount of travel by airport are available in the supplementary materials.

The goal here is to derive comparable corpora that represent each local population using geo-referenced tweets. The challenge is that social media data represents many topics and sub-registers, so that there is a possible confound presented by geographically-structured variations in topic or sub-register. For instance, if tweets from Chicago are sports-related and tweets from Christchurch are business-related, then observed variation is likely to be partially register-based as well as dialect-based. To control for topic, we create samples by aggregating tweets which contain the same set of keywords.

First, we select 250 common words which are neither purely topical nor purely functional, for example: *girl*, *know*, *music*, and *project*. These keywords are chosen by inspecting the most frequent lexical items in the larger collection of tweets and discarding those which are too place-specific (like sports teams) or too grammaticalized. Second, for each local metro area we create samples containing one tweet for each keyword; each sample thus contains 250 individual tweets, for a total size of approximately 3,900 words. Importantly, the distribution of keywords is uniform across all samples from all local areas. This allows us to control for variations in topic or sub-register which might otherwise lead to non-dialectal sources of variation. The keywords used for selection for shown in Appendix 1; this corpus has been previously used for studies of variation (Dunn 2023; Dunn et al. 2024).

Table 1. Number of Countries, Cities, and Samples Per Region

Region	**N. Countries**	**N. Cities**	**N. Samples**
Africa, North	7	8	748
Africa, Southern	3	17	3,303
Africa, Sub-Saharan	13	40	4,834
America, Central	10	23	4,189
America, North	2	96	11,331
America, South	9	17	850
Asia, Central	6	8	981
Asia, East	4	18	2,138
Asia, South	6	73	13,777
Asia, Southeast	10	36	6,333
Europe, East	14	43	4,960
Europe, West	18	94	19,377

Middle East	7	11	1,761
Oceania	2	21	8,146
TOTAL	**111**	**505**	**82,728**

The method used for comparing local populations requires sampling 100 unique pairs of sub-corpora representing each metro area (c.f., Section 4). As a result, we only retain metro areas which have at least 25 unique samples (each a sub-corpus of approximately 3,900 words). This provides a total of 505 local populations across 111 countries, as shown in Table 1. More samples come from Western Europe (19k), South Asia (13k), North America (11k), and Oceania (8k) than from other regions. Among inner-circle countries, Canada has 4,261 samples across 24 cities; the United States has 7,070 samples across 72 cities; and the UK has 5,071 samples across 25 cities. The complete inventory of cities along with the number of samples from each is available in the supplementary materials.

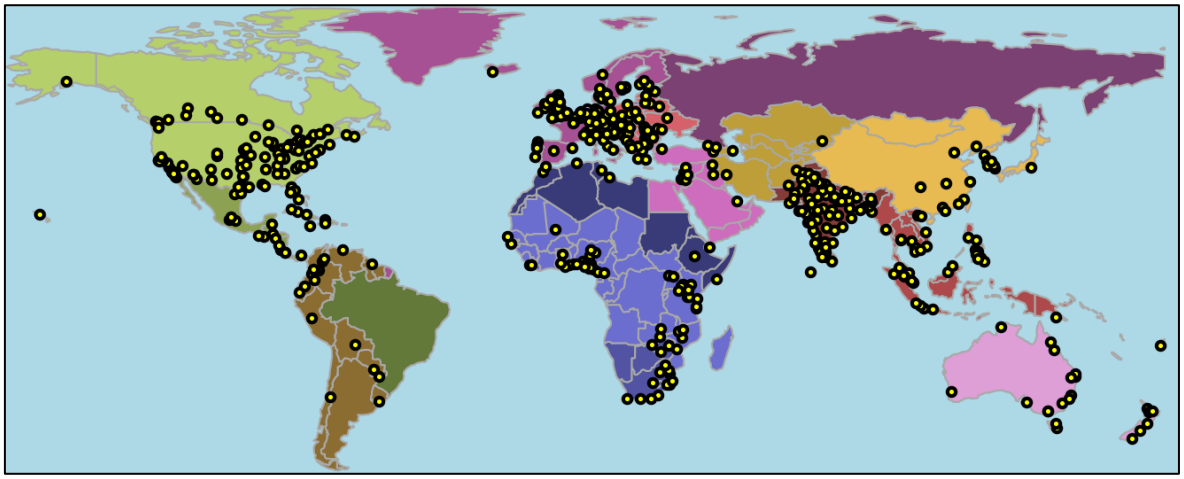

Figure 1. Map of Cities by Country and Region.

The distribution of local populations is mapped in Figure 1. Each region is represented by a distinct color; for example, North American countries are shaded in a pale green. Each point represents a local metro area, giving an indication of the geographic distances represented in the corpus. Given that this study only represents English usage, there are more metro areas in historically English-speaking countries simply because these locations produce more corpora in English.

This corpus represents variation over local dialect communities within a single written register, social media. The advantage of this approach is that, by focusing on a single register, we do not mistake register variation for dialectal variation. And yet even within a single digital register there are many sub-registers that are difficult to distinguish (Egbert et al. 2015; Biber et al. 2020). This difference in nuanced sub-registers within a single source (like political announcement tweets vs personal communication tweets) is controlled for by selecting for specific keywords. The side-effect of controlling for register, however, is that we are only observing dialectal variation within a single register. We could imagine, for instance, that certain abstract constructions are more complex in other registers than social media so that our view of the diffusion of highly abstract

constructions is not representative of, say, abstract constructions in scientific discourse. This is an unavoidable result of focusing on a single register in order to control for register variation itself.

3. Constructions as Syntactic Features

The basic approach in this paper is to use an unsupervised grammar derived from the Construction Grammar paradigm (CxG) as a feature space for measuring the distance between local dialects of English. This section describes the grammar which is used as a feature space and the next section describes the distance measures. The network structure of the grammar is made up of inheritance relations (mother-child) and similarity relations (sibling-sibling) between constructions. A *construction* in this context is a symbolic mapping between form and meaning, where an individual construction is unique either syntactically or semantically. For example, there is an inheritance relationship between the schematic ditransitive construction, with examples like (1a), and idiomatic constructions, with examples like (1b) and (1c). While some of the properties of the ditransitive are inherited by these idiomatic children, they also retain unique and non-compositional meanings.

(1a) “write the store a check”
(1b) “give me a hand”
(1c) “give me a break”

A construction grammar is a network of form-meaning mappings at various levels of schematicity (Goldberg 2006) which are built from learned rather than innate ontologies of slot constraints (Croft 2013). The grammar is a network with inheritance relationships and similarity relationships between pairs of constructions (Diessel 2023). CxG is a usage-based approach to syntax which, in practical terms, means that more item-specific constructions are learned first and then generalized into more schematic constructions (Nevens et al. 2022; Doumen et al. 2023). Variation in the usage of constructions is related to both the level of abstraction (schematicity) and the degree of centrality (how core or peripheral a construction is; Hollmann and Siewierska 2011). While most work in CxG has focused on differences in language-internal meaning, we can also think about the indexicality or social meaning of constructions as a way in which constructions are differentiated (Leclercq and Morin 2023). A final dimension of the grammar network is entrenchment: some constructions are core and others are peripheral.

Why use CxG for dialectology? The main reason is that the flexibility of usage-based representations works better for describing syntactic variation across dialects (Dunn 2019). Constructions are also flexible enough to describe individual differences within communities (Anthonissen 2020). The fact that CxG has been successful in previous work in describing syntactic variation on a large scale does not mean that other paradigms like dependency grammar or head-driven phrase structure grammar might not be adapted to perform equally well on this task. Thus, the performance of CxG in this area is an argument in support of a constructional approach rather than an argument against other syntactic theories. In order to use variation as a task for evaluating competing syntactic paradigms, it would be necessary to adequately operationalize each paradigm and test it against the same data. Given that the current paper is focused on a constructional approach

to variation, however, we leave for other researchers the task of determining whether some other syntactic paradigm is also capable of describing variation.[3]

The grammar learning algorithm used in this paper is taken from previous work (Dunn 2017, 2018b, 2019c, 2022; Dunn and Nini 2021, Dunn and Tayyar Madabushi 2021). The specific grammar used is trained from tweets, the same register as the dialectal data. Rather than describe the computational details of this line of work, this section instead analyzes constructions within the grammar as examples of the kinds of features used to model syntactic variation. The complete grammar together with examples is available in the supplementary material and the codebase for computational CxG is available as a Python package.[4]

There is a long history of computational models of the emergence of construction grammars, from specific forms like argument structure (Alishahi and Stevenson, 2008; Barak and Goldberg 2017) to shallow but wide-coverage models of holophrase constructions (Wible and Tsao 2010; 2020) to template-based methods that build on annotations and introspection (Perek and Patten 2019). Some research has taken a more natural agent-based approach in which individual learners acquire constructions during communication until all observed utterances can be successfully interpreted (Beuls and Van Eecke 2023). This body of work as a whole can be viewed as a discovery-device grammar (Goldsmith 2015) which formulates syntactic representations given exposure to a corpus; thus, CxG as a theory can be viewed as a mapping between exposure and an emergent grammar.

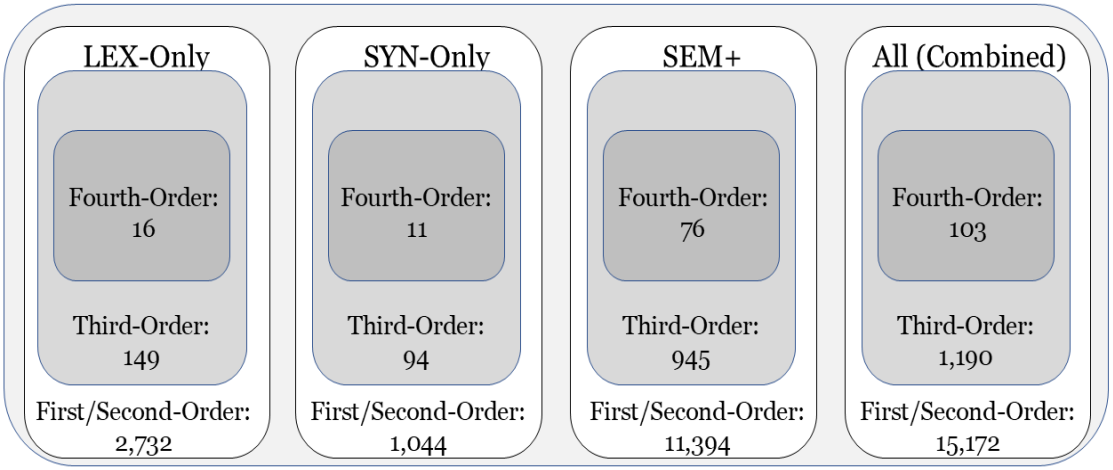

Figure 2. Break-Down of Grammar into Nodes by Type of Representation and Level of Abstraction

A break-down of the grammar used in the experiments is shown in Figure 2, containing a total of 15,172 individual constructions. Constructions are represented as a series of slot-constraints and the first distinction between constructions involves the types of constraints used. Computational CxG uses three types of slot-fillers: lexical (LEX, for item-specific constraints), syntactic (SYN, for form-based or local co-occurrence constraints), and semantic (SEM, for meaning-based or long-distance co-occurrence constraints). As shown in (2), slots are separated by dashes in the notation used here.

[3] The test corpora are available at: https://doi.org/10.17605/OSF.IO/SUZY9
[4] https://github.com/jonathandunn/c2xg/tree/v2.0

Thus, SYN in (2) describes the type of constraint and *determined-permitted* provides its value using two central exemplars of that constraint. Examples or tokens of the construction from a test corpus of tweets are shown in (2a) through (2d).

(2) [SYN: *determined-permitted* -- SYN: to -- SYN: *pushover-backtrack*]
(2a) “refused to play”
(2b) “tried to watch”
(2c) “trying to run”
(2d) “continue to drive”

Thus, the construction in (2) contains three slots, each defined using a syntactic constraint. These constraints are categories learned at the same time that the grammar itself is learned, formulated within an embedding space. An embedding that captures local co-occurrence information is used for formulating syntactic constraints (a continuous bag-of-words fastText model with a window size of 1) while an embedding which instead captures long-distance co-occurrence information is used for formulating semantic constraints (a skip-gram fastText model with a window size of 5)[5]. Constraints are then formulated as centroids within that embedding space. Thus, the tokens for the construction in (2) are shown in (2a) through (2d). For the first slot-constraint, the name (*determined-permitted*) is derived from the lexical items closest to the centroid of the constraint. The proto-type structure of categories is modeled using cosine distance as a measure of how well a particular slot-filler satisfies the constraint. Here the lexical items “reluctant”, “ready”, “refusal”, and “willingness” appear as fillers sufficiently close to the centroid to satisfy the slot-constraint. The construction itself is a complex verb phrase in which the main verb encodes the agent's attempts to carry out the event encoded in the infinitive verb. This can be contrasted semantically with the construction in (3), which has the same form, but instead encodes the agent's preparation for carrying out the social action encoded in the infinitive verb**.** Constraints do not need to conform to traditional syntactic categories; for instance, here the governing entity can be either an adjective (3a-b) or a noun (3c-d).

(3) [SYN: *determined-permitted* -- SYN: *to* -- SYN: *demonstrate-reiterate*]
(3a) “reluctant to speak”
(3b) “ready to exercise”
(3c) “refusal to recognize”
(3d) “willingness to govern”

An important idea in CxG is that structure is learned gradually, starting with item-specific surface forms and moving to increasingly schematic and productive constructions. This is called scaffolded learning because the grammar has access to its own previous analysis for the purpose of building more complex constructions (Dunn, 2022). In computational CxG this is modelled by learning over iterations with different sets of constraints available. For example, the constructions in (2) and (3) are learned with only access to the syntactic constraints, while the constructions in (4) and (5) have access to lexical and semantic constraints as well. This allows grammars to become more complex while not assuming basic structures or categorizations until they have been learned. In the

[5] This window size of 5 is a typical default setting because it provides enough context to model semantics but remains narrow enough to be tractable. Examinations of windows size are available in other work (Levy, et al., 2015).

measurements of dialect similarity (Sections 5 and 6), we distinguish between different nodes within the grammar based on the stage of learning and the level of abstraction (c.f., Figure 2).

(4) [LEX: “the” -- SEM: *way* -- LEX: “to”]
(4a) “the chance to”
(4b) “the way to”
(4c) “the path to”
(4d) “the steps to”

Constructions have different levels of abstractness or schematicity. For example, the construction in (4) functions as a modifier, as in the X position in the sentence “Tell me [X] bake yeast bread.” This construction is not purely item-specific because it has multiple types or examples. But it is less productive than the location-based noun phrase construction in (5) which will have many more types in a corpus of the same size. A modifier in this case does not need to be a traditional constituent but rather a catenae (Osborne and Gross 2012) which provides additional specification about its head. CxG is a form of lexico-grammar in the sense that there is a continuum between item-specific and schematic constructions, exemplified here by (4) and (5), respectively. The existence of constructions at different levels of abstraction makes it especially important to view the grammar as a network with similar constructions arranged in local nodes within the grammar. This approach focuses on the form of the construction; much like words, a construction could have different functions depending on context.

(5) [LEX: “the” -- SEM: *streets*]
(5a) “the street”
(5b) “the sidewalk”
(5c) “the pavement”
(5d) “the avenues”

A grammar or constructicon is not simply a set of constructions but rather a network with both taxonomic and similarity relationships between constructions. In computational CxG this is modelled by using pairwise similarity relationships between constructions at two levels: (i) representational similarity (how similar are the slot-constraints which define the construction) and (ii) token-based similarity (how similar are the examples or tokens of two constructions given a test corpus). Matrices of these two pairwise similarity measures are used to cluster constructions into smaller and then larger groups. For example, the phrasal verbs in (6) through (8) are members of a single cluster of phrasal verbs. Each individual construction has a specific meaning: in (6), focusing on the social attributes of a communication event; in (7), focusing on a horizontally-situated motion event; in (8), focusing on a motion event interpreted as a social state. These constructions each have a unique meaning but a shared form. The point here is that, at a higher-order of structure, there are a number of phrasal verb constructions which share the same schema. These constructions have sibling relationships with other phrasal verbs and a taxonomic relationship with the more schematic phrasal verb construction. These phrasal verbs are an example of a *third-order construction* referenced in the dialect experiments below (c.f. Dunn 2024). Thus, a similarity measure based on third-order constructions views all of these specific phrasal verbs as instances of the same structure. In this case,

each first-order construction has a unique meaning while all first-order constructions in this third-order family share the same form.

(6) [SEM: *screaming-yelling* -- SYN: *through*]
(6a) “stomping around”
(6b) “cackling on”
(6c) “shouting out”
(6d) “drooling over”

(7) [SEM: *rolled-turned* -- SYN: *through*]
(7a) “rolling out”
(7b) “slid around”
(7c) “wiped out”
(7d) “swept through”

(8) [SEM: *sticking-hanging* -- SYN: *through*]
(8a) “poking around”
(8b) “hanging out”
(8c) “stick around”
(8d) “hanging around”

An even larger structure within the grammar is based on groups of these third-order, structures which we will call *fourth-order*. Such a fourth-order construction is much more abstract because it contains many sub-clusters which themselves contain individual constructions. Viewed as a taxonomy, first-order and second-order constructions are children which belong to different families; third-order constructions are immediate families and fourth-order constructions are more distantly-related families.

An example of a fourth-order construction is given with five constructions in (9) through (13) which all belong to same neighborhood of the grammar. The partial noun phrase in (9) points to a particular sub-set of some entity (as in, “parts of the recording"). The partial adpositional phrase in (10) points specifically to the end of some temporal entity (as in, “towards the end of the show"). In contrast, the partial noun phrase in (11) points a particular sub-set of a spatial location (as in, “the edge of the sofa"). A more specific noun phrase in (12) points to a sub-set of a spatial location with a fixed level of granularity (i.e., at the level of a city or state). And, finally, in (13) an adpositional phrase points to a location within a spatial object. The basic idea here is to use these third-order and fourth-order constructions as features for dialect similarity in order to find out the level of abstraction at which diffusion operates.

(9) [SEM: *part* -- lex: “of” -- SYN: *the*]
(9a) “parts of the”
(9b) “portion of the”
(9c) “class of the”
(9d) “division of the”

(10) [SYN: *through* -- SEM: *which-whereas* -- LEX: “end” -- LEX: “of” -- SYN: *the*]
(10a) “at the end of the”
(10b) “before the end of the”
(10c) “towards the end of the”

(11) [SEM: *which-whereas* -- SEM: *way* -- LEX: “of”]
(11a) “the edge of”
(11b) “the side of”
(11c) “the corner of”
(11d) “the stretch of”

(12) [SEM: *which-whereas* -- SYN: *southside-northside* -- SYN: *chicagoland*]
(12a) “in north texas”
(12b) “of southern california”
(12c) “in downtown dallas”
(12d) “the southside chicago”

(13) [LEX: “of” -- SYN: *the* -- SYN: *courtyard-balcony*]
(13a) “of the gorge”
(13b) “of the closet”
(13c) “of the room”
(13d) “of the palace”

The examples in this section have illustrated some of the fundamental properties of CxG and also provide a discussion of some features which are used in the dialect similarity measures. A more detailed linguistic examination of the contents of a grammar like this is available elsewhere (Dunn 2023b). A break-down of the contents of the grammar is shown in Figure 2 and in Table 2. The 15,172 total constructions are first divided into different scaffolds depending on the type of representation (LEX-only, SYN-only, SEM+, and All Constructions). Earlier stages, with fewer types of representation, tend to have fewer simpler constructions. LEX-Only constructions are the earliest stage because they make no reference to word classes, only to individual word-forms. SYN-Only constructions are a middle stage which make reference to word classes which are based only on local co-occurrences. And, finally, SEM+ are late-stage constructions which make reference to word classes which also have access to non-local relationships. This division by stage of emergence is based directly on computational experiments. Thus, one way of exploring the network structure of the grammar has to do with the type of representation and scaffolded learning, reflected in Table 2 by row.

Table 2. Breakdown of Number of CxNs in the 12 Feature Sets

	First/Second-Order	**Third-Order**	**Fourth-Order**
LEX-Only	2,732	149	16
SYN-Only	1,044	94	11
SEM+	11,394	945	76
All Constructions	15,170	1,188	103

The other way of navigating the network structure of the grammar relies on level of abstraction. First-order constructions are individual items (i.e., children sharing the same mother), while third-order and fourth-order constructions are families of related constructions (i.e., a mother with many children). This is reflected by column in Table 2,

with fewer mothers than children: for SYN-only constructions, there are 1,044 children branching from 94 mothers in 11 families. We can thus look at variation across the entire grammar, across different levels of scaffolded structure, and across different levels of abstraction. The main reason for doing this is to determine whether all nodes within the grammar are subject to the same processes of diffusion. While we could view the grammar as a complete network, with each construction as a node, we simplify by instead dividing the grammar into larger nodes (collections of constructions) based on the level of abstraction (e.g., first-order vs fourth-order) and the order of emergence (e.g., lexical vs syntactic). This level of detail is sufficient for testing our hypotheses.

In other words, the grammar produces twelve distinct feature sets depending on (i) the type of representation and (ii) the level of abstractness. These twelve feature-driven conditions are shown in Table 2 together with the number of constructions in each. This approach allows us to examine different portions of the grammar. For instance, the SYN-only grammar contains 1,044 first-order and second-order constructions. A distance measure with access to these features could rely on differences in the usage of this particular portion of the grammar network, without reference to its family structure or to other portions of the grammar. The third-order SYN-only grammar is a more abstract level of the network, which makes distinctions between only 94 families of constructions. A distance measure with access to these features would thus not be able to discern differences in lower-level constructions which have the same mother. Essentially, this means that we can ignore variation below a certain level of abstraction by focusing on higher-order constructions.

4. Measuring Dialect Similarity

Given this grammar, we divide the network into twelve distinct sets of features in order to determine whether similarity relations between local populations are consistent depending on which part of the grammar we are observing. This section describes the method of measuring similarity/distance. The basic idea is to first parse each sample for each construction and then to use the type frequency or the token frequency as specific measures of the usage of that construction. Parsing here is automated using the c2xg package.[6] *Token* frequency is more common in computational work, providing the number of uses of a construction regardless of its manifestation. *Type* frequency, on the other hand, only counts unique manifestations of a construction and thus provides a better view of productivity: an unproductive construction would have a high token frequency (with many instances observed) but a low type frequency (with all those instances taking the same form). For each sample, then, we extract the type and token frequencies for each construction in each of these twelve subsets of the grammar.

Drawing from work in authorship analysis, we compare the distance between samples in these grammatical feature spaces using Burrow's Delta (Evert et al. 2017). This distance metric has four steps: First, we find the expected frequency of each construction in the corpus as a whole; here we estimate the expected frequency by parsing 20k samples. Second, we use the expected frequency to convert each constructional feature into its z-score normalized form. Now we are viewing each feature in terms of whether it is more common or less common than expected. The advantage of standardizing here is that some

[6] https://github.com/jonathandunn/c2xg

constructions are much more frequent overall and thus would dominate the similarity measure. Burrow's Delta is a method of removing this frequency bias without relying on trained feature weights from a supervised classifier. Third, we calculate Euclidean distance between two samples in this standardized space. Euclidean distance as a metric accumulates differences in the (standardized) frequency across features; this means that the average distance in a larger feature space will be higher. For example, the full grammar contains 15,172 constructions while the higher-order SYN-only grammar contains only 94 more abstract constructions. This means that the average distance for the one grammar will be much higher than the other. Thus, the fourth step is to standardize the distances within each feature space across all samples using the z-score. This means that each node-specific distance measure is directly comparable, essentially ranking pairs of samples from least similar to most similar.

We have many samples from each local metro area. Rather than calculate a single distance between two populations, then, we instead use many individual samples to compute a distribution of distances. For each pair (for example, Christchurch and Chicago) we randomly select 100 unique pairs of samples. We then calculate Burrow's Delta between each of these pairs. This gives us a distribution of distance measures so that we can also observe variation in distances. In some of the analysis we work with this distribution directly (e.g., in the homogeneity plots in Figure 6). In other parts we calculate the mean distance using a Bayesian estimate with a 95% confidence interval. But the basic idea is that we use 100 observations for each comparison in order to provide a more robust view of the grammatical distance between two local metro areas.

We use this basic measurement in order to capture two distinct properties: *homogeneity* and *pairwise distance*. The first measures variation within a single metro area. For example, a place like Chicago has a large and diverse population and thus we would expect speakers from Chicago to differ from one another. We capture this by comparing 100 unique pairs from the same city, to provide a view towards of the internal variation or homogeneity of that local population. The second, pairwise distance, measures the difference in grammar between two distinct populations; for example, we can use this to find out how different Chicago and Christchurch are across different nodes of the grammar.

Table 3. Breakdown of Measures

Representations	***Abstraction***	***Frequency***	***Measure***
LEX-Only	First/Second-Order	Tokens	Homogeneity
SYN-Only	Third-Order	Types	Pairwise Distance
SEM+	Fourth-Order		
All Constructions			

The range of measures used is summarized in Table 3. The first two columns represent different parts of the grammar: divided by representation and divided by level of abstraction. This provides the twelve distinct feature sets. The third column captures the

way that frequency is calculated in a given sample, either by the number of construction tokens or by the number of construction types. And, finally, the fourth column captures what is being measured, the amount of internal variation within a single local population or the difference between two distinct populations. These are the basic observables which we use to look at the diffusion of grammatical variants.

5. Hypothesis 1: Diffusion Differs Across the Grammar

Our first hypothesis is that, if the grammar is inherently structured as a network, then processes of diffusion should operate on different parts of that network in very different ways. Viewed in synchronic terms, this hypothesis predicts that both homogeneity (within local populations) and pairwise distance (between local populations) will vary according to the portion of the grammar they represent. This predicts both that constructions do not spread as a single unit (the entire grammar) and also that constructions do not spread randomly (one at a time). We expect that nodes within the grammar are subject to diffusion, so that processes of diffusion spread according to the links within the network. We operationalize this hypothesis using the twelve segments of the grammar described in Section 3.

If diffusion takes place without reference to the network structure of the grammar, then there should be a very close relationship between distances calculated on different nodes of grammar. We operationalize this using the Pearson correlation between distance measures representing different portions of the grammar. A high correlation means that the two nodes within the grammar agree on which local dialects are more similar, which in turn means that these two dialects have experienced the same processes of diffusion. But a low correlation means that two nodes in the grammar disagree on which local dialects are more similar, and they disagree because diffusion has occurred in one part of the grammar but not the other. We focus here on a linear measure of correlation, leaving more complex non-linear models as a matter for future work.

5.1. Internal Variation Depends on the Grammatical Neighborhood

We start by looking at the global correlation between homogeneity scores for all 505 cities using token frequencies across different portions of the grammar. For all cities, we calculate self-similarity using 100 unique pairs of samples from the same place. Each distance measure here is Burrow's Delta given a different portion of the grammar; this measure essentially ranks cities from the most homogeneous to the most heterogeneous (with one ranking for each sub-set of the grammar). More homogeneous cities have less internal variation.

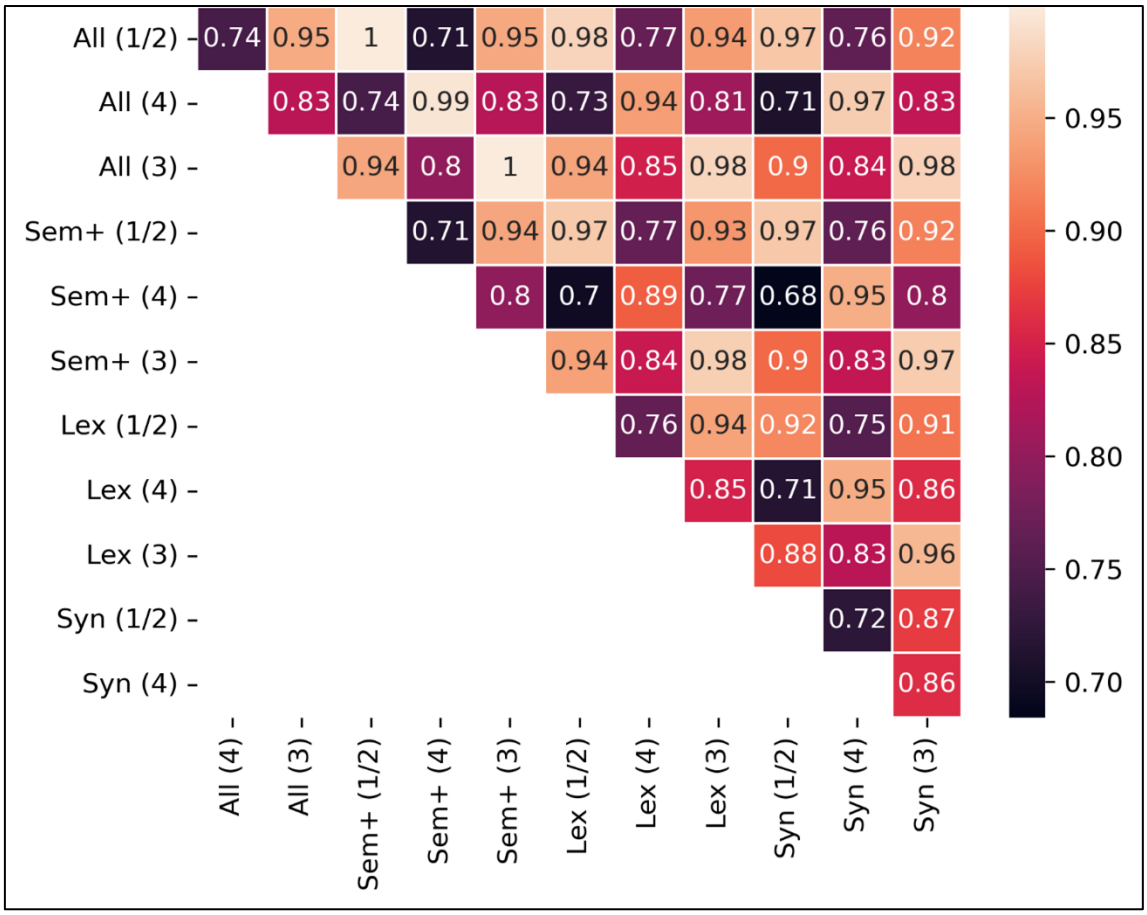

Figure 3. Correlations Between Homogeneity Ranks by Feature Type for All Regions using Token Frequency. A high correlation here means that two parts of the grammar agree on ranks of most homogeneous to least homogenous cities.

The heatmap in Figure 3 shows the correlation across homogeneity ranks given different portions of the grammar. The lower the correlation, the more two parts of the grammar deviate. For instance, there is a very close relationship (0.99) between the fourth-order levels of the grammar containing all three types of slot-constraints (SEM+) and the grammar containing all constructions (All). This is not surprising because there is a great deal of overlap in these constructions themselves. On the other hand, there is a surprising amount of divergence between low-level syntactic-only constructions (*SYN* 1/2) and high-level syntactic-only constructions (SYN 4) -- a correlation of only 0.72. This indicates that there is internal variation within local populations which is captured by lower-order constructions but not by higher-order constructions. There is a higher degree of overlap (0.92) between lexical-only and syntactic-only constructions at the same level of abstraction, indicating the abstraction is more important here than the type of representation.

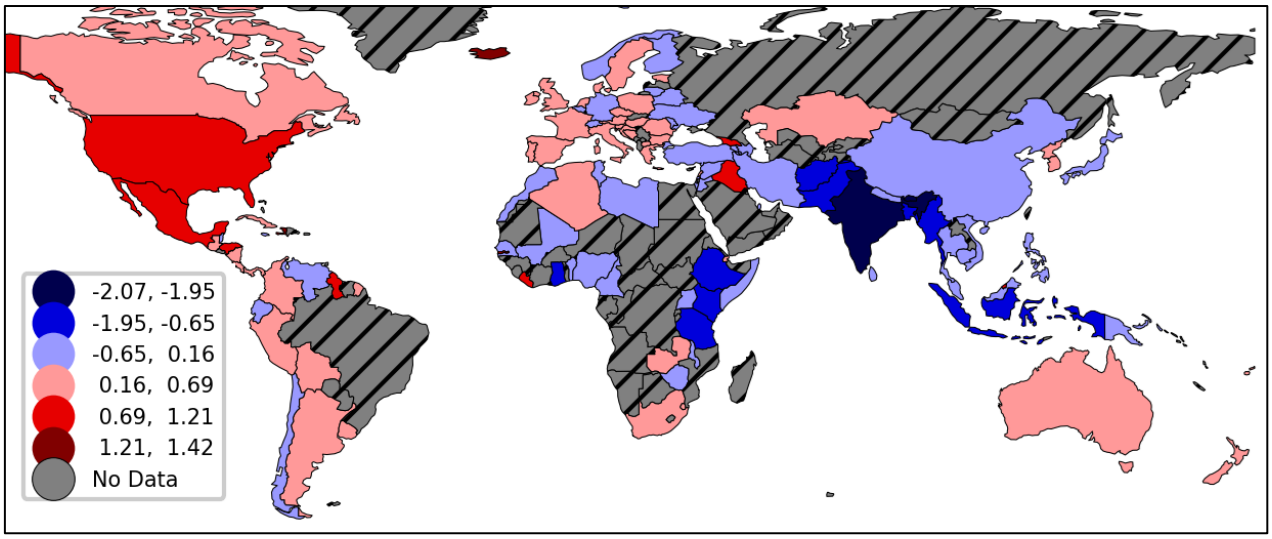

Figure 4a. Map of homogeneity by country, SYN (1/2). Higher values are more heterogeneous which indicates more internal variation within cities.

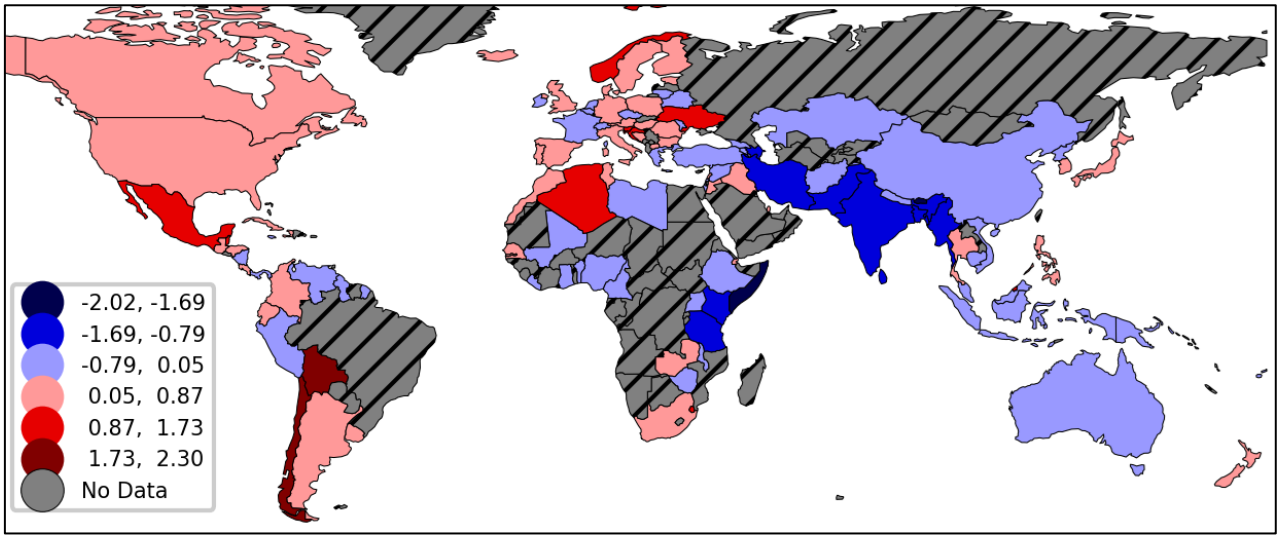

Figure 4b. Map of homogeneity by country, SYN (4). Higher values are more heterogeneous which indicates more internal variation within cities.

The speech community is a network in the same way that the grammar is a network. Thus, we take a closer look at *where* homogeneity differs by level of abstraction in Figures 4a and 4b. The first shows standardized homogeneity for the first-order SYN(1/2) set of features; these are less abstract constructions. The second shows the same measure for fourth-order SYN(4) features; these are more abstract families. This measure of homogeneity is based on a distance measure (Burrow's Delta) which means that values above 1 are a standard deviation above the mean, thus more heterogeneous than average. In the lower-order features, the United States is especially subject to internal variation and India is especially devoid of internal variation. But in the more abstract features, each country matches its surrounding neighbors, showing a regional pattern of homogeneity. In other words, if we rank countries by how much internal variation they have, the US is at the top of the rank with lower-order features but not with higher-order features. This means that the variation is concentrated in one part of the grammar.

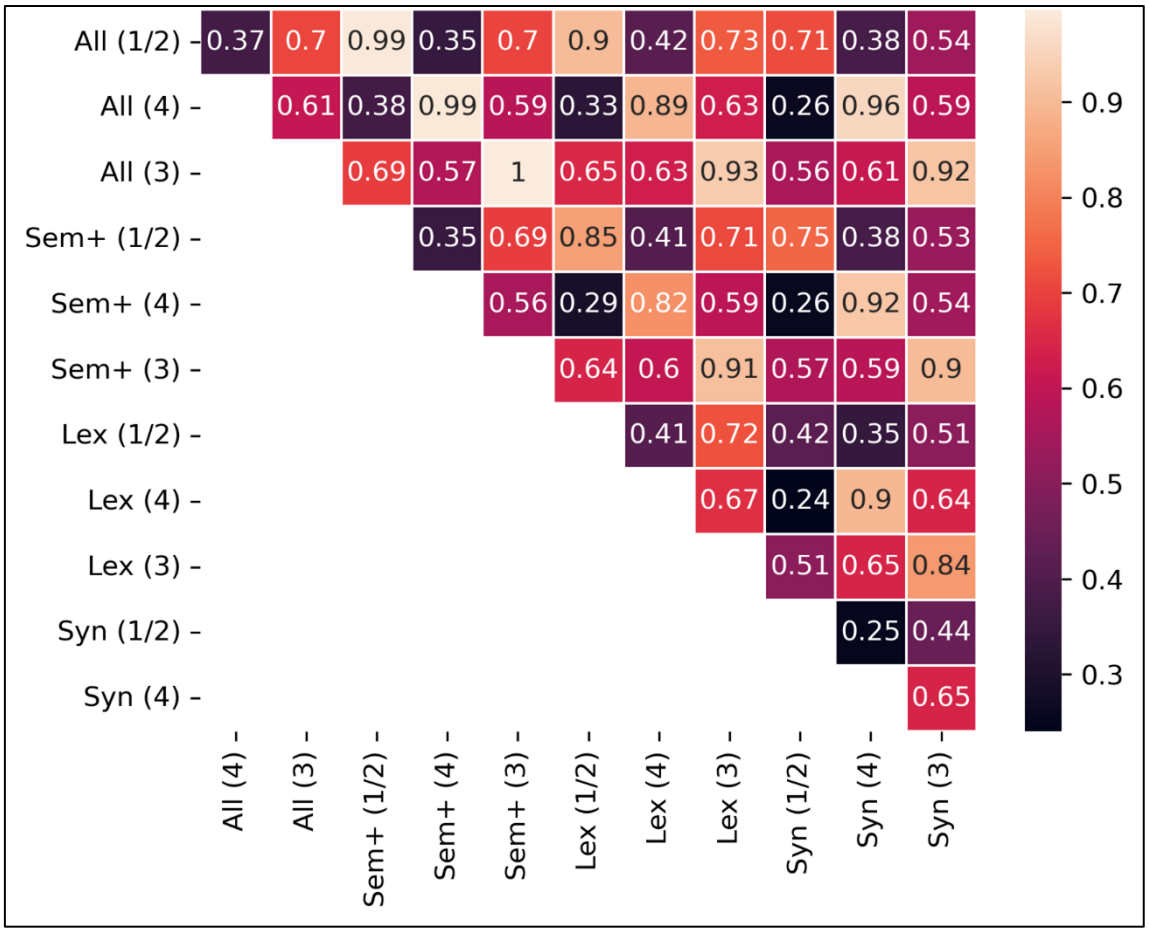

Figure 5. Correlations Between Homogeneity Ranks by Feature Type for only North American cities using Token Frequency. A high correlation here means that two parts of the grammar agree on ranks of most homogeneous to least homogenous cities.

Since North America is a source of high internal variation, we look at correlations across feature sets in Figure 5, focusing only on the 96 cities in that region. Here the correlations are markedly lower; for instance, the same syntactic-only constructions have only 0.25 correlation across levels of abstraction in terms of how they rank cities by homogeneity. There is a similar low correlation of 0.37 between the non-abstract All (1/2) feature set and the abstract All (4) feature set. As before, a low correlation indicates that these sub-sections of the grammar show different patterns of internal variation which, in turn, indicates that only some variants have spread within local populations. Thus, these low correlations match the prediction that diffusion follows the network structure of the grammar, creating different similarity relations between grammar depending on which portion of the network we observe.

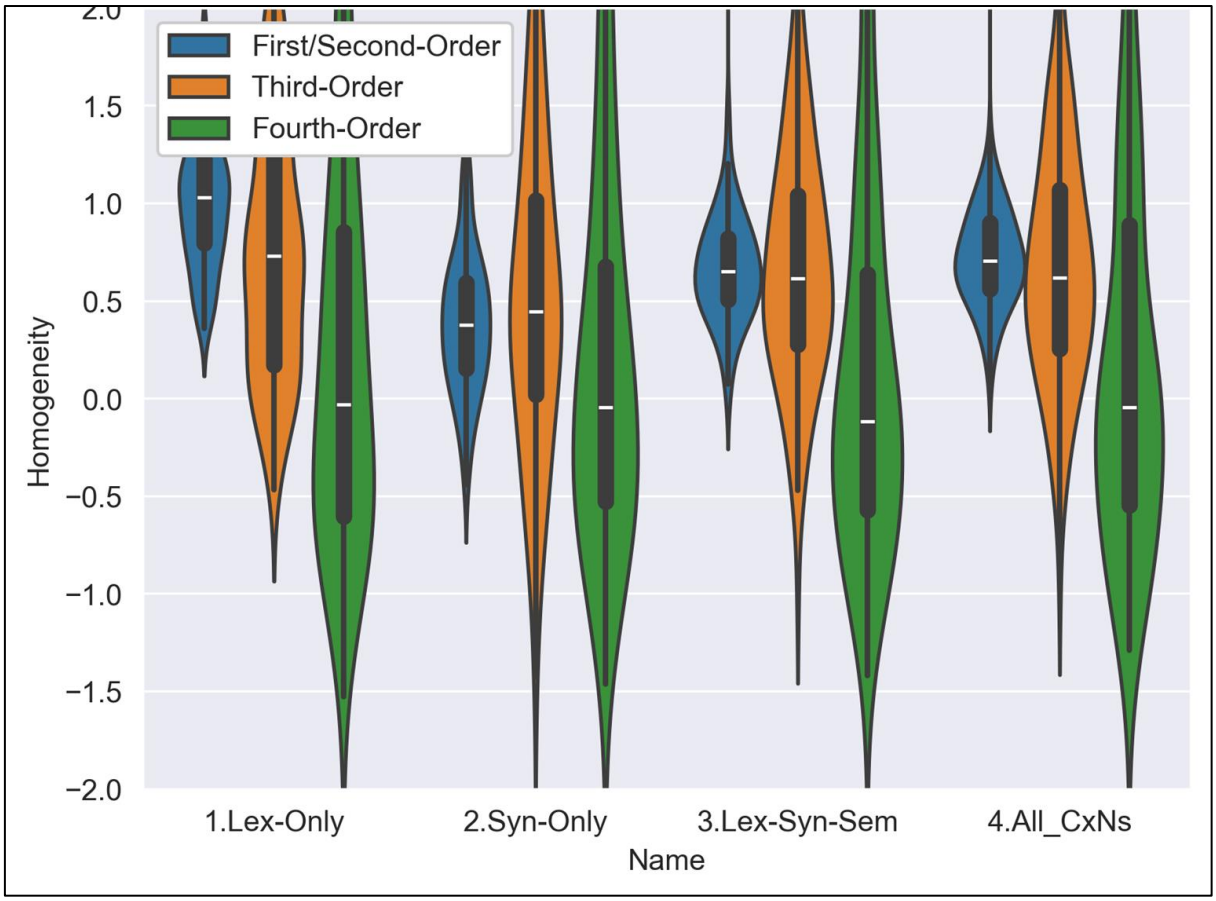

Figure 6. Full Distribution of Homogeneity Across Feature Sets for Chicago, using Token Frequency

Another view of this is shown in Figure 6, which shows a violin plot of the distribution of self-similarity (homogeneity) values for one local population in the US, Chicago. Each representation-based segmentation of the grammar is represented as a column, starting with lexical-only constructions and ending with all constructions. Within each column, the least abstract first-order constructions are in blue and the most abstract fourth-order constructions are in green. Higher values (e.g., 1.0) represent greater distance within the same city and thus a more heterogenous local population.

This figure thus shows a more detailed look at the distribution of homogeneity scores within a single location. First, we see that less abstract constructions (such as lexical and first-order constructions) have more internal variation. Here, this means that these constructions have not spread within the local population, some members of the population using the constructions and other members not using them. More abstract portions of the grammar (for instance, syntactic-only fourth-order constructions) are more homogenous on average (higher in the figure) but also have a greater standard deviation (a wider spread of values). This supports an item-specific theory of diffusion, in which item-specific representations spread first and are later generalized into higher-order representations. This is because greater heterogeneity within a part of the grammar indicates that this portion is under-going change: not everyone is using the same constructions. The more abstract a construction the more it is universal across dialects or languages. Here, however, our representations are not abstract to the level of typology and thus we can use them to understand sociolinguistic processes of diffusion.

5.2. Diffusion Depends on Network Structure of the Grammar

The previous sub-section focused on internal variation within local populations to show that there are differences in homogeneity values depending on which sub-part of the grammar is being observed. This means that some parts of the grammar have spread across the entire local population while others have not. This section takes another view of diffusion across the grammar by looking at all 150,000 pairs of local metro areas. The standardized Burrow's Delta essentially sorts pairs of local metro areas from most similar to least similar. By repeating this ranking across different portions of the grammar, we can determine whether two portions of the grammar have the same ranking. Because similarity here is an after-effect of diffusion, this investigation of similarity allows us to uncover the previous pathway of diffusion taken by these constructions. If the grammar is organized as a network, we would see differences in pairwise distances depending on the node of the grammar being observed.

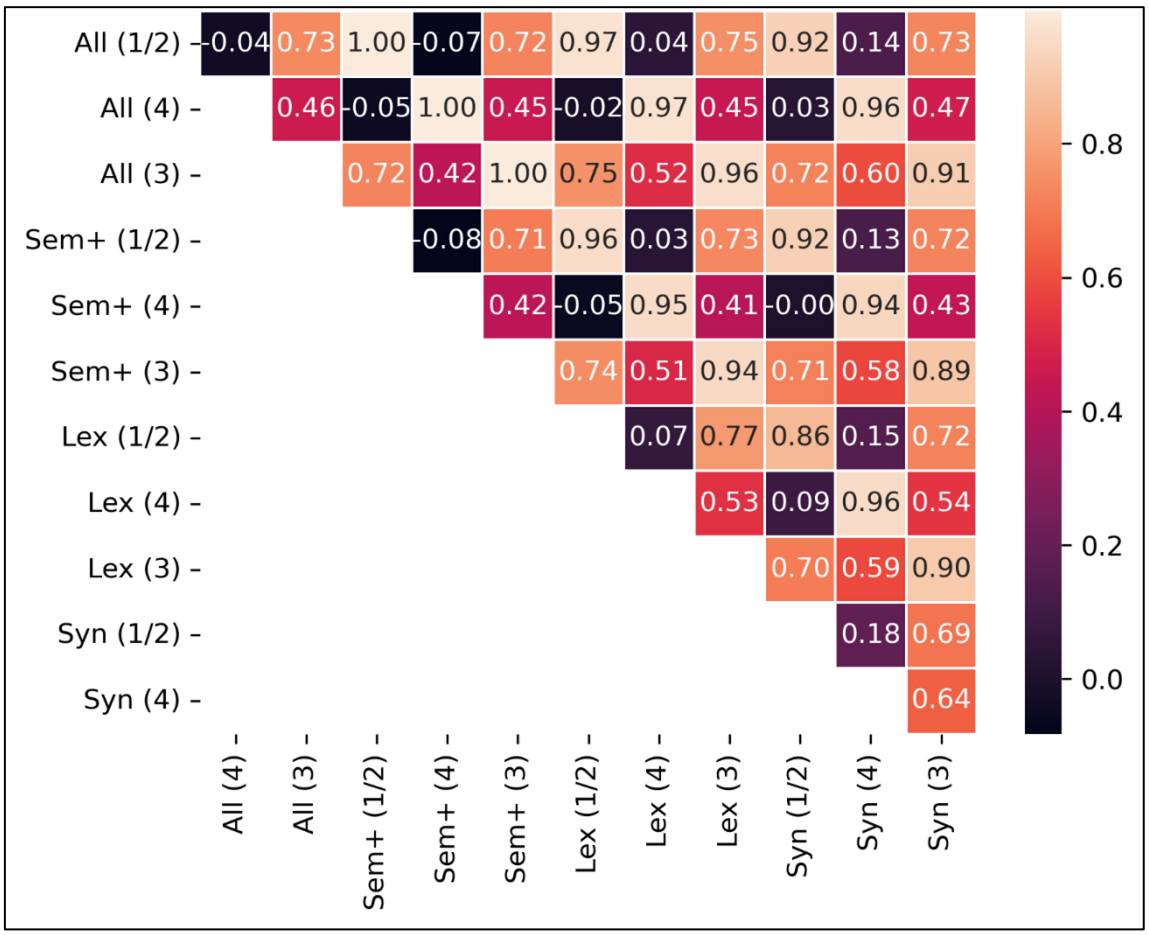

Figure 7. Correlations between Pairwise Distances of Local Populations for All Regions using Token Frequency. Higher correlations mean that two parts of the grammar agree on which local dialects are the most similar and which are the least similar.

Figure 7 above shows the global correlations in pairwise similarities between local dialects across portions of the grammar. As we would expect, the differences are much stronger here than when local metro areas were compared to themselves: there is greater distance between rather than within cities and thus diffusion processes reflected are easier to observe. Some parts of the grammar are more closely related; for instance, fourth-order constructions across the whole grammar and across the syntactic-only constructions have

a correlation of 0.96. In most cases, however, the correlations are much lower. This is true across levels of abstraction: All(3) vs All(4) has a correlation of only 0.46. And it is true across types of representation: Syn(3) vs Sem+(3) has a correlation of 0.89. Agreement across level of abstraction is lower than agreement across type of representation, which again indicates that variants spread first at an item-specific level.

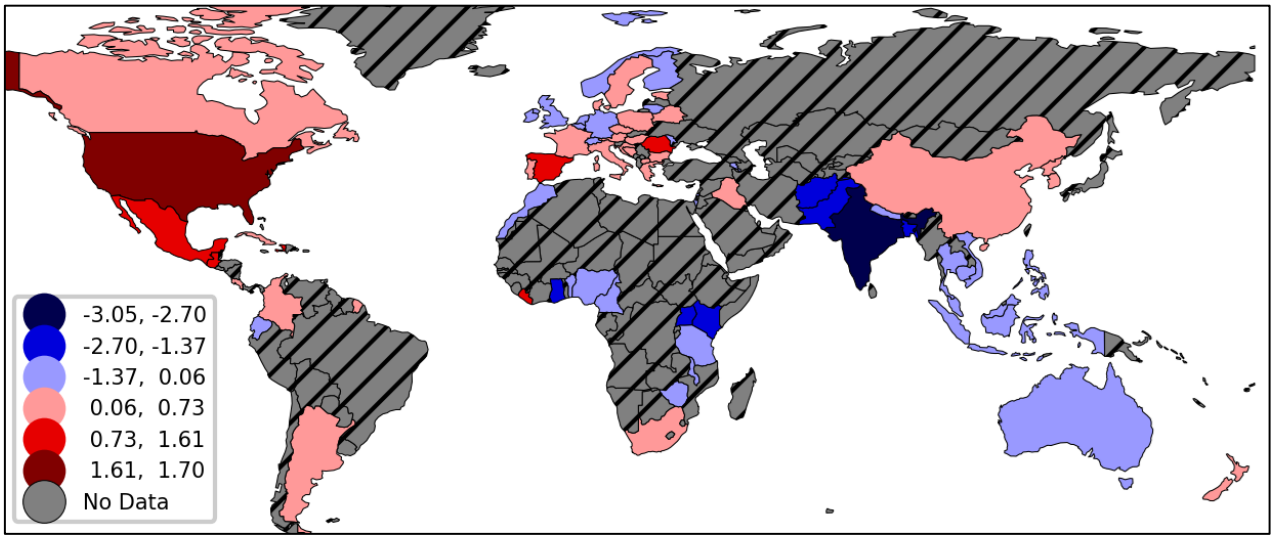

Figure 8. Similarity of Local-Populations Within a Country; All First-Order Constructions. Higher values mean that there is more variation across cities within a country.

The map in Figure 8 shows countries by the amount of similarity between cities within that country, using all first-order constructions. As before, this is a standardized Burrow's Delta so that high values (like the US in dark red) indicate lower similarity between cities. As with homogeneity within local areas, there is much more variation between cities in the US (dark red) than in India (dark blue). We thus take a closer look at correlations between distance rankings across portions of the grammar in the US in the heatmap in Figure 9. Note that because Burrow's Delta is standardized within each setting, the absolute values for pairwise distance vs homogeneity are not comparable.

Areas of the grammar with a high correlation agree on which local dialects are the most similar. This means that active, on-going processes of diffusion are concentrated in those pairs with a lower correlation; in essence, this means that less-correlated parts of the grammar have not spread from one local dialect to another. Within the United States, as shown in Figure 9, some parts of the grammar even have negative correlations: for instance, all first-order constructions (All 1/2) and fourth-order constructions with all types of representation (Sem+ 4) have a correlation of –0.64, showing opposing similarity ranks. These results imply that diffusion takes place differently across both type of representation and degree of abstraction as parts of the grammar which differ on both of these dimensions have the lowest agreement in similarity rankings.

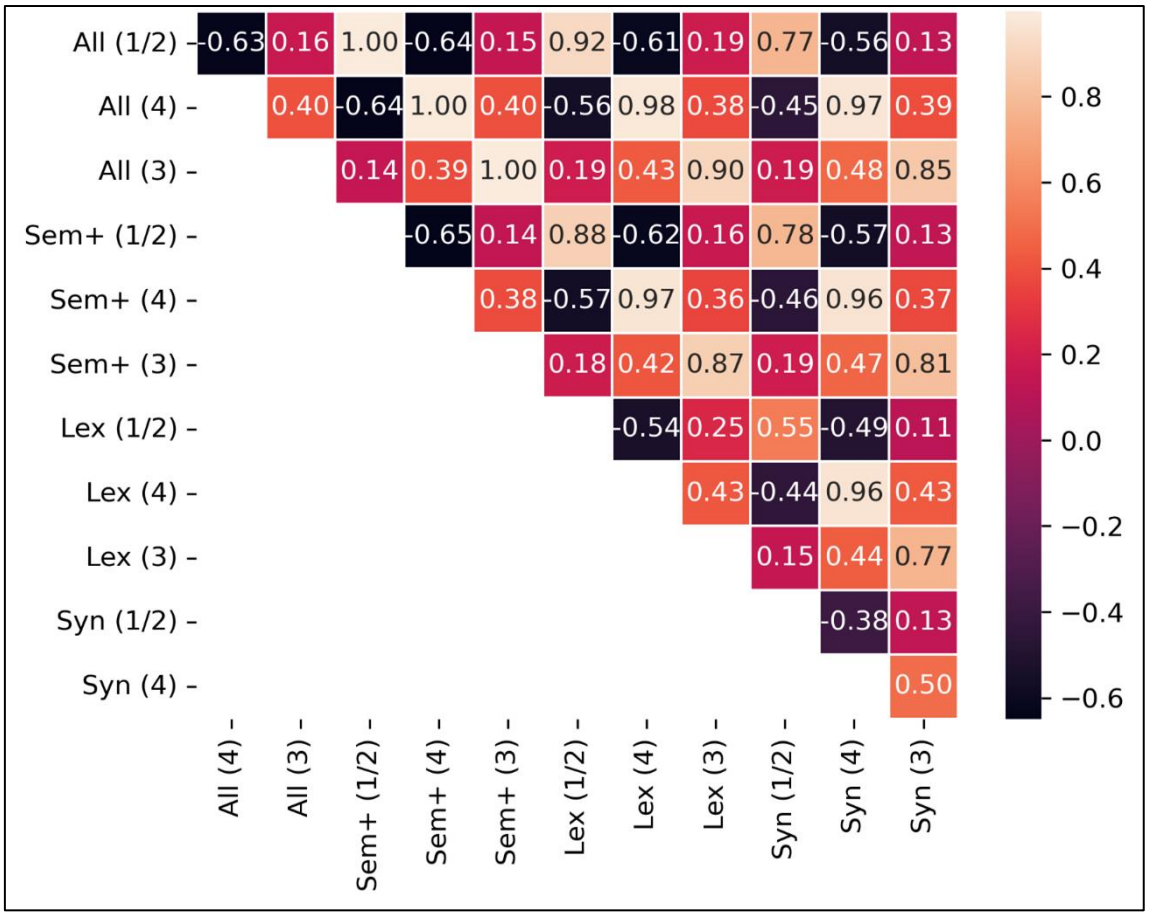

Figure 9. Correlations between Pairwise Distances of Local Populations for Cities in the USA using Token Frequency. Higher correlations mean that two parts of the grammar agree on which local dialects are the most similar and which are the least similar.

5.3. Discussion of Correlation Results

This section has consistently shown that different portions of the grammar have different rankings of both (i) variation within metro areas and (ii) variation between metro areas. To the degree that high similarity reflects a process of diffusion which has impacted both dialects, this indicates that diffusion spreads within the grammar as network rather than either (i) across all constructions at the same time or (ii) randomly across specific constructions. In this sub-section we further explore two issues which are important for contextualizing these results.

First, in Figure 10 below we look at the correlation across features for token frequency and type frequency. Token frequency reflects how many times any instance of a construction is used; this means that high token frequency could reflect many instances of the same specific utterance. Type frequency, on the other hand, reflects unique instances or manifestations of a construction; this means that type frequency better represents productivity in terms of how many novel utterances are made with each construction. We look at global correlation between type and token frequency within specific parts of the grammar in Figure 10. A high correlation means that there is no difference within this part of the grammar between the two measures; a lower correlation

means that type frequency provides a somewhat different view of dialect similarity than does token frequency.

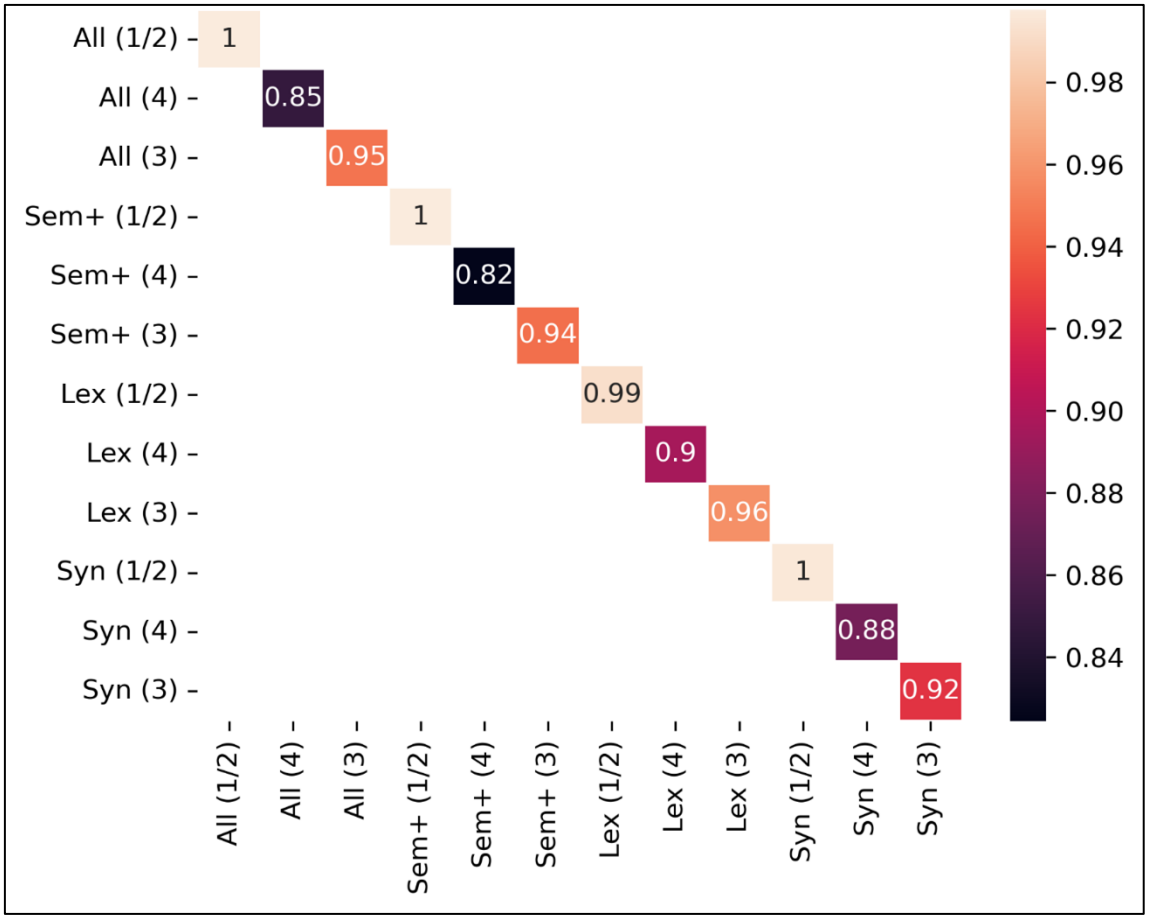

Figure 10. Correlations between Pairwise Distances of Local Populations for All Regions for the Same Feature Sets with Token-based Frequency vs Type-based Frequency. Higher correlations mean that the two frequency measures agree on which local dialects are the most similar.

The correlations here are much higher overall, as expected since we are comparing different frequency measures within the same sets of constructions. Less abstract layers of the grammar (i.e., first and second-order constructions) have the highest correlation. But the more abstract our representations become (reflecting larger bundles of related families of constructions), the more type frequency deviates from token frequency. In other words, these larger families also differ in the productivity of each construction in a way that lower-level constructions do not. This is important for a discussion of diffusion because it suggests that less abstract item-specific constructions are spread first, initially with limited productivity to new forms and new meanings. As these constructions are generalized, however, they become more productive, so that two local dialects differ not only in which constructions they use but also in the general productivity or diversity in usage of those constructions.

Second, we have been comparing all 505 cities, some of which are better represented than others. For instance, all cities have at least 25 samples (with 100 unique pairwise comparisons), but the largest cities have thousands of samples. We thus repeat the experiments above but with a sub-set of only the 112 cities with at least 200 samples; this allows us to see whether difference in the number of samples per city has an impact on the correlations being discussed. This additional experiment examines whether the

amount of data per city has an impact on the relation between feature-specific similarity ranks.

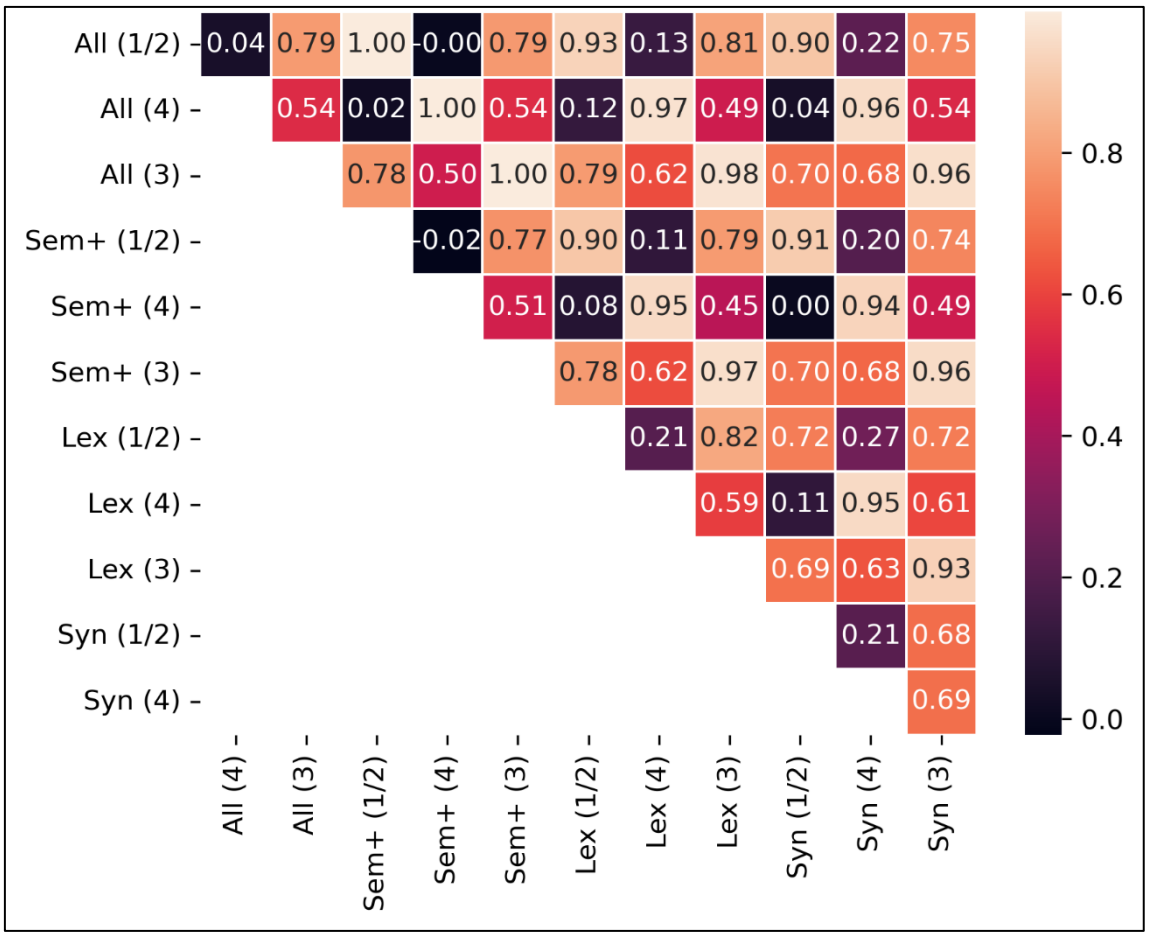

Figure 11. Correlations between Pairwise Distances of Local Populations for All Regions using Token-based Frequency. **Only the cities with the largest numbers of samples are included**. Higher correlations mean that the two frequency measures agree on which local dialects are the most similar.

A comparison of Figure 11 (only the most common cities in the data set) with Figure 7 (all cities), shows that there are only minor differences between the correlations with this smaller set of observations. In other words, the same conclusions about diffusion being structured according to the network structure of the grammar, with more constructions being spread and thus lower similarity for less abstract constructions, remains valid regardless of the number of samples observed per city.

6. Hypothesis 2: English Has a Global Speech Community

Our second hypothesis is that, because users of English are spread around the world, English forms a global speech community -- especially in digital settings like social media. While long-distance contact may be a weak influence in other settings, in digital contexts this could be a significant means of diffusion. The basic idea is that users everywhere are exposed to users from all countries and, likewise, cause users from all countries to be exposed to them. Some countries will be more central, because of their size or wealth or historical status. But we expect that the population network is a graph which connects all populations around the world that use English in digital settings.

The question, then, is whether the pathway of diffusion across the grammar differs across regional populations. As a metaphor, imagine that the spread of constructions is like water running down a sandy beach. The network structure of the grammar in this metaphor is like the terrain of the beach, small trenches or hills which influence the flow of the water. The experiments above have shown that the spread of constructions is, in fact, influenced by the network structure of the grammar. But one possibility is that all speech communities have the same terrain. If this were the case, then the spread of constructions across the grammar would take the same (complicated) pathway regardless of the population being observed. In this metaphor, the network structure of the population is like the beach where diffusion is taking place. If there is one pathway through the grammar, then we should observe the same similarity relationships between nodes of the grammar regardless of the structure of the population. We could also expect that a single process of diffusion operating at different rates of change would appear to be different pathways of change while actually representing differences in state. The data here do not have sufficient time depth to determine if this is the case; however, given that each generation of learners exists in a single time period, even this difference in rate would have significant impacts over time on the grammar of each local dialect.

A Mantel Test is used to determine the relationship between different sets of distance measures in the aggregate. We use it here to essentially repeat the experiments above while constraining the speech community to only cities within a particular region (where regions are shown by color in Figure 1). The Mantel Test tells us how much difference there is between North American diffusion and Western European diffusion, for instance. In reference to the metaphor above, the question is whether there is one beach or whether the terrain is different for each population. If the second is true, then the pathway along which constructions spread through the grammar is unique to each speech community. This would result in Mantel Test correlations that are lower between two region-specific populations. So our first question, in essence, is whether diffusion operates differently in grammatical terms across regional speech communities.

Figure 12 shows the estimated correlations between region-specific correlations of distance measures across the grammar. This is a rather abstract test, so it is worth breaking down what information this represents. *First*, we observe many samples from each city and calculate pairwise distance between cities by drawing 100 unique random pairs of samples. *Second*, we calculate the distance between each of these 100 pairs in order to estimate the distance between each local area: higher values for Burrow's Delta represent more differences between these local dialects which thus represents less diffusion. *Third*, we divide the grammar network into twelve sub-sets based on the type of representation and the level of abstraction; then we calculate the correlation of similarity ranks of cities across these different parts of the grammar. Thus, each part of the grammar ranks the similarities of cities and a high correlation means that two parts of the grammar have similar ranks. If the grammar is subject to a single, central process of diffusion then the correlation will always be high; otherwise, different parts of the grammar are subject to competing pressures of diffusion. The results in Section 5 show that the second is the case: each part of the grammar is subject to different diffusion processes. *Fourth*, representing each region (such as North America) as the set of correlations of distance measures within that region, we calculate the Mantel test to determine whether, for example, North America and Western Europe have the same pattern of agreement

between grammar-specific similarity measures. This replicates the study of differences across the grammar for each regional population. When the correlations are low, it means that diffusion has taken different pathways across the grammar.

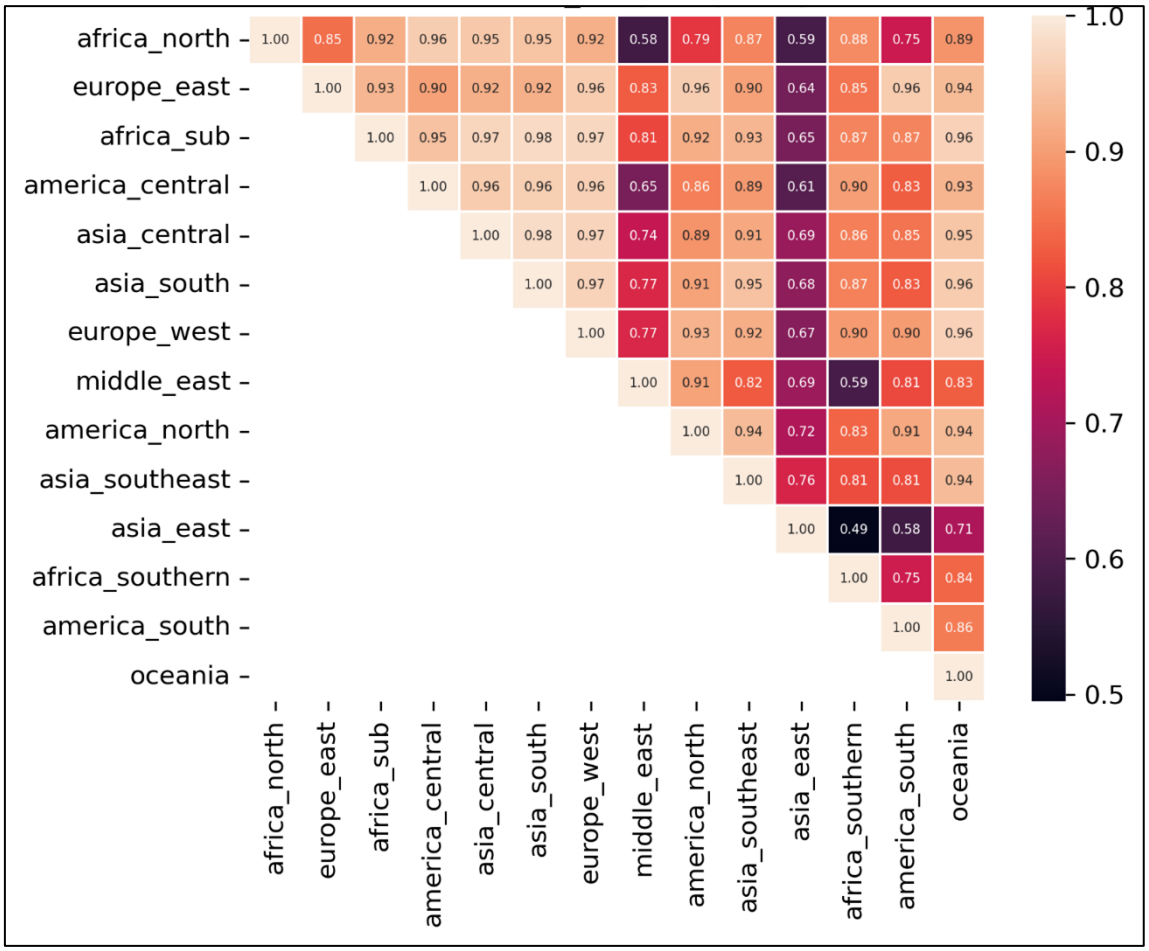

Figure 12. Mantel Test Correlations between Region-Specific Distances Between Nodes within the Grammar using Token Frequency. The Mantel Test indicates the overall difference between a set of distance measures, here with comparisons divided by region. Lower agreement indicates that the pathway of diffusion across the grammar differs between regions.

Thus, Figure 12 shows us that, although diffusion operates on the network structure of the grammar, there is not one single pathway along which constructions are spread. For instance, the Mantel correlation estimate between North Africa and East Asia is quite low (0.59). In fact, East Asia has low relationships with all other regions. North America, on the other hand, has relatively close relationships with some regions (Oceania: 0.94 and Western Europe: 0.93), but low relationships with other regions (North Africa: 0.79 and Southern Africa: 0.83). In and of itself, this does not show that English has a global speech community. But it does show that the patterns we would see for diffusion across the grammar depend on the population we observe, which is the first justification for taking a global view rather than restricting ourselves to inner-circle varieties or a few select outer-circle varieties, as is typically done.

While these results suggest that further work is merited on modelling diffusion across a global speech community, there are three challenges that keep us from definitively answering the second hypothesis here: *First*, there are wide differences in the number of

cities per region and the number of samples per city: for instance, North Africa and South America and Central Asia are all poorly represented. An approach which limited itself to inner-circle and outer-circle varieties (cf. Dunn 2023) would avoid this problem.

Second, the methodology used here samples many unique pairs of corpora for each comparison and works with the distribution of the resulting similarity values. However, the adequacy of these estimations can vary: we used a Bayesian estimate of the mean with a 95% confidence interval, but in expanding circle varieties especially the upper and lower boundaries can be quite different. For instance, the distribution of similarities between San Francisco (USA) and Bhavnagar (India) ranges from 0.839 to 1.447 in standardized space – a very large difference. The challenge is that a global approach based on pairwise distances creates many unexpected and distant pairs which have less precision than more local comparisons. In this paper we also worked with within-city differences (c.f., Figure 3), within-country differences (c.f., Figure 8), and within-region differences (c.f., Figure 9), all of which show that this comparison method works sufficiently well if the local populations are sufficiently similar. But very distant pairs remain a challenge.

Third, the method here involves first learning a single umbrella grammar to represent all varieties of English (cf. Dunn 2024) and then comparing local populations within that fixed feature space. But one side-effect of this approach is that less common (peripheral) varieties will be under-represented in the grammar and thus have more false negatives or missing constructions (cf. Dunn 2019b). Previous work relied on a supervised classification model, which could work around such missing constructions. But the impact is higher in an unsupervised distance-based approach. For instance, let's say that South Asian varieties have more missing constructions; this means that their frequency vectors will be smaller overall. Since Burrow's Delta is based on a comparison of frequency vectors, varieties with more false negatives will be closer together (i.e., Euclidean distance of low frequencies will be low overall). This is acceptable for comparisons within countries and regions, where the impact of false negatives is consistent. But it creates a challenge for directly comparing inner-circle and expanding-circle varieties, for instance.

These challenges keep us from directly answering the second hypothesis about global patterns of diffusion. However, the comparison between regions in Figure 12 tells us that the pathway of diffusion differs across regions when looking at only region-specific similarities, which are more reliable. Thus, we see a hint that there are non-local influences in diffusion, but it remains a problem for future research to address the methodological challenges above.

7. Conclusions

The basic idea in this paper was to examine processes of diffusion while taking into account both (i) the network structure of the grammar and (ii) the network structure of the speech community on a macro-scale. To capture processes of diffusion in purely synchronic data we rely on similarity measures, both within local populations and between local populations. Two dialects are similar, in this framework, as a result of constructions that have spread previously. Thus, greater similarity between dialects is taken to represent greater amounts of diffusion.

The first hypothesis is that, if the grammar is organized as a network (cf. Section 3), then the spread of constructions will take place across that network structure. The results show quite clearly that different sub-sets of the grammar produce different similarity ranks. This means that synchronic similarity is organized relative to the network structure of the grammar. The implication is that diffusion operates on specific nodes within the grammar, creating different patterns of similarity depending on the node.

The second hypothesis is that, if the speech community of English has a global organization, then we should observe non-local pressures in diffusion. The challenge is to capture global relationships given the methodological problems described above. At the very least, the results in this study and in related work (Dunn 2023) show that there are non-local relationships between dialects within regions and countries. The presence of non-local influence itself is not surprising; the larger question is the distance at which two populations continue to influence each other: how much exposure and what type of contact is required for long-distance diffusion to take place? This study has not been able to resolve this question but contributes to its answer by looking at diffusion pathways over distinct regions to show that there is not one single grammatical pathway for diffusion (a plausible but incorrect hypothesis).

Most approaches to syntactic variation and change focus on a few isolated constructions within largely local populations. The results in this study show that no such study is complete without taking into account both (i) the network structure of the grammar and (ii) non-local connections in the wider speech community. This finding reinforces previous work that relied on classification models rather than similarity measures (Dunn, 2018a, 2023). However, many questions remain unanswered: What is the relative pressure from local and non-local exposure situations? What is the nature of global digital speech communities? Does exposure and variation in one register (i.e., social media) have an influence on production in other registers? Is there a consistent cross-linguistic relationship between dialectal variation and register variation (Li et al. 2023; Eida et al. 2024) and how would such a difference influence diffusion across dialects? What types of exposure create diffusion pressures and do different portions of the grammar respond differently to different exposure situations? What is the relationship between entrenchment processes which make a construction more fixed in the grammar and diffusion processes which make a construction more variable? These important questions remain to be answered.

Appendix 1. Keywords

Each *sample* is an aggregation of 250 unique tweets, one tweet selected for each keyword below. This means that each sample as the same distribution of these keywords regardless of where it collected from.

actually	check	friend	human	might	point	soon	trump	women
after	city	friends	into	mind	police	sorry	try	won
against	come	full	job	money	post	start	trying	work
already	coming	future	join	morning	power	state	under	working
always	congratulations	game	keep	most	president	stay	understand	world
amazing	country	getting	kind	much	problem	still	until	wrong
another	covid	girl	know	music	project	stop	use	yeah
anyone	day	give	last	name	put	story	used	year
anything	days	go	least	need	read	such	using	years
around	different	god	left	needs	ready	support	very	yes
ask	doing	going	let	never	real	sure	via	yet
away	done	gonna	life	new	really	take	video	
back	down	got	literally	news	remember	talk	vote	
bad	during	government	little	nice	right	talking	wait	
beautiful	end	great	live	night	said	team	waiting	
because	enough	group	long	nothing	same	tell	want	
believe	ever	guy	look	off	saw	than	watch	
best	everyone	guys	looking	old	say	thank	watching	
better	everything	happy	looks	once	saying	thanks	way	
between	face	hard	lot	open	says	then	week	
big	family	hate	love	other	school	things	well	
black	far	head	make	over	season	think	where	
both	feel	health	makes	part	see	though	which	
business	few	heart	making	party	seen	thought	while	
buy	find	help	man	people	set	through	white	
call	first	here	many	person	shit	time	whole	
called	follow	high	maybe	place	show	times	why	
care	food	home	mean	play	since	today	win	
cause	found	hope	media	playing	something	tomorrow	wish	
change	free	house	men	please	song	top	without	

References

Anthonissen, Lynn. 2020. Cognition in construction grammar: Connecting individual and community grammars. *Cognitive Linguistics* 31(2). 309-337. https://doi.org/10.1515/cog-2019-0023

Alishahi, Afra & Suzanne Stevenson. 2008. "A computational model of early argument structure acquisition." *Cognitive Science* 32(5). 789-834. https://doi.org/10.1080/03640210801929287

Bamman, David, Jacob Eisenstein, & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135-160. https://doi.org/10.1111/josl.12080

Barak, Libby and Adele Goldberg. 2017. Modeling the Partial Productivity of Constructions. 131-138. In *Proceedings of AAAI 2017 Spring Symposium on Computational Construction Grammar and Natural Language Understanding.* Association for the Advancement of Artificial Intelligence. https://cdn.aaai.org/ocs/15297/15297-68208-1-PB.pdf

Beckner, Clay, Richard Blythe, Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. 2009. Language Is a Complex Adaptive System: Position Paper. *Language Learning*, 59(1). 1-26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Beuls, Katrien and Paul Van Eecke. 2023. Fluid Construction Grammar: State of the Art and Future Outlook. 41-50. In *Proceedings of the First International Workshop on Construction Grammars and NLP*. Association for Computational Linguistics. https://aclanthology.org/2023.cxgsnlp-1.6

Biber, Douglas, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3). 581-616. https://doi.org/10.1515/cllt-2018-0086

Carvalho, Ana Maria. 2004. I Speak like the Guys on TV: Palatalization and the Urbanization of Uruguayan Portuguese. *Language Variation and Change*, 16(2). 127–151. https://doi.org/10.1017/S0954394504162030

Cook, Paul and Laurel Brinton. 2017. Building and Evaluating Web Corpora Representing National Varieties of English. *Language Resources and Evaluation*, 51(3). 643-662. https://doi.org/10.1007/s10579-016-9378-z

Croft, William. 2013. Radical Construction Grammar. 211-232. In *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780195396683.013.0012

Dąbrowska, Ewa. 2021. How Writing Changes Languages. 75-94. In Mauranen, A. and Vetchinnikova, Svetlana (eds.), *Language Change: The Impact of English as a Lingua Franca.* Cambridge University Press. https://doi.org/10.1017/9781108675000

Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). BYU Corpora. https://www.english-corpora.org/glowbe/

Davies, Mark and Robert Fuchs. 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. https://doi.org/10.1075/eww.36.1.01dav

Diessel, Holger. 2023. *The Constructicon: Taxonomies and Networks*. Elements in Construction Grammar. Cambridge University Press. https://doi.org/10.1017/9781009327848

Donoso, Gonzalo and David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. 16-25. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1202

Doumen, Jonas, Katrien Beuls, and Paul Van Eecke. 2023. Modelling Language Acquisition through Syntactico-Semantic Pattern Finding. 1347-1357. In *Findings of the Association for Computational Linguistics at EACL 2023*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.99

Dunn, Jonathan. 2017. Computational Learning of Construction Grammars. *Language & Cognition*, 9(2). 254-292. https://doi.org/10.1017/langcog.2016.7

Dunn, Jonathan. 2018a. Finding Variants for Construction-Based Dialectometry: A Corpus-Based Approach to Regional CxGs. *Cognitive Linguistics*, 29(2). 275-311. https://doi.org/10.1515/cog-2017-0029

Dunn, Jonathan. 2018b. Modeling the Complexity and Descriptive Adequacy of Construction Grammars. 81-90. In *Proceedings of the Society for Computation in Linguistics*. Association for Computational Linguistics. https://doi.org/10.7275/R59P2ZTB

Dunn, Jonathan. 2019a. Global Syntactic Variation in Seven Languages: Towards a Computational Dialectology. *Frontiers in Artificial Intelligence: Computational Sociolinguistics*. https://doi.org/10.3389/frai.2019.00015

Dunn, Jonathan. 2019b. Modeling Global Syntactic Variation in English Using Dialect Classification. 42-53. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1405

Dunn, Jonathan. 2019c. Frequency vs. Association for Constraint Selection in Usage-Based Construction Grammar. 117-128. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2913

Dunn, Jonathan. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54. 999-1018. https://doi.org/10.1007/s10579-020-09489-2

Dunn, Jonathan. 2022. Exposure and Emergence in Usage-Based Grammar: Computational Experiments in 35 Languages. *Cognitive Linguistics*, 33(4). 659-699. https://doi.org/10.1515/cog-2021-0106

Dunn, Jonathan. 2023a. Syntactic variation across the grammar: Modelling a complex adaptive system. *Frontiers in Complex Systems: Complexity in Language Variation and Change*. https://doi.org/10.3389/fcpxs.2023.1273741

Dunn, Jonathan. 2023b. Exploring the Constructicon: Linguistic Analysis of a Computational CxG. 1-11. In *Proceedings of the Workshop on CxGs and NLP @*

the Georgetown University Round Table on Linguistics / SyntaxFest. Association for Computational Linguistics. https://aclanthology.org/2023.cxgsnlp-1.1

Dunn, Jonathan. 2024. *Computational Construction Grammar: A Usage-Based Approach*. Elements in Cognitive Linguistics. Cambridge University Press. https://doi.org/10.1017/9781009233743

Dunn, Jonathan, Benjamin Adams, and Harish Tayyar Madabushi. 2024. Pre-Trained Language Models Represent Some Geographic Populations Better Than Others. 12966–12976. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation.* https://doi.org/10.48550/arXiv.2403.11025

Dunn, Jonathan and Wikke Nijhof. 2022. Language Identification for Austronesian Languages. 6530-6539. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*. European Language Resources Association. https://aclanthology.org/2022.lrec-1.701

Dunn, Jonathan and Andrea Nini. 2021. Production vs Perception: The Role of Individuality in Usage-Based Grammar Induction. 149-159. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.19

Dunn, Jonathan and Harish Tayyar Madabushi. 2021. Learned Construction Grammars Converge Across Registers Given Increased Exposure. 268-278. In *Proceedings of the Conference on Computational Natural Language Learning*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.21

Dunn, Jonathan and Sidney Wong. 2022. Stability of Syntactic Dialect Classification Over Space and Time. 26-36. In *Proceedings of the International Conference on Computational Linguistics. International* Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.3

Eida, Mai, Mayar Nassar, and Jonathan Dunn. 2024. How well do tweets represent sub-dialects of Egyptian Arabic? 41-55. In P*roceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.vardial-1.4

Eisenstein, Jacob, Brendan O'Connor, Noah Smith, and Eric Xing. 2014. Diffusion of lexical change in social media. *PloSOne,* 10:1371. https://doi.org/10.1371/journal.pone.0113114

Egbert, Jesse, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9). 1817-1831. https://doi.org/10.1002/asi.23308

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(2). ii4–ii16 https://doi.org/10.1093/llc/fqx023

Fagyal, Zsuzsanna, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. 2010. Centers and peripheries: Network roles in language change. *Lingua*, 2061-2079. https://doi.org/10.1016/j.lingua.2010.02.001

Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, Oxford.

Goldsmith, John. 2015. Towards a new empiricism for linguistics. 58-105. In *Empiricism and Language Learnability*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780198734260.003.0003

Grafmiller, Jason and Benedikt Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *Language Variation and Change*, 30(3). 385-412. https://doi.org/10.1017/S0954394518000170

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

Gonçalves, Bruno and David Sánchez. 2014. Crowdsourcing Dialect Characterization through Twitter. *PLOS ONE*, 9(11). 1-6. https://doi.org/10.1371/journal.pone.0112074

Gonçalves, Bruno, Lucía Loureiro-Porto, José Ramasco, and David Sánchez. 2018. Mapping the Americanization of English in space and time. *PLOS ONE*, 13(5). 1-15. https://doi.org/10.1371/journal.pone.0197741

Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge University Press, Cambridge, UK.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence* 2:11. https://doi.org/10.3389/frai.2019.00011

Hollmann, Willem and Anna Siewierska. 2011. The status of frequency, schemas, and identity in cognitive sociolinguistics: A case study on definite article reduction. *Cognitive Linguistics*, 22(1). 25-54. https://doi.org/10.1515/cogl.2011.002

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

Kodner, Jordan. 2020. Modeling Language Change in the St. Louis Corridor. *Language Variation and Change*, 32(1). 77-106. https://doi.org/10.1017/S0954394519000255

Laitinen, Mikko, Masoud Fatemi, and Jonas Lundberg. 2020. Size Matters: Digital Social Networks and Language Change. *Frontiers in Artificial Intelligence: Computational Sociolinguistics*. https://doi.org/10.3389/frai.2020.00046

Leclercq, Benoît and Cameron Morin. 2023. No Equivalence: A new principle of no synonymy. *Constructions* 15(1). https://doi.org/10.24338/cons-535

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3. 211-225. https://aclanthology.org/Q15-1016/

Li, Haipeng, Jonathan Dunn, and Andrea Nini. 2023. Register variation remains stable across 60 languages. *Corpus Linguistics and Linguistic Theory*, 19(3). 397-426. https://doi.org/10.1515/cllt-2021-0090

Lucy, Li and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9. 538-556. https://doi.org/10.1162/tacl_a_00383

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLOSOne*, 10. 1371. https://doi.org/10.1371/journal.pone.0061981

Nevens, Jens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language Acquisition through Intention Reading and Pattern Finding. 15-25. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.2

Osborne, Timothy and Thomas Gross. 2012. Constructions are catenae: Construction Grammar meets Dependency Grammar. *Cognitive Linguistics*, 23(1). 165-216. https://doi.org/10.1515/cog-2012-0006

Perek, Florent and Amanda Patten. 2019. Towards an English Constructicon using patterns and frames. *International Journal of Corpus Linguistics*, 24(3). 354-384. https://doi.org/10.1075/ijcl.00016.per

Schneider, Edgar. 2020. Calling Englishes As Complex Dynamic Systems: Diffusion and Restructuring. 15-43. In *Language Change: The Impact of English as a Lingua Franca*. Cambridge University Press. https://doi.org/10.1017/9781108675000.004

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press: Cambridge, UK.

Szmrecsanyi, Benedikt and Jason Grafmiller. 2023. *Comparative Variation Analysis: Grammatical Alternations in World Englishes*. Cambridge University Press: Cambridge, UK.

Trudgill, Peter. 2014. Diffusion, drift, and the irrelevance of media influence. *Journal of Sociolinguistics*, 18(2). 213-222. https://doi.org/10.1111/josl.12070

Wible, David and Nai-Lung Tsao. 2010. StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. 25-31. In *Proceedings of the Workshop on Extracting and Using Constructions in Computational Linguistics*. Association for Computational Linguistics. https://aclanthology.org/W10-0804

Wible, David and Nai-Lung Tsao. 2020. Constructions and the problem of discovery: A case for the paradigmatic. *Corpus Linguistics and Linguistic Theory*, 16(1). 67-93. https://doi.org/10.1515/cllt-2017-0008

Wieling, Martijn; John Nerbonne, and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS One*, 6:9. https://doi.org/10.1371/journal.pone.0023613

Würschinger, Quirin. 2021. Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence: Computational Sociolinguistics*. https://doi.org/10.3389/frai.2021.648583