Exploring the Constructicon: Linguistic Analysis of a Computational CxG

Jonathan Dunn
Department of Linguistics and
New Zealand Institute for Language, Brain and Behaviour
University of Canterbury
Christchurch, New Zealand
jonathan.dunn@canterbury.ac.nz

Abstract

Recent work has formulated the task for computational construction grammar as producing a constructicon given a corpus of usage. Previous work has evaluated these unsupervised grammars using both internal metrics (for example, Minimum Description Length) and external metrics (for example, performance on a dialectology task). This paper instead takes a linguistic approach to evaluation, first learning a constructicon and then analyzing its contents from a linguistic perspective. This analysis shows that a learned constructicon can be divided into nine major types of constructions, of which *Verbal* and *Nominal* are the most common. The paper also shows that both the token and type frequency of constructions can be used to model variation across registers and dialects.

1 Introduction

Construction Grammar (CxG) is a usage-based approach to language which views grammatical structure as a set of form-meaning mappings called a *constructicon* (Langacker, 2008). From this usage-based perspective, a *construction* could belong in the grammar either (i) because it is sufficiently entrenched (i.e., frequent) that it is stored and processed as a unique item or (ii) because it is sufficiently irregular (i.e., idiomatic) that it requires a unique grammatical description (Goldberg, 2006). The advantage of CxG from this perspective is that it focuses on explaining the creativity, the flexibility, and the idiosyncrasy of actual language use in real-world settings (Goldberg, 2019).

Given this focus of CxG as a linguistic theory, the ideal computational implementation must be data-driven and unsupervised. For example, approaches which rely on manual annotations derived from individual introspection (Steels, 2017) fail to capture the usage-based foundations of CxG, in addition to being unreproducible and difficult to scale. For this reason, most recent work on computational CxG has taken an unsupervised learning approach to forming constructicons (Dunn, 2017, 2022). Such an unsupervised approach has its own challenges, however, especially the challenge of evaluation. Grammars from other syntactic paradigms can be evaluated by annotating a gold-standard corpus and then measuring the ability of both supervised and unsupervised models to predict those same sets of annotations (cf., Zeman et al. 2017, 2018). Given its usage-based foundations, this approach to evaluation is simply not feasible for computational CxG because the standard for what counts as a construction depends to some degree on the corpus or the community of speaker-hearers that is being observed.

For this reason, recent work on computational CxG has undertaken both internal and external evaluations for determining which one of a set of posited constructicons is better. An internal metric measures the fit between a grammar and a given corpus to determine which alternative constructicon offers a better description (Dunn, 2018b, 2019a). This work has drawn on Minimum Description Length (Goldsmith, 2001, 2006) as an evaluation metric because it combines both descriptive adequacy (i.e., the fit between the grammar and the test set) and model complexity (i.e., the number and the type of constructions in the grammar).

An external metric evaluates and compares constructicons using their performance when applied to a specific prediction task. Recent work has focused on the use of computational CxG for modelling individual differences (Dunn and Nini, 2021), register variation (Dunn and Tayyar Madabushi, 2021), and population-based dialectal differences (Dunn, 2018a, 2019c,b; Dunn and Wong, 2022). Because CxG is a usage-based paradigm, the definition of a construction that is referenced above depends on both entrenchment and idiomaticity. Both of these are properties of a corpus of usage rather than properties of a language as a whole.

In other words, it is only meaningful to describe *entrenchment* relative to a particular individual, dialectal community, or context of production. These external tasks have therefore focused on the degree to which computational CxG can in fact account for differences in usage across these dimensions.

The contribution of this paper is to undertake a detailed qualitative and quantitative evaluation of a learned grammar. While it is not possible to start with gold-standard linguistic annotations of constructions, it is possible to apply a linguistic analysis to the output of an unsupervised, usage-based framework. We start by describing the model and the data which are used to learn the constructicon (Section 2) before presenting examples of types of constructions that it contains (Section 3). We then proceed to a quantitative analysis of the grammar (Section 4). Finally, we end with a discussion of the challenge of parsing a nested and hierarchical grammar which contains representations at different levels of abstraction (Section 5).

2 Methods and Data

Computational CxG is a theory in the form of a grammar induction algorithm that provides a reproducible constructicon given a corpus of exposure (Dunn, 2017, 2022). The theory is divided into three components, each of which models a particular aspect of the emergence of constructicons given exposure to a corpus of usage.

First, a psychologically-plausible measure of association, the ΔP, is used to measure the entrenchment of potential constructions (Ellis, 2007; Dunn, 2018c). These potential constructions are sequences of lexical, syntactic, and semantic slot-constraints. The problem of *category formation* is to define the inventory of fillers that are used for slot-constraints. In this implementation, lexical constraints are based on word-forms, without lemmatization. Syntactic constraints are formulated using the universal part-of-speech tagset (Petrov et al., 2012) and implemented using the Ripple Down Rules algorithm (Nguyen et al., 2016). Semantic constraints are based on distributional semantics, with k-means clustering used to discretize fastText embeddings (Grave et al., 2018). The semantic constraints in the examples in this paper are formulated using the index of the corresponding clusters, a simple notational convention.

Second, an association-based beam search is used to identify constructions of arbitrary length by finding the most entrenched representations in reference to a matrix of ΔP values (Dunn, 2019a). The beam search parsing strategy allows the grammar to avoid relying on heuristic frames and templates for producing potential constructions.

Third, a measure of fit based on the Minimum Description Length paradigm is used to balance the increased storage of item-specific constructions against the increased computation of more generalized constructions (Dunn, 2018b). The point is that any construction could become entrenched but more idiomatic constructions come at a higher cost.

The contribution of this paper is to evaluate this existing model of CxG (Dunn, 2022) rather than to alter its overall method of learning a constructicon. We therefore apply the model without further discussion of its implementation and focus instead on a linguistic analysis of the resulting constructicon. The data used to learn grammars is collected from three sets of corpora: social media (Twitter), non-fiction articles (Wikipedia), and web pages (from the Common Crawl) drawn from the *Corpus of Global Language Use* (Dunn, 2020). This training corpus contains 2 million words per register for a total of 6 million words.

From a usage-based perspective, exposure to language continues after the grammar has been acquired and such exposure might change the entrenchment of particular constructions. The model thus undertakes a second pruning stage which updates the constructicon given an additional 2 million words of exposure (Dunn, 2022). The model observes sub-corpora from each of the three registers in increments of 100k words. Each construction in the grammar receives an activation weight with an initial value of 1. For each sub-corpus in which a construction is not observed, its weight decays by 0.25. For each sub-corpus in which a construction is observed, its weight is returned to 1. When a construction's weight falls below 0, it is forgotten and removed from the grammar.

This is a simple model of the way in which continued exposure leads to the forgetting of previously entrenched constructions. While somewhat arbitrary, the decay rate (0.25) is chosen to ensure that a construction is not forgotten simply because it occurs primarily in a specific register: this decay rate means that a construction must be absent from four successive sub-corpora, thus ensuring that each of the three registers has been observed. Thus, this pruning method removes unproductive

constructions given additional exposure while ensuring that all three registers remain represented. A package for reproducing this grammar induction algorithm is available[1] as well as the specific grammars used in this study.[2]

This method produces a constructicon that contains 12,856 constructions. The analysis in this paper is based on using this constructicon to annotate samples of 1 million words from 12 independent corpora: Project Gutenberg (Rae et al., 2019), Wikipedia (Ortman, 2018), European Parliament proceedings (Tiedemann, 2012), news article comments (Kesarwani, 2018), product reviews (Zhang et al., 2015), blogs (Schler et al., 2006), and tweets from six countries (with 1 million words representing each country; Dunn 2020). This range of corpora allows us to consider both register (different contexts of production) and dialect (different populations using the same register) when measuring the frequency and the productivity of individual constructions in the grammar.

3 Categorizing Constructions

In this section we categorize the learned constructions to aid our quantitative analysis of the contents of the constructicon. We annotate a random sample of 20% of the constructions using the categorization described below, thus allowing an estimate of the overall composition of the grammar. The primary categories are *Verbal*, *Nominal*, *Adjectival*, *Adpositional*, *Transitional*, *Clausal*, *Adverbial*, *Sentential*, and *Fixed Idioms*. These categories are defined and exemplified in this section.

The first category consists of VERBAL constructions. As shown in (1), we notate the construction using its slot-constraints, with each slot separated by dashes. Lexical constraints are shown in italics; syntactic constraints are shown in small caps; and semantic constraints are shown using the index of their distributional cluster (e.g., <521>). Using this notation, the construction in (1) is a simple passive verb phrase in a continuous aspect, defined using primarily syntactic constraints.

(1) [AUX – *being* – VERB]
  (1a) were being proposed
  (1b) was being spread
  (1c) is being invaded
  (1d) am being kept

The verbal construction in (2) now contains a semantic constraint (<521>). This domain contains lexical items like *house* and *carriage*, all locations that can be moved into or out of. The construction thus captures a meaning-based pattern of movement in relation to some area.

(2) [VERB – ADP – DET – <521>]
  (2a) come to this house
  (2b) leaped into a carriage
  (2c) seated at that window
  (2d) hurried across the room
  (2e) lying on the floor

A lexical constraint for the main verb is shown in the construction in (3). This leads to an idiomatic usage of *play*, a set of utterances whose behaviour differs from the basic transitive verb phrase. The construction in (4) shows the influence of a lexical constraint in a different position, here *time* as a noun introducing the verb phrase. This again results in idiomatic utterances with behaviour more specific than a construction with only syntactic constraints. Finally, the lexical constraint in (5) defines a particle verb, again with idiomatic semantics resulting for the utterances in (5a) through (5e). This series of examples shows how a lexical constraint in different locations within a verb phrase leads to different types of idiomatic verbal constructions.

(3) [*play* – DET – NOUN]
  (3a) play the game
  (3b) play the part
  (3c) play the coquette
  (3d) play the king

(4) [*time* – *to* – VERB]
  (4a) time to plead
  (4b) time to write
  (4c) time to tell
  (4d) time to consider
  (4e) time to worry

(5) [*to* – VERB – *down*]
  (5a) to sit down
  (5b) to put down
  (5c) to settle down
  (5d) to bring down
  (5e) to strike down

While these examples are relatively simple verbal constructions, a more complex example is shown in (6). This construction contains a main

[1] https://www.github.com/jonathandunn/c2xg
[2] https://doi.org/10.18710/CES0L8

verb with an infinitive complement followed by an argument that takes the form of a noun phrase. The entrenchment of these more complex constructions shows the flexibility of computational CxG as well as the infeasibility of relying on the introspection of individual linguists.

(6) [VERB – *to* – *be* – <830> – ADP – DET – NOUN]
(6a) seem to be unaware of the fact
(6b) came to be known as the *Newcastle*
(6c) have to be supplied from that source
(6d) is to be found in the world
(6e) expect to be ushered into the temple

Moving to NOMINAL constructions, the first examples show the influence that a semantic constraint in one slot exerts across the entire construction. We focus here on complex nominal constructions, with both of these first examples containing a subordinate adpositional phrase within the noun phrase. In each case, the noun in the adpositional phrase is constrained to a specific semantic domain. In (7), this leads to lexical items like *empire* and *palace* and, in (8), like *ground* and *road*. Not all examples of a construction are perfect matches; an example of this is shown in (8e), marked with an asterisk, in which the first word is actually a mistagged verb rather than a noun.

(7) [NOUN – *of* – DET – <587>]
(7a) part of the empire
(7b) inmates of the palace
(7c) guardianship of the wanderer
(7d) pursuit of a chimera
(7e) circuit of the citadel

(8) [NOUN – ADP – *the* – <484>]
(8a) feet on the ground
(8b) side of the road
(8c) law of the land
(8d) entrance of the path
(8e) journey through the forest
(8e) *wanders around the forest

(9) [*one* – ADP – *the* – *best* – NOUN]
(9a) one of the best paintings
(9b) one of the best apologies
(9c) one of the best examples
(9d) one of the best books

More idiomatic noun phrases, with lexical constraints, are shown in (9) and (10). In the first, an adpositional phrase *one of the best* functions as a single adjective. In the second, a superlative adjective frames the core noun phrase. In both cases, these constructions provide additional flexibility to describe unique nominal phrases, made into constructions by their entrenchment and their idiosyncrasy in this set of usage.

(10) [*the* – *most* – ADJ – NOUN]
(10a) the most amusing instance
(10b) the most violent writhings
(10c) the most astounding instances
(10d) the most important generalizations
(10e) the most unfavourable circumstances

A single example of an ADJECTIVAL construction is shown in (11). While the previous nominal constructions included adjectival material within them, this construction as a whole provides a modifier for a noun phrase. For example, (11e) as an abstract adjective could be combined with a variety of nouns like *immigrants*, *the elderly*, or *house sparrows* to form a larger nominal construction.

(11) [*huge* – NOUN – *of*]
(11a) huge pair of
(11b) huge influx of
(11c) huge clumps of
(11d) huge piece of
(11e) huge population of

The next category is ADPOSITIONAL constructions, as shown in (12) through (14). As before, a semantic constraint leads to a meaning-based group of utterances, as with the terms specific to legal language in (12). In other words, this adpositional construction is specific to the category of nouns contained within it. A potentially problematic case is shown in (12e), here with what is likely a fixed idiom, where *case* is not used in the legal sense. A lexical constraint for the head noun in (13) leads to idiosyncratic adpositional phrases with *beginning*. Other adpositional constructions are more syntactically complex. For example, the phrase in (14) transitions from a noun into a relative clause which describes that noun.

(12) [ADP – DET – <959>]
(12a) in the case
(12b) of the provisions
(12c) as a rule
(12d) from the petitioners
(12e) ?in which case

(13) [ADP – *the* – *beginning*]
(13a) towards the beginning
(13b) at the beginning
(13c) from the beginning
(13d) in the beginning
(13e) for the beginning

(14) [ADP – *the* – NOUN – *where*]
(14a) in the world where
(14b) at the spot where
(14c) from the point where
(14d) near the ceiling where

The example of an adpositional phrase that transitions into a relative clause in (14) introduces another category of constructions, those which capture TRANSITIONAL material connecting other types of constructions. In particular, the constructions in this category capture different types of transitions without containing the substance of the involved structures themselves. For example, in (15) there is the introduction of a new main clause with a first-person verb phrase. In (16) there is the introduction of a subordinate clause. In (17) there is a comparison between two nominal constructions. The final example in (17e) represents a problematic parse: the phrase is likely *at least* rather than *least* alone. These examples show how this category serves to link other constructions together.

(15) [*but* – *i* – VERB]
(15a) but i think
(15b) but i knew
(15c) but i regret
(15d) but i noticed

(16) [SCONJ – VERB – *to*]
(16a) without seeming to
(16b) because according to
(16c) as opposed to
(16d) while listening to
(16e) in resorting to

(17) [ADV – <917> – *than*]
(17a) far deeper than
(17b) considerably better than
(17c) now more than
(17d) much smaller than
(17e) *least better than

While transitional constructions focus mainly on the connecting element, CLAUSAL constructions are those which contain a significant portion of a subordinate clause. For instance, (18) is an example of a relative clause embedded within a larger noun phrase and (19) of a relative clause in which the subject is defined by the proceeding element. A problematic example is shown in (19e), where the phrase *a lot* is treated as two separate slots. The complex subordinate clause in (20) consists of a gerund within an adpositional phrase, where the verb is further defined by a semantic constraint. Finally, a reduced relative clause is captured by (21), again with a semantic constraint on the verb. This series of examples shows the way in which subordinate clauses are captured in the grammar.

(18) [NOUN – ADP – *those* – *who*]
(18a) hearts of those who
(18b) arguments of those who
(18c) side of those who
(18d) minds of those who
(18e) tactics of those who

(19) [*which* – VERB – *a* – NOUN]
(19a) which formed a snare
(19b) which occasioned a detour
(19c) which presented a problem
(19d) which contained a letter
(19e) ? which looked a lot

(20) [SCONJ – <113> – DET – NOUN – *of*]
(20a) by taking the life of
(20b) in sacrificing the rights of
(20c) after collecting the remains of
(20d) by applying a drop of
(20e) in neglecting the cultivation of

(21) [DET – NOUN – *he* – <830>]
(21a) the loan he solicited
(21b) the temple he discovered
(21c) the words he used
(21d) the life he led
(21e) the flask he carried

While these clausal constructions are connected into the main clause itself, the category of ADVERBIAL constructions contain clauses which are more independent of the structure of the main clause. For example, in (22) there is a gerund clause within an adpositional phrase, now with a semantic constraint. In (23) there is an adposition introducing a finite verb. And in (24), with a lexical constraint,

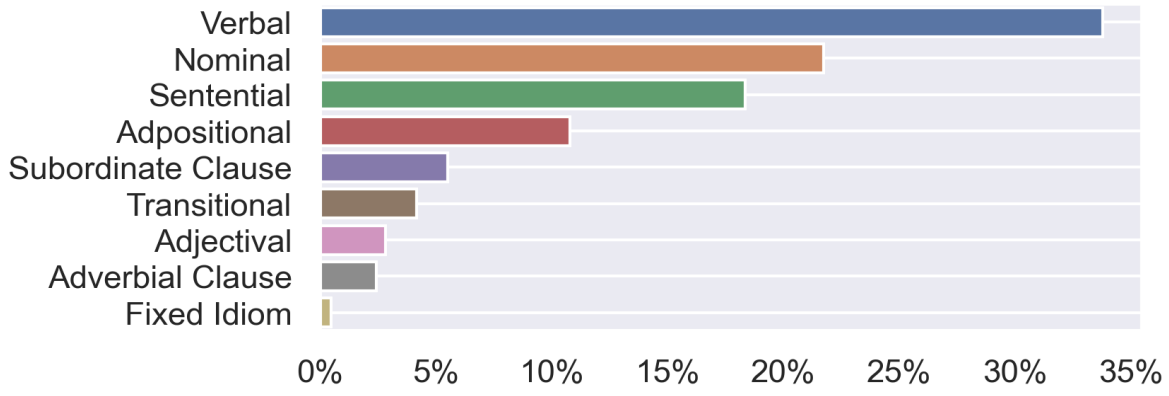

Figure 1: Distribution of Construction Types in the Grammar

there is a similar construction again with a finite verb. While similar to the clausal category, this class of constructions is less integrated with the main clause structure.

(22) [SCONJ – VERB – ADP – DET – <512>]
(22a) in dealing with that section
(22b) after referring to the matter
(22c) as bearing on the question
(22d) without glancing within the volume
(22e) by bringing up the subject

(23) [SCONJ – PRON – AUX – VERB – *to*]
(23a) that it would come to
(23b) if he had lived to
(23c) as they were trying to

(24) [*when* – DET – NOUN – *is*]
(24a) when the end is
(24b) when a man is
(24c) when the heart is
(24d) when the patient is
(24e) when the temperature is

(25) [PRON – *were* – VERB – ADP]
(25a) we were accosted by
(25b) they were employed by
(25c) these were succeeded by
(25d) they were drilled by
(25e) ? who were barred from

SENTENTIAL constructions contain the structure of the main clause. This category overlaps to some degree with verbal constructions; the key difference is that the sentential constructions contain the subject while verbal constructions do not. A simple passive clause is shown in (25), together with an adpositional argument. In many examples, this adpositional argument specifies the agent, but the example in (25e) differs in specifying a location. An active clause introducing an indirect speech clause is shown in (26), constrained to the subject *he*. Finally, a sequence of main verb and infinitive is shown in (27), with the final verb defined using a semantic constraint.

(26) [*he* – VERB – *that*]
(26a) he remembered that
(26b) he said that
(26c) he realised that
(26d) he discovered that
(26e) he promised that

(27) [*they* – VERB – PART – <583>]
(27a) they began to draw
(27b) they threatened to destroy
(27c) they chose to assert
(27d) they wanted to persuade
(27e) they began to look

A more complex passive construction is shown in (28), containing both a semantic constraint on the main verb as well as an adpositional argument. Finally, a main clause with an existential *there* as subject is shown in (29). As with the clausal constructions, these sentential constructions overlap with verbal constructions, thus illustrating the problem of parsing as clipping (c.f., Section 5).

(28) [NOUN – *are* – ADV – <830> – ADP]
(28a) villages are thickly scattered about
(28b) recruits are never measured for
(28c) substances are universally regarded as
(28d) lines are then drawn from

	Blogs		Comments		Parliament		Gutenberg		Reviews		Wikipedia	
	Freq	*Type*	*Freq*	*Type*	*Freq*	*Type*	*Freq*	*Type*	*Freq*	*Type*	*Freq*	*Type*
Adjectival	57	36	69	43	66	40	79	59	80	45	73	43
Adpositional	207	141	222	150	433	215	401	272	221	145	327	181
Adverbial	118	87	107	80	117	79	95	80	127	88	56	45
Idiom	32	3	33	2	54	13	12	4	27	3	13	2
Nominal	95	82	128	109	261	184	189	163	123	101	179	138
Sentential	199	115	144	103	176	107	144	110	195	111	109	77
Clausal	156	99	157	112	182	117	154	112	152	97	70	58
Transitional	102	75	96	77	103	72	107	89	108	82	49	43
Verbal	137	104	143	116	188	142	139	122	144	108	116	86

Table 1: Mean Frequency and Productivity of Constructions by Category and Register

(29) [*there* – VERB – *a* – NOUN – ADP]
(29a) there was a kind of
(29b) there is a habit of
(29c) there were a number of
(29d) there were a couple of
(29e) there came a sort of

The final category of constructions are FIXED IDIOMS, which here are mainly lexical constructions. These have a very limited number of types for each construction because the constraints are lexical: *in favor of*, *seems to be*, *all the best*, or *no matter* ADV. Taken together, the categories illustrated in this section describe the contents of the learned constructicon. A quantitative analysis of the distribution of construction types and their properties follows in the next section.

3.1 Marginal Examples of Categories

Not all constructions that are classified as belonging to a given category are equally good examples of that category. This section provides a few examples of such marginal tokens in order to provide a more transparent picture of the grammar as a whole. Starting with a construction categorized as adjectival in (30), we could also see this being categorized as a nominal construction. The reason behind this annotation decision is that the overall unit is used to describe a part of some piece of writing.

(30) [*beginning* – ADP – DET – NOUN]
(30a) beginning of this note
(30b) beginning of the article

A marginal example of a nominal construction is shown in (31). Here, this sequence of noun and adpositional phrase, when taken in context, is quite likely to be two separate arguments of a double object verb phrase: for example, "They [ran [this country] [with the help...]]. However, the construction itself only includes the two arguments on their own. At the same time, (31) would clip together nicely with a verbal construction (c.f., Section 5).

(31) [*this* – NOUN – ADP – *the* – NOUN]
(31a) this country with the help
(31b) this morning to the surprise

(32) [VERB – *by* – DET – <88>]
(32a) occcupied by a foreign
(32b) used by the american

A final marginal example is shown in (32), here within the verbal category. This example is a passive verb together with a prepositional phrase that expresses the agent. The issue here is that only part of the noun phrase specifying the agent is explicitly defined, and the slot constraint is semantic. From the perspective of clipping constructions, many noun phrases could be merged here but would not experience the same emergent relationships between slot-constraints. In other words, the impact of the semantic constraint would not transcend the construction boundary. These examples are meant to show some weaknesses of both the categorization scheme and the constructions themselves.

4 Distribution of Construction Types

The first step in quantifying the contents of the constructicon is to estimate the relative distribution across these nine categories. This is shown in Figure 1 using annotations of 20% of the grammar to estimate the overall distribution. The y-axis contains a bar chart for each category of construction

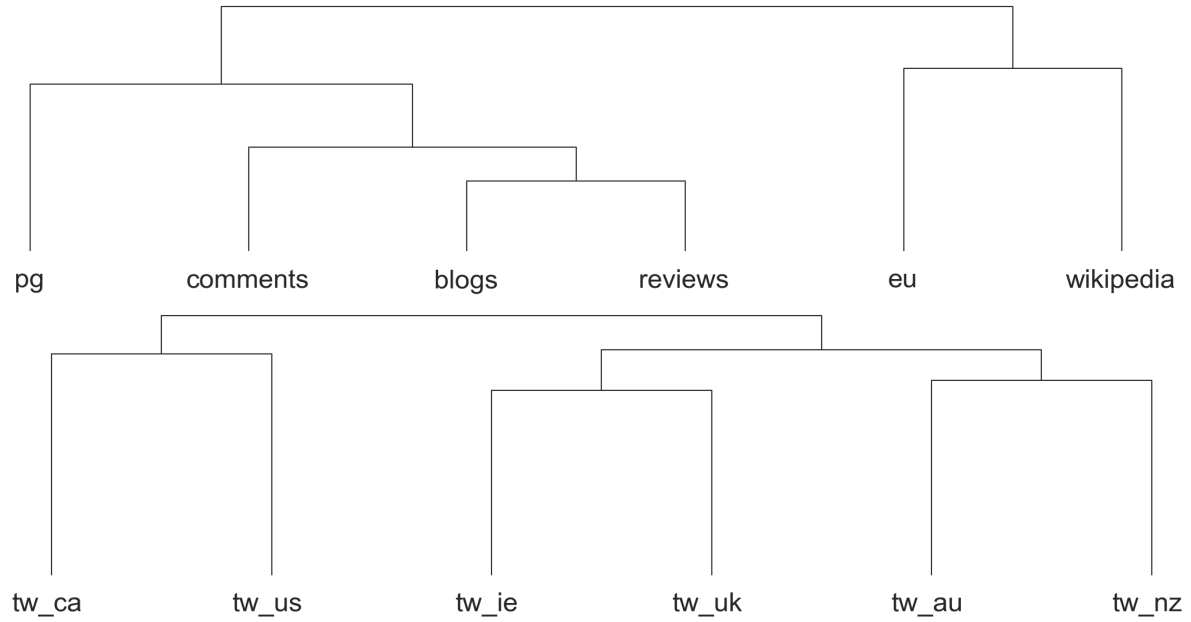

Figure 2: Clustering of Corpora Using Burrow's Delta, Register (Above) and Dialect (Below)

and the x-axis shows the percent of the constructicon which falls into that category.

Thus, for example, the most frequent type of construction is *verbal* at 33.7% of the grammar, followed by *nominal* at 21.7% and *sentential* at 18.3%. This distribution is not surprising given that verbs and nouns are the most common open-class lexical items and that sentential clauses form the basic structure of sentences.

The next step is to measure the frequency of each construction and the number of its unique types, thus capturing its productivity. These measures of frequency and productivity are corpus-specific in the sense that different constructions are more likely to be used in specific contexts or by specific populations. We thus consider 12 distinct corpora of 1 million words each, six representing distinct registers and six representing distinct populations within the same register.

Starting with a comparison across registers, Table 1 shows the mean frequency of tokens and the mean number of types for each class of constructions in each register-specific corpus. For example, the Project Gutenberg corpus has significantly more types per adpositional construction than the corpus of blogs. While some categories of construction are more common in the grammar, the measures in Table 1 take the average for each category. While there are more verbal constructions in the grammar, for example, adpositional and sentential constructions have more tokens per construction.

The frequency of each category of construction (i.e., the mean number of tokens) also provides a view of the grammatical differences between these six registers. For instance, blogs contain fewer adpositional constructions than other registers while published books and speeches in parliament contain approximately twice as many overall. Wikipedia articles contain many fewer cases of clasual and transitional constructions, indicating a register with fewer embedded clauses. Further, blogs have nearly twice as many sentential constructions (i.e., base main clauses) as Wikipedia, but many fewer adpositional phrases. This would indicate that information can be packaged in short sentences or in additional adpositional constructions, depending on the register. Note that another set of Wikipedia corpora was available during the grammar learning process, so that the reduced frequencies of these types are not simply a matter of under-fitting the register.

The next question is whether the differences in frequency of individual constructions across corpora are random or whether they reveal underlying relationships between the corpora themselves. In other words, given the frequencies of each construction in the grammar, we would expect a meaningful grammar to create meaningful relationships between conditions. A *condition* in this case refers to the register or the population represented by the corpus. This is shown in Figure 2 using Burrow's Delta to calculate the distances between corpora

and then hierarchical clustering to visualize relationships based on these distances.

The figure shows relationships between registers on the top. The two core clusters are with modern formal documents (EU and WIKIPEDIA) and digital crowd-sourced documents (COMMENTS and BLOGS and REVIEWS). The books from Project Gutenberg, from a different historical period, are an outlier. On the bottom the figure shows relationships between different dialects within the same register (tweets). The core pairs are the countries which are closest in geographic terms: Ireland and the UK together with Australia and New Zealand, with Canada and the US as a distant pair. In both cases, we see that the frequencies of constructions in the grammar provide meaningful relationships between both registers and dialects. This is important because it shows that the differing frequencies of constructions are not simply arbitrary patterns from this particular model but also reproduce two sets of real-world relationships.

5 Clipping: The Problem of Parsing

The analysis in this paper has categorized and described the kinds of constructions that are contained in a learned constructicon, has quantified the frequency and productivity of each kind, and has shown that the usage of these constructions can reconstruct meaningful relationships between corpora. The analysis of construction types in Section 3, however, reveals a major challenge in this approach to computational CxG: the unification or *clipping together* of these constructions into complete utterances during parsing (Jackendoff, 2013).

The idea in CxG is that word-forms are not the basic building blocks of grammar. Rather, the types of constructions analyzed in this paper form the basic units, themselves built out of slot-constraints that depend on basic category formation processes. With the exception of short utterances, however, no single construction provides a complete description of a linguistic form. These constructions must be clipped together: a sentential construction, for example, joined with a verbal construction and then a nominal construction. CxG posits a continuum between the lexicon and the grammar, so that the constructicon contains basic units at different levels of abstraction. We must distinguish, however, between **first-order constructions** of the type discussed in this paper and **second-order constructions** which are formed by clipping together these lower constructions. A complete constructicon would thus also contain emergent structures formed from multiple first-order constructions.

As a desideratum for future developments, we can conceptualize two types of second-order constructions: First, SLOT-RECURSION would allow a higher-order construction to contain first-order constructions as slot-fillers. For example, the set of sentential constructions could be expanded by allowing verbal constructions to fill verbal slots. Second, SLOT-CLIPPING would allow two overlapping constructions to be merged, for instance connecting a transitional construction with a verbal construction. An overlapping shared slot-constraint would license such slot-clipping unifications.

6 Conclusions

The main contribution of this paper has been to provide a qualitative linguistic analysis of a learned construction grammar, providing a new perspective on grammars which have previously been evaluated from a quantitative perspective. We presented a division of construction types into nine categories such as *Verbal* and *Nominal*, with those two open-class categories the most common. The discussion of examples shows both the range and the robustness of computational construction grammar.

This linguistic analysis does point to two current weaknesses: First, not all constructions fit nicely into the categories used for annotation (c.f., Section 3.1). A truly usage-based grammar does not necessarily align with introspection-based analysis, especially in regards to boundaries between constructions. Introspection often focuses on constructions which are complete or self-contained units, while the computational constructions place common pivot points at boundaries. Second, these constructions do not generally describe entire utterances, so that we must consider a form of clipping to provide complete parses (c.f., Section 5).

From a quantitative perspective, the analysis of register and dialectal differences shows that the productivity of these constructions also reproduces expected relationships between corpora. This is important for providing an external evaluation of the grammar: the differences between registers, for example, show how functions which are salient in a given communicative situation ultimately drive constructional frequencies. In other words, the frequencies of different types of constructions reflect meaningful patterns in real-world usage.

References

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. 2018c. Multi-Unit Association Measures: Moving beyond pairs of words. *International Journal of Corpus Linguistics*, 23(2):183–215.

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*, page 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 Sixth 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. 2022. Exposure and Emergence in Usage-Based Grammar: Computational Experiments in 35 Languages. *Cognitive Linguistics*, 33:659–699.

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 *Conference on Natural Language Learning*, pages 268–278.

J. Dunn and S. Wong. 2022. Stability of Syntactic Dialect Classification over Space and Time. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 26–36.

N. Ellis. 2007. Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, 27(1):1–24.

A. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford.

A. Goldberg. 2019. *Explain Me This*. Princeton University Press.

J. Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

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. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3483–3487.

R. Jackendoff. 2013. Constructions in the Parallel Architecture. In *The Oxford Handbook of Construction Grammar*, pages 70–92. Oxford University Press.

A. Kesarwani. 2018. New York Times Comments. Kaggle.

R. Langacker. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press, Oxford.

Dat Quoca Dai Quocb Dat Quoca Dai Quocb Nguyen, Dat Quoca Dai Quocb Dat Quoca 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.

M. Ortman. 2018. Wikipedia Sentences. Kaggle.

S. Petrov, D. Das, and R. McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth Conference on Language Resources and Evaluation*, pages 2089–2096. European Language Resources Association.

J. Rae, A. Potapenko, S. Jayakumar, and T. Lillicrap. 2019. Compressive Transformers for Long-Range Sequence Modelling.

J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

L. Steels. 2017. Basics of Fluid Construction Grammar. *Constructions and Frames*, 9(2):178–255.

J. Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218.

D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

D. Zeman, M. Popel, M. Straka, J. Hajič, and Others. 2017. CoNLL 2017 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In

Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–19.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification.