Constructions 18 (2026)
doi10.24338/cons-755

How register and region shape the language network: Evidence from Computational Construction Grammar

Cameron Morin[1], Steven Coats[2] & Jonathan Dunn[3]
[1] Université Paris–Cité [2] University of Oulu [3] University of Illinois Urbana-Champaign

Abstract

While Construction Grammar has proven effective at modelling regional and register variation separately, it has seldom been used to explore the interaction between the two. The present paper fills this gap by applying a Computational Construction Grammar framework to a collection of large English corpora, including two digital registers (written tweets and spoken YouTube transcripts) and five inner-circle varieties (US, UK, Canada, Australia, and New Zealand). We show that constructionist principles successfully capture a range of register- and region-based distinctions across the grammar, and we report the novel finding that both sources lead to systematic, largely independent patterns of variation. Specifically, register effects are more pervasive and concentrated in abstract, high-level constructions, while regional effects are relatively sparser and manifest most prominently in lower-level, surface constructions. To account for these results, we hypothesise that register and regional associations operate along a continuum of constructional 'salience': while the former require the explicit learning of variants for communicative functions, the latter begin as products of exposure before they can acquire indexicality. We conclude with implications of our study for a more comprehensive model of variation in the language network, as well as for future endeavours towards intersecting Construction Grammar and sociolinguistic theory.

1 Introduction

It is often claimed that Construction Grammar is an adequate theory for modelling sociolinguistic variation (Östman & Trousdale 2013; Colleman & Noël 2025; Morin in press), although efforts to put this to the test at scale have been few and far between. From an empirical perspective, most studies investigating social dimensions of constructions have examined individual variants directly (e.g. double modals in Scots, Morin et al. 2024), constructional alternations reminiscent of the variationist 'sociolinguistic variable' (e.g. definite article reduction in Lancashire, Hollmann & Siewierska 2011), or specific families of constructions predetermined on formal and functional grounds (e.g. possessive noun phrases in Lancashire, Hollmann & Siewierska 2007). The main limitation of these approaches

Corresponding author(s): Cameron Morin, Université Paris–Cité, cameron.morin@u-paris.fr

is that 'the grammar, a network which connects thousands of structures at different levels of abstraction, is reduced to a few disconnected variables' (Dunn 2023a: p. 1). In other words, the impact of variation is captured in a small and necessarily incomplete portion of language, whereas the ideal search space for this phenomenon would be the entire grammar, befitting a framework of 'constructions all the way down' in a complex-adaptive system (Goldberg 2006; Beckner et al. 2009: p. 18).

Achieving a holistic analysis of variation in the 'constructicon' (Diessel 2023) would provide strong evidence for the compatibility between Construction Grammar and sociolinguistics. However, such an objective runs us against a substantial methodological challenge. The linguist would need to observe an entire grammar derived from naturalistic language use, whose nodes are derived from nothing more than the emergent structural principles of Construction Grammar (Goldberg 2013). Then, they would need to identify all instances of variation occurring across the grammar, relying on socially conditioned subsets of the data, and with no pre-specified variable features. While this comprehensive approach would be highly desirable for constructionists, variationists, and empirically-minded linguists alike, its undertaking would have seemed monumentally difficult, if not technically impossible, as recently as a few decades ago.

In this article, we leverage recent advances in Computational Construction Grammar (henceforth CCG) to conduct precisely this kind of systematic study. CCG is an unsupervised machine learning paradigm that acquires, represents, and arranges constructions in a grammatical network from corpus data alone (Dunn 2017). It simulates the usage-based learning of language, starting from surface-level chunks 'which are then increasingly generalized given distributional information to form more and more abstract representations' (Dunn 2024: p. 3). Beyond offering a replicable and falsifiable method of constructional analysis with structures of unprecedented detail, CCG has enabled the first bottom-up accounts of variation in grammar. Notably, it has been deployed as a tool for construction-based dialectometry, automatically detecting variants in regional subcorpora (Dunn 2018a). More recently, it has also been used to assess how register differences (Biber & Conrad 2019) affect grammatical varieties and their convergence patterns under increased exposure (Dunn & Madabushi 2021). Assuming that register constitutes a sociocultural dimension of probabilistic grammars alongside categories of social identity (Szmrecsanyi 2019; Leclercq & Morin 2025), CCG thus provides unique opportunities to deepen our understanding of inherent variability (Labov 1969) and to refine models of contextual dynamics in the language network (Diessel 2019).

So far, however, the two types of variation mentioned above have only been examined in isolation rather than in tandem, as would be appropriate in a truly usage-based framework of language where all factors intersect.[1] Our contribution aims to better engage with this expectation by studying the effects of register and region simultaneously using CCG. Through the inclusion of these two predictors

[1] This observation applies specifically to Construction Grammar; corpus-based variationist studies have jointly incorporated register and region in multivariate designs (see, e.g., Bohmann 2019).

in a single computational experiment, with corpus data stratified along both dimensions, we propose a more complete account of constructional variation than has been previously available. In addition to the descriptive benefits expected from this endeavour, our goal is to elucidate crucial theoretical questions. What is the exact relationship between register and region in the formation of linguistic structure? Does one source of variation lead to stronger differentiations than the other? Are the two factors independent or interdependent? Do they yield similar kinds of impact across domains of the constructicon, or do they exhibit distinct, clearly circumscribed areas of influence at various levels of abstraction? If differences emerge on any of these planes, are they categorical or gradient? Finally, and perhaps most significantly, what might be the causes of the relative patterns that we attest?

With these research objectives in mind, we apply the inductive method of CCG to a collection of large English corpora, ranging from several million to several billion words. Section 2 presents the data sources in detail, including two digital registers (written Tweets and spoken YouTube transcripts) and five represented inner-circle varieties (United States, United Kingdom, Canada, Australia, and New Zealand). In Section 3, we outline how CCG learns and organises grammars from this natural language data, focusing on the operationalisation of constructions according to 'level of abstraction', 'order of emergence', and 'degree of centrality'. Section 4 then describes the experiment setup: a general grammar is derived from the entire set of corpora, and classifiers are trained to predict register and regional variation in unseen samples of the same corpora. Section 5 reports the main results of the study and underlines a surprising new finding: two distinct, largely independent patterns of variation emerge, whereby register effects cluster in more abstract, schematic constructions with excellent prediction accuracy, while regional effects are distributed in lower-level, surface constructions with less consistent classification scores. In Section 6, we discuss this prominent finding, and hypothesise that it is attributable to degrees of sociocognitive 'salience' – the extent to which constructional associations 'stand out' in the mind (Morin in press: Chapter 6). Specifically, we argue that (i) register variants tend to be explicitly learned as communicative norms for societal functions (Biber & Conrad 2019), making them highly salient and entrenched from the get go, while (ii) dialect variants self-organise more continuously, beginning as non-salient, surface-level exposure traces before they can become 'markers' and 'stereotypes', acquiring further indexicality (Johnstone et al. 2006; Jaeger & Weatherholtz 2016). In addition, we consider some implications of this study for theorising work on the 'social meaning' of constructions (Leclercq & Morin 2025). Finally, we conclude on the paper's main contributions and their potential for the future of construction-based sociolinguistics.

2 Corpus data

The dataset used for this study is overviewed in Table 1 below. It covers two broad English registers, consisting of more than 2 billion words from written social media posts (Tweets) and over 430 million words from video speech tran-

scripts (YouTube). These registers are further stratified across five 'inner-circle' Anglophone countries (Kachru 1985), namely the United States, United Kingdom, Canada, Australia, and New Zealand. Each of the subcorpora contains data for multiple individual cities, ranging from 50 (NZ) to 97 (US), and totalling 392.

Country	N. Cities	Tweets		YouTube	
		Samples	Words	Samples	Words
Australia (AU)	84	218,815	547 mil	11,390	28 mil
Canada (CA)	68	142,591	356 mil	11,445	28 mil
New Zealand (NZ)	50	39,371	98 mil	7,969	19 mil
United Kingdom (UK)	93	186,755	466 mil	10,494	26 mil
United States (US)	97	234,784	586 mil	133,163	332 mil
TOTAL	392	822,316	2.05 bil	174,461	436 mil

Table 1: Overview of corpora

Individual observations (i.e. tweets from the same place) are aggregated into larger samples of approximately 2,500 words, which are produced in the same location but not by the same person. For instance, a sample may represent an excerpt of Tweets written by several users in Chicago, or a snippet of speech by distinct content creators and recorded individuals in Christchurch. Although the total number of samples is variable both across registers and varieties (e.g. 822,000 for Tweets and 174,000 for YouTube), the sampling method enables us to introduce symmetry in the dataset, so as to replicate differences in subcorpora over many different populations. For example, the results will allow us to detect potential signals that spoken and written English differ between rural Australia and urban Britain, regardless of the perturbations caused by raw data amounts for these two locales.

The tweet-based data is extracted from the Corpus of Global Language Use (Dunn 2020), which was compiled through geographic searches within 25km of 10,000 cities around the world. For this study, only those cities within our five countries of interest are included, and the samples represent a period from 2018 to 2023. We do not have individual or demographic metadata for each post, but the aggregation of tweets serves to average out such differences within populations, so long as we can assume that the platform (Twitter/X) has the same demographics across each of these 392 cities (Dunn 2025b; see also Morin & Grieve 2024).

Meanwhile, the spoken data used in this paper is extracted from three corpora of automatic speech recognition transcripts, covering content uploaded to the

YouTube channels of local government entities in the US, Canada, UK, Australia, and New Zealand: the Corpus of North American Spoken English (CoNASE, Coats 2023), the Corpus of British Isles Spoken English (CoBISE, Coats 2022b), and the Corpus of Australia and New Zealand Spoken English (CoANZSE, Coats 2022a). Because the transcripts are mostly recordings of local government meetings, the speech content sampled from these locations is largely comparable in terms of register, communicative contexts, and discursive topics. However, the corpora also include a range of spoken genres, including interviews, vlog-style commentaries, public service announcements, and news reports. For our current purposes of symmetrical comparison, only the transcripts matching the 392 cities defined in the Tweet collection were included in the dataset. The samples represent a period from 2004 to 2023, with the majority found in the years 2019–2022.[2]

3 Operationalising constructions

Rather than exploring variation in a small number of manually-chosen features, this paper uses the framework of Computational Construction Grammar (CCG; Dunn 2024) to identify a large number of 'constructions', which are then evaluated for register and regional effects. While we do not claim that the learned grammar constitutes the definitive description of the English constructicon, it does offer a more comprehensive account than previously available methods, and crucially, one that is both replicable and falsifiable.

3.1 Bottom-up learning

The CCG system induces constructions through a multi-stage, bottom-up process that moves from raw text to hierarchically organised grammatical patterns, without requiring pre-specified structure. This process begins with learning distributed representations of words, proceeds through the formation of 'slot constraints', and ultimately yields a network of constructions at multiple levels of abstraction.

The first stage involves creating two complementary sets of word representations that capture different aspects of distribution. Continuous Bag-of-Words (CBOW) models with narrow context windows (1 word in each direction) learn which words appear in immediately adjacent positions, capturing syntagmatic relationships—for instance, *can*, *might*, and *should* pattern similarly before base-form verbs. Meanwhile, skip-gram models with wider context windows (5 words in each direction) learn which words co-occur within the same topics or meaning frames, capturing paradigmatic relationships—e.g. the clustering of *oak*, *elm*, and *chestnut* in natural settings. Both models employ fastText character-based embeddings, allowing the system to handle morphological variation (e.g., *flow*, *flows*, *flowing*). The result is a continuous vector space where each word is located

[2] As pointed out by a reviewer (p.c), the potential for diachronic effects within this time window is an intriguing question. CCG has not yet been systematically applied to the study of language change, though the framework's bottom-up approach to grammar induction could offer novel insights into constructional change over time. We leave this for future research.

based on its distributional properties.

Once these representations are established, words are grouped into categories that will serve as slot constraints in constructions. Using k-medoids clustering on the embedding spaces, three types of constraints emerge: lexical constraints (LEX) for specific word forms (e.g., *the*, *under*), syntactic constraints (SYN) based on local distributional patterns, and semantic constraints (SEM) based on broader co-occurrence patterns. Each category is defined by a prototypical exemplar (Bybee 2013) and exhibits gradient membership (Ungerer 2023), with goodness-of-fit measured by cosine distance to other items. For example, (1) shows a lexical slot filler with *suppose* as its exemplar, and similarity scores for the remaining members ranging from 0.92 for *pretend* to 0.88 for *expect*.

(1) Exemplar: *suppose*
   a. *pretend* (0.92)
   b. *think* (0.89)
   c. *misunderstand* (0.89)
   d. *believe* (0.88)
   e. *expect* (0.88)

3.2 Constructing the grammatical network

With slot constraints in place, the system then identifies sequences that qualify as constructions based on three entrenchment criteria: (i) token frequency above a parts-per-million threshold, (ii) internal coherence between slots through the directional association metric of ΔP (Ellis 2007; Dunn 2018b), and (iii) balance between storage of specific patterns and productive generalisations through the metric of Minimum Description Length (Grünwald 2007). The Minimum Description Length (MDL) metric is a sort of loss function that is used to evaluate the quality of the grammar as a set of constructions rather than individual constructions on their own. Thus, the storage of redundant or overlapping representations may be allowed, so long as those overlapping representations are useful for describing idiomatic usage in the testing corpora. In the MDL paradigm, balance is achieved when the cost of encoding redundant constructions (i.e., storing two overlapping constructions instead of a single more schematic construction) is less than the improvement gained when encoding a test corpus (i.e. the complexity of the grammatical description of that corpus).

This process initially yields what we term 'first-order' constructions—9,504 distinct types in our dataset—representing the lowermost layer of the grammatical network.[3] For instance, (2) and (3) are first-order constructions, exhibiting similar structures as prepositional phrases, but encoding distinct meaning domains:

[3] Note that it is possible for different stages of the grammar to relearn the same construction independently. A close examination of the features in the supplementary material will reveal that 11 constructions are learned twice in the mid-stage and late-stage grammars and 421 in the early-stage and late-stage grammars. These repeated constructions will not influence the classification models used to describe regional and register variation because the particular model is robust to redundant or correlated features.

spatial relations in a 'botanical' context versus abstract relations in a 'legal' context.

(2) [SEM:6 ⟨above_under⟩ — "the" — SYN:136 ⟨cottonwood⟩]
   a. "above the oak"
   b. "under the chestnut"
   c. "among the cedars"
   d. "under the trellis"

(3) [SEM:6 ⟨above_under⟩ — "the" — SYN:206 ⟨prosecution⟩]
   a. "of the act"
   b. "under the law"
   c. "of the offense"
   d. "under the interdict"

Building from these first-order constructions, the grammar continues to emerge through scaffolded learning as increasingly abstract patterns are discovered. Second-order constructions are induced by 'clipping' together first-order constructions that share overlapping slot constraints (Dunn 2024: pp. 56–59). Once formed, these second-order constructions are included alongside first-order constructions for future analyses. The abstraction process continues further upward: families of second-order constructions that share formal and functional properties are gathered into third-order constructions (n = 654 in our dataset), while fourth-order constructions represent highly schematic patterns encoding broad grammatical generalisations (n = 54). Table 2 summarises the distribution of constructions across these levels and other organisational dimensions that we will use to study variation (see Section 3.3).

Level of Abstraction	First-Order	Third-Order	Fourth-Order
	9,504	654	54
Order of Emergence	Early-Stage	Mid-Stage	Late-Stage
	1,737	983	6,628
Degree of Centrality	Peripheral	Middle	Core
	9,004	1,045	163

Table 2: Number of constructions by way of dividing the grammar (overlapping)

To illustrate how these levels relate to one another, consider examples (4) and (5), which showcase two first-order constructions belonging to a single third-order construction. Both encode motion event verbs with adpositional phrases indicating trajectory. As with many learned constructions, the head adposition is included as a slot but the dependents on that head are not, since these would be connected later in the construction parsing process, when representations are

further clipped together to represent larger utterances. While structurally similar, these constructions differ in aspectual properties: the verb slot constraint in (4) appears to be more open than in (5), allowing a broader range of motion verbs not restricted to activities (Vendler 1957).

(4) [SYN:10 ⟨tapped_stumped⟩ — SYN:165 ⟨through⟩ — SYN:165 ⟨through⟩ — SEM:2285 ⟨another_this⟩]
   a. “run away from a”
   b. “skipped out on the”
   c. “buzzed down on a”
   d. “blown away by the”

(5) [SYN:239 ⟨cruising_gliding⟩ — SYN:165 ⟨through⟩ — SYN:165 ⟨through⟩]
   a. “drive up from”
   b. “dancing around with”
   c. “coming out from”
   d. “coming out onto”

Another third-order family is illustrated through constructions (6) and (7), which represent noun phrases introducing quantity expressions. Though both constructions share the general meaning of defining some quantity or subset of another nominal, and thus function much like an adjective phrase to be clipped onto a second nominal, they encode different scales: (6) focuses on larger quantities, while (7) encodes smaller quantities.

(6) [SEM:2285 ⟨another_this⟩ — SYN:13 ⟨stubble_coarse⟩ — SEM:1418 ⟨these_those⟩]
   a. “a hint of”
   b. “a load of”
   c. “a pile of”
   d. “a heap of”

(7) [SEM:2285 ⟨another_this⟩ — SEM:2344 ⟨bit_little⟩ — SEM:1418 ⟨these_those⟩]
   a. “a bit of”
   b. “single bit of”
   c. “a smidgen of”
   d. “a tad of”

From a variation perspective, individual dialects might prefer one first-order variant over another while maintaining the same third-order construction, which would suggest that such effects manifest at surface levels of knowledge rather than throughout the network. Fourth-order constructions are more abstract and difficult to analyse; we leave detailed discussion of these for elsewhere (Dunn 2024: Section 3.3).

3.3 Three dimensions for analysing variation

To develop a fine-grained analysis of how register and regional factors impact different areas of the language network, we organise the learned grammar along three complementary dimensions (Table 2). The first dimension, 'Level of Abstraction', differentiates between first-, third-, and fourth-order constructions as described above, operationalising a gradient from specificity to schematicity. The second dimension, 'Order of Emergence', indicates the point when constructions become learnable during the acquisition process: early-stage constructions exclusively comprise lexically specific patterns (n = 1,737), mid-stage constructions extend to include local syntactic categories (n = 983), while late-stage constructions require both syntactic and semantic categories for their identification (n = 6,628). The third dimension, 'Degree of Centrality', quantifies entrenchment[4] based on frequency, radiating from a core of highly frequent constructions (n = 163) outwards to moderately frequent constructions (n = 1,045) and low-frequency, 'peripheral' constructions (n = 9,004). We define core constructions as those greater than one standard deviation above the mean frequency, peripheral constructions as those falling below the mean, with all others constituting the middle frequency band.

Importantly, Table 2 shows that these dimensions overlap in their partitioning of the grammar: for instance, a single construction may simultaneously be classified as first-order, late-stage, and peripheral. Overall, this multidimensional organisation enables us to meticulously examine whether register and regional effects concentrate in particular neighborhoods of the constructicon. With this framework established, we now present our method for detecting variation across the full breadth of linguistic structure.

4 Comparing register and region

Having identified over 10,000 constructions organised across multiple dimensions of the grammar, we turn to the central challenge of this study: how do we systematically examine register and regional variation across such a vast feature space?

Classification models offer a principled solution to this aggregation problem. Rather than manually examining each construction for social conditioning, we employ machine learning classifiers to discover which combinations of constructions systematically distinguish between usage contexts. Our pipeline, illustrated in Figure 1, works as follows. First, we take the learned grammar and use it to annotate corpus samples (see Section 2), calculating the token frequency of each construction per sample. Second, we split these annotated samples into training and test sets. Third, we train two parallel classifiers on the same data: one learns

[4] Strictly speaking, corpus frequency data reflect conventionalisation at the community level rather than entrenchment in individual minds (Schmid 2010). However, following standard assumptions in usage-based linguistics, we treat aggregated frequency as a reasonable proxy for cognitive entrenchment, given the well-documented relationship between the two (Stefanowitsch & Flach 2017).

to distinguish between registers (spoken vs. written) regardless of region, while the other learns to distinguish between regions (the five countries) regardless of register. Crucially, both classifiers operate on identical feature sets, differing only in the category labels they predict. Finally, we evaluate both models on the held-out test samples, measuring how accurately they can predict the register or dialect of unseen language use based on constructional features.

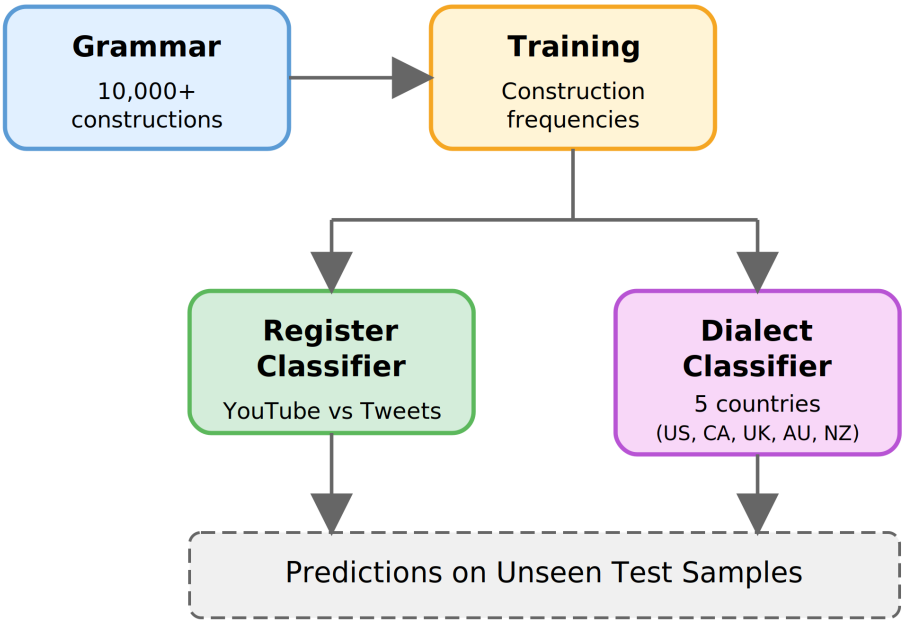

Figure 1: Parallel classification pipeline for assessing register and regional variation

We employ Linear Support Vector Machines (SVMs) for classification, which are particularly effective with correlated features (as grammatical constructions naturally co-occur and overlap). In addition, SVMs feature the advantage of outputting individual linear feature weights, interpretable as each construction's predictive power for a given category. For example, constructions used predominantly in spoken English will show high weights for the register classifier, while constructions specific to New Zealand English will show high weights for the dialect classifier.

To ensure robust results, we use 10-fold cross-validation: our samples are partitioned into ten equal parts, with nine parts used for training and one part reserved for testing, rotating so that each part serves as the test set once. This means that each model is trained and tested ten times on different data splits, allowing us to estimate performance variance and ascertain the independence of findings from any particular division of the data.

We test on over 150,000 samples for dialect classification and over 130,000 for register classification, providing strong foundations for the models' ability to generalise. The validity of our approach rests on classification accuracy: high accuracy (measured using the f-score, with 1.00 indicates perfect prediction) would demonstrate that the model successfully maps patterns to category labels, thereby

confirming the presence of constructional variation. Conversely, low accuracy would show that the model struggles to map the patterns to labels, suggesting a relative absence of variation. While the number of classes differs here between dialect and register classification, recent work has shown that the performance of register classification would remain at this level even with many more classes (Dunn 2026).

As we will see in Section 5, analysing the differential performance of register versus dialect in terms of predictive accuracy (of classifiers) and predictive power (of constructions) will offer important evidence for the impact of both types of variation on the constructicon.

5 Results

This section examines whether constructions exhibit distinct register and regional profiles in various areas of the grammar. Our approach proceeds in four stages. First (Section 5.1), we validate the classification models using prediction accuracy, to make sure that they provide reliable descriptions of constructional variation. If the models achieve high accuracy for the contextual conditioning of constructions based solely on frequencies, we can be confident that they capture systematic linguistic differences rather than mere noise or artifacts. Second (Section 5.2), we analyse the predictive power (i.e. feature weights) assigned to each construction, in order to determine (i) where register and regional effects concentrate in the network and (ii) whether the two dimensions of variation are independent. Third (Section 5.3), we explore the consistency of register distinctions across regional varieties, to further elucidate the exact relationship between these factors. Finally (Section 5.4), we conduct robustness checks to ensure that our findings are not driven by topical confounds from cultural regions.

5.1 Validating the model of constructional variation

Before assessing the impact of variation in the grammar, we must first establish that our classification models capture 'orderly heterogeneity' (Weinreich et al. 1968; Labov 2014) rather than chance associations. This validation is crucial because our analytical pipeline involves multiple stages where errors could accumulate: grammar learning, sample annotation, and category attribution. If any stage introduced substantial noise (i.e. misidentifications of constructions), the resulting model would not represent an adequate generalisation for the analysis of new language samples.

In this case, a grammar learning error would mean that the constructions do not capture the grammar sufficiently to enable the later classifier to detect variants. A sample annotation error would be cases where the grammar represented a relevant construction but the parser consistently missed that construction in these two sets of corpora (Tweets and YouTube transcripts), thus making the construction unavailable for the classifier. Finally, a category attribution error would be a case where, for instance, an abstract legal term is not included in the constraint for a domain-specific construction like (3) above. These three kinds of errors

would accumulate across the pipeline and result in lower prediction accuracy on the evaluation samples.

Table 3 presents the validation results as F-scores, reporting the minimum and maximum F1 at 95% confidence intervals from the 10-fold cross-validation iteration. The F-score combines precision (how many predictions were correct) and recall (how many actual cases were detected) into a single metric ranging from 0 (complete failure) to 1 (perfect prediction). The final scores here are macro-averaged, i.e. they represent the combination of separate F-scores calculated for each class. This constitutes a more stringent measure than simple accuracy, as the model must perform well on all classes to reach a high score (i.e. the five regions or the two registers), not just on the ones with the most data. For example, correctly predicting New Zealand English (our smallest regional subcorpus) counts as much toward the final score as correctly predicting American English (our largest).

	First-Order		Third-Order		Fourth-Order	
	Min F1	Max F1	Min F1	Max F1	Min F1	Max F1
By Level of Abstraction						
Dialect	0.97	0.97	0.72	0.72	0.48	0.48
Register	1.00	1.00	1.00	1.00	1.00	1.00
	Early-Stage		Mid-Stage		Late-Stage	
	Min F1	Max F1	Min F1	Max F1	Min F1	Max F1
By Order of Emergence						
Dialect	0.90	0.90	0.81	0.81	0.97	0.97
Register	1.00	1.00	1.00	1.00	1.00	1.00
	Peripheral		Middle		Core	
	Min F1	Max F1	Min F1	Max F1	Min F1	Max F1
By Degree of Centrality						
Dialect	0.97	0.97	0.79	0.79	0.59	0.59
Register	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Validation results by region in the grammar

The results in Table 3 reveal two striking patterns. First, register classification achieves perfect accuracy (F-score = 1.00) across all portions of the grammar, indicating that our spoken and written varieties of English are strongly and pervasively distinct in their constructional profiles. Second, dialect classification exhibits a clear continuum in two of the three constructional dimensions, with similar intervals. For abstraction level, accuracy decreases from first-order (0.97) through third-order (0.72) to fourth-order constructions (0.48). For centrality, accuracy decreases from peripheral (0.97) through middle-frequency (0.79) to core constructions (0.59). Overall, these gradient trends indicate that region is a relatively sparser and weaker source of variation than register in our learned grammar. In addition, they suggest that regional differences concentrate in the

more specific and less-entrenched areas of the network, an interpretation to be evaluated in more detail shortly.

With nearly all predictive accuracy scores exceeding 0.7, and many approaching or achieving perfect classification, the general results provide robust validation of the CCG approach for studying constructional variation. Furthermore, they offer an initial glimpse into how register and region may lead to systematically different impacts on grammatical structure, as studied below.

5.2 Exploring the differences between register and region

Having established the reliability of the classifiers for observing constructional variation, we now analyse the feature weights directly to explore register and region as organising forces in the grammar. Our estimation procedure again employs 10-fold cross-validation to guarantee statistical consistency, and the results below are based on the averaged values over all folds.

We begin by comparing the predictive power of each construction between both tasks (register versus regional classification), taking the absolute value of each feature weight regardless of directionality. For example, a construction with a register weight of $+0.8$ (strongly predicting spoken) and one with -0.8 (strongly predicting written) will both have a general predictive power of 0.8 for register after the transformation. This procedure allows us to ask: are the constructions that strongly predict register differences also the ones that strongly predict regional differences? For example, we might assume that a first-order lexical construction such as *I reckon* could to some extent be associated with the regional varieties of Australian or New Zealand English, and perhaps UK English, at a national level. Indeed, in our data, when considering the entire set of features, the lexical feature [LEX:i $>$ LEX:reckon] has weights that are much greater in magnitude for Australia, New Zealand, and the UK, compared to the US and Canada (AUS: -0.0034, NZ: -0.0026, UK: -0.0022, CA: 0.0002, US: -0.0001). Because the dialect classifier contains five classes, we represent the average prediction power of each construction using the average of each class-specific weight.

We use Pearson correlations to analyse the predictiveness of the same features for register and regional variety, aggregating across all constructions by level of abstraction, order of emergence, and degree of centrality. Table 4 reports the results, revealing persistently low coefficients in all cases. The very highest scores are 0.41 for fourth-order constructions and 0.24 for core constructions, both of which were shown to be the least predictively accurate of their respective categories (‘abstraction’ and ‘centrality’, Table 3). Everywhere else, the correlations approach zero, indicating virtually no relationship between register and dialect predictiveness. Importantly, this confirms that the two sources of variation operate independently: constructions that distinguish between registers are generally not the same as those that distinguish between regions.

To better understand how this independence manifests across the grammar, we visualise the distribution of individual constructions along both dimensions of variation in the scatterplots of Figures 2 and 3. Each point represents a construction, with register predictiveness on the y-axis and dialect predictiveness on

Abstraction		Emergence		Centrality	
First-Order	0.12	Early	0.06	Peripheral	0.14
Third-Order	0.04	Mid	0.09	Middle	0.13
Fourth-Order	0.41	Late	0.11	Core	0.24

Table 4: Pearson correlations between register and dialect variability of constructions

the x-axis. Point size reflects overall predictiveness (the sum of both dimensions), with larger, darker points indicating constructions that strongly distinguish categories on at least one dimension. High predictiveness points towards sharp stratification: for instance, a construction used exclusively in one dialect would show maximum predictiveness for that dimension.

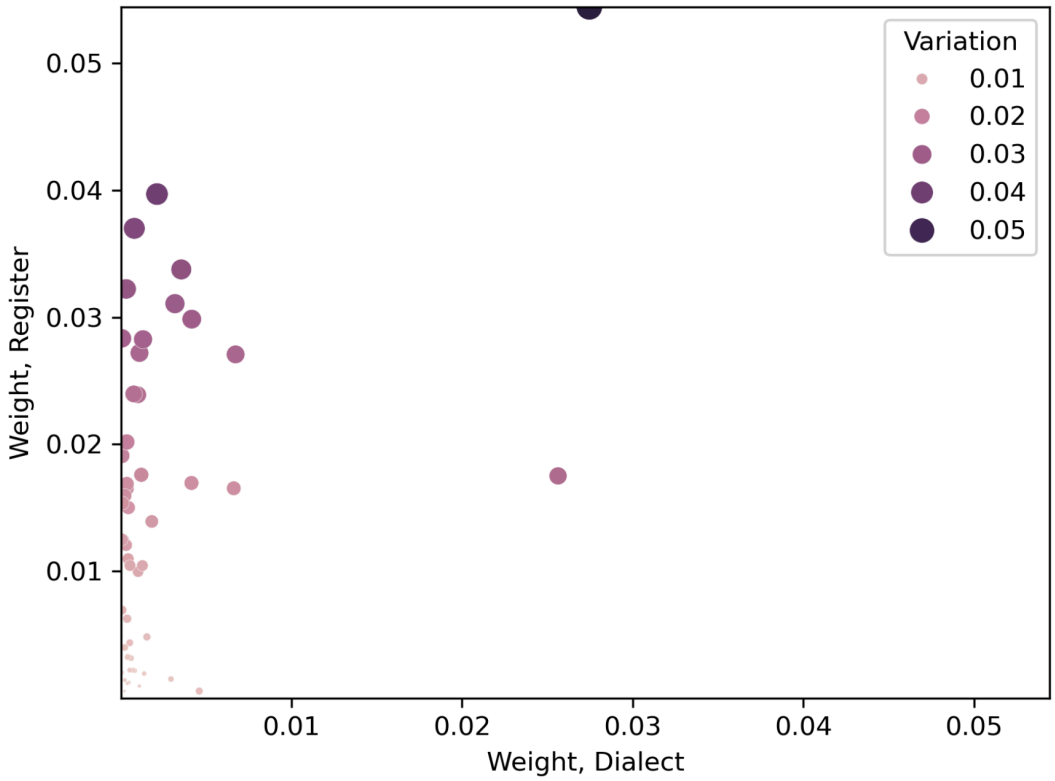

Figure 2: Register vs. dialect predictiveness by construction (fourth-order)

On the one hand, Figure 2 displays the fourth-order constructions, which are the most abstract and schematic patterns in the grammar. These constructions cluster near zero on the dialect axis, confirming their minimal predictive power for regional variation (as suggested in Table 3). In stark contrast, they spread widely along the register axis, with many showing high predictiveness. This indicates that while abstract constructions are shared across regional varieties, they sharply differentiate English recorded on YouTube from English written on Twitter. Two constructions appear as clear outliers for dialect prediction; we address

the impact of such extreme values in Section 5.4.

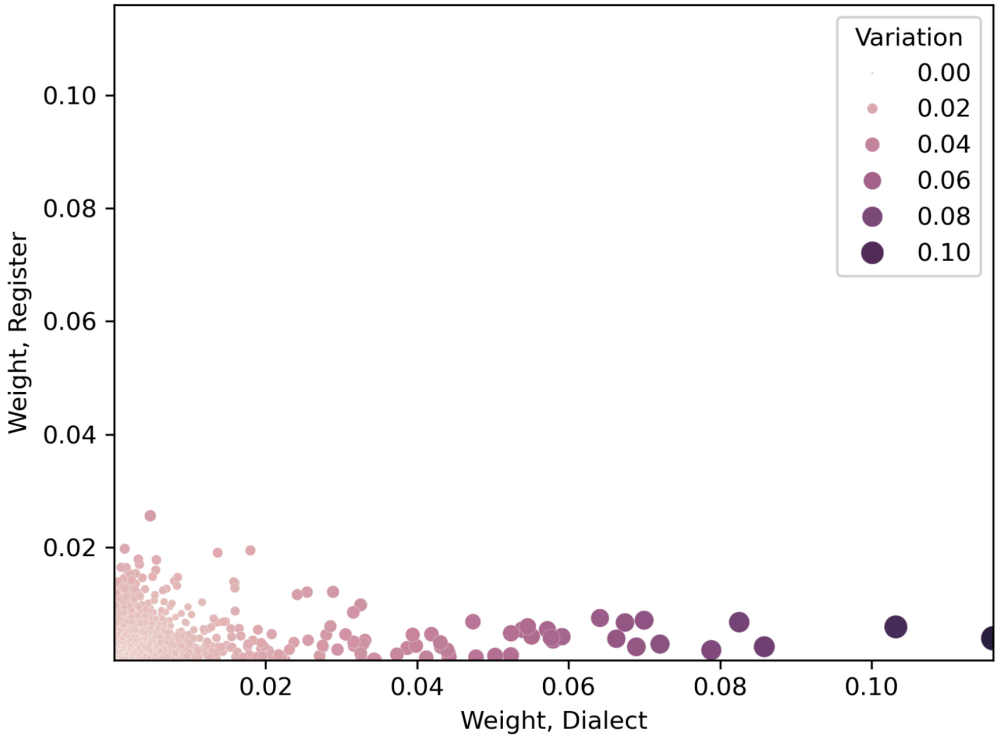

Figure 3: Register vs. dialect predictiveness by construction (first-order)

On the other hand, Figure 3 exhibits a very different pattern for first-order constructions, which constitute the least abstract level of the grammar. Despite achieving high predictive accuracy for both classification tasks (Table 3), these constructions exhibit markedly different predictive power distributions across the two dimensions of variation. For the y-axis, all constructions cluster around very low scores: notably, no individual first-order construction emerges as strongly predictive of register. Conversely, dialect predictiveness spreads towards a subset of constructions with high feature weights, indicating the presence of genuinely region-specific patterns at this surface level of grammatical representation.

Overall, this differential patterning across levels of abstraction foreshadows crucial insights into the nature of grammatical variation, which we will return to in Section 6. Fourth-order constructions indeed represent schematic families, encompassing numerous first- and third-order children that share core functional properties. Through usage-based processes of generalisation during acquisition, we would expect that learners abstract away from surface variation to identify these common functional patterns. It is therefore notable that our results seem to align particularly well with this expectation. Regional differences, which would originate as traces of differential exposure to specific lexical sequences, appear to become attenuated at higher levels of abstraction. By contrast, register differences, which would encode functional distinctions between communicative situations (Biber & Conrad 2019; Li et al. 2023), appear to persist across all levels of grammatical representation, including the most schematic ones. The results

of our CCG approach support this striking asymmetry, and bear important implications for a Construction Grammar model of socially-conditioned language variation, as we will further discuss.

To see if this generalisation holds in the rest of the data, Figure 4 plots third-order constructions, which occupy an intermediate position in the abstraction hierarchy between the surface-level first order (Figure 3) and schematic fourth order (Figure 2). Here, we observe a more balanced distribution of predictiveness across both dimensions, though constructions still specialise in one or the other, as previously shown by the negligible overall correlation ($r = 0.04$, Table 4). The constructions most predictive of dialect are 'SYN_Type_11_Token_5.0', with a value of 0.067080, and 'SYN_Type_14_Token_0.0', with a value of 0.056676. The first category includes items of the form MODAL + *be* + GERUND/PAST PARTICIPLE, such as *will be taking* or *shall be replaced*, among many others. The relatively large weight for this construction may reflect dialect differences in the use of *will* and *shall* in different English varieties. Meanwhile, the second category includes a variety of constructions, including those with the form *to be* + PAST PARTICIPLE (e.g. *to be mitigated*), constructions such as *to* + ADJ (*to bravely*, *to slowly*), MODAL/SEMI-MODAL + *be* (*will be*, *shall be*, *gonna be*), as well as a great many other types, representing various grammatical configurations or lexical bundles.

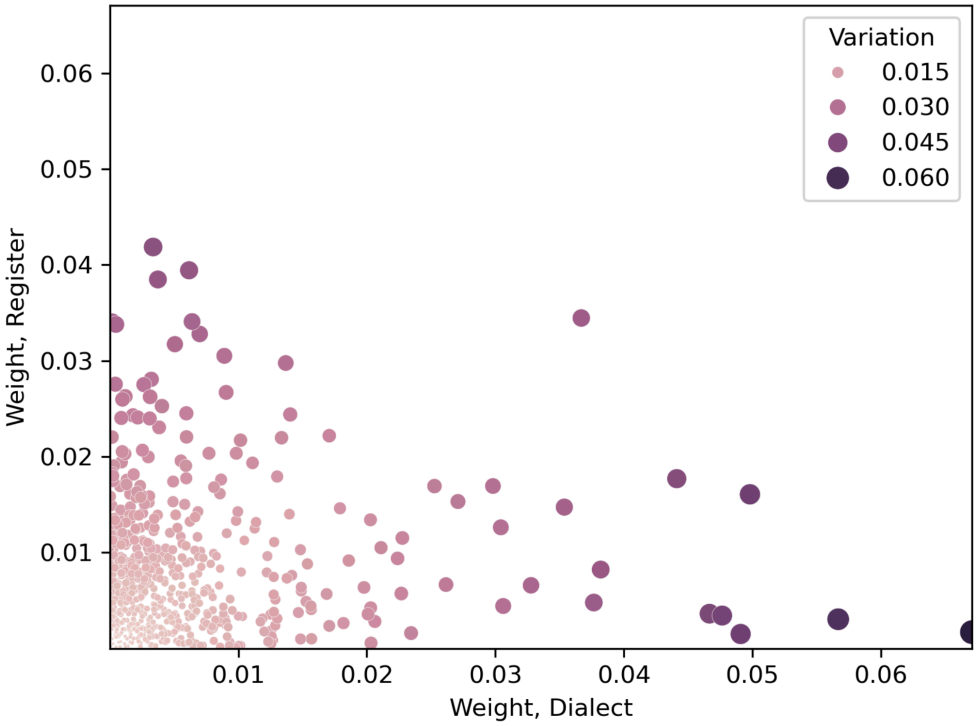

Figure 4: Register vs. dialect predictiveness by construction (third-order)

For the first time in our analysis, we also see that certain constructions are predictive for both dimensions simultaneously. This is the case, for instance, of the construction labelled LEX_Type_2_Token_12.0, which in our data corresponds to three lexical sequences: *the new*, *the news*, and *the ones*. Determining why these constructions co-vary in both register and region merits further investigation, but

the point remains that such cases are exceptional, and that in the aggregate, the two sources of variability appear to be almost entirely independent phenomena.

5.3 Register variation by country

Our next research question is whether the constructional differences between spoken (YouTube) and written (Twitter) registers remain constant across our speech communities of interest (Australia, Canada, New Zealand, United Kingdom, and United States). This will further clarify whether register and regional variation have neatly separate effects on the language network. To this end, we replicate five parallel experiments (one per national variety) that describe register variation in samples annotated by the same inventory of representations, i.e. the learned grammar of 10,000+ constructions.

More specifically, we reproduce the binary register classification task presented in Section 4, but with the key difference that we train five separate models exclusively on data from the respective individual countries. Similarly to our previous approaches, each country-specific model undergoes 10-fold cross-validation for the sake of robust estimates, and we analyse the averaged results over all folds. In essence, our goal is to test whether register variation manifests uniformly across populations: if it does, we should obtain comparable descriptions regardless of which variety we examine, further supporting the idea that it is a particularly enduring, independent source of constructional variation. Conversely, population-specific patterns should produce divergent models, revealing a previously underappreciated conditioning of register associations on regional context.

We first validate the country-specific models by examining their prediction accuracy. Remarkably, all five of them achieve perfect classification (F-score = 1.00) across the 10-fold cross-validation, demonstrating that register distinctions are equally pervasive in the constructicons of each national variety. Next, we examine the consistency of these distinctions across dialects by analysing the country-specific feature weights assigned to each construction. Figure 5 reports pairwise Pearson correlations for first-order constructions across all countries: each coefficient compares the predictive power for register of the same 9,504 constructions between two given countries. A high coefficient indicates that similar constructions distinguish spoken from written registers regardless of national variety, pointing towards the uniformity of this variation type. Conversely, a lower coefficient would suggest dialect-specific patterns in the entrenchment of register-based knowledge.

Figure 5: Heatmap of correlations of register weights by dialect (first-order)

Figure 5 reveals striking patterns of cross-national alignment in register-based constructional variation. The correlations here are markedly stronger than those observed between register and dialect predictiveness (cf. Table 4), with even the weakest relationships exceeding 0.60. This high baseline of agreement confirms that register conventions operate similarly across our five English-speaking nations. However, we also make an intriguing finding when we observe fluctuations in the values between 0.6 and 0.8: namely, the correlation structure appears to reflect established social and historical heritage relationships between varieties. For example, the United States is the closest variety to Canada (0.74), while being the most divergent from the other three. Alternatively, the United Kingdom, Australia, and New Zealand form a tighter register cluster ($>$0.75), with the latter two exhibiting the highest correlation overall (0.8). In other words, although register distinctions are highly consistent across the five countries considered, we still observe subtle regional effects on its entrenchment and conventionalisation (Schmid 2020), which appear to reflect diachronic trajectories. This suggests a small degree of porosity between the two sources of variation that was not detected heretofore.

Importantly, however, these results must not be interpreted as register associations being directly indicative of regional variety: Section 5.2 showed that constructions predicting register and those predicting dialect are sets that generally do not overlap. Rather, the present models examine register differences within each dialect separately, achieving perfect prediction accuracy in every case. What is most likely captured here, then, consists in subtle exposure effects on the acquisition of register: the same functional distinctions between spoken and written communication may be realised by slightly different surface-level constructions, perhaps dependent on country-specific media, institutions, and discourse com-

munities. This represents a valuable new hypothesis to fully consider in future usage-based studies of variation.

To further explore the fine-grained cross-national differences identified in Figure 5, we conducted a regression analysis examining how register predictiveness varies across individual constructions, the results of which are plotted in Figure 6. This analysis compared feature weights for register prediction between the UK (x-axis) and each of the other four varieties (y-axis, differentiated by color and shape). If register distinctions operated identically across all varieties, points would cluster along the diagonal with all regression lines showing the same slope. By contrast, deviations from this pattern reveal country-specific register effects: regression slopes that are shallower or steeper than the diagonal indicate greater divergence from UK register patterns. The distance of constructions from the origin reflects their overall predictiveness: those further from the bottom-left corner distinguish more strongly between spoken and written registers.

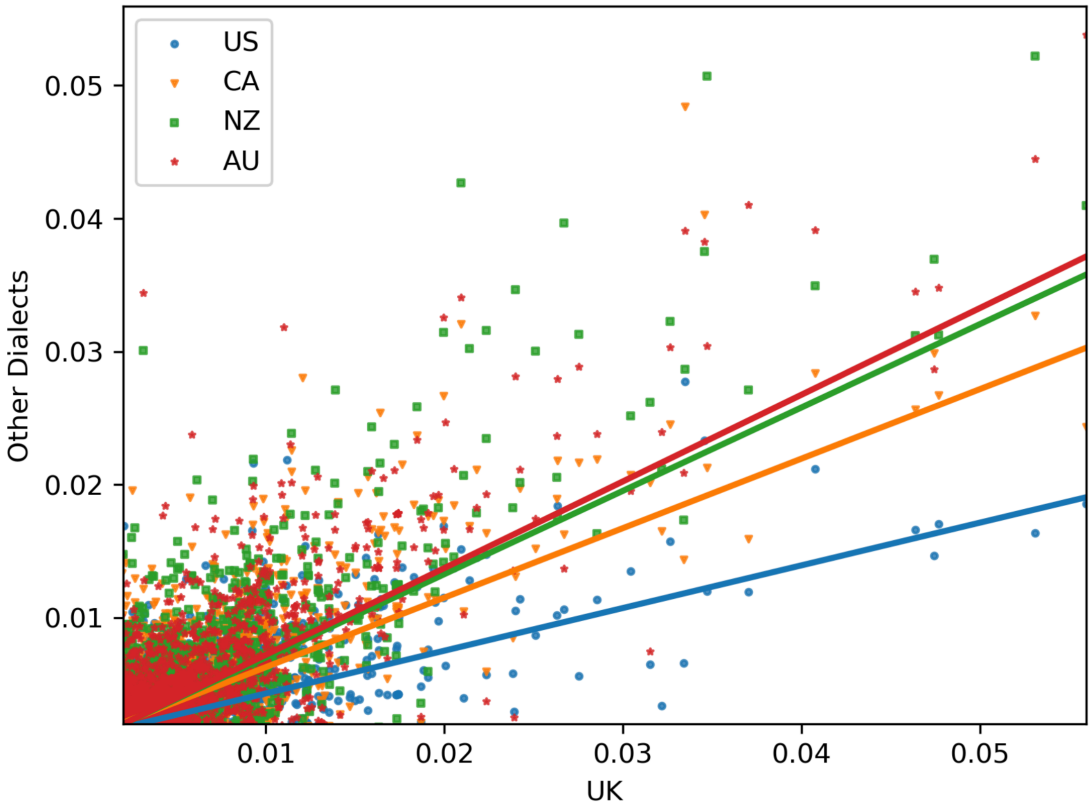

Figure 6: Regression analysis of register predictiveness across dialects (first-order)

Figure 6 shows that the US exhibits the shallowest regression slope among all varieties, confirming its status as an outlier suggested in Figure 5. Importantly, we find that the distinction between spoken and written registers, as manifested in Youtube transcripts and Tweets, is substantially less pronounced in American English than in British, Australian, and New Zealand English, but also in Canadian English, its closest heritage variety. One possible explanation for this pattern is the tendency for grammatical innovations in speech to more quickly be adopted in written language in American English, compared to British English (Leech et al.

2009). Such levelling would result in greater convergence between spoken and written forms, reducing the overall range of register variation in the American constructicon. Although this explanation awaits confirmation in future studies, it may represent an important instantiation of the phenomenon that we have postulated above, i.e. a subtle exposure effect of region on register in the linguistic system.

5.4 Robustness analysis of topical confounds

Following our detailed analysis of constructional variation, we must now verify that the register and regional effects we find are not artifacts of topical confounds in our data. A potential concern is indeed that our strong classification results, especially the high dialect predictiveness of first-order constructions, might be driven by place-specific content skewing our picture of grammatical variation. This is especially relevant given the CCG approach to learning constructions as form-meaning mappings along a continuum of abstractness. Because constructions can be both meaningful and item-specific, some representations inevitably capture place-based information. For instance, rather than learning only generic schematic noun phrases, CCG identifies many noun phrases that encode specific, geographically determined semantic domains: these become incorporated into larger constructions as substantive elements, not as abstract slot-constraints such as NP.

To illustrate this issue, consider construction (8), which encodes spatial relationships specifically calibrated to city-level entities. These prepositional phrases carry meaning distinctions that would be inappropriate for larger geographical units such as countries, or less defined spaces such as forests: they are genuine constructions precisely because they encode this specialised meaning, distinguishing them from more generic spatial expressions such as *to the store* or *into the book*. While CCG correctly identifies such meaning-based, item-specific constructions as part of the grammatical network, their high predictiveness for Australian English might be considered a topical artifact rather than a classic instance of dialect variation. Our robustness analysis must therefore ascertain that such place-based confounds do not excessively inflate the model's performance.

(8) [SEM:164 ⟨then_before⟩ — SEM:74 ⟨sydney_melbourne⟩]
   a. "in Sydney"
   b. "from Canberra"
   c. "at Brisbane"
   d. "to Melbourne"
   e. "into Darwin"

Previous work has dealt with this potential problem in two different ways. The first involves unmasking methods that iteratively remove the most predictive constructions (Dunn 2019). If the dialect model were heavily dependent on a small number of topic-based representations as in (8), we would observe a sharp decline in performance as such constructions are removed. This approach, applied to

web and social media data across seven languages, has demonstrated that dialect models remain robust to such highly predictive confounds. The second method involves constructing corpora where every sample represents the same distribution of topics. Ensuring the same topical distribution across locations would eliminate topic variation as a potential source of model accuracy (e.g., that speakers in Australia refer more frequently to Australian cities). This alternative has similarly shown that dialect models remain robust even with topic-balanced samples (Dunn 2023b, 2025a).

Another line of research on dialect versus register variation has identified the same profile reported in this paper: minimal overlap between the two dimensions, with dialect effects distributed more evenly across the grammar (Dunn 2025b). Crucially, this work relies exclusively on written corpora, thereby eliminating ASR transcriptions as potential explanations for the observed patterns. Building on these convergent findings, we adopt the unmasking method to ensure that the prediction accuracy reported above does not depend on a small number of place-based constructions.

The most predictive constructions are, in a sense, outliers because they are used predominantly in one or another country. For example, of the 9,504 first-order constructions in the grammar, only 170 have a mean feature weight above 0.009. Our unmasking analysis removes constructions above this threshold, including cases such as (8), and then repeats the classification experiments described in the previous subsections. If model performance depends heavily on place-based content, accuracy should plummet when these most predictive constructions are removed. Conversely, if regional differences represent highly systematic patterns, overall performance should remain largely unaffected by pruning these constructions. Upon replicating the original experiments with the trimmed data, we can determine the extent to which our findings depend on this relatively small subset of representations (170 out of 9,504, or less than 2%).

Table 5 shows that the register model remains completely unaffected by their removal, maintaining perfect accuracy. This demonstrates that topic-based representations do not drive our model's ability to distinguish spoken from written usage. For dialect classification, however, we observe moderate impacts in specific areas (highlighted in bold): first-order constructions drop from 0.97 to 0.90 F-score, late-stage constructions from 0.97 to 0.87, and peripheral constructions from 0.97 to 0.89. While these reductions are notable, all scores remain well above the chance baseline of 0.2, confirming that regional variation extends far beyond the subset of pruned constructions. Importantly, only some of the removed constructions contain obvious confounds such as implicit place-names; others simply exhibit strong regional associations without artifacts, hence genuine instances of dialect variation. With these features removed, we nonetheless retain excellent models of variation, which provide compelling support for the findings of this study. We can now confidently proceed to a discussion of their theoretical implications and conclude the article.

	First-Order		Third-Order		Fourth-Order	
	Min F1	Max F1	Min F1	Max F1	Min F1	Max F1
By Level of Abstraction						
Dialect	**0.90**	**0.90**	0.66	0.66	0.45	0.46
Register	1.00	1.00	1.00	1.00	1.00	1.00
	Early-Stage		Mid-Stage		Late-Stage	
	Min F1	Max F1	Min F1	Max F1	Min F1	Max F1
By Order of Emergence						
Dialect	0.80	0.80	0.74	0.74	**0.87**	**0.87**
Register	1.00	1.00	1.00	1.00	1.00	1.00
	Peripheral		Middle		Core	
	Min F1	Max F1	Min F1	Max F1	Min F1	Max F1
By Degree of Centrality						
Dialect	**0.89**	**0.89**	0.72	0.73	0.57	0.57
Register	1.00	1.00	1.00	1.00	1.00	1.00

Table 5: Classification accuracy after pruning the most predictive constructions

6 Discussion and conclusion

In this paper, we drew on the CCG paradigm to analyse register and regional variation simultaneously across a large network of constructions rooted in natural language use. Over the course of the experiments, we uncovered three important findings. First, register variation is generally stronger and more pervasive in the grammatical system than regional variation, as demonstrated by the predictive accuracy of classifiers. Second, register variation concentrates in higher-level, abstract units, while regional variation manifests primarily in lower-level, concrete units, as shown by the predictive power of constructions. Third, our country-specific classifications reveal that the two sources of variation have largely independent impacts on linguistic structure, notwithstanding slight regional effects in the exact magnitude of otherwise stable register distinctions across varieties.

We now discuss the significance of these findings by weaving two main perspectives. First, what are the social and cognitive mechanisms behind such categorically different behaviours between register and regional associations? Second, how do our results aid recent efforts towards intersecting Construction Grammar and sociolinguistic theory through questions of 'social meaning' in constructions (Morin in press)?

To begin with, our computational investigations clearly point towards the idea that register and region are full-fledged categories of both linguistic variation and linguistic representation. The fact that some constructions are predictive of the former category, while others are predictive of the latter, has substantial implications for usage-based theories of language as a complex-adaptive system (Beckner

et al. 2009). Indeed, Schmid has argued (2020: p. 309) that 'variation must not be regarded as an add-on to linguistic structure, but instead part and parcel of it'. Accordingly, our results strongly suggest that register and region should be conceptualised as genuine dimensions of constructional meaning (Goldberg 2019: p. 67), where meaning is understood as memory traces of embodied (linguistic) experience (Johnson 2017).

From this perspective, meaning is an inherently non-modular facet of linguistic knowledge (Hudson 2007; Goldberg 2013): register and region-based representations should be viewed as contiguous with traditional categories of semantics and pragmatics, differing from them only in terms of the conceptual content they denote (Leclercq 2020). In particular, register and region can be reasonably postulated as aspects of 'social context' in a usage event (Mikkelsen & Morin 2025), making them separate from the 'referential' purpose of a linguistic expression (Labov 1978). To formalise this assumption, Leclercq & Morin (2025: p. 43) put forward an explicitly 'social' component alongside semantics and pragmatics in a multidimensional model of constructional meaning (middle band of Figure 7). The authors further define the category of 'social meaning' as encompassing two notable sub-types: 'interactional' and 'sociocultural' meaning. They argue that register associations, which denote functions associated with communicative situations, belong to the first sub-type, while regional associations, which denote concepts of social identity in a community (i.e. geographical provenance), belong to the second sub-type (see also Morin in press, Chapter 3 for an extended discussion).

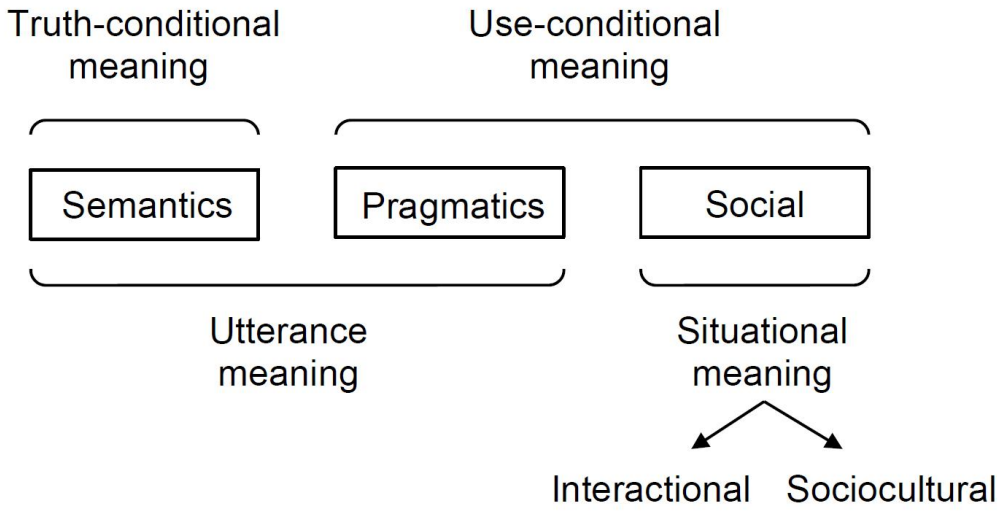

Figure 7: Model of the meaning of constructions (Leclercq & Morin 2025)

Importantly, although this account considers register and region as related aspects of social meaning, it is also true that the two notions instantiate distinct subcategories. Figure 7 iconically signals their relative distance from semantics and pragmatics: while region would be classified at the far right-hand side of the diagram, register would sit further to the left, hence closer to pragmatic meaning. Our CCG-based study offers crucial evidence for upholding such a demarcation, given the stark differences that register and regional effects exhibit across the lan-

guage network. Namely, the former are generally more pervasive, more abstract, and more stable, while the latter are sparser, more concrete, and more fluctuating. The key ensuing questions are (i) why is this the case?, and (ii) what does this entail for enriching the above model of (social) meaning in constructions?

To answer the first question, we believe it important to interpret meaning relations in constructions as measurable on a cline of sociocognitive 'salience' (Schmid & Günther 2016; Morin in press). Salience is a notoriously polysemous and controversial concept in usage-based cognitive linguistics, variationist sociolinguistics, and the broader field of linguistics alike (e.g. Rácz 2013; Jaeger & Weatherholtz 2016; Boswijk & Coler 2020). One of the most widely shared approaches in the first of these disciplines is to equate salience with mental activation (Schmid 2007), a phenomenon that correlates positively with entrenchment and frequency of use. Under this view, the results of our study compellingly suggest that register associations are highly salient across the constructicon, while regional associations are relatively less salient on the whole. This finding alone represents a valuable theoretical contribution, though it leads us to a more fundamental issue: what factors cause register distinctions to be systematically more salient than regional ones?

Our hypothesis to explain this pattern is inspired by a passage in Biber & Conrad (2019: p. 12), who use their own terminology to describe the difference between register and dialect variation:

> Although most sociolinguistic studies have traditionally focused on linguistic differences among dialects, the linguistic differences among registers are actually more extensive. When speakers switch between registers, they are doing different things with language—using language for different communicative purposes and producing language under different circumstances. The associated linguistic differences are functionally motivated, related to these differing purposes and situations, and thus the linguistic differences among registers are often dramatic. In contrast, dialect differences are largely conventional, expressing a person's identity within a social group. Regardless of any dialect differences, speakers using the same register are doing similar communicative tasks; therefore, in most basic respects the characteristic language features used in a given situation are similar across speakers from different dialects.

In light of this quote, it is especially striking to see how well the results of our unsupervised empirical method confirm the predictions. Our very large, geolocated corpora of YouTube transcripts and Tweets indeed show that register variation is more sharply stratified and more stable than regional variation, and that the two phenomena are largely independent. Furthermore, Biber & Conrad foreground an essential trait of register that may account for its high degrees of 'salience'. While register is taken to be 'functionally motivated' and to denote 'communicative purposes', dialect variation expresses more peripheral differences in social identity, which are less 'dramatic' and 'extensive'. This is exactly what we observe across

the levels of structure in our learned constructicon (i.e. abstraction, emergence, and centrality).

Combined with these remarks, our results lay important groundwork for new theoretical proposals in usage-based Construction Grammar. On the one hand, we believe that register distinctions are more salient in the constructional network because their specific functions—the achievement of communicative purposes in social situations—require them to be explicitly learned by users. On the other hand, we submit that dialect distinctions are less salient overall because they originate as exposure effects on linguistic structure, for instance as small perturbations in register variation itself (Section 5.3). In other words, we hypothesise that regional meaning in constructions, at least at the incipient stage, is passively acquired at the surface level, rather than actively learned at high levels of abstraction; hence, it does not necessitate the same degree of salience as register from the get-go.

Of course, it is not our intention to suggest that regional differences are unable to acquire additional salience, in low or in high increments, over time and the repetition of socially significant linguistic interactions. Variationist sociolinguists have long established that dialect variants readily undergo shifts in indexical orders (Johnstone et al. 2006; Eckert 2008), becoming increasingly salient, meta-conscious, and imbued with social meaning along a continuum from 'indicators' to 'markers' and 'stereotypes' (Labov 1972; Sherwood et al. 2023). However, what our results do tell us is that while register distinctions are categorically salient across the grammar, regional distinctions exhibit this property more gradiently, operating most notably at the concrete, item-specific level of the linguistic system.

If this hypothesis is confirmed in future research, it would represent a novel finding with significant implications beyond the Construction Grammar community. Not only would it corroborate standard predictions formulated in register studies (Biber & Conrad 2019), but it would also inform the current debate in sociolinguistics as to whether social meaning is mainly found in surface-level generalisations, such as sound structures, rather than in more abstract generalisations, such as syntactic structures (e.g. Eckert & Labov 2017; see MacKenzie & Robinson's (In press) critical discussion of the 'Grammatical Invisibility Principle'). Overall, this study illustrates the field-wide potential of CCG as a tool for analysing language variation bottom-up and holistically, with no theoretical priors other than the maxim of 'usage affecting grammar' (Bybee 2010; Leclercq et al. 2025).

Turning to the second question, our results also bear important consequences for current Construction Grammar theorising, particularly constructionist models of sociolinguistic phenomena. Most prominently, they validate the CCG paradigm for observing variation across the entire constructicon, without the need to predetermine a single sociolinguistic variable. The method is therefore wholly consistent with the framework's core assumptions of language as (i) a complex-adaptive system, and (ii) a large network of interrelated form-meaning pairs. The flexibility of CCG, which can be applied to any collection of corpora, also means that it could be used for many exciting research questions in the future. For example,

does the register stability that we find across varieties persist when we zoom in on patterns in intra-regional areas, such as cities? What effects might we find in corpora stratified by other social categories such as age, gender, or even ethnicity? Could this approach shed light on the interaction between sociolinguistic variation and diachronic change spanning the constructicon? These represent only some of the avenues revealed by the technical abilities of usage-based machine learning.

In addition, the generalisations yielded by CCG can help refine models of constructional meaning such as the one shown in Figure 7. For instance, Leclercq & Morin (2025) point out that much of the constructionist literature, though implicitly endorsing the gradient assumption of linguistic knowledge (Ungerer 2023), often does not systematically engage with its implications for representing levels of constructional meaning. Furthermore, although preliminary efforts have been made to conceptualise the 'social meaning' of constructions (Morin in press), it remains to be determined how rich, intersectional, and enduring this category is across levels of abstraction in constructions.

In these respects, our study yields several important insights. For one thing, the results confirm that 'interactional' and 'sociocultural' should remain separate subcategories of social meaning, given the difference in behaviour empirically shown by two of their most typical instantiations (i.e. register and region). Specifically, we can more confidently postulate that register-based meanings should be positioned closer to semantic and pragmatic meaning, as in Figure 7, which would reflect its stronger connection to communicative functions than regional meanings. In fact, we might even suggest that our results support the 'expendability cline' hypothesis by Leclercq & Morin (2025). Namely, in default communicative exchanges where the core function is semantic reference, register provides constitutive information that region generally lacks, making sociocultural meaning more 'expendable' than interactional meaning in optionality contexts (see Morin in press, Chapter 5, however, for a reappraisal).[5]

Furthermore, this study provides compelling evidence for a systematic distinction in how social meaning is distributed across grammatical structure, filling another important gap in the literature (Leclercq & Morin 2025). Our results suggest that interactional social meaning, in the shape of 'register', permeates all levels of the constructicon, from concrete first-order constructions to abstract fourth-order schemas. By contrast, sociocultural meaning, in the shape of 'region', appears to be more constrained, becoming progressively bleached as one moves beyond first-order constructions toward higher levels of abstraction. If confirmed in the future, this would constitute a significant generalisation about the architecture of constructional meaning in linguistic knowledge. However, such an account would need to engage with another important instantiation of sociocultural meaning known as 'interlingual social meaning' (Morin in press), i.e.

[5] A reviewer aptly notes that this generalisation may not hold in contact situations, where sociocultural meaning can become highly salient, and that 'expendability' must be reconciled with the fact that we communicate through concrete, region-marked utterances rather than abstract schemas. These important considerations merit systematic investigation in future research.

the association of constructions with entire languages (Höder 2012). Language distinctions may pattern differently from 'intralingual' regional varieties, potentially showing even greater stratification than register across the constructicon. Notably, CCG has already demonstrated its effectiveness at analysing interlingual differences through grammar induction in a wide typological range of languages (Dunn 2022). Such investigations would further clarify the complex relationships between all types of social meaning in constructionist theory, perhaps requiring revisions of the model in Figure 7 for more adequate descriptions.

Through this computational experimentation, we showcased the productive feedback loop that emerges when Construction Grammar and sociolinguistics are brought together via CCG. We hope to have shown that our approach generates testable predictions about constructional variation in the language network, lays the ground for new hypotheses and research topics, and generally strengthens the explanatory power of Construction Grammar as a comprehensive theory of the linguistic system (Cappelle 2024).

Data availability

Supplementary material for this paper is available at https://doi.org/10.17605/OSF.IO/24J5E.

References

Beckner, Clay, Richard A. Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language Learning* 59(Suppl. 1). 1–26.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge: Cambridge University Press 2nd edn.

Bohmann, Axel. 2019. *Variation in English worldwide: Registers and global varieties*. Cambridge: Cambridge University Press.

Boswijk, Vincent & Matt Coler. 2020. What is salience? *Open Linguistics* 6. 713–722.

Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan L. 2013. Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 49–69. Oxford: Oxford University Press.

Cappelle, Bert. 2024. *Can Construction Grammar be proven wrong?* Elements in Construction Grammar. Cambridge: Cambridge University Press.

Coats, Steven. 2022a. The Corpus of Australian and New Zealand Spoken English: A new resource of naturalistic speech transcripts. In Prabha Parameswaran, Jesse Biggs & David Powers (eds.), *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, 1–5.

Coats, Steven. 2022b. The Corpus of British Isles Spoken English (CoBISE): A new resource of contemporary British and Irish speech. In Karl Berglund, Matti La

Mela & Inge Zwart (eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries*, 187–194. CEUR.
Coats, Steven. 2023. Dialect corpora from YouTube. In Beatrix Busse, Natalija Dumrukcic & Ingo Kleiber (eds.), *Language and linguistics in a complex world*, 79–102. Berlin: De Gruyter.
Colleman, Timothy & Dirk Noël. 2025. Constructions and lectal variation. In Mirjam Fried & Katerina Nikiforidou (eds.), *The Cambridge handbook of construction grammar*, 497–518. Cambridge: Cambridge University Press.
Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.
Diessel, Holger. 2023. *The constructicon: Taxonomies and networks*. Cambridge: Cambridge University Press.
Dunn, Jonathan. 2017. Computational learning of construction grammars. *Language and Cognition* 9(2). 254–292.
Dunn, Jonathan. 2018a. Finding variants for construction-based dialectometry: A corpus-based approach to regional CxGs. *Cognitive Linguistics* 29(2). 275–311.
Dunn, Jonathan. 2018b. Multi-unit directional measures of association moving beyond pairs of words. *International Journal of Corpus Linguistics* 23(2). 183–215.
Dunn, Jonathan. 2019. Global syntactic variation in seven languages: Toward a computational dialectology. *Frontiers in Artificial Intelligence* 2. Article 15.
Dunn, Jonathan. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation* 54. 999–1018.
Dunn, Jonathan. 2022. Exposure and emergence in usage-based grammar: Computational experiments in 35 languages. *Cognitive Linguistics* 33(4). 659–699.
Dunn, Jonathan. 2023a. Syntactic variation across the grammar: Modelling a complex adaptive system. *Frontiers in Complex Systems* 1. Article 1273741.
Dunn, Jonathan. 2023b. Variation and instability in dialect-based embedding spaces. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects*, 67–77. Association for Computational Linguistics.
Dunn, Jonathan. 2024. *Computational construction grammar: A usage-based approach*. Cambridge: Cambridge University Press.
Dunn, Jonathan. 2025a. Diffusion across the grammar: Complexity in areal interactions between dialects of English. In Andrés Enrique-Arias, Carlota de Benito Moreno & Florencio del Barrio de la Rosa (eds.), *The spatial diffusion of linguistic changes: New methods and theoretical perspectives* (Studies in Language Change 26), Berlin: De Gruyter.
Dunn, Jonathan. 2025b. *Syntactic variation from individuals to populations: Language as a complex system*. Cambridge: Cambridge University Press.
Dunn, Jonathan. 2026. *Syntactic variation from individuals to populations: Language as a complex system* Elements in Construction Grammar. Cambridge University Press. doi:10.1017/9781009420280.
Dunn, Jonathan & Harish Tayyar Madabushi. 2021. Learned construction grammars converge across registers given increased exposure. In Anna Bisazza & Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, 268–278. Association for Computational Linguistics.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.
Eckert, Penelope & William Labov. 2017. Phonetics, phonology and social meaning. *Journal of Sociolinguistics* 21(4). 467–496.
Ellis, Nick C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.
Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Goldberg, Adele E. 2013. Constructionist approaches. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 15–31. Oxford: Oxford University Press.
Goldberg, Adele E. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton, NJ: Princeton University Press.
Grünwald, Peter D. 2007. *The minimum description length principle*. Cambridge, MA: MIT Press.
Hollmann, Willem B. & Anna Siewierska. 2007. A construction grammar account of possessive constructions in Lancashire dialect: Some advantages and challenges. *English Language and Linguistics* 11(2). 407–424.
Hollmann, Willem B. & 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.
Hudson, Richard. 2007. English dialect syntax in Word Grammar. *English Language and Linguistics* 11(2). 383–405.
Höder, Steffen. 2012. Multilingual constructions. In Kurt Braunmüller & Christoph Gabriel (eds.), *Multilingual individuals and multilingual societies*, 241–258. Amsterdam: John Benjamins.
Jaeger, T. Florian & Kodi Weatherholtz. 2016. What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology* 7. Article 1115.
Johnson, Mark. 2017. *Embodied mind, meaning, and reason: How our bodies give rise to understanding*. Chicago, IL: University of Chicago Press.
Johnstone, Barbara, Jennifer Andrus & Andrew E. Danielson. 2006. Mobility, indexicality, and the enregisterment of 'Pittsburghese'. *Journal of English Linguistics* 34(2). 77–104.
Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: English language in the outer circle. In Randolph Quirk & Henry Widdowson (eds.), *English in the world: Teaching and learning the language and literatures*, 11–36. Cambridge: Cambridge University Press.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4). 715–762.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William. 1978. Where does the linguistic variable stop? A response to Beatriz Lavandera. *Working Papers in Sociolinguistics* 44. 1–17.
Labov, William. 2014. What is to be learned: The community as the focus of social cognition. In Martin Pütz, Justyna Robinson & Monika Reif (eds.), *Cognitive*

sociolinguistics: Social and cultural variation and language use, 23–52. Amsterdam: John Benjamins.
Leclercq, Benoît. 2020. Semantics and pragmatics in construction grammar. *Belgian Journal of Linguistics* 34. 225–234.
Leclercq, Benoît & Cameron Morin. 2025. *The meaning of constructions*. Cambridge: Cambridge University Press.
Leclercq, Benoît, Cameron Morin & Dirk Pijpops. 2025. The principle of no equivalence: An agent-based model. *Cognitive Linguistics* 36(4). 633–672.
Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.
Li, Haipeng, Jonathan Dunn & Andrea Nini. 2023. Register variation remains stable across 60 languages. *Corpus Linguistics and Linguistic Theory* 19(3). 397–426.
MacKenzie, Laurel & Mary Robinson. In press. Spelling out grammatical variation. In Daniel Duncan & Mary Robinson (eds.), *Sociosyntax: Current approaches* Topics in English Linguistics, Berlin: De Gruyter Mouton.
Mikkelsen, Olaf & Cameron Morin. 2025. Register as a source of non-equivalent contracted constructions: going to and gonna in British English. *English Language and Linguistics* 29(3). 609–630.
Morin, Cameron. in press. *Construction grammar and sociolinguistic theory: A case study of social meaning*. Oxford: Oxford University Press.
Morin, Cameron, Guillaume Desagulier & Jack Grieve. 2024. A social turn for construction grammar: Double modals on British Twitter. *English Language and Linguistics* 28(2). 275–303.
Morin, Cameron & Jack Grieve. 2024. The semantics, sociolinguistics, and origins of double modals in American English: New insights from social media. *PLOS ONE* 19(1). Article e0295799.
Rácz, Péter. 2013. *Salience in sociolinguistics: A quantitative approach*. Berlin: Mouton de Gruyter.
Schmid, Hans-Jörg. 2007. Entrenchment, salience, and basic levels. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*, 117–138. Oxford: Oxford University Press.
Schmid, Hans-Jörg. 2010. Does frequency in text instantiate entrenchment in the cognitive system? In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 101–134. Berlin: De Gruyter Mouton.
Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford: Oxford University Press.
Schmid, Hans-Jörg & Fritz Günther. 2016. Toward a socio-cognitive framework for salience in language. *Frontiers in Psychology* 7. 1110.
Sherwood, Simon, Jason Shaw, Shigeto Kawahara, Robert Mailhammer & Mark Antoniou. 2023. Variation, gender, and perception: The social meaning of Japanese linguistic variables. *Linguistics* 61(4). 959–995.
Stefanowitsch, Anatol & Susanne Flach. 2017. The corpus-based perspective on entrenchment. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology*

of language learning: How we reorganize and adapt linguistic knowledge, 101–127. Washington, DC and Berlin: American Psychological Association and De Gruyter.

Szmrecsanyi, Benedikt. 2019. Register in variationist linguistics. *Register Studies* 1(1). 76–99.

Ungerer, Tobias. 2023. A gradient notion of constructionhood. *Constructions* 15(1).

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2). 143–160.

Weinreich, Uriel, William Labov & Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics: A symposium in Austin, TX*, 95–188. Austin, TX: University of Texas Press.

Östman, Jan-Ola & Graeme Trousdale. 2013. Dialects, discourse, and construction grammar. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 476–490. Oxford: Oxford University Press.