Cognitive Linguistics Meets Computational Linguistics:

Construction Grammar, Dialectology, and Linguistic Diversity

(Draft)

Jonathan Dunn

University of Canterbury, jonathan.dunn@canterbury.ac.nz

1 Data-Driven Cognitive Linguistics

Computational linguistics and cognitive linguistics come together when we use data-driven methods to conduct linguistic experiments on corpora. This chapter uses usage-based construction grammar to model geographic variation in language. The basic challenge is to show how grammatical structure emerges given exposure to usage and then how grammatical structures change given exposure to different sub-sets of usage. We first show how computational methods can be used to experiment with language learning by training a usage-based model of construction grammar. We then show how computational methods can be used to experiment with language variation by training a construction-based model of dialectology. To make these two experiments possible, we must also consider the validity of the corpora that we use for the experiments and how well they represent specific populations. Taken together, the work described here constitutes a computational theory of usage-based grammar that covers seven languages (English, French, German, Spanish, Portuguese, Russian, Arabic) and 79 distinct national dialects of these languages. Each part of the theory is an implemented computational model that can be evaluated using its predictions on held-out testing data.

How does a computational experiment work? The illustration in Figure 9.1 shows the three main components: First, language usage is represented using a corpus. Second, a computational model represents our linguistic theory. This means that all theories must be fully implemented. Third, we validate our theories using their predictions on held-out evaluation data. For example, in Section 3 we experiment with two variants of usage-based construction grammar, using the same data for training and testing across both theories. In this paradigm, whichever theory makes better predictions provides better generalizations.

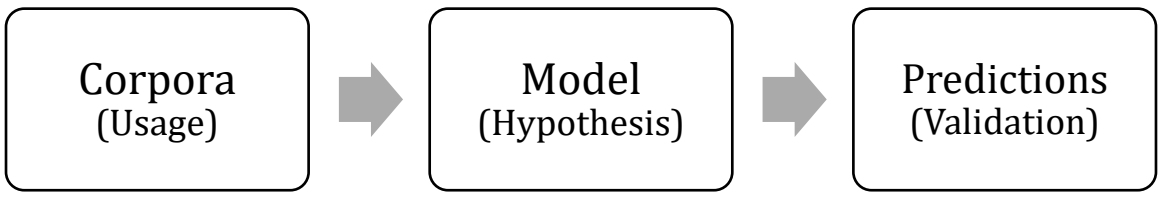


Figure 9.1. The Computational Paradigm

The challenge for a computational experiment is that each component must be fully implemented. In other words, no part of the theory can be left under-specified. In this chapter we thus consider all aspects of this experiment in usage-based grammar, from data collection to language learning to language variation. Our first question, in Section 2, is how to represent constructions in a computational manner. In other words, how do we define slot constraints? What are the relationships between slots? The important point here is that usage-based representations must be as unsupervised as possible, not relying on our own introspections to form syntactic generalizations. The basic idea is that the more we rely on introspection the less we rely on observed usage.

Our second question, in Section 3, is how to learn a grammar of constructions. In other words, construction grammar provides a rich framework for describing form-meaning mappings. But this rich framework creates an enormous number of *possible* constructions, most of which the learner is exposed to but does not actually learn. A usage-based theory of grammar must be able to balance memory (the ability to store frequent constructions) and computation (the ability to combine more abstract constructions on-the-fly). The important point here is that we can experiment with implementations of usage-based grammar that are not limited to just one narrow selection problem, like the dative vs. the ditransitive.

Our third question, in Section 4, is how to create corpora from purely digital sources like the web and social media. In other words, data-driven computational methods require massive amounts of data, so that interview-based or survey-based corpora are simply not sufficient. How do we create this kind of data? The section gives an overview of the 420 billion word *Corpus of Global Language Use* (Dunn 2020). While collections of language data of this size are what make computational methods possible, their scale requires different methods for using and validating the data.

Our fourth question, in Section 5, is how to evaluate digital corpora against local population demographics. In other words, how well do tweets, for example, represent actual language use by actual populations? We look at the relationship between data production, population size, and population demographics. Then, we use the difference-in-differences method to find out whether there is a significant presence of non-local populations in digital corpora. The important point in Sections 4 and 5 is that we are able to systematically evaluate digital corpora before we use them for linguistic experiments. Because we rely on corpora as representations of usage, it is essential to understand what populations these corpora represent.

Our fifth question, in Section 6, is how a construction grammar varies across national dialects of a language. In other words, we model geographic variation in the usage of constructions, assuming political boundaries as a constant. How is variation distributed across an entire construction grammar? What does geographic variation look like from the perspective of usage-based grammar? The important point here is that a computational approach to dialectology can make accurate predictions about which dialect a particular sample comes from, providing a measure of how well the model characterizes a particular dialect.

Thus, this paper provides an overview of one approach to computational cognitive linguistics, from language learning to language variation. The basic problem is to understand the relationship between exposure to usage and both (i) the emergence of grammatical structure and (ii) variation in grammatical structure. Each of these sections draws on a specific Python package

used to implement the details in question.[1] Because visualization is important for understanding computational models, we also provide an interactive visualization for the geographic data.[2]

2 Representing Constructions

This section reviews recent work on multi-lingual construction grammars (CxGs) that are learned directly from observed usage, as represented in a corpus (Dunn 2017, 2018a, 2019b, Dunn & Nini 2021; Dunn & Tayyar Madabushi 2021). The goal of the section is to show how we can approach the problem of using *unsupervised learning* to create a CxG. The term unsupervised learning refers to algorithms which do not start with ground-truth annotations. Such an approach is the culmination of usage-based hypotheses in linguistics, where constructions are based on the data and not filtered through introspections.

The Construction Grammar paradigm (CxG: Langacker 2008; Goldberg 2006) represents grammar using a hierarchical inventory of constraint-based *constructions*. In computational terms, a construction is a possibly non-continuous sequence in which each unit satisfies some combination of lexical, syntactic, and semantic constraints (e.g., Chang et al. 2012; Steels 2004, 2012, 2017). This section uses unsupervised methods to represent slot-constraints and their fillers.

To understand why this is important, consider implementations of CxG such as Fluid Construction Grammar (FCG) and Embodied Construction Grammar (ECG) that require the manual specification of constraints using a knowledge representation framework like FrameNet (e.g., Laviola et al. 2017; Matos et al. 2017; van Trijp 2017; Ziem & Boas 2017; Dodge et al. 2017). While these approaches can provide high-quality representations of a few constructions, they cannot model the emergence of slot-constraints: their constraints are *defined* rather than *learned*. We instead follow work that models CxG from a usage-based perspective: first, generating potential constructions given a corpus (Wible & Tsao 2010; Forsberg et al. 2014); second, selecting the optimal set of constructions, where optimality is measured against a test corpus. This provides a model of how syntactic constraints are learned. The point is that there is a significant difference between a linguistic *annotation* of a construction and a linguistic *theory* of usage-based grammar.

Following previous work, constructions are represented as a sequence of slot-constraints, as in (1a). Slots are separated by dashes and constraints are defined by both type (Syntactic, Joint Semantic-Syntactic, Lexical) and filler (for example: *noun*, a part-of-speech or *animate*, a semantic domain).

(1a) [SYN:*noun* --- SEM-SYN:*transfer*[V] --- SEM-SYN:*animate*[N] --- SYN:*noun*]

(1b) “He gave Bill coffee.”

(1c) “He gave Bill trouble.”

(1d) “Bill sent him letters.”

(2a) [SYN:*noun* --- LEX: “give” --- SEM-SYN:*animate*[N] --- LEX: “a hand”]

[1] https://github.com/jonathandunn/

[2] https://www.earthlings.io

(2b) “Bill gave me a hand.”

The construction in (1a) contains four slots: two with joint semantic-syntactic constraints and two with simple syntactic constraints. The examples in (1b) to (1d) are tokens of the construction in (1a). Lexical constraints, as in (2a), represent idiomatic sentences like (2b). These constructions are context-free because any sequence that satisfies the slot-constraints becomes a token or instance of that construction.

The difficulty of modelling slot-constraints is that constructions can overlap: multiple constructions in the grammar are allowed to represent a single phrase. For example, (2b) is actually a token of both (1a) and (2a). This makes identifying constructions more difficult because reaching the representation in (1a) does not rule out also reaching the representation in (2a). Both could be part of a single speaker's grammar.

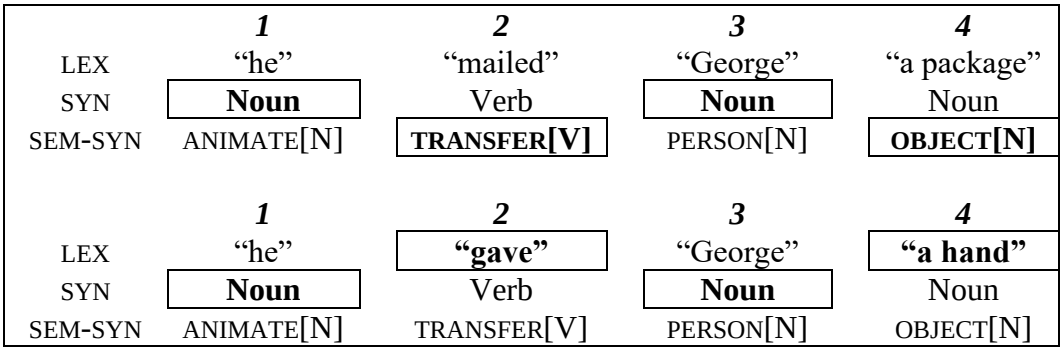


Figure 9.2. Slot-Constraints as Transitions

To illustrate the problem of construction parsing, we can view each slot as a node, with the beginning of a construction the root node (c.f., transition parsing for dependency grammars: Zhang and Nivre 2012; Goldberg et al. 2013). A construction's root can occur anywhere in a sentence. Each slot-constraint is a state, as visualized in Figure 9.2 with two forms of the ditransitive. There are four possible transitions: *LEX*, *SYN*, *SEM-SYN*, *STOP*. In the first example, the slot-constraints are generalized to any *transfer* verb and any *object* noun. In the second example, the verb and object slots require idiomatic lexical items. The problem is to find the sequence of slot-constraints that best represents the construction. From a usage-based perspective, the choice between these two representations is an empirical problem and cannot be resolved by introspection.

We first have to develop a pipeline for representing all the possible constraints shown in Figure 9.2. Such a pipeline provides our hypothesis space: any sequence of constraints that is observed in the training data is a potential construction. First, lexical constraints use word-forms separated at whitespace; no morphological analysis is included in the pipeline. The lexicon of allowed word-forms is drawn from a background corpus. An example of a lexical slot-constraint is given in (2a), where this particular construction requires the specific words “give” and “a hand”, as in (2b).

Second, syntactic representations are drawn from the part-of-speech categories in the Universal POS tag set using the RDRPOS tagger (Petrov et al. 2012; Nguyen et al. 2016); this is

a pre-defined syntactic ontology. An example of a syntactically-defined slot-constraint is given in (2a), in which any noun can fill the subject position.

Third, semantic constraints are defined using a domain dictionary in which each word-form is assigned to a cluster of word-forms. Clusters are based on word embeddings. First, a background corpus for each language is pos-tagged. No word sense disambiguation is used but word-forms are separated by syntactic category (i.e., *table_verb* is distinct from *table_noun*). A skip-gram embedding with 500 dimensions is trained for each language. Clusters are then formed by applying x-means to these embeddings (Pelleg and Moore 2000). These clusters are heterogenous syntactically. Each output cluster is further divided by syntactic category so that each semantic cluster only contains words from a single part-of-speech, allowing joint semantic-syntactic constraints.

Examples of construction representations that are learned in this unsupervised manner for shown in Table 9.1. At the top of each example is the construction itself, represented using slot-constraints. The idea is that any observed utterance which satisfies these constraints counts as a token of that construction. Below each representation, then, are tokens or examples that show which linguistic material represents that more abstract representation. There is a range of constructions here, from very abstract syntactic templates to item-specific phrases. This mixture of levels of abstraction is an important feature of usage-based construction grammar.

Table 9.1. Examples of Constructions and Their Tokens

["very" -- ADJ -- NOUN] (a) *very strong link* (b) *very powerful tool* (c) *very favorable image* (d) *very good results*	*Partially-Idiomatic Adjective Phrase* The first example is a partially productive adjective phrase which is somewhat idiomatic because of the lexical constraint requiring "very".
[DET -- ADJ -- <335>] (a) *the vertical organization* (b) *a general consensus* (c) *the European Union* (d) *the local resources*	*Semantically-Defined Noun Phrase* This example shows a noun phrase that is defined by semantic class; the number <335> refers to a group of nouns which we can see includes "organization" and "union".
["prepared" -- "to" -- VERB] (a) *prepared to accept* (b) *prepared to assist* (c) *prepared to support* (d) *prepared to act*	*Complex Verb Phrase* This shows a complex verb phrase that is lexically defined to contain "prepare" in addition to an infinitive verb phrase.
[NOUN -- "funded" -- <335>] (a) *EU funded project* (b) *state funded organization*	*Verb-Specific Semantically-Defined Object* This example shows a lexically-defined verb together with an object that is defined

(c) *ARPA funded consortium* (d) *EU funded research*	by semantic class; for simplicity, this is the same semantic class used above.
[NOUN -- AUX -- VERB -- "below"] (a) *measures are described below* (b) *data is found below* (c) *documents are given below* (d) *framework is suggested below*	*Semantic Verb Phrase* This example shows a verb phrase that picks up a semantic class even though it contains only syntactic and lexical constraints; this illustrates the idea of competing slot-constraints.
["who" -- AUX -- VERB] (a) *who are involved* (b) *who are inconvenienced* (c) *who is paying* (d) *who had forgotten*	*Relative Clause* This example shows a relative clause, defined using a lexical constraint for "who"; there are no semantic constraints so that this is a highly abstract construction.
[ADJ -- NOUN -- "are" -- VERB -- "by"] (a) *nutritive requirements are covered by* (b) *alcoholic drinks are characterized by* (c) *veterinary registrations are completed by* (d) *internal policies are influenced by*	*Partial Passive Main Clause* This example shows a passive main clause, not complete in the sense that the passivized agent after "by" is not included in the construction representation itself.

This section has reviewed work on representing constructions from an unsupervised and usage-based perspective. Rather than use introspection to define slot constraints, we instead start by generating these potential constructions: *potential* because not every possible representation has been entrenched for any given speaker. The idea here is to capture exposure from corpus data: what potential constructions has the learner been exposed to? But this still leaves us with a problem: now we need to select some of the representations (as entrenched) and discard others (as not entrenched). The next section considers how we can model the selection or competition between constraints from a computational perspective.

3 A Computational Theory of Usage-Based Grammar

Given a very large number of potential constructions like these, how do we model which specific ones best generalize the usage that we observe in a corpus? In other words, what is the relationship between exposure to potential constructions in a corpus and the emergence of grammatical structure? This section is where we implement a theory of usage-based grammar that faces the same challenge that language learners face: selecting which constructions to remember and use for generalization. This section draws on previous work that implements multiple hypotheses about usage-based grammar and compares them experimentally. For the sake of space, we focus on one particular theory, an algorithm that uses association values (specifically, the ΔP: Ellis 2007; Gries 2013; Dunn 2018b) to measure relationships between slots fillers. The basic idea is that representations with higher association values are more entrenched in the grammar.

An overview of our model of usage-based construction grammar is shown in Figure 9.3. The first step is to search through potential constructions, using ΔP as part of an algorithm to evaluate and discard poor representations. This creates a large but manageable pool of plausible constructions. The second step is to search through potential CxGs, where each potential CxG is a constructicon made up of constructions acquired in the first step. We use the Minimum Description Length paradigm (MDL: Rissanen 1978, 1986; Goldsmith 2001, 2006) to model usage-based grammar as part of this search. The MDL metric quantifies the trade-off between memory (operationalized as the encoding size of a grammar) and computation (operationalized as the encoding size of a test corpus given that grammar). In other words, any item-specific or idiomatic construction could be memorized, but that kind of storage comes at a cost. The final step is to evaluate the best grammars on held-out data. In this case, because we are working with large corpora, we retain five independent out-of-sample evaluation sets. This kind of design ensures that we do not over-fit a particular segment of a corpus.

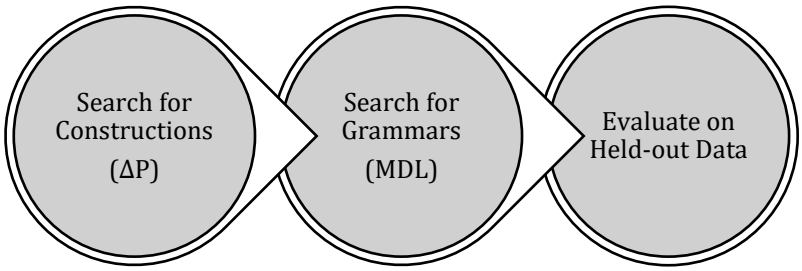


Figure 9.3. Overview of Computational CxG Model

3.1 Searching for Constructions

The first part of the model, the association-based algorithm in Table 9.2, uses the total directional ΔP (a sum across all transitions) to evaluate potential sequences of constraints. To implement this idea, the search follows transitions from one slot-constraint to the next, proceeding left-to-right through the sentence. Any transition below a threshold ΔP stops that line of the search. This algorithm references local association values when choosing a transition from the current state. It also references global (i.e., construction-wide) association for selecting different paths, rather than using the frequency of specific templates (c.f., the frequency-based algorithm described in Dunn 2019b).

Table 9.2. Association-Based Selection Algorithm

Variables
node = unit (i.e., word) in line *startingNode* = start of potential construction *state* = type of slot-constraint for node *path* = route from root to successor states [c] = list of immediate successor states c_i, c_{i+1} = transition to successor constraint *candidateStack* = plausible constructions *evaluate* = maximize sum(ΔP for c_i, c_{i+1} in *path*)
Main Loop
for each possible *startingNode* in *line*: RecursiveSearch(path = *startingNode*) evaluate *candidateStack*
Recursive Function

RecursiveSearch(*path*):
for c_i, c_{i+1} in [*c*] from *path*:
if ΔP of c_i, c_{i+1} > *threshold*:
add c_i, c_{i+1} to *path*
RecursiveSearch(*path*)
else if *path* is long enough:
add to *candidateStack*

Any series of constraints identified by this search whose transitions exceed the ΔP threshold is added to the candidate stack. At the end of the search, this stack is scored using each candidate's total ΔP across all transitions. While primarily a transition-based parse, this approach thus incorporates some global evaluation methods (c.f., Nivre and McDonald 2008; Zhang and Clark 2008). A grid search for the best ΔP threshold per language is performed using independent test data (the corpora used for these experiments is described further in Sections 4 and 5).

This association-based algorithm is less influenced by the assumption that co-located slots govern one another's constraints. For example, in reference to Figure 9.2, the slot filled by a *noun* in 3 and the slot filled by “a hand” in 4 have a local transition that is measured using the association between these two representations. Should we instead ignore the relationship between these two objects and focus on the relationship between each object and the verb slot? This algorithm tries to avoid specifying particular templates like this (i.e., a verb-centered frame) by using the global ΔP evaluation and the thread of associations to draw out these relationships.

But this raises an interesting empirical question: does the entrenchment of the ditransitive construction predict a higher association between the two object slots whether or not the verb itself is included? Is there a shared effect across all double-object constructions? A beam-search dependency parser could resolve this in a practical sense by simply evaluating more non-local relationships. But does CxG itself predict that such local relationships will be more entrenched because they are present within a single construction? This is the kind of question that becomes important when we develop a fully specified theory of construction grammar.

3.2 Searching for grammars

The second part of the algorithm uses Minimum Description Length and a tabu search to explore the space of possible CxGs. The process of searching over selected slot-constraints using a tabu search (Glover 1989, 1990) is adopted from previous work (Dunn 2018a). A tabu search is a meta-level heuristic search that evaluates a number of possible local moves for each turn and then makes the move which produces the best grammar. Importantly, a tabu search allows moves which make the grammar worse in the short-term (with a restricted set of tabu moves) so that the learner can climb out of local optima. Here, each state is a grammar that contains a specific set of constructions (i.e., a constructicon). The search works by taking a series of turns. During each turn, some constructions are *learned* (added to the constructicon) or *unlearned* (removed from the constructicon).

A grammar that provides better generalizations will allow the test corpus to be encoded using a smaller number of bits. The metric combines three encoding-based terms: L_1 (the cost of encoding the grammar), $L_2\{C\}$ (the cost of encoding pointers to constructions in the grammar), and $L_2\{R\}$ (the cost of encoding linguistic material that is not in the grammar and thus cannot be encoded using a pointer). A pointer here is a partial parse of an utterance that refers to a construction that is already contained in the grammar.

These terms represent the grammar, the data as described by the grammar, and the data that is not described by the grammar; note that both L_2 terms are combined below. In other words, $L_2(D|G)$ is the sum of both $L_2\{C\}$ and $L_2\{R\}$. *D* in this equation refers to the data set which is used to evaluate the model. The point is that the MDL metric is trying to minimize the combination of memory (L_1) and descriptive adequacy (L_2).

$$MDL = \min_{G}\{L_1(G) + L_2(D \mid G)\}$$

Encoding size, in turn, is based on probability: the encoding size of an item, *X*, is measured in bits, below, using the negative log of its probability. We describe how probabilities are estimated later in this section. The basic idea is that more probable constraints should have smaller encoding sizes. In other words, more entrenched items should be easier to retrieve.

$$L_C(X) = -log_2 P(X)$$

According to this model, a construction is only worth remembering if its contribution to decreasing the overall encoding size of the test corpus is smaller than its contribution to the encoding size of the grammar. This is important for CxGs because similar constructions overlap, describing the same sentences in the corpus. Each overlapping construction must be individually represented in the grammar, adding to the L_1 term: similar constructions must be encoded separately in L_1 but do not improve the encoding of L_2. For example, the two constructions in (1a) and (2a) describe the same utterance in (2b). Both of these constructions need to be encoded in the grammar, increasing L_1. But encoding only one of them would not increase the regret portion of L_2 because the utterance itself can still be encoded using a pointer to the construction that is in the grammar.

The encoding size of a grammar, L_1, is the sum of the encoding size of all constructions in that grammar. Each construction is a series of slot-constraints that must be satisfied for a linguistic utterance to be an instance of that construction. For each constraint, two items must be encoded: (i) the constraint type (lexical, semantic, syntactic) and (ii) the filler which defines that constraint.

The cost of (i) is fixed because each representation is considered equally probable: the grammar is not explicitly biased towards syntactic constraints. But the cost of (ii) depends on the type of representation: syntactic units come out of a much smaller inventory, so that any given part-of-speech is more probable and thus easier to encode. For example, if there are 14 parts-of-speech, then the probability of observing one of them is *1÷14 = 0.0714* bits. On the other hand, because there are more lexical items, each word is less probable and thus more expensive to encode. For example, if there are 50k lexical items, then the probability is *1÷50,000 = 0.00002*. In this way, the grammar is allowed to employ item-specific slot-constraints, but doing so increases the encoding cost of the grammar. Here, a syntactic constraint contributes 3.8 bits but a lexical constraint contributes 15.6 bits. The total encoding size of a construction is the accumulated bits required to encode each slot-constraint, where N_R represents the number of representation types (here, 3) and T_R represents the number of possible slot-fillers for that type.

$$\sum_{i}^{N_{SLOTS}} -log_2(\frac{1}{N_{R_i}}) + -log_2(\frac{1}{T_R})$$

The encoding size of the test corpus, L_2, contains two quantities: first, the cost of encoding pointers to constructions in the grammar; second, the cost of encoding on-the-fly any parts of the corpus that cannot be described by the grammar. The cost of encoding pointers is also based on probabilities, so that more probable or common constructions require fewer bits to encode. For example, a construction that occurs 100 times in a corpus of 500k words has a pointer encoding size of 12.28 bits, but a construction that occurs 1,000 times costs only 8.96 bits per use. In this way, the probability of potential constructions influences encoding size. The regret portion of the L_2 term is the cost of words which are not covered by constructions in the current grammar. Each of these is encoded on-the-fly (i.e., not remembered): the more unencoded words accumulate, the more each one costs.

There is a close relationship between MDL and Bayesian inference methods (c.f., Barak et al. 2016; Barak and Goldberg 2017; Goldwater et al. 2009). Information theory describes the relationship between the log probabilities of representations and their encoding size. But it does not estimate the probability of the grammar itself, which here is handled in two ways: First, there is a choice in CxG between different types of representation (LEX, SYN, SEM). This model does not enforce one type, but syntactic constraints are more likely because there are fewer categories. Second, pointers to constructions are assigned probabilities based on their observed frequency; this means that more likely constructions are cheaper to encode and implicitly favored by the model.

It is worth pausing at this point to think about what we have done here. Most linguistic theories are under-specified, in the sense that there are important details missing. What we have presented is a theory of usage-based construction grammar in which every necessary detail is made falsifiable and replicable. Our implementation of the MDL metric calculates the

relationship between a given grammar and a given corpus. The data[3] and the code[4] for these experiments are both available for replication. This level of detail is required for a fully-specified linguistic theory. At the same time, the details of the model are subject to empirical evaluation and improvement. This cycle of rapid and direct empirical evaluation is what makes the computational paradigm so promising.

3.3 Evaluating Grammars

At this point we turn to the evaluation of these usage-based grammars. We evaluate the association-based model that we have described here with an alternate frequency-based model (Dunn 2019b). The basic idea is that we evaluate these different hypotheses on the same test data, using the same representation pipeline, using the same implementation of the MDL metric. While we have not evaluated counter-factuals for every development decision made within the pipeline, both competing models rely on the same decisions. This gives us a measure of the relative quality of each hypothesis. The measure is relative in the sense that we can only compare implemented models.

MDL provides a single metric of a grammar's fit relative to a particular data set. This metric itself is dependent on each data set; we thus calculate a baseline encoding score that represents the encoding of the data set without a grammar and use this to derive a compression metric: MDL_{CxG}/MDL_{Base}. The lower this compression metric, the greater the generalizations provided by the CxG. Compression as used in MDL is similar to perplexity within language modelling.

The evaluation uses all seven languages in order to provide a cross-linguistic counter-factual: do the generalizations agree across languages? Additionally, we evaluate the theories against five independent sets of 10 million words for each language. Table 9.3 shows the average compression by model for each language across these five test sets. We also report the p-values for a paired t-test (paired by data set) to ensure that the difference in compression between theories is significant.

Table 9.3. Compression Rates by Language with Significance of Difference Between Models

Language	Frequency	Association	P
Arabic	44.08%	***29.45%***	0.0001
German	52.49%	***18.69%***	0.0001
English	51.80%	***23.11%***	0.0001
French	43.28%	***40.52%***	0.0037
Portuguese	45.13%	***38.91%***	0.0137
Russian	54.14%	***13.93%***	0.0001
Spanish	60.34%	***26.36%***	0.0001

Lower compression scores reflect better generalizations; as shown in Table 9.3, the association-based model out-performs the frequency-based model for every language. In each case the difference between models is significant. The gap and the significance level, however, vary across languages. For Russian, there is a gap of 40.21% compression that is significant

[3] https://publicdata.canterbury.ac.nz/Research/NZILBB/jonathandunn/CxG_Data_FixedSize/

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

below the p = 0.0001 level. But for French and Portuguese that gap is only 2.76% and 6.22%, with larger p-values to match. Association always provides a better model of the emergence of slot-constraints, but for French and Portuguese the two models are much closer together than for other languages.

What do these experiments tell us about usage-based construction grammar? First, it could have been the case that there is variation across languages and across data sets. In other words, maybe a frequency-based grammar best describes one language (i.e.,. English) but not another (i.e., German). Instead, we see a very robust result in which each of five independent evaluation sets for each of seven languages shows the same result. This scale of experimentation holds our theories to a high standard and gives us confidence that we are making generalizations about *language* rather than a simple description of one language's constructicon. Second, this gives us strong evidence that frequency-alone is not sufficient for usage-based grammar: infrequent constructions can still be acquired. Thus, a theory of usage-based grammar that depends on frequency as its main descriptive mechanism is incorrect.

4 Working with Digital Language Data

This section examines sources of demographic bias in gigaword corpora and how these biases can be corrected. This is important because computational experiments rely on large digital data sets. In other words, it is possible that although the scale and precision of these experiments is very robust, the findings do not reflect actual language use. The goal of this section is to justify the validity of these data sets as a source of linguistic experiments.

We are working with the *Corpus of Global Language Use* (CGLU: Dunn 2020), a collection of over 420 billion words across 295 languages and 189 countries. The goal of this corpus is to systematically gather comparable language samples from every country in the world. The expectation is that some languages (e.g., Swahili) will be found only in certain regions of the world. Other languages (e.g., English and French) will be found in all regions and, as a result of their geographic distribution, will participate more widely in different language mixing situations. Countries are grouped into sixteen larger geographic regions to simplify the analysis of language distribution. The distribution of the corpus across regions by number of words and by percentage of words is shown in Table 9.4. The corpus draws on web data and social media data, two different forms of digital language use. The inventory of regions is relatively straight-forward. It is worth noting, however, that Brazil and Russia are large enough and produce enough language data that they are separated from surrounding countries.

Table 9.4. Words Per Region

	CGLU v.4.2		Twitter	
Region	**Words**	**%**	**Words**	**%**
Africa, North	1,223,532,000	0.29%	311,577,000	2.38%
Africa, Southern	26,868,000	0.01%	261,431,000	2.00%
Africa, Sub	5,938,870,000	1.39%	786,718,000	6.01%
America, Brazil	2,265,386,000	0.53%	291,254,000	2.23%
America, Central	8,877,634,000	2.08%	1,249,076,000	9.55%
America, North	51,921,657,000	12.15%	756,306,000	5.78%
America, South	22,441,384,000	5.25%	1,508,749,000	11.53%
Asia, Central	17,069,517,000	4.00%	311,615,000	2.38%
Asia, East	49,521,933,000	11.59%	579,847,000	4.43%
Asia, South	15,147,872,000	3.55%	937,978,000	7.17%
Asia, Southeast	21,386,781,000	5.01%	678,805,000	5.19%
Europe, East	65,413,609,000	15.31%	898,885,000	6.87%
Europe, Russia	15,363,644,000	3.60%	247,415,000	1.89%
Europe, West	143,748,386,000	33.65%	2,928,220,000	22.39%
Middle East	1,721,856,000	0.40%	800,238,000	6.12%
Oceania	1,743,571,000	0.41%	530,804,000	4.06%
TOTAL	**423 billion**	**100%**	**13 billion**	**100%**

The number of words for a given region depends on more than simply the population of the region: (i) the number of sites indexed by the Common Crawl; (ii) the population's degree of access to internet technologies; (iii) data cleaning decisions for this project that are subject to future improvements (i.e., identifying words across different writing systems). Although the relationship between words in the corpus and individuals in the regions is imperfect, in the aggregate this data set can still be used to infer many things about language use around the world. The relationship between populations and digital language data is explored further in Section 5.

A computational approach to building digital corpora has three main steps, as shown in Figure 9.4 below: finding the data (i.e., crawling or using an API), cleaning the data (e.g., to remove duplicate text), and sorting the data by language (i.e., language identification). These steps are discussed in more detail elsewhere (Dunn 2020; Dunn and Adams 2020), but an

overview is given here. A Python package is available for cleaning[5] and for language identification.[6] An interactive visualization for exploring the corpus is also available.[7]

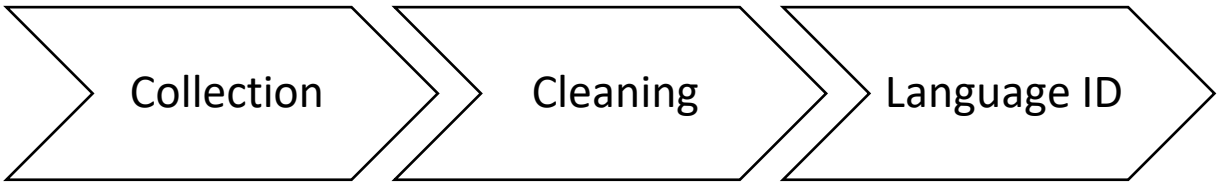


Figure 9.4. Steps in Creating Digital Corpora

This section presents the decisions made for processing the raw web data, as an example of what is required for working with this kind of data. Language samples are geo-located using country-specific top-level domains: we assume that a sample from a web-site under the ".ca" domain is from Canada. This approach does not assume that whoever produced that sample was born in Canada or represents a traditional Canadian dialect group. Some countries are not available because their top-level domains are used for non-geographic purposes (i.e., ".ai", ".fm", ".io", ".ly", ".ag", ".tv"). Domains that do not contain geographic information are also removed from consideration (e.g., ".com" sites). An important improvement in CGLU v.4.2 is the inclusion of geographic TLDs that are not in a Latin script; this significantly increases the amount of data from languages like Hindi, Urdu, and Chinese that is collected. A complete list of TLDs is contained in the codebase. We evaluate the correspondence between this data and population demographics (in Section 5) as well as the linguistic similarity between geographic data drawn from different sources (in Section 6). The basic idea is that we can validate this kind of corpus by triangulating multiple sources to measure linguistic and demographic similarity. For example, if dialectal features in Twitter data correspond with dialectal features in traditional survey-based studies, this helps to validate the collection of Twitter data as a representation of local language use (Grieve et al. 2019).

The raw portions of the Common Crawl data set used to build the corpus are shown in Table 9.5, for the purpose of showing the scale of the task. The corpus uses every portion of the crawl from March 2014 to June 2019, totaling 147 billion web pages in total. No temporal divisions are included in the corpus because these dates represent the time of collection rather than the time of production: web data does not expire and there is a long-tail in which the same samples are observed multiple times across different periods. Deduplication can remove this long-tail but cannot add accurate time information.

[5] https://github.com/jonathandunn/common_crawl_corpus
[6] https://github.com/jonathandunn/idNet
[7] https://www.earthlings.io

Table 9.5. Common Crawl Raw Data Size

Year	Period Represented (Months)	Pages
2014	March to December (8)	22.53 billion
2015	January to December (10)	17.98 billion
2016	January to December (9)	16.91 billion
2017	January to December (12)	37.28 billion
2018	January to December (12)	36.30 billion
2019	January to June (6)	16.05 billion
Total	**64 months**	**147.05 billion**

In isolation, web-crawled data provides a single observation of digital language use. Another common source of data is from Twitter (e.g., Grieve et al. 2019). We can use a baseline Twitter corpus as a point of comparison: does the Common Crawl agree with Twitter data? For example, recent work has shown that there is systematic agreement between geo-referenced corpora from the web and from Twitter across nine languages (Dunn 2021). In other words, the more precise geo-location of tweets enables us to confirm the less-precise geo-location of web data. We use a spatial search to collect Tweets from within a 50km radius of 10k cities taken from the GeoNames project.[8] This search method avoids biasing the selection of languages by relying on language-specific keywords or hashtags. The same deduplication and text cleaning methods are used as for the main web-crawled corpus. Because the language identification component only has reliable predictions for samples with at least 50 characters, a threshold of 50 characters is enforced after cleaning has taken place. The break-down of this cleaned comparison corpus by region is shown in Table 9.1 in Section 1; this represents two years of collection (July 2017 to July 2019).

Figure 9.5. Map of Web Corpus

The geographic distribution of the web corpus (by number of words per country) is shown in Figure 5; the distribution of the Twitter corpus is shown in Figure 9.6. The purpose of these maps is to provide a first pass at understanding *where* digital language data comes from. Why is this important? Recent work (Dunn and Adams 2019, 2020) has shown that a naïve corpus from these sources will over-represent North America and Western Europe. Thus, the danger is that our experiments are replicating the same geographic bias that is found in traditional dialectology studies (e.g., focusing on the US, the UK, France, etc.). Further, this work has shown that there is a significant linguistic difference between models trained on data from different countries, which means that geographic bias in our corpora could lead to bias in the experiments that we conduct using these corpora (c.f., Section 3). We examine this question further in the next section.

[8] https://www.geonames.org

Figure 9.6. Map of Twitter Corpus

The problem of language identification is often overlooked in linguistics, where language identity is taken as a given. First, what languages are distinct enough to require their own label? Second, how do we identify each language with sufficient accuracy? This corpus depends on *idNet*, a multi-layer perceptron model for language identification that covers 464 languages. Importantly, this model draws evaluation samples from over a dozen different registers. Previous work has focused on registers like Bible translations, which allow parallel data across many languages. But register variation within languages means that language use in a Bible translation may be significantly different than language use in other contexts. For our purposes, a rigorous held-out evaluation of *idNet* (Dunn 2020) shows that it is able to make highly accurate predictions about language labels across many registers.

5 Population Demographics and Digital Language Data

As soon as we try to use computational linguistics to tell us about *people* or *languages* we need to evaluate how well the data that we are using actually represents our object of study. The computational experiments in this chapter use digital corpora to study the role of exposure in language learning and language variation. But the data that we use to represent usage needs to be validated. In other words, the more we use digital corpora for scientific purposes, the more we need to control for *bias* in that data. In Section 6 we use digital corpora to represent geographic variation, so that it is essential to understand the relationship between this language data and the underlying communities we are trying to represent. There are three sources of bias that we need to take into account.

First, *production bias* occurs when one location (like the US) produces so much digital data that most corpora over-represent that location (Jurgens et al. 2017). For example, by default a corpus of English from the web or Twitter will mostly represent the US and the UK. It has been shown that this type of bias can be corrected using population-based sampling (Dunn and Adams 2020) to enforce the representation of all relevant populations.

Second, *sampling bias* occurs when a subset of the population produces a disproportionate amount of the overall data. This type of bias has been shown to be closely related to economic measures: more wealthy populations produce more digital language per capita (Dunn and Adams 2019). By default, a corpus will contain more samples representing wealthier members of the population. Thus, this is similar to production bias, but with a demographic rather than a geographic scope.

Third, *non-local bias* is the problem of over-representing those people *in* a place who are not *from* that place: tourists, aid workers, students, short-term visitors, etc. For example, in countries with low per-capita GDP (i.e., where local populations often lack internet access) digital language data is likely to represent outsiders like aid workers. On the other hand, in countries with large numbers of international tourists (e.g., New Zealand), data sets are likely to instead be contaminated with samples from these tourists.

Of these three sources of bias, non-local bias is the most difficult to uncover. We can identify production bias when the amount of data per country exceeds that country's share of the

global population. In this sense, the ideal corpus of English would equally represent each country according to the number of English speakers in that country. Within a country, we can measure the amount of sampling bias by looking at how economic measures like GDP and rates of internet access correspond with the amount of data per person. Thus, we could use median income by zip code to ensure that the US is properly represented. But non-local bias is more challenging because we need to know which samples from a place like New Zealand come from those speakers who are only passing through for a short time. Such speakers would not be representative of New Zealand English as a dialect.

Only with widespread restrictions on international travel during the COVID-19 pandemic do we have access to a collection of digital language from which non-local populations are largely absent (Gössling et al. 2020; Hale et al. 2020). This section uses changes in linguistic diversity during these travel restrictions, against a historical baseline, to calibrate the collection of digital corpora. This is a part of the larger problem of estimating population characteristics from digital language data and removing the bias that could impact our use of computational experiments.

The first question is the degree to which the production of this data is driven by underlying populations (potential production bias) and by demographic factors like GDP (potential selection bias). These experiments are based on the Twitter portion of the data described above, because this data comes with more reliable temporal meta-data. We start, in Figure 9.7, by looking at the relationship between each country's population and share of the corpus. Each country is an observation that is represented by its average monthly data production and several demographic factors. Overall, there is a very significant correlation (Pearson) between population and the amount of data from each country (0.46). Thus, the number of people in a country is an important factor explaining how much data that country produces. While this is significant, however, it also means that there are many other factors that influence the geographic distribution of the data.

Figure 9.7. Demographic Factors and Digital Corpora

To better understand the factors influencing the geographic distribution of the data, we work with three variables: *population*, the number of people in each country; *internet population*, the number of internet users in each country; and *GDP*, a measure of each country's economic output (United Nations 2011, 2017a, 2017b). Figure 9.7 shows three regression plots in which these variables (on the y axis) are compared with the average monthly data production per country (given in number of tweets per month on the x axis).

In each case, there is a close relationship between data production and demographics, with several extreme outliers. For *population*, the outliers are China and India. Both are highly populated countries with significantly lower than expected data production (especially China). Both countries have relatively low rates of internet access: 38% for China and 11% for India; this lowers the total population in each country. Thus, although the populations are quite large, most of the population is not able to produce digital language data. For the influence of GDP, the outliers are the US and China. For the US, in particular, the GDP is quite high: there seems to be a ceiling after which increased GDP is unlikely to influence digital behaviors. Further, that GDP

is not evenly distributed across the entire population. For the influence of internet access, the outliers are again China and the US. With a few notable exceptions there is a relatively close relationship between data production and the demographic factors of each country.

With these three outliers removed (the US, China, India), there are very significant correlations between these three variables and the geographic distribution of the data: 0.46 (population), 0.61 (population with internet access), and 0.59 (GDP). This leaves some unexplained production factors. The most obvious missing factor here is social media platforms specific to given countries (e.g., Sina Weibo). These alternative platforms will siphon away enough users to distort the representation of a population given access only to other platforms. Further, Twitter is banned in China: because only some companies are allowed to use it through specific VPNs, the text is not representative of language use in China. Casual users of Twitter will use a VPN through another country which would distort this method of data collection.

Regardless, this shows that we can explain a significant portion of the geographic distribution of the data. This is important because we want to describe *populations* by observing *digital corpora*. If there is no relationship between the two in terms of distribution, it is difficult to make such inferences. What we have seen, however, is that there is a very significant relationship. What is the required threshold for establishing a relationship like this? We should think about this as a metric for evaluating digital corpora: data with a stronger relationship to demographic variables are more representative.

We measure linguistic diversity as a probability distribution over languages for each country. Given this probability distribution for each country, we compare countries using the Herfindahl-Hirschman Index (HHI). The HHI was developed in economics to measure market concentration: the more of a given industry is dominated by a small number of companies, the higher the HHI (Hirschman 1945). The measure is derived using the sum of the square of shares, in this case the share of each language in each country.

Table 9.6. Sample Language Distributions by Country

	Israel	***India***	***United States***
HHI	0.207	0.356	0.852
Language #1	27.3%	50.8%	92.3%
Language #2	25.9%	30.8%	02.6%
Language #3	23.5%	03.4%	00.6%
Language #4	07.5%	02.5%	00.6%
Language #5	05.3%	01.4%	00.4%

Thus, the HHI is higher when the distribution is centered around just a few languages. For example, in Table 9.6 we focus on three countries that show a range of linguistic diversity: Israel, India, and the US. Israel has the lowest HHI (0.207). Looking at the share of the top five languages, we see roughly equal usage of three languages (in the 20s) followed by two significant minority languages. This lower HHI reflects the fact that a number of languages are being used together: no language has a monopoly. On the other extreme, the US has one of the highest values for HHI (0.852). There is one very dominant language (92%), one significant minority language (2.6%), and a number of very insignificant languages. English has a metaphoric monopoly on the linguistic landscape of the US. Global variation in linguistic diversity on Twitter is shown in Figure 9.8.

Figure 9.8. Map of Linguistic Diversity on Twitter, using the HHI

The point of measuring linguistic diversity is to evaluate changes over time: to what degree do countries change during travel restrictions resulting from COVID-19? The point here is that, if Twitter is representing non-local populations, we should see a shift in diversity during travel restrictions. Models of this kind of bias can then be used to correct for that bias and make digital corpora align more closely with population demographics. We have a measure of diversity (the HHI) and data collected by month. The basic approach is to create two groups of samples: first, months during the pandemic (March through August, 2020); second, months not during the pandemic (March through August, 2019). These two groups are aligned by month so that seasonal fluctuations are taken into account (e.g., tourism high season in February for New Zealand and in July for Italy). Given these two groups of samples, we use a t-test for two independent samples to determine whether these groups are, in fact, different. If we reject the null hypothesis, it means that linguistic diversity during travel restrictions is significantly different than the seasonally-adjusted baseline.

Figure 9.9. Countries with Changes to Linguistic Diversity During Travel Restrictions, By P-Value

The results show that 70 countries have a changed linguistic landscape during the pandemic. This is visualized in Figure 9.9, with p-values classed into highly significant (under 0.001), very significant (under 0.01), and significant (under 0.05). We see, for example, that the US and Canada undergo significant change, but not Mexico and South America. There are clear geographic patterns in linguistic change: North but not Central or South America; East Africa but not West Africa; South/east Asia but not East Asia; Europe but not Russia.

These significant changes during international travel restrictions show that our measure (the HHI) and our data (tweets) offer a meaningful representation of underlying populations. If the data did not represent populations, we would not see the relationships examined above. There are no random fluctuations in the distribution of the data across countries or in the distribution of languages within countries (Dunn et al. 2020). At the same time, given a massive social change (i.e., the COVID-19 pandemic), the measure clearly identifies changes in the linguistic landscape. Thus, the measure is both precise (not disguised by noise) and accurate (observing change where we expect it). The key point is that the change in diversity during the COVID-19 period is identifiable against the background noise.

So far we have seen that there is a significant change in the linguistic diversity of many countries *during* the travel restrictions caused by COVID-19. But to what degree are these changes *related* to the travel restrictions themselves? For example, we could imagine a population that is changing over time which we just happen to observe in mid-change. It could be the case that a country has been becoming less diverse over the past decade because of fewer incoming immigrants; the approach taken so far in this paper would misinterpret such macro-trends to be a direct result of COVID-19.

We use a difference-in-differences method (Card and Krueger 1994) to correct for this. The basic idea behind a difference-in-differences approach is to conduct a *natural experiment* with a control group (here, data from 2018) and an effect group (here, data from 2020) differentiated by time. We have three months (July, August, September) that are shared across 2018, 2019, and 2020. So, using the same methods described above, we find out which countries have a significant change between 2019 and 2020. This is the period that takes place during travel restrictions. If travel restrictions influence linguistic diversity, we would expect such influence to take place during this period. We then find out if the countries which show a significant change in 2020 also show a significant change from 2018 to 2019. This provides a baseline: removing any country whose linguistic diversity was already in the process of changing.

Over this three-month period (July through September), 58 countries show a change in linguistic diversity during the pandemic. This is a smaller number than the main results reported above for two reasons: (i) the time span is shorter, giving less robust results and (ii) this particular time span came after some travel had resumed. Of these 58 countries that show a significant change in diversity, most (38) show no difference at all in the baseline period before the pandemic. Another eight show a much greater difference during the COVID-19 period (e.g., p-values of 0.03 vs 0.004 for baseline and COVID-19, respectively). This means that the pandemic has either created or has significantly contributed to 79.3% of the cases of changing linguistic diversity. The remaining 20.7% of changes, then, must have been created by macro-trends like immigration or changes in bilingual behaviour. The main conclusion from this difference-in-differences examination, however, is that most of these changes can be specifically connected to COVID-19.

The important point in this section has been that, like all sources of language data, digital corpora are subject to certain biases. In other words, there is not a perfect relationship between the data that our experiments are using and the populations that we want to study. As with all data, we need to systematically measure and remove this kind of bias in order to improve how well our experiments generalize across global populations. The study presented here is an example of what it means to validate this kind of data to take into account production bias, sampling bias, and non-local bias. Another approach, based on register variation, is to determine if digital language shows the same grammatical and lexical usage as non-digital language. Recent work has shown that traditional survey-based methods can be replicated using digital corpora (Grieve et al. 2019). Other recent work has triangulated different sources of digital corpora to show that they are closely related (Dunn 2021). This body of work is important for validating the corpora that our computational experiments depend on.

6 Global Dialectology and Computational Construction Grammar

So far we have seen how we can conduct computational experiments on theories of usage-based construction grammar. This section goes a step further and describes recent computational experiments on variation in construction grammars (Dunn 2018c, 2019a, 2019c). The goal is to show that a theory of usage-based grammar can also account for variation. In other words, a theory of grammar must be tested on its predictions for both language learning and language variation because these are essential aspects of language that any theory needs to describe. Here the difference between dialects is modelled as the preference for some

constructions over others given a single umbrella-grammar. We experiment with the same association-based and frequency-based CxGs (c.f., Section 3), this time using their ability to make predictions about geographic variation. Thus, we previously evaluated these grammars and the theories they represent using internal measures like goodness-of-fit. Here we evaluate these grammars and the theories they represent using an external measure based on geographic variation: how well is each theory capable of capturing the difference between national dialects of a language?

Previous work on syntactic dialectology has depended on the idea that a grammar is an inventory of specific structures: the double-object construction versus the prepositional dative, for example. Under this view, there is no language-independent feature set for syntax in the way that there is for phonetics. But we can also view syntax from the perspective of a discovery-device grammar: in this case, our theory of grammar is not a specific description of a language like English but rather a function for mapping between observations of English and a lower-level grammatical description of English: $G=D(CORPUS)$. Thus, a discovery-device grammar (G) is an abstraction that represents what the grammatical description would be if we applied the learner (D) to a specific sample of the language ($CORPUS$). A discovery-device grammar allows us to generalize syntactic dialectology: we are looking for a model of syntactic variation, V, such that when applied to a grammar, $V(G)$, the model is able to predict regional variation in the grammar. But G is different for each language, so we generalize this to $V(D(CORPUS))$. In other words, we use an independent corpus for each language as input to a discovery-device grammar and then use the resulting grammar as a feature space for studying syntactic variation. This approach, then, produces an inventory of syntactic features for each language in a reproducible manner. From the perspective of cognitive linguistics, a usage-based grammar is ideally a discovery-device grammar. In other words, there is no individual grammar that is not driven by observed usage.

This section uses data-driven language mapping (c.f., Sections 4 and 5) to choose which languages in which countries need to be included as national dialects. The seven languages we consider account for 59.2% of the web-crawled corpus and 74.6% of the social media corpus. The corpora are regionalized to countries. Thus, the assumption is that any country which frequently produces data in a language has a national dialect of that language. For example, whether or not there is a distinct variety of New Zealand English depends entirely on how much English data is observed from New Zealand in these data sets. The models then have the task of determining how distinct New Zealand English is from other national dialects of English.

A Linear Support Vector Machine classifier is used to model dialects. This is a supervised method that observes a number of samples (i.e., vectors of construction frequencies representing samples from a given country) and estimates a function for mapping that vector into a hyperplane maximizing the separation between classes (i.e., national dialects). A Linear SVM is preferable to other linear classifiers with inspectable feature weights, such as Naïve Bayes, because it can better handle redundant representations. This is important because constructions vary in their level of abstraction so that a single utterance may have several constructions describing it, producing correlated features.

Constructions are quantified using their raw frequency; since all samples are the same size, this is relative frequency. Thus, the grammar is turned into a vector that contains the frequency of each construction in each observed sample. Morphosyntactic dialectometry in this paradigm depends on the fact that speakers have a large number of grammatical structures

available to them but can only choose a small sub-set of these structures in actual usage. Positive evidence for a speaker's preference is provided by each observed structure and negative evidence by each unobserved structure. In terms of cognitive sociolinguistics, an entire CxG can perform all of the functions that language is used for. Studying only a few constructions in isolation limits the functions that are represented. Thus, even if constructions are chosen because they have overlapping functions, this approach (i) may miss constructions that fulfil those same functions in other contexts or (ii) may miss some functions that are covered by those constructions in other contexts.

So long as the total choice space is relatively well covered (i.e., so long as the CxG has descriptive adequacy), the amount of negative evidence will be much higher than the amount of positive evidence. Corpus-based dialectology thus does not require the active elicitation of either specific variants or specific minimal pairs: given enough passively observed language use, the observed frequency of each structure (the input to the model) supports the estimation of each region's preferences for that structure against its competition (the output of the model).

True positives occur when the model assigns unseen samples to the correct dialect and false positives occur when the model incorrectly assigns a sample to a given dialect. The standard measures used to evaluate such an experiment are precision (the proportion of predictions for region X that actually belong to region X) and recall (the proportion of samples from region X that were correctly classified). The F-Measure reported here is the harmonic mean of these two measures averaged across all classes. The overall prediction accuracy across languages is shown in Table 9.7 (with the web corpus on the left and the Twitter corpus on the right). These scores are computed using cross-validation to protect against over-fitting. Within each data set, we compare the prediction accuracy using two different grammars: a frequency-based theory of CxG and an association-based theory of CxG. This is the same experimental comparison that we saw previously.

Table 9.7. F1 of Classification of Regional Varieties by Language and Grammar Type

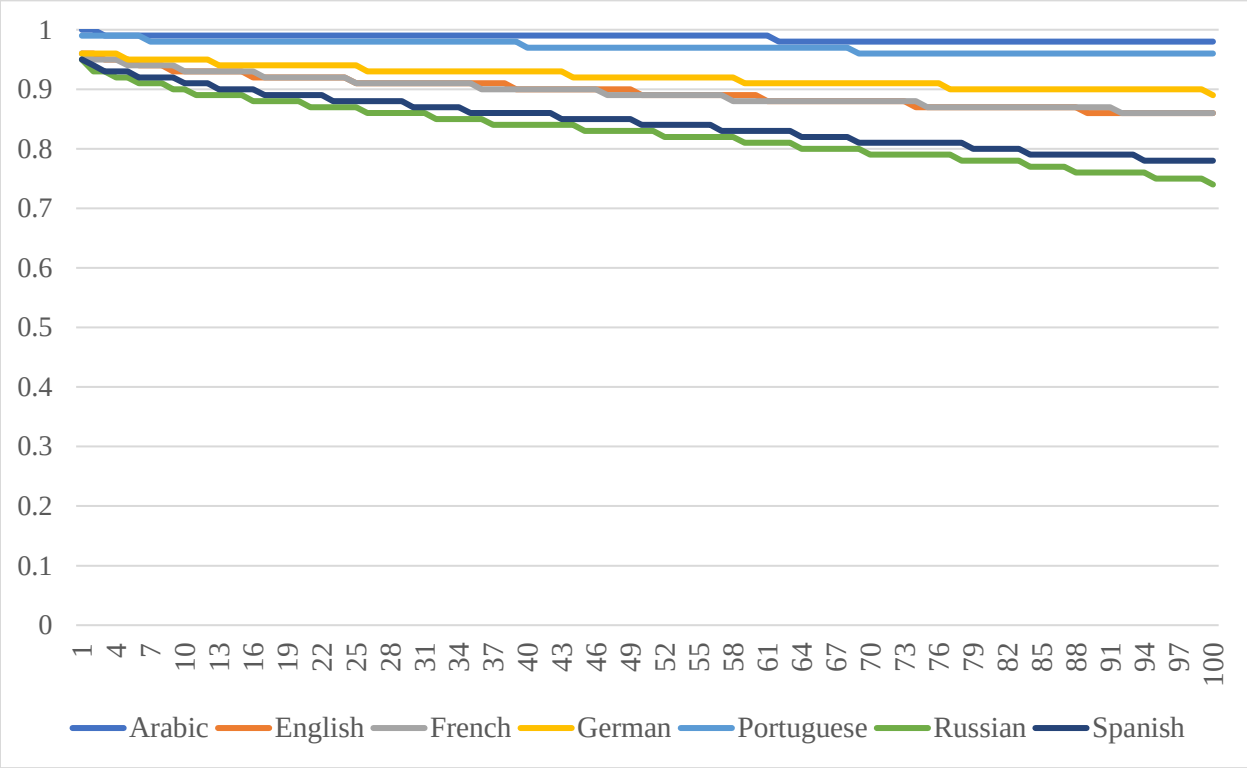
	Web Corpus		***Twitter Corpus***	
	Frequency CxG	*Association CxG*	*Frequency CxG*	*Association CxG*
Arabic	0.90	**1.00**	0.88	**0.98**
English	0.80	**0.96**	0.76	**0.92**
French	0.78	**0.96**	0.98	0.98
German	0.89	**0.96**	0.90	**0.95**
Portuguese	0.98	**0.99**	0.99	**1.00**
Russian	0.79	**0.95**	0.83	**0.93**
Spanish	0.78	**0.95**	0.82	**0.94**

A classification-based approach has the goal of distinguishing between national dialects. We would expect, then, that the task of distinguishing between a small number of dialects is easier than distinguishing between a larger number of dialects. For example, there are only two dialects of German and Portuguese in the Twitter corpus. Models on the web corpus (left) have higher predictive accuracy than models on the Twitter corpus (right). This is true except in cases, such as Portuguese, where there is a wide difference in the number of national varieties represented (for Portuguese, two vs. four). For reasons of data availability, only English and Spanish have strictly aligned varieties; in both of these languages, the grammars perform better

on the web corpus than the Twitter corpus, although the gap is wider for English than for Spanish.

What does the F-Measure tell us about models of syntactic variation? First, the measure is a combination of precision and recall that reflects the predictive accuracy while taking potentially imbalanced classes into account: how many held-out samples can be correctly assigned to their actual region-of-origin? On the one hand, this is a more rigorous evaluation than simply finding a significant difference in a syntactic feature across varieties within a single-fold experimental design: not only is there a difference in the usage of a specific feature, but we can use the features in the aggregate to characterize the difference between national dialects. On the other hand, it is possible that a classifier is over-fitting the training data so that the final model inflates the difference between varieties. For example, let's assume that there is a construction that is used somewhat frequently in Pakistan English but is never used in other varieties. In this case, the classifier could achieve a very high prediction accuracy while only a single construction is actually in variation. Before we interpret these models further, we evaluate whether this sort of confound is taking place.

Figure 9.10. Unmasking on Web Corpus



If a classification model depends on a small number of highly predictive features, thus creating a confound for dialectology, the predictive accuracy of that model will fall abruptly as such features are removed (Koppel et al. 2007). Within authorship verification, *unmasking* is used to evaluate the robustness of a text classifier: First, a linear classifier is used to separate documents; here, a Linear SVM is used to classify national dialects of a language. Second, for each round of classification, the features that are most predictive are removed: here, the highest

positive and negative features for each national variety are pruned from the model. Third, the classifier is retrained without these features and the change in predictive accuracy is measured: here, unmasking is run for 100 iterations using the association-based grammar as features, as shown in Figure 9.10 (with the web-based corpus). For example, this removes 28 constructions from the model of English each iteration (two for each national dialect), for a total of approximately 2,800 features removed. The figures show the F-Measure for each iteration. On the left-hand side, this represents the performance of the models with all features are present; on the right-hand side, this represents the performance of the models after many features have been removed. This provides a measure of the degree to which these models are subject to a few highly predictive features.

There is a relationship between the rate of decline and the number of national dialects included in the model. What we see, however, is that the performance is not showing the steep decline that we would expect if these results were an artifact. The purpose of this evaluation is to show that a classification approach to dialectology is not subject to the confound of a small number of highly predictive features.

The point of this section has been to extend our theory of usage-based grammar to geographic variation. The work discussed here shows that the same computational grammars learned in Sections 2 and 3 can be used to identify dialect membership on held-out testing data with a high degree of accuracy. In other words, we are evaluating predictions not only about what constructions are learned but also which constructions are favored by each national dialect. This is significant because the scale of these experiments covers seven languages and dozens of national dialects, so that our theory of usage-based grammar is tested in many different contexts. A rigorous experimental paradigm shows, again, that an association-based grammar makes better predictions than a frequency-based grammar, across all languages and data sets. These models do not depend on a few highly predictive constructions. This set of experiments is important as yet another piece of evidence that we can use to test our linguistic theories: a fully specified theory of usage-based grammar that covers both language learning and language variation.

7 Computational Cognitive Linguistics

This paper has presented work that shows how a computational model of usage-based grammar provides a fully replicable and falsifiable theory that can be evaluated against corpora. Our experiments show that association is more important than frequency for learning generalizations. What makes this work important is its scale: these findings are robust on out-of-sample experiments across seven languages. One potential weakness in the computational paradigm is the kind of language data that we are forced to rely on (written digital texts). We know, however, that all sources of data are subject to bias; Sections 4 and 5 have worked to measure and correct for the bias present in digital corpora. The result is a robust and fully-specified (i.e., falsifiable) theory of usage-based grammar that extends from language learning to language change.

What does this line of work mean for cognitive linguistics? First, it is clear that this theory of usage-based grammar looks somewhat different from existing theories (Langacker 2008; Goldberg 2006). The main difference is that a fully falsifiable linguistic theory must be expressed with much greater mathematical precision. Every concept must be defined in a

computable manner, rather than using human intuition and metaphoric terminology. Second, the level of abstraction here is significantly higher than in traditional linguistic argumentation. In other words, specific constructions like the dative or the ditransitive are not, themselves, directly specified or enumerated. This theory of usage-based grammar is by necessity a discovery-device grammar, in the sense that any given grammar (i.e., description of a language) exists only in relationship to the corpus that it is describing. Thus, the theory in fact covers all languages and all dialects, although here it is evaluated on only seven languages.

This is the beginning and not the final formulation of computational cognitive linguistics. There are many remaining weaknesses and many missing components. These can be addressed by continued rigorous empirical evaluation. For example, there have also been computational theories of metaphor that implement and evaluate predictions about metaphor in the same way that this paper has worked with construction grammar (Dunn 2013a, 2013b, 2013c, 2014a, 2014b, 2015a, 2015b). However, there is currently no overlap between a computational theory of construction grammar and a computational theory of metaphor, an obvious area for future research. After all, constructions are form-meaning mappings which interact with metaphor (Sullivan 2013). A better theory of computational cognitive linguistics would make predictions about (1) the entrenchment of constructions, (2) geographic variation in construction usage, and (3) where and how metaphors can be expressed in specific constructions. But the underlying idea is the same: to formalize linguistic theory as a computational model and evaluate the theory's predictions on held-out testing data.

References

Barak, Libby & Adele Goldberg. 2017. Modeling the Partial Productivity of Constructions. In *Proceedings of the Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 131-138.

Barak, Libby, Adele Goldberg & Suzanne Stevenson. 2017. Comparing Computational Cognitive Models of Generalization in a Language Acquisition Task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 96-106.

Card, David & Alan Krueger. 1994. Minimum Wages and Employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84.

Chang, Nancy, Joachim De Beule & Vanessa Micelli. 2012. Computational construction grammar: Comparing ECG and FCG. In Steels, L. (ed.), *Computational Issues in Fluid Construction Grammar*. Berlin: Springer. 259-288.

Dodge, Ellen, Sean Trott, Luca Gilardi & Elise Stickles. 2017. Grammar Scaling: Leveraging FrameNet Data to Increase Embodied Construction Grammar Coverage. In *Proceedings of the Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 154-162.

Dunn, Jonathan. 2013a. How Linguistic Structure Influences and Helps To Predict Metaphoric Meaning. *Cognitive Linguistics*, 24(1): 33-66. doi: 10.1515/cog-2013-0002

Dunn, Jonathan. 2013b. Evaluating the Premises and Results of Four Metaphor Identification Systems. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, Vol. 1*. Heidelberg: Springer. 471-486. doi: 10.1007/978-3-642-37247-6_38

Dunn, Jonathan. 2013c. What Metaphor Identification Systems Can Tell Us About Metaphor-in-Language. In *Proceedings of the First Workshop on Metaphor in Natural Language Processing*. Association for Computational Linguistics. 1-10.

Dunn, Jonathan. 2014a. Measuring Metaphoricity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 745-751. doi: 10.3115/v1/P14-2121

Dunn, Jonathan. 2014b. Multi-Dimensional Abstractness in Cross-Domain Mappings. In *Proceedings of the Second Workshop on Metaphor in Natural Language Processing*. Association for Computational Linguistics. 27-32. doi: 10.3115/v1/W14-2304

Dunn, Jonathan. 2015a. Modeling Abstractness and Metaphoricity. *Metaphor & Symbol*, 30(4): 259-289. doi: 10.1080/10926488.2015.1074801

Dunn, Jonathan. 2015b. Three Types of Metaphoric Utterances That Can Synthesize Theories of Metaphor. *Metaphor & Symbol*, 30(1): 1-23. doi: 10.1080/10926488.2015.980694

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

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

Dunn, Jonathan. 2018b. Multi-Unit Directional Measures of Association: Moving Beyond Pairs of Words. *International Journal of Corpus Linguistics*, 23(2): 183-215. doi: 10.1075/ijcl.16098.dun

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

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

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

Dunn, Jonathan. 2019c. Global Syntactic Variation in Seven Languages: Towards a Computational Dialectology. In *Frontiers in Artificial Intelligence*, 2. doi: 10.3389/frai.2019.00015

Dunn, Jonathan. 2020. Mapping Languages: The Corpus of Global Language Use. *Language Resources and Evaluation*. doi: 10.1007/s10579-020-09489-2

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

Dunn, Jonathan & Benjamin Adams. 2019. Mapping Languages and Demographics with Georeferenced Corpora. In *Proceedings of Geocomputation 2019.* doi: 10.17608/k6.auckland.9869252.v2

Dunn, Jonathan & Benjamin Adams. 2020. Geographically-Balanced Gigaword Corpora for 50 Language Varieties. In P*roceedings of the Language Resources and Evaluation Conference*. European Language Resources Association. 2528–2536.

Dunn, Jonathan & Andrea 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*. Association for Computational Linguistics. 149-159.

Dunn, Jonathan & Harish Tayyar Madabushi. 2021. Learned Construction Grammars Converge Across Registers Given Increased Exposure. In *Proceedings of the Conference on Computational Natural Language Learning*. Association for Computational Linguistics. 268-278.

Dunn, Jonathan, Tom Coupe & Benjamin Adams. 2020. Measuring Linguistic Diversity During COVID-19. *Proceedings of the Workshop on NLP and Computational Social Science*. Association for Computational Linguistics. 1-10. doi: 10.18653/v1/P17

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

Forsberg, Markus, Richard Johansson, Linnéa Bäckström, Lars Borin, Ben Lyngfelt, Joel Olofsson & Julia Prentice. 2014. From Construction Candidates to Constructicon Entries: An experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1): 114-135.

Glover, Fred. 1989. Tabu Search, Part 1. *ORSA Journal on Computing*, 1(3): 190-206.

Glover. Fred. 1990. Tabu Search, Part 2. *ORSA Journal on Computing*, 2(1): 4-32.

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

Goldberg, Yoav, Kai Zhao & Liang Huang. 2013. Efficient Implementations of Beam-Search Incremental Parsers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 628-633.

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

Goldsmith, John. 2006. An Algorithm for the Unsupervised Learning of Morphology. *Natural Language Engineering*, 12(4): 353-371.

Goldwater, Sharon, Thomas Griffiths & Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Gössling, Stefan, Daniel Scott & C. Michael Hall. 2020. Pandemics, tourism and global change: a rapid assessment of COVID-19. *Journal of Sustainable Tourism*, 1–20.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next. *International Journal of Corpus Linguistics*, 18(1): 137-165.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami & Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence*, 2:11.

Hale, Thomas, Anna Petherick, Toby Phillips & Samuel Webster. 2020. Variation in government responses to COVID-19. *Blavatnik School of Government: Working Paper*, 31.

Hirschman, Albert. 1945. *National power and the structure of foreign trade*. University of California Press.

Jurgens, David, Yulia Tsvetkov & Dan Jurafsky. 2017. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 51–57. Association for Computational Linguistics.

Koppel, Moshe, Jonathan Schler & Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8: 1261–1276.

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

Laviola, Adrieli, Ludmila Lage, Nalália Marção, Tatiane Tavares, Vánia Almeida, Ely Matos & Tiago Torrent. 2017. The Brazilian Portuguese Constructicon: Modeling Constructional Inheritance, Frame Evocation and Constraints in FrameNet Brasil. In *Proceedings of the Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 193-196.

Matos, Ely, Tiago Torrent, Vánia Almeida, Adrieli Laviola, Ludmila Lage, Nalália Marção & Tatiane Tavares. 2017. Constructional Analysis Using Constrained Spreading Activation in a FrameNet-Based Structured Connectionist Model. In *Proceedings of the Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 222-229.

Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham & Son Bao Pham. 2016. A Robust Transformation-Based Learning Approach Using Ripple Down Rules for Part-Of-Speech Tagging. *AI Communications*, 29(3): 409-422.

Nivre, Joakim & Ryan McDonald. 2008. Integrating Graph-Based and Transition-Based Dependency Parsers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 950-958. Association for Computational Linguistics.

Pelleg, Dau & Andrew Moore. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In Proceedings of the Seventeenth International Conference on Machine Learning*. 727-734.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. European Association for Language Resources.

Rissanen, Jorma. 1978. Modeling by the Shortest Data Description. *Automatica*, 14: 465-471.

Rissanen, Jorma. 1986. Stochastic Complexity and Modeling. *Annals of Statistics*, 14: 1,080-1,100.

Steels, Luc. 2004. Constructivist development of grounded construction grammar. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 9-16.

Steels, Luc. 2012. Design methods for fluid construction grammar. In Steels, L. (ed), *Computational Issues in Fluid Construction Grammar*. Berlin: Springer. 3-36.

Steels, Luc. 2017. Requirements for Computational Construction Grammars. In *Proceedings of the Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 251-257.

Sullivan, Karen. 2013. *Frames and Constructions in Metaphoric Language.* Constructional Approaches to Language 14. Amsterdam & Philadelphia: John Benjamins.

United Nations. 2011. *Economic and Social Statistics on the Countries and Territories of the World with Particular Reference to Children's Well-Being*. United Nations Children's Fund.

United Nations. 2017a. *National Accounts Estimates of Main Aggregates. Per Capita GDP at Current Prices in US Dollars*. United Nations Statistics Division.

United Nations. 2017b. *World Population Prospects: The 2017 Revision, DVD Edition*. United Nations Population Division.

van Trijp, Remi. 2017. A Computational Construction Grammar for English. In *Proceedings of Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 266-273.

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

Zhang, Yue & Stephen Clark. 2008. A Tale of Two Parsers: Investigating and Combining Graph-based and Transition-based Dependency Parsing using Beam-search. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 562-571.

Zhang, Yue & Joakim Nivre. 2012. Analyzing the Effect of Global Learning and Beam-search on Transition-based Dependency Parsing. In *Proceedings of the International Conference on Computational Linguistics.* 1391-1400.

Ziem, Alexander & Hans Boas. 2017. Towards a Constructicon for German. In *Proceedings of the Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Association for the Advancement of Artificial Intelligence. 274-277.