

Stability of Syntactic Dialect Classification Over Space and Time

Jonathan Dunn¹ and Sidney Wong²

Department of Linguistics

University of Canterbury

¹New Zealand Institute for Language, Brain and Behaviour

²Geospatial Research Institute Toi Hangarau

jonathan.dunn@canterbury.ac.nz

sidney.wong@pg.canterbury.ac.nz

Abstract

This paper analyses the degree to which dialect classifiers based on syntactic representations remain stable over space and time. While previous work has shown that the combination of grammar induction and geospatial text classification produces robust dialect models, we do not know what influence both changing grammars and changing populations have on dialect models. This paper constructs a test set for 12 dialects of English that spans three years at monthly intervals with a fixed spatial distribution across 1,120 cities. Syntactic representations are formulated within the usage-based Construction Grammar paradigm (CxG). The decay rate of classification performance for each dialect over time allows us to identify regions undergoing syntactic change. And the distribution of classification accuracy within dialect regions allows us to identify the degree to which the grammar of a dialect is internally heterogeneous. The main contribution of this paper is to show that a rigorous evaluation of dialect classification models can be used to find both variation over space and change over time.

1 Geographic Variation Over Time

This paper experiments with the stability of dialect classification models over space and time in order to determine the degree to which they capture language variation and change. The assumption in previous work has been that a geo-referenced corpus (Davies and Fuchs, 2015; Cook and Brinton, 2017; Dunn, 2020) captures the linguistic behaviour of specific populations. This paper experiments with the spatial and temporal stability of dialect models by systematically constructing monthly test sets spanning a three-year period. This allows us to evaluate the continuing effectiveness of dialect models over time, an important criteria for determining their validity. Because different locations represent different populations, we use spatial sampling to construct test sets which represent different local

populations within each country. This allows us to determine the degree to which a dialect like New Zealand English adequately represents the varied populations within New Zealand.

Dialect classification is the task of predicting the location of origin for the individual who produced a given sample (Dunn, 2019b; Chakravarthi et al., 2021; Gaman et al., 2020). Thus, dialect classification, by focusing on the latent properties of geo-referenced samples, differs from geo-location (Rahimi et al., 2017) which focuses on predicting the location of the sample itself and from geo-characterization (Adams and McKenzie, 2018) which focuses on predicting attributes of the location. While all three tasks rely on geographic information, dialect classification is unique in modelling variations in the linguistic system. Beyond this, dialect classification is part of ensuring that NLP represents the world’s population, including non-standard and non-western populations.

The temporal evaluation (Section 6) shows that most dialects share the same performance decay rate. This indicates a general effect of model decay rather than cases of change over time within individual dialects. The spatial evaluation, however, shows that prediction accuracy for all dialects is spatially-conditioned within countries (Section 7). This indicates that, while dialect models capture proto-typical populations within each country, they do not equally describe all local populations.

The experiments in this paper use construction grammar (CxG: Goldberg 2006; Langacker 2008; Croft 2013) to represent syntactic structure for the purpose of observing dialectal variation. CxG is a *usage-based* approach to syntax, a bottom-up theory of language in which frequent exposure is hypothesized to lead to the emergence of grammatical units (Hopper, 1987; Bybee, 2006). The use of syntactic representations for dialect classification ensures that the model does not rely on extraneous information like place names or local topics of in-

terest. From this perspective, a GRAMMAR is a set of constructions that together represent the structure of a language. A DIALECT MODEL is a matrix of spatial weights in which the number of rows corresponds to the number of constructions in the grammar and the number of columns corresponds to the number of dialects. These weights, learned using a Linear SVM, support dialect classification and also represent spatial variation in the grammar.

In order to undertake a spatio-temporal evaluation, we collect a balanced corpus of tweets to represent 12 varieties of English around the world. The basic experimental paradigm is to train models on a fixed period (July through December 2018) and then test those models at monthly intervals from 2019 to 2021. Each monthly test set maintains the same geographic distribution as the training data, so that fluctuations in performance are not caused by changes in the locations represented.

After considering related work on dialect classification and other geographic models (Section 2), we consider the corpora used in these experiments (Section 3). We then present the syntactic representations used (Section 4) and the basic experimental methods (Section 5). The performance of dialect models over time is presented in Section 6 and the performance over space in Section 7. The main contribution of this paper is to show that the performance of dialect classification models remains stable over time but that there is significant spatial variation in performance within dialect areas.

2 Related Work

Early work showed that part-of-speech trigrams are able to distinguish between some regional dialects (Sanders, 2007), a method that continues to appear in recent work (Kreutz and Daelemans, 2018). Similar methods have been used for authorship analysis (Hirst and Feiguina, 2007) and for characterizing immigrant populations (Nerbonne and Wiersma, 2006). In other contexts, non-syntactic features can out-perform syntactic features for modelling dialects (Kroon et al., 2018), so that many approaches to distinguishing between dialects are similar to language identification models (Ali, 2018).

More recent work has modelled geographic syntactic variation by combining grammar induction with geospatial text classification (Dunn, 2018a, 2019b,c). The use of grammar induction to learn a syntactic feature space mitigates the fact that most grammars represent standard varieties (Jørgensen

et al., 2015), thus poorly representing many dialects around the world. In this paradigm, the learned grammar provides a feature space (c.f., Section 4) and the frequency of grammatical constructions in each sample is used to model dialects: a bag-of-constructions approach to text classification.

Most work on geographic variation is focused on lexical variation (Eisenstein et al., 2010) and change (Eisenstein et al., 2014). Recent work has shown a close correspondence between lexical variation in tweets and lexical variation in a dialect survey (Grieve et al., 2019). This work is important for showing that digital usage mirrors face-to-face usage. Other work has shown that geographic variation can be taken into account during language identification to ensure the inclusion of non-standard varieties (Jurgens et al., 2017). Models of lexical variation have generally failed to account for polysemy, so that competition between senses is not captured (Zenner et al., 2012), but more recent work has been able to account for polysemy in this context (Lucy and Bamman, 2021).

A related line of work uses language data to model non-linguistic properties of populations and places. For example, the problem of geo-location is to predict the location of a user given properties of a document (Wing and Baldrige, 2014; Alex et al., 2016; Rahimi et al., 2017). This task differs from dialect classification in that named entities and topic features can provide significant information. A related task is to model the characteristics of a particular place rather than the population of that place (Adams, 2015; Adams and McKenzie, 2018; Hovy and Purschke, 2018; Villegas et al., 2020). While there is a close connection between a place and its population, this line of work remains focused on characterizing non-linguistic attributes.

This paper makes two main contributions: *First*, it experiments with geographic syntactic variation over time and within dialect regions, significantly expanding our understanding of geographic variation in syntax. *Second*, from a more practical perspective, this paper evaluates the degree to which geographic models remain robust over space and time, an evaluation not previously available.

3 Geographic Language Data

This paper draws on social media data from the *Corpus of Global Language Use* (CGLU), using geo-referenced tweets that are identified for language using the idNet package (Dunn, 2020). The

Circle	Region	Country	N. Cities	N. Words
Inner-Circle	Oceania	Australia	98	3.9 mil
Inner-Circle	Oceania	New Zealand	99	2.0 mil
Inner-Circle	North American	Canada	95	4.9 mil
Inner-Circle	North American	United States	86	4.5 mil
Inner-Circle	European	Ireland	100	3.6 mil
Inner-Circle	European	United Kingdom	89	5.5 mil
Total Inner-Circle	3	6	567	24.4 mil
Outer-Circle	African	Ghana	69	1.1 mil
Outer-Circle	African	Kenya	98	1.8 mil
Outer-Circle	South Asian	India	96	2.5 mil
Outer-Circle	South Asian	Pakistan	100	1.0 mil
Outer-Circle	Southeast Asian	Malaysia	99	0.8 mil
Outer-Circle	Southeast Asian	Philippines	91	1.1 mil
Total Outer-Circle	3	6	553	8.37 mil

Table 1: Inventory of Regions, Countries, and Cities for Data Collection (One Month)

collection method for social media in the CGLU involves geographic searches from co-ordinates of individual cities. Here we sample from 1,120 cities representing 12 countries and six regions, as shown in Table 1. This table shows the amount of data by place by month. The total data set contains six months for training and 36 months for testing. The data set as a whole is visualized at earthLings.io.

This corpus is designed to provide a balanced representation of different varieties of English over time. The colonial history of English has led to a distinction within the World Englishes paradigm (Kachru, 1990) between *inner-circle* varieties that represent the first diaspora (e.g., Canada) and *outer-circle* varieties that represent the second diaspora (e.g., India). We include six dialects/varieties each from the inner-circle and outer-circle groups.

Within each group we include three regions, each with two country-level varieties. As shown in Table 1, the inner-circle group contains three regions: Oceania (Australia and New Zealand), North America (Canada and the US), and Europe (the UK and Ireland). The collection of data from these countries is distributed across 567 cities, where each city represents a 50km radius from the city center. For each month, we sample 24.4 million words representing these inner-circle varieties.

The outer-circle group also contains three regions: Africa (Ghana and Kenya), South Asia (India and Pakistan), and Southeast Asia (Malaysia and the Philippines). The collection of data from these countries is distributed across 553 cities, with a comparable sample of 8.37 million words for

each month across the training and testing periods.

To maintain a comparable geographic distribution over time, we maintain the same number of samples from each city. This means, for example, that the relative influence of Brisbane and Perth in Australia remain constant over time. A *sample* for the purposes of this paper is an aggregation of individual tweets from the same place and time until the sample reaches 500 words. These larger samples provide more syntactic information for each dialect than do individual tweets. While previous work has used samples of 1,000 words (Dunn, 2019b), here we use smaller samples in order to increase the capacity for error analysis. As with many tasks, there is a trade-off between the higher accuracy provided by larger sample sizes and the flexibility provided by smaller sample sizes.

The distribution of samples across cities is taken from the training period (2018). Thus, the density of data by location across time is fixed to represent the density during the training period. This allows us to control for changes in the collection: for example, if Wellington began to produce more data in 2021, this change in distribution within New Zealand would appear to be syntactic variation while actually reflecting a change in the means of observation. Data collection spans from 07-2018 until 12-2021, a period of 42 months. The training period is 2018 and the testing period is 2019 through 2021. The geographic distribution across countries, as shown in Table 1, is held constant across this period, controlling for other sources of variation that might impact dialect models.

4 Syntactic Representations

This section details the main ideas of construction grammar (CxG), including both (i) the grammar induction algorithm used to learn syntactic representations here and (ii) examples of constructions used in the dialect models. The basic approach here is, first, to use grammar induction to learn a grammar and, second, to use the frequency of the constructions in that grammar to undertake geospatial text classification (Dunn, 2019b,c).

CxG can be distinguished from other approaches to syntax given its three core ideas: First, CxG posits a continuum between the lexicon and the grammar rather than a strict separation (for example, into a vocabulary and a set of phrase structure rules). This CONSTRUCTICON contains both lexical items and traditional syntactic structures. For example, a grammar-and-lexicon approach would analyze (a) below as an intransitive sentence by labelling the verb *laugh* as intransitive. The problem is that verb valency is quite fluid, as shown in (b) and (c). The CxG analysis of this fluidity is that (a) represents an INTRANSITIVE construction into which *laugh* is merged and (b)/(c) represent a CAUSED-MOTION construction into which *laugh* is merged. Thus, the fluidity of the argument structure here is explained by an underlying construction, itself meaningful, which interacts with specific lexical items. (Note that the grammar used for modelling dialects does not contain any individual lexical items as constructions).

- (a) Peter laughed.
- (b) The audience laughed Peter off the stage.
- (c) His marriage laughed Peter into rehab.
- (d) Peter laughed all the way to the bank.

A second main idea in CxG is that syntactic structure varies in its level of abstractness, with some representations being quite item-specific. The constructicon is an inheritance hierarchy in which fully-productive constructions like the CAUSED-MOTION construction in (b)/(c) have item-specific children like the idiom in (d). Essentially, (d) is a non-compositional and idiomatic version of the construction in (b)/(c) with some of the slots constrained to require a fixed phrase.

- (e) [SYN:NP – SYN:VP]
- (f) [SYN:NP – SYN:VP – SEM:*object* – SEM:*loc*]
- (g) [SYN:NP – SYN:VP – LEX:*all the way to the bank*]

A third main idea in CxG is that constructions are constraint-based representations in which slot-fillers are drawn from lexical, syntactic, and semantic categories. Each unit in a construction is a *slot*, separated by dashes in (e)/(f)/(g) above. Each slot is defined using a *slot-constraint*. For example, the INTRANSITIVE construction in (e) can be represented using only syntactic constraints. In contrast, the CAUSED-MOTION construction in (f) has two semantic constraints; these are labelled for purposes of exposition as *object* and *location*. The construction in (g) is item-specific and idiomatic, so that it can only be described using lexical constraints. The point, then, is that different levels of abstraction are captured in CxG using different types of slot-constraints.

This paper draws on previous approaches to the unsupervised learning of constructions (Dunn, 2017, 2018b). The first challenge is to build the inventory of lexical, syntactic, and semantic constraints that constructions are built on. Here we use the most frequent 100k words across the entire corpus of tweets as the lexicon. The syntactic constraints are drawn from the Universal Part-of-Speech tagset (Petrov et al., 2012) as implemented by the Ripple-Down-Rules tagger (Nguyen et al., 2016). The semantic constraints are drawn from fastText embeddings (Grave et al., 2019) clustered into discrete semantic domains using k-means. A complete inventory of these semantic domains is provided in the supplementary material; this approach ignores polysemy in lexical items when defining semantic constraints, using a single representation for each word-form.

From the perspective of varying levels of abstractness, syntactic constraints are the most general because they are divided into the smallest inventory of labels (only 14). Lexical constraints are the least general, with a lexicon of 100k words. And semantic constraints are in the middle, with an inventory of 1,000 domains. This parameter choice (i.e., using 1,000 semantic domains) results from the desired granularity in domains, falling between the very general syntactic constraints and the very specific lexical constraints. Thus, constructions are a sequence of slots, each of which is defined by a slot-constraint. Each type of slot-constraint (lexical, semantic, and syntactic) differ in their level of abstractness. For instance, lexically-defined constructions are more idiomatic and item-specific than syntactically-defined constructions.

This work relies on a loss function based on Minimum Description Length (Goldsmith, 2006; Grünwald and Rissanen, 2007) and a construction parser with a beam-search strategy (Dunn, 2019a) that operates on top of a psychologically-plausible association measure, the ΔP (Ellis, 2007). The contribution of this paper is to analyze syntactic variation across space over time using previous work on computational CxG; thus, we do not provide a fuller description of the framework here. Previous work has shown that these grammars converge onto stable representations as the amount of training data is increased (Dunn and Tayyar Madabushi, 2021), that grammars of individuals are significantly different than grammars of groups of individuals (Dunn and Nini, 2021), and that transformer-based language models can be fine-tuned using constructional information (Tayyar Madabushi et al., 2020).

The grammar used in these experiments is learned from the training period (2018) but includes a wider pool of 18 English-speaking countries in order to provide a global grammar of English. This larger training corpus for grammar induction contains 478 million words. The fastText embeddings are trained on this same extended corpus, but covering the entire period in order to increase the amount of data available for training; this larger corpus contains 4.2 billion words. This results in a single grammar that contains 6,119 individual constructions, some of which are shown in (h) through (n) below. Dialect models are learned by parsing each sample using this grammar, counting the frequency of each construction in each sample, and using the resulting feature space for dialect classification. The complete grammar, along with examples from the training data for each construction, is available in the supplementary material.

The following examples illustrate the nature of constructions; both constructions like (h) and examples like (h1) are drawn from the grammar used in the experiments. Each slot in (h) is separated by dashes and each slot-constraint is defined using lexical (LEX), syntactic (SYN), or semantic (SEM) categories. Lexical constraints are words given in italics; syntactic constraints are drawn from part-of-speech tags; and semantic constraints are formulated using numbers that refer to clustered embeddings, such as <443> in (k). For dialect classification, each construction (h) provides a feature and the frequency of that construction (h1 through h3) provides a sample-specific quantification.

(h) [LEX:*it* – SYN:AUX – SYN:V]

- (h1) ‘it is set’
- (h2) ‘it was shut’
- (h3) ‘it can go’

The first example, in (h), shows a simple clause with an expletive *it* as subject and a variable auxiliary verb. The example in (i) is a lexically-constrained noun phrase with *ability* as the head of an infinitival verb. A further lexically-constrained noun phrase in (j) shows the importance of a tweet-specific grammar: *ur* replaces the more traditional *your* as the pronoun.

(i) [LEX:*ability* – LEX:*to* – SYN:V]

- (i1) ‘ability to focus’
- (i2) ‘ability to live’
- (i3) ‘ability to wait’

(j) [LEX:*ur* – SYN:ADJ – SYN:N]

- (j1) ‘ur new journey’
- (j2) ‘ur own money’
- (j3) ‘ur mad tunes’

The adposition phrase in (k) contains a semantic constraint on the complement noun, in this case a type of location. As an example of how constructions themselves can be meaningful, (l) shows a copula construction with an ending conjunction. But the construction as a whole marks a caveat on the evaluation that is expressed by the copula.

(k) [SYN:ADP – SYN:N – <443>]

- (k1) ‘along airport road’
- (k2) ‘in union station’
- (k3) ‘into police station’

(l) [SYN:N – LEX:*was* – SYN:ADJ – SYN:CC]

- (l1) ‘bike was awesome but’
- (l2) ‘birthday was great and’
- (l3) ‘movie was better but’

The more complicated verb phrase in (m) contains a main verb, *myself* as a direct object, and an infinitival verb. This implicitly constrains the main verb to verbs of thinking like *compare* and *tell*, showing that implicit semantic constraints arise from interactions between slots. Finally, the complex noun phrase in (n) reflects a specific template of NP + ADP. In this way, constructions capture grammatical units of varying size and abstractness.

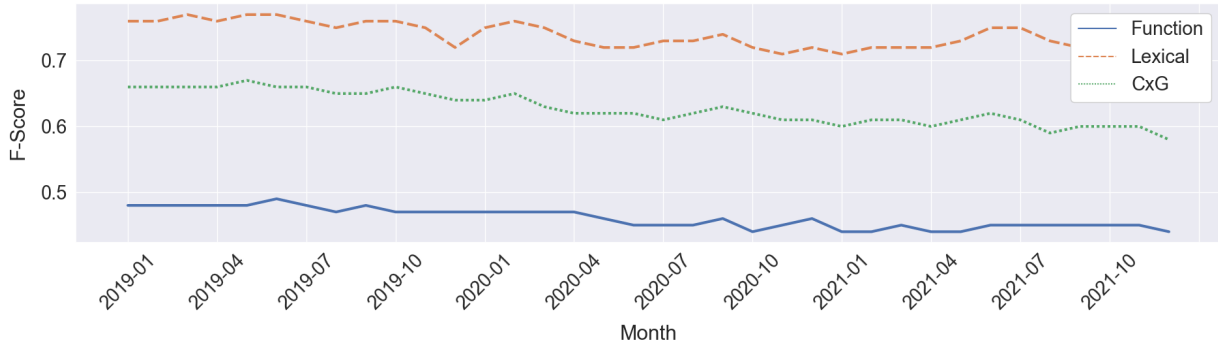


Figure 1: F-Score Against Baselines Over Time, All Varieties

- (m) [SYN:V – LEX:*myself* – LEX:*to* – SYN:V]
 (m1) ‘allowing myself to hope’
 (m2) ‘forcing myself to sleep’
 (m3) ‘tell myself to stop’
- (n) [LEX:*the*–SYN:N – LEX:*of* – SYN:DET – SYN:N]
 (n1) ‘the happiness of another person’
 (n2) ‘the owner of the station’
 (n3) ‘the masters of the game’

This section has presented CxG as a paradigm for usage-based syntax and reviewed previous work on computational CxG. An unsupervised construction grammar is learned from the training period, providing an adaptable feature space that contains structures from many different dialects. As the discussed examples show, these learned constructions provide a rich syntactic feature space for modelling geographic variation in production over time.

5 Dialect Models

The task of dialect classification or identification is to predict the location of origin for the author of a sample given some set of linguistic features. The classification here predicts country-level dialects like New Zealand English or Australian English. From the perspective of linguistics, dialect classification allows us to study variation in a high-dimensional space: variation across an entire grammar (Dunn, 2019b) rather than variation in individual and independent features (Grieve et al., 2019). From the perspective of NLP, dialect classification is part of the general problem of ensuring that language technology represents the world’s population rather than privileged sub-sets of the world’s population (Dunn and Adams, 2020).

Because part of the goal is to model spatio-temporal variation in the grammar, a dialect model takes the form of a matrix in which each feature

(a construction in the grammar) is a row and each dialect (a country-level label) is a column. This matrix represents the degree to which a given part of the grammar is subject to geographic variation. Taken row-wise, this matrix provides a measure of whether a particular construction varies across space. And, taken column-wise, this matrix provides a description of each dialect that, for example, can be compared with every other dialect to determine which are the most similar. As discussed below, dialect models are implemented as Linear SVMs that are trained using a bag-of-constructions approach in which the parser counts how many times each construction occurs in each sample.

Using the data from 2018 for training, we compare three models: *First*, a syntactic model based on the frequencies of the constructional features described above. *Second*, a baseline model that uses the frequency of function words like *of* or *was*, a common baseline for problems in authorship analysis (Grieve, 2007; Stamatatos, 2009; Argamon, 2018) when content words need to be avoided. *Third*, for the purpose of comparison, we include a unigram lexical model with TF-IDF weighting and function words removed so that it contains no syntactic information. Each of these models are implemented as a Linear SVM. Within this task, SVMs remain competitive, as shown by recent shared tasks on Romanian dialect identification (Gaman et al., 2020) and on identifying similar Uralic languages (Chakravarthi et al., 2021). In each case, we use a development set to determine parameters.

In each case, we train three models: INNER contains only inner-circle varieties like American English; OUTER contains only outer-circle varieties like Indian English; and ALL contains all 12 varieties. These are trained on the data from 2018 and tested on data from 2019, 2020, and 2021. The reason for maintaining separate models in some

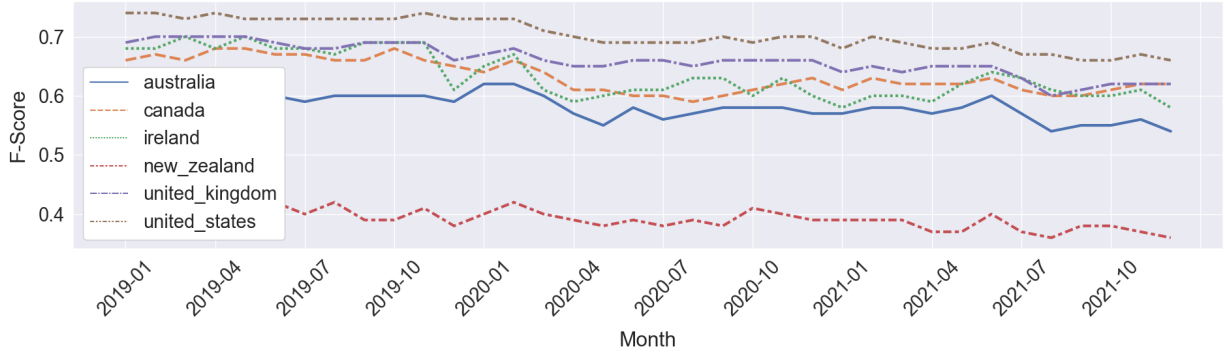


Figure 2: F-Score by Country Over Time for CxG Model, Inner-Circle Varieties

conditions is that inner-circle varieties have significantly more training and testing data available, which could lead to higher performance as an artifact. Thus, for example, the inner-circle condition contains only training and testing data from the six countries listed as *inner-circle* in Table 1.

As an initial analysis, the f-scores of each of these three models over time is shown in Figure 1, with the y-axis indicating the weighted f-score and the x-axis indicating time. All three classifiers are well above the majority baseline. The lowest performing is the function word model, a weak approximation for syntactic variation. The highest performing is the lexical model. This hierarchy remains stable across the three year testing period.

AU	CA	IE	NZ
australia	canada	ireland	nz
australian	canadian	irish	zealand
mate	ontario	dublin	auckland
melbourne	trump	cork	jacinda
sydney	toronto	limerick	te
abc	vancouver	galway	kiwi
brisbane	trudeau	lads	liked
labor	km	hurling	lincoln
nsw	kpa	county	hamilton
turnbull	alberta	final	kph

Table 2: Top Lexical Features By Country

Given the results in Figure 1, could we use the lexical model to examine dialects? The issue, as in previous work, is that the information contained in this model does not represent linguistic variation. Table 2 shows the top lexical items for four inner-circle countries: Australia, Canada, Ireland, and New Zealand. Most of these terms are place-names (like *australia*), place-specific named-entities (like *abc*), or people associated with these countries (like

jacinda). Only a few terms would qualify as dialectal variants, for example *mates* vs *lads*. As a representation of latent linguistic variation, the lexical model is not relevant; we thus focus on the syntactic models in the remaining analysis.

6 Syntactic Variation Over Time

We begin the analysis by looking at the weighted average f-score by model for the beginning of the test period (2019-01) and the end (2021-12), as shown in Table 3. This represents the impact of time on the overall accuracy. First, we see that outer-circle models have better performance. The most likely reason for this is that outer-circle varieties are more distinct from one another, in part because these varieties exist in more linguistically-diverse settings. For example, the US is less linguistically diverse than India in digital settings (Dunn et al., 2020). Although outer-circle varieties have a higher average f-score, they also have a greater change in f-score. This indicates more variability over time.

	Function	Grammar
Inner-Only, 2019-01	0.44	0.66
Inner-Only, 2021-12	0.40	0.59
<i>Inner-Only Decline</i>	<i>0.04</i>	<i>0.07</i>
Outer-Only, 2019-01	0.75	0.83
Outer-Only, 2021-12	0.66	0.75
<i>Outer-Only Decline</i>	<i>0.09</i>	<i>0.08</i>
All Dialects, 2019-01	0.48	0.66
All Dialects, 2021-12	0.44	0.58
<i>All Dialects Decline</i>	<i>0.04</i>	<i>0.08</i>

Table 3: Change in Performance Over Time by Model

Second, we notice in Table 3 that the relative performance of function words and the CxG model remain similar across the testing period. The full grammar model always out-performs the function

word baseline. We use a regression analysis to model the decay rate for each dialect in the CxG models, examining the amount of change in precision and recall over time (c.f., Figure 2). The basic idea here is that a consistent decay rate indicates model error while a faster rate of decay for individual dialects indicates change in those dialects themselves. Among inner-circle varieties, only NZ has a significant difference from the others, for recall but not for precision. A decline in precision would mean that samples from other dialects have become more similar to NZ; this does not happen. The observed decline in recall means that samples from NZ have become more similar to other dialects. This indicates that there has been a significant change in NZ but not in other dialects. No outer-circle varieties have a different decay rate, so that only NZ shows this type of change.

	AU	CA	IE	NZ	UK	US
AU	0	0	0	0	0	0
CA	0	0	0	0	0	0
IE	0	0	0	0	0	-.04
NZ	.16	0	0	0	.28	0
UK	0	0	0	0	0	0
US	0	0	0	0	0	0
	GH	IN	KE	MY	PK	PH
GH	0	-1.01	0	0	-.28	-.40
IN	0	0	0	-.91	0	0
KE	0	0	0	0	0	0
MY	0	0	0	0	0	0
PK	-.20	0	0	0	0	.13
PH	0	0	0	0	0	0

Table 4: Changing Relationships Between Dialects Using a VECM Analysis of False Positive Errors

The decay rate represents the overall trend for a given dialect but it does not take into account the specific errors made. The confusion matrix for each dialect provides a monthly representation of the distribution of false positive errors. For example, in the CxG model that includes all dialects, Canadian English has 1,488 false positives as American English in the first test period, but only 48 with India and 7 with Pakistan. This distribution of false positive errors over time provides a more detailed view of the classifier’s performance. Because the classification model itself does not change after training, changes in the distribution of errors reflect changes that have arisen in a given dialect after training.

The question here is whether the relationship be-

tween dialects (geographic variation) changes over time. We model this using a Vector Error Correction Model (VECM: Lütkepohl and Krätzig 2004). This model checks for relationships between multiple time series, which in this case reflect changing error patterns between dialects. The data represents a non-stationary time series because the number of errors in all dialects increases over time (i.e., there is a decline in performance as shown in Table 3). To partially control for the increase in errors over time, we examine the relative frequency of false positives by country by month. The VECM model allows us to determine if there is a significant long-term trend in the distribution of errors from a given dialect, robust to short-term variations.

We examine the significant changes by country for the inner-circle and outer-circle models with CxG features in Table 4. Only significant changes are shown; negative values indicate that samples for the row have become more frequently mistaken for the column. Thus, for the inner-circle varieties, NZ becomes more similar over time to Australia and the UK. This means that, in addition to lower classification performance, NZ is also subject to the most change in the way it is situated among other dialects. Outer-circle varieties on the whole are subject to more change in error distribution over time than inner-circle varieties. The analysis of decay rates also shows that NZ was subject to change over time; the difference is that this analysis takes into account the distribution of errors rather than viewing the error rate as a black box. The outer-circle varieties have a changing error distribution, but not a changing error rate.

7 Syntactic Variation Within Countries

While previous work has viewed a dialect area as a homogeneous entity, here we have sampled from approximately 100 points for each country and maintained a consistent sample over time. To what degree is the performance of dialect classifiers driven by geographic trends within a country? If a country like Australia has a single dominant grammar, then the performance of the syntax-based classifier should be relatively consistent within that country. To test this hypothesis, we look at the average accuracy over time for samples collected from each point within a country.

This is shown in Table 5 with a global Moran’s I used as a measure of spatial autocorrelation within a country (Anselin, 1988). A common method in

	Moran's I	Mean Acc.	Min	Max
AU	0.30	61%	18%	83%
CA	0.54	65%	07%	100%
IE	0.17	58%	35%	89%
NZ	0.20	36%	08%	62%
UK	0.22	73%	41%	82%
US	0.18	79%	53%	97%

	Moran's I	Mean Acc.	Min	Max
GH	0.30	86%	42%	94%
IN	0.38	84%	27%	95%
KE	0.24	89%	62%	97%
MY	0.70	79%	50%	95%
PK	0.42	70%	15%	87%
PH	0.20	77%	37%	88%

Table 5: Geographic Variation in Performance

geospatial statistics, Moran's I measures the correlation in a single variable (here, prediction accuracy for dialect classification) across different locations. This measure has values closer to 1 when the variable is highly spatially organized and closer to 0 when there is no spatial organization. Given that there are different numbers of samples from each location, it is possible that a generic Moran's I would view sparse locations as outliers; thus, we use the Empirical Bayes rate adjustment to control for the level of precision in each location as well (Xia and Carlin, 1998; Anselin et al., 2006).

The table also shows the mean accuracy across cities and the min and max accuracy. These results show that there is an effect for location: the dialect models work well in some places and not so well in others. The Moran's I determines whether this variation in performance is spatially structured. Because different locations represent different populations, these are measures of how well the dialect models work for the entire population of a country. Full maps and spatial results are available in [the supplementary material](#).

All countries have a significant spatial pattern to their accuracy distribution. Within inner-circle countries, Canada has the highest deviation, with a wide range in accuracy and a significant spatial structure to that variation. The US and UK have the highest accuracy, while NZ performs much worse than other dialects, perhaps because of the change over time discussed above. To explore this further, we visualize the internal variation for NZ, the inner-circle dialect with the lowest performance and the most change over time, in Figure 3. Each collection

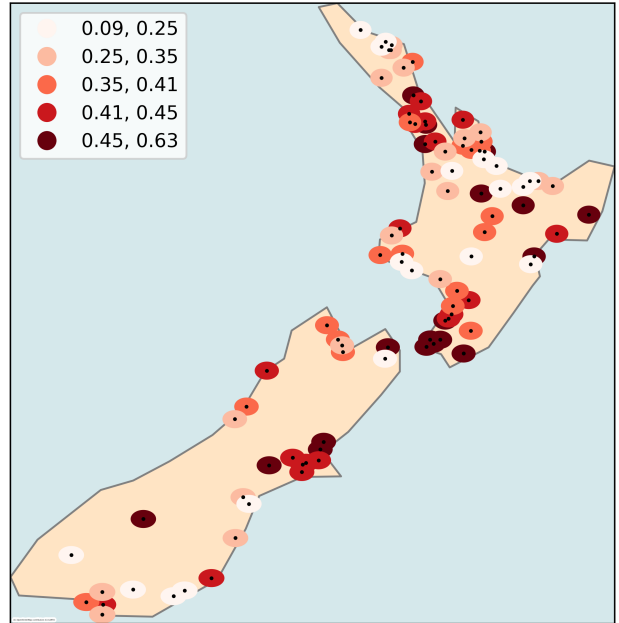


Figure 3: Map of Average City-Level Accuracy, NZ

point is a dot and the shading in the surrounding radius represents the accuracy for that collection area. Darker colors represent higher accuracy. The main cities (Auckland, Wellington, Christchurch) have the most consistent performance. But areas with known distinct linguistic landscapes like Northland (far north) and Southland (far south) have much lower accuracy. More rural areas around the country have consistently lower accuracy as well. The main point in this spatial error analysis is that, because different locations represent different populations, the observed variations in accuracy show that these dialect models do not equally represent all populations within the country.

8 Conclusions

This paper has shown that syntax-based dialect classifiers can reveal both spatial and temporal patterns in linguistic variation. We find that the models remain robust over time, with a fixed decay rate, with the exception of change observed in NZ. This means that, while classification performance does decline, the rate of decline is predictable and evenly distributed. Within dialect regions, however, there is a significant spatial effect on performance. This evaluation is important for establishing an understanding of how dialect models and other geographic models function in the face of on-going linguistic change and population change over space and time. Here, even the best dialect models do not equally represent all speakers of a dialect.

References

- B. Adams. 2015. Finding Similar Places using the Observation-to-Generalization Place Model. *Journal of Geographical Systems*, 17(2):137–156.
- B. Adams and G. McKenzie. 2018. Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification. *Transactions in GIS*, 22(2):394–408.
- B. Alex, C. Llewellyn, C. Grover, J. Oberlander, and R. Tobin. 2016. Homing in on Twitter users: Evaluating an enhanced geoparser for user profile locations. *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3936–3944.
- M. Ali. 2018. Character Level Convolutional Neural Network for Arabic Dialect Identification. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 122–127.
- L. Anselin. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.
- L. Anselin, N. Lozano-Gracia, and J. Koschinsky. 2006. Rate transformations and smoothing. Technical report, Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana, IL.
- S. Argamon. 2018. Computational Forensic Authorship Analysis: Promises and Pitfalls. *Language and Law*, 5(2):7–37.
- J. Bybee. 2006. From Usage to Grammar: The mind’s response to repetition. *Language*, 82(4):711–733.
- B. Chakravarthi, G. Mihaela, R. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, and M. Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11.
- P. Cook and J. Brinton. 2017. Building and Evaluating Web Corpora Representing National Varieties of English. *Language Resources and Evaluation*, 51(3):643–662.
- W. Croft. 2013. Radical Construction Grammar. In *The Oxford Handbook of Construction Grammar*, pages 211–232. Oxford University Press.
- M. Davies and R. 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.
- J. Dunn. 2017. Computational Learning of Construction Grammars. *Language & Cognition*, 9(2):254–292.
- J. Dunn. 2018a. Finding Variants for Construction-Based Dialectometry: A Corpus-Based Approach to Regional CxGs. *Cognitive Linguistics*, 29(2):275–311.
- J. Dunn. 2018b. Modeling the Complexity and Descriptive Adequacy of Construction Grammars. In *Proceedings of the Society for Computation in Linguistics*, pages 81–90.
- J. Dunn. 2019a. Frequency vs. Association for Constraint Selection in Usage-Based Construction Grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 117–128.
- J. Dunn. 2019b. Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology. *Frontiers in Artificial Intelligence*, 2:15.
- J. Dunn. 2019c. Modeling Global Syntactic Variation in English Using Dialect Classification. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 42–53.
- J. Dunn. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54:999–1018.
- J. Dunn and B. Adams. 2020. Geographically-Balanced Gigaword Corpora for 50 Language Varieties. In *Proceedings of the International Language Resources and Evaluation Conference*, pages 2528–2536.
- J. Dunn, T. Coupe, and B. Adams. 2020. Measuring Linguistic Diversity During COVID-19. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science*, pages 1–10.
- J. Dunn and A. Nini. 2021. Production vs Perception: The Role of Individuality in Usage-Based Grammar Induction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 149–159.
- J. Dunn and H Tayyar Madabushi. 2021. Learned Construction Grammars Converge Across Registers Given Increased Exposure. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 268–278.
- J. Eisenstein, B. O’Connor, N. Smith, and E. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 221–227, 287.
- J. Eisenstein, B. O’Connor, N. Smith, and E. Xing. 2014. Diffusion of lexical change in social media. *PloSOne*, 10:1371.
- N. Ellis. 2007. Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, 27(1):1–24.
- M. Gaman, D. Hovy, R. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, C. Purschke, Y. Scherrer, and M. Zampieri. 2020. A Report on the VarDial Evaluation Campaign 2020. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14.

- A. Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, Oxford.
- J. Goldsmith. 2006. [An Algorithm for the Unsupervised Learning of Morphology](#). *Natural Language Engineering*, 12(4):353–371.
- E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2019. [Learning word vectors for 157 languages](#). In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3483–3487.
- J. Grieve. 2007. [Quantitative authorship attribution: An evaluation of techniques](#). *Literary and Linguistic Computing*, 22(3):251–270.
- J. Grieve, C. Montgomery, A. Nini, A. Murakami, and D. Guo. 2019. [Mapping Lexical Dialect Variation in British English Using Twitter](#). *Frontiers in Artificial Intelligence*, 2:11.
- P. Grünwald and J. Rissanen. 2007. *The Minimum Description Length Principle*. MIT Press.
- G. Hirst and O. Feiguina. 2007. [Bigrams of syntactic labels for authorship discrimination of short texts](#). *Literary and Linguistic Computing*, 22(4):405–417.
- P. Hopper. 1987. Emergent Grammar. In *Proceedings of the Berkeley Linguistics Society*, pages 139–157.
- D. Hovy and C. Purschke. 2018. [Capturing regional variation with distributed place representations and geographic retrofitting](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394.
- A. Jørgensen, D. Hovy, and A. Søgaard. 2015. [Challenges of studying and processing dialects in social media](#). *Proceedings of the Workshop on Noisy User-Generated Text*, pages 9–18.
- D. Jurgens, Y. Tsvetkov, and D. Jurafsky. 2017. [Incorporating Dialectal Variability for Socially Equitable Language Identification](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 51–57.
- B. Kachru. 1990. *The Alchemy of English The spread, functions, and models of non-native Englishes*. University of Illinois Press, Urbana-Champaign.
- T. Kreutz and W. Daelemans. 2018. [Exploring Classifier Combinations for Language Variety Identification](#). *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 191–198.
- M. Kroon, M. Medvedeva, and B. Plank. 2018. [When Simple n-gram Models Outperform Syntactic Approaches Discriminating between Dutch and Flemish](#). *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 225–244.
- R. Langacker. 2008. *Cognitive Grammar A basic introduction*. Oxford University Press.
- L. Lucy and D. Bamman. 2021. [Characterizing English variation across social media communities with BERT](#). *Transactions of the Association for Computational Linguistics*, 9:538–556.
- H. Lütkepohl and M. Krätzig. 2004. *Applied Time Series Econometrics*. Cambridge University Press.
- J. Nerbonne and W. Wiersma. 2006. [A measure of aggregate syntactic distance](#). In *Proceedings of the Workshop on Linguistic Distances*, pages 82–90.
- Dat Quoca Nguyen, Dai Quocb Nguyen, Dang Ducc Pham, and Son Baod Pham. 2016. [A robust transformation-based learning approach using ripple down rules for part-of-speech tagging](#). *AI Communications*, 29(3):409–422.
- S. Petrov, D. Das, and R. McDonald. 2012. [A universal part-of-speech tagset](#). In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2089–2096.
- A. Rahimi, T. Baldwin, and T. Cohn. 2017. [Continuous representation of location for geolocation and lexical dialectology using mixture density networks](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 167–176.
- N. Sanders. 2007. [Measuring syntactic difference in British English](#). In *Proceedings of the ACL Student Research Workshop*, pages 1–6.
- E. Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- H. Tayyar Madabushi, L. Romain, D. Divjak, and P. Milin. 2020. [CxGBERT: BERT meets Construction Grammar](#). In *Proceedings of the International Conference on Computational Linguistics*, pages 4020–4032.
- D. Villegas, D. Preoțiu-Pietro, and N. Aletras. 2020. [Point-of-Interest Type Inference from Social Media Text](#). In *arXiv*, 2009.14734.
- B. Wing and J. Baldrige. 2014. [Hierarchical Discriminative Classification for Text-Based Geolocation](#). In *Proceedings of the Conference on Empirical Methods in NLP*, pages 336–348.
- H. Xia and B. Carlin. 1998. [Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality](#). *Statistics in Medicine*, 17:2025–2043.
- E. Zenner, D. Speelman, and D. Geeraerts. 2012. [Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch](#). *Cognitive Linguistics*, 23(4):749–792.