Modeling Constructional Prototypes with Sentence-BERT

Yuri V. Yerastov
Cornerstone OnDemand, 4120 Dublin Blvd, Dublin, California
yyerastov@csod.com

Abstract

This paper applies Sentence-Bert embeddings to the analysis of three competing constructions in Canadian English: *be* perfect, predicate adjective and *have* perfect. Samples are drawn from a Canadian news media database. Constructional exemplars are vectorized and mean-pooled to create constructional centroids, from which top-ranked exemplars and cross-construction similarities are calculated. Clause type distribution and definiteness marking are also examined. The embeddings-based analysis is cross-validated by a traditional quantitative study, and both lines of inquiry converge on the following tendencies: (i) prevalence of embedded – and particularly adverbial – clauses in the *be* perfect and predicate adjective constructions, (ii) prevalence of matrix clauses in the *have* perfect, (iii) prevalence of definiteness marking in the direct object of the *be* perfect, and (iv) greater statistical similarities between *be* perfects and predicate adjectives. These findings support the argument that *be* perfects function as topic-marking constructions within a usage-based framework.

1 Introduction

Canadian English has a *be* perfect construction, e.g. *I'm done my homework*, with the range of participles varying by dialect, but normally restricted to *done, finished* and occasionally *started*. The construction is similar in form and function to the *have* perfect, e.g. *I've done my homework*, as well as the predicate adjective construction, e.g. *I'm done with my homework*. For the sake of terminological clarity, the constructions in question are exemplified in Table 1; their abbreviations, listed parenthetically in the first column, are used throughout the paper for brevity.

While superficially similar, these constructions differ systematically in semantics, syntax, and discourse function. These differences are examined in this study, based on samples collected from a Canadian news media database. The analysis combines an embeddings-based approach with traditional quantitative methods. Using Sentence-BERT (SBERT) embeddings, constructional exemplars are vectorized, mean-pooled, and aggregated into constructional centroids. Exemplar similarity to centroids provides a measure of prototypicality, while cross-construction comparisons yield a similarity matrix. Clause type distributions are then analyzed and statistically validated against these embeddings-based prototypes. The analysis is further supported by quantitative evidence from direct object marking.

The study addresses two questions: 1.) How do the *be* perfect, predicate adjective, and *have* perfect compare in terms of semantic density, clause distribution, and definiteness marking? 2.) Can embeddings-based prototypes capture constructional tendencies in ways consistent with traditional corpus analysis?

The findings converge on three points: (i) the *be* perfect patterns most closely with the predicate adjective, (ii) it contrasts sharply with the *have* perfect, which predominantly codes new information in main clauses, and (iii) its pragmatic specialization lies in topic marking. The analysis contributes both to the description of Canadian English variation and to the methodological toolkit of construction grammar, showing how embeddings can model constructional prototypes within a usage-based framework.

2 Theoretical background

2.1 Sentence embeddings and construction prototypes

Theoretical work in construction grammar has commonly relied on acceptability judgments and corpus-based statistics for argumentation. An

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 24–33

Cxn	Example
be perfect (*be*)	I am done my homework I am finished my homework I am started my homework
predicate adjective (*be-with*)	I am done with my homework I am finished with my homework I am started on my homework
have perfect (*have*)	I have done my homework I have finished my homework I have started my homework

Table 1: Examples of the three constructions analyzed

embeddings-based approach might enhance and guide traditional methods, as well as amplify statistical signal through the strengths of deep learning models.

Advances in distributional semantics have enabled us to capture linguistic meaning in vectorized representations of words, phrases, and sentences. Early approaches such as word2vec (Pennington et al., 2014), GloVe (Mikolov et al., 2013), and fastText (Mikolov et al., 2018) produced static embeddings that reflected global co-occurrence patterns. While effective for short sequences and lexical slot analysis, these methods are limited in modeling pragmatic nuance.

Transformer models – and SBERT (Reimers and Gurevych, 2019) in particular – address this limitation. This family of models is better suited for modeling discourse-level relationships and sentence-level meaning. These models work best on longer stretches of text such as a multi-clausal sentence or a sequence of sentences and perform better in capturing pragmatic relationships than do static embeddings. The specific version of SBERT used in this study for inferencing is all-mpnet-base-v2 (Song et al., 2020); it was fine-tuned by Microsoft with semantic similarity tasks on a corpus of 1 billion sentence pairs.

Sentence embeddings are particularly effective for modeling constructional prototypes because they capture a mix of semantic, syntactic, and pragmatic information within a dense vector space. Mean pooling constructional exemplars allows us to abstract away a construction's most prototypical properties and create an idealized representation that defines its central meaning. The result is constructional centroids that represent abstract prototypes relative to its member exemplars.

In order to compute sentence embeddings, this study employed an inferencing technique based on mean pooling of tokens over complete sentential spans, rather than isolated clausal domains. This choice reflects the principles of the usage-based paradigm, which posits a gradient continuum from syntax through semantics to pragmatics. For instance, Hopper and Thompson (1980) show that discourse context influences grammatical choices, and Goldberg (2005, 129-165) demonstrates how information structure can constrain syntax. Given that lexical retrieval activates a web of semantic associations, it is crucial to analyze the semantic signal extending beyond the immediate clausal domain. Because neighboring clauses can provide vital semantic associations, the broader contextual analysis enables a precise characterization of a construction's placement along a continuum of lexical specificity and schematic generality.

2.2 *be* perfect in North American English

Occurrences of the *be* perfect have been documented in Canada (Hinnell, 2012; Yerastov, 2017; Murphy, 2018) and Philadelphia (Fruehwald and Myler, 2015). These attestations have been generally restricted to aspectual participles: *done, finished, started*, although other transitive participles in the *be* perfect have been documented in Southern Atlantic states and Pennsylvania (Atwood, 1953, 26-27), in Lumbee English in the US (Wolfram, 1996), and in Bungi English in Canada (Gold, 2007).

The *be* perfect is not fully abstract. It behaves like a prefab with some fixed material, in the meaning of Bybee (2006), subject to a number of constraints. Thus, the subject slot is restricted to animate referents, the participle slot favors three items only, and the direct object slot tends to be marked for definiteness, showing sensitivity to lexical idiosyncrasy (Yerastov, 2012, 2015), and requiring exhaustivity (Hinnell, 2012, 74-77).

The *be* perfect is quite distinct from its relatives and not reducible to an elliptical or surface instantiation of any other structure. For instance, semantically, it resembles the *have* perfect when it yields resultative interpretations, e.g. *I'm done dishes* "I've finished washing the dishes". In contrast to *I'm done with dishes*, the *be* perfect cannot have a stative entailment such as "I do not want to do dishes ever again". More to the point, consider the contrast between *I'm never finished with my home-*

work on time and **I'm never finished my homework on time*, where a stative interpretation of the *be* perfect fails. Finally, in dialects that do allow *start* in the construction, it is hard to induce a stative reading on an inceptive verb; *I'm started my homework* can only be interpreted resultatively.

Yet in other environments the *be* perfect behaves like its predicate adjective relative: both constructions share stative adjectival properties. For instance, the two constructions allow extent adverbs, while the *have* perfect does not, e.g. *I'm all done (with) my chores*, c.f. **I've all done my chores* (Yerastov, 2012, 444) and *I'm all ready*. Further, the *be* perfect and predicate adjective constructions cannot accept adverbial modifiers of manner, e.g. **I am carefully done my homework* (Fruehwald and Myler, 2015), c.f. **I'm carefully done with my homework*. Semantic similarities between *be* perfects and predicate adjectives can be further seen in the fact that they both generally disallow continuitive, hot-news and experiential readings.

These stative similarities have led linguists working within the generative tradition to resolve the status of the *be* perfect to a stative passive (Fruehwald and Myler, 2015; Murphy, 2018). While the *be* perfect undeniably exhibits stative passive properties in some environments, its resultative semantics and behavior are equally apparent in others. Such functional duality does not pose theoretical problems for a usage-based approach to language, adopted here, which allows for gradience of morphosyntactic categories (Barlow, 2000).

3 Methods

3.1 Data collection

Geographically, the present study is restricted to Canada. The data used in the study originated in Canadian Newsstream (formerly Canadian Newsstand), a news media database, available through many North American academic and public libraries. This choice is motivated by the low to non-existent frequency of the *be* perfect in general linguistic corpora; as an example, Yerastov (2017) provides a review of scarce search results from the Corpus of American English, the Corpus of Historical American English, the Strathy Corpus, the Bank of Canadian English, the Scottish Corpus of Texts and Speech, and Project Gutenberg – the documented attestations in these sources, while valuable, are insufficient for statistical generalizations.

Because the *be* perfect is a low-frequency, dialectally marked construction, an exhaustive search was feasible, yielding 1719 tokens. For comparison, stratified probability samples were collected for the *have* perfect (603 tokens) and predicate adjective constructions (702 tokens). Stratification ensured balanced representation across participles and tense permutations. Post-processing of the samples led to the filtering-out of sequences that did not meet target morphosyntactic criteria (e.g., misparsed complements). The end result was the difference in sample size for the three constructions. However, the resulting samples were large enough to afford meaningful statistical generalizations.

Two exclusions from the study should be noted: 1.) the participle *started*, and 2.) stand-alone and interposed adverbial clauses. Only 6 *started* exemplars were found in the *be* perfect sample. Because the baseline for the comparison was the *be* perfect, these exemplars were excluded from the study. Only 1 interposed and 3 stand-alone adverbial clauses were found in the three samples; they were omitted in the adverbial analysis due to their scarcity.

3.2 Analytical procedure

Constructional exemplars were represented by mean-pooling of token embeddings from the last hidden layer, with inference performed over the entire sentential span. While most exemplars were complete sentences, some were truncated search engine results; however, even in these cases, constructional slots and clause status information were fully preserved. Centroids of the exemplar embeddings were then computed for the three samples, using mean pooling as well.

The constructional prototypes were modeled by computing cosine similarity scores between each constructional exemplar and their respective centroid in order to rank all members of a distribution relative to its center. To ensure a focus on the most representative data, the 10 highest-ranking exemplars from each distribution were selected for further analysis. While a more extensive analysis involving longer rank lists or clustering of the exemplars would be beneficial, it was beyond the scope of the present study due to space limitations.

4 Results

4.1 Centroid similarity

The cosine similarity matrix in Table 2 shows that the *be* perfect is more similar to the predicate adjective (0.83) than to the *have* perfect (0.75). Distributional analyses of centroid-to-exemplar scores confirmed this pattern. The statistical properties of each of the *be, be-with* and *have* score distributions are visually summarized in Figure 1. Median similarity was found to be highest for the *be* perfect ($\tilde{x} = 0.3564$), followed by the predicate adjective ($\tilde{x} = 0.3418$) and the *have* perfect ($\tilde{x} = 0.3268$). However, applying a statistical test to assess central tendency is problematic in this case because centroid-to-exemplar similarities might violate the assumption of intra-sample independence: a centroid is defined by all vectors in a set.

As an alternative, approximate independence was achieved by summarizing per-exemplar similarities. To re-assess differences in the semantic density of each sample, cosine similarities were computed between all pairs within each sample. For each exemplar, its average similarity was calculated relative to all other exemplars in the same sample. These per-exemplar averages were then treated as approximately independent observations. In order to select an appropriate test of central tendency for these observations, their intra-sample normality was evaluated using the Shapiro-Wilk test. Because the *be* sample was found to deviate from normality ($p = 0.002$), the non-parametric Kruskas-Wallis test was applied to compare the medians across the three samples. The test indicated a statistically significant difference among the samples, $H(2) = 176.26, p < 0.001$; post-hoc pairwise comparisons were performed with Dunn's test using the Holm correction – all pairwise differences remained significant after adjustment ($p < 0.001$). The *be* sample exhibited the highest median of intra-sample similarity means ($\tilde{\tilde{x}} = 0.1257$), followed by *be-with* ($\tilde{\tilde{x}} = 0.1165$) and *have* ($\tilde{\tilde{x}} = 0.103$) – the same ranking as was observed in the centroid-based analysis.

The differences in the medians of cosine similarity distributions are not readily explained by variations in information quantity among the samples. All three samples were tokenized using spaCy (Honnibal et al., 2020), and their tokens counted per sentence. The *be* sample was found to have a higher median token count ($\tilde{x} = 25$) than the *have* sample ($\tilde{x} = 22$), yet the *be* sample exhibited the highest centroid-to-exemplar and per-exemplar similarity medians, contrasting with the *have* sample's lowest values. More to the point, the *be* sample has the largest number of exemplars, while the *have* sample the lowest. These relationships suggest that increased token length and exemplar count do not necessarily equate to greater semantic diversity in this comparison.

	be	be-with	have
be	−−	0.8318	0.7519
be-with	0.8318	−−	0.7517
have	0.7519	0.7517	−−

Table 2: Cosine similarity scores between constructional centroids

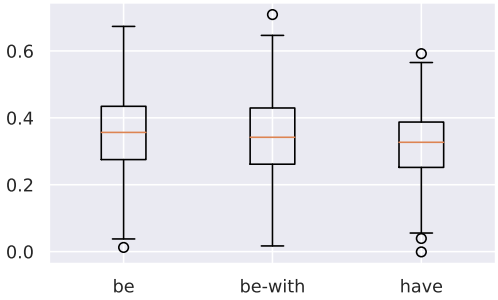

Figure 1: Distribution of centroid-to-exemplar similarity scores

The findings with respect to cosine similarity distributions allow us to evaluate the constructional samples in terms of semantic homogeneity. The more exemplars in a sample are like the center – and, more broadly, the more they are like each other – the more homogeneous the sample is overall. Therefore, it can be concluded that the *be* perfect is most semantically homogeneous, while the *have* perfect is least homogeneous (conversely, the *have* perfect is most semantically diverse); and the predicate adjective occupies a position in the middle of this continuum. From the viewpoint of construction grammar, semantic homogeneity can be interpreted as an indicator of lexical specificity, while semantic diversity – as an indicator of abstraction.

4.2 Clause type distribution

The top-ranked exemplars for the *be* perfect construction are presented in (1) through (10) sorted by cosine similarity in descending order. We observe that there are only 2 main clauses (3), (9) in

this subset, while the rest of the exemplars occur in embedded clauses. Within the embedded subset, there are 6 preposed adverbials (1), (4), (5), (6), (7), (10), 1 nominal clause (2), and 1 postposed non-finite adverbial clause of reason (8).

(1) When employees are finished that we'll send them home.

(2) I thought I'd be done school by now.

(3) By the time we got up there on Monday afternoon, they were done that part of it [...]

(4) Now my friends are done school, they're doing what they really want to do [...]

(5) Once those teachers are finished their last practicum, and they're eligible for graduation

(6) In Vancouver, when people are finished work, they're finished work.[1]

(7) When they are finished their work, they will bring it forward to us.

(8) He'll be glad to be done the homework and on to the holidays [...]

(9) I'll be done university two years from now, hopefully,

(10) Once the kids are finished school in June 1999, we'll be looking at going down.

The tendency toward embedding can also be observed in the top-ranked exemplars of the predicate adjective construction, sorted by descending similarity in (11) through (20). There are 3 preposed adverbial (13), (14), (17), 1 postposed adverbial (12), 1 relative (15), and 2 nominal (16), (18) clauses. The remaining 3 exemplars occur in main clauses (11), (19), (20).

(11) Now we are done with them.

(12) People here have made lifelong decisions because we were finished with this, Mr. Coma said.

(13) When I was finished with Mitch and Abby, I was, you know, as a creator, I was done with them, he said.

(14) When he was finished with the game, that's it, period, Gravelle said from his home in Maniwaki.

(15) It was time to have another and be done with it.

(16) If I knew I was done with this sport, it'd have been over, [Ahman Green] said.

(17) When we are finished with them, they are not finished with us.

(18) They said, for themselves, when they retired, they knew in their heart they were finished with the amateur sport world, said [Jennifer Robinson], a native of Windsor, Ont.

(19) I am done with them.

(20) We are finished here, we are done with this transaction, Einhorn, 42, told reporters on a conference call.

A distinct distributional shift is observed in the top ranked exemplars for *have* perfects, which are sorted by similarity in (21) through (30). We observe that main clauses (21), (23), (24), (25), (27), (28) have a slight edge over nominal ones (22), (26), (29), (30), with no incidence of adverbials.

(21) They have done a wonderful job and they are to be congratulated

(22) To have finished construction and started up the GTG well ahead of our schedule is an extraordinary achievement, said Derrick Kershaw, general manager of the Aurora Project.

(23) And we have done a tremendous amount of work improving our [...]

(24) This committee has done a lot of great work in the past two years [...]

(25) Our associates have done a fantastic job making sure we're ready to [...]

(26) I would have liked to have finished a little bit stronger, but to me what's important is next weekend and I'm pretty happy with today in a lot of ways, Nesbitt said in a conference call before going for a recovery massage.

(27) We have done our best and presented our best.

(28) We have finished the basic work of organizing Arts Alive in Kneehill as a registered Society in Alberta, and clarified our

[1] Here and elsewhere in the examples, when the construction of interest occurs in more than one clause within the same sentence, the more marked variant becomes the focus of analysis. Thus, this particular exemplar is treated as adverbial.

clause type	be	be-with	have
main	2	2	6
nominal	1	2	4
relative	0	1	0
adverbial	7	5	0

Table 3: Clause type distribution in top-ranked exemplars.
$G(6, n = 30) = 17.23, p = .008$

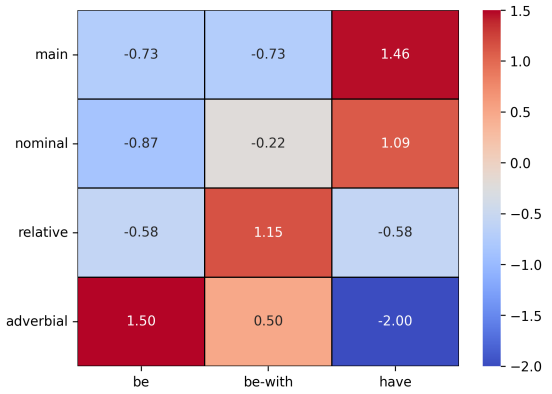

Figure 2: Standardized residuals from G-test of clause type distribution in top-ranked exemplars

mandate and goals.

(29) By now, nearly five years after he took over, it is evident that Gainey has finished the job of bringing the Canadiens out of the abyss, that awful trough where the club languished between 1999 and 2001.

(30) Hunt said council and staff have done a lot of good things over the past four [...]

To assess the distinctness of the three distributions within the centroid-based subsamples, a G-test (likelihood ratio statistic) was performed, as detailed in Table 3; this test was selected due to the limited number of observations. The statistically significant outcome ($p = .008$) supports the conclusion that the distributions differ. Further examination of the standardized residuals, shown in Figure 2, highlights a particularly strong deviation from the expected values for *have* and *be* adverbial clauses. Because this test is based on a small non-random sample, the result should be interpreted as illustrative rather than confirmatory.

5 Discussion

5.1 Semantic density

The embeddings-based prototype of the *be* perfect indicates that this construction is pragmatically rooted in recurring topical domains, particularly education and work. These patterns become especially important given that the *be* perfect displays the highest median values for both centroid-to-exemplar and per-exemplar similarity.

From a usage-based perspective – which does not draw strict boundaries between syntax, semantics, and pragmatics – such topical concentration is best interpreted as an intrinsic attribute of a construction. Specific discourse topics activate related semantic networks, thereby shaping and constraining syntactic choices. Accordingly, the informational density within a construction serves as a quantifiable measure of its lexical specificity and degree of markedness. The characteristic context of a construction is not incidental, but essential; it primes the selection of both lexical items and constructional schemas.

5.2 Adverbial clause distribution

The constructional prototype analysis reveals a strong preference for adverbial clauses among *be* perfects, and a slightly weaker adverbial tendency among predicate adjectives. In contrast, *have* perfects occur primarily in main clauses and rarely in adverbials. Since centroid exemplars represent the most prototypical instances, the absence of adverbial clauses among the top-ranked *have* perfects suggests that adverbial uses are peripheral to this construction.

The results of the prototype analysis were confirmed by a full quantitative analysis of clause type distribution across the three constructions. Table 4 presents the counts of clause types within each sample. A chi-square test of independence on these distributions revealed a statistically significant relationship between clause type and construction ($p < 0.001$), with a moderate effect size ($V = 0.294$). Based on the standardized residuals from the test, presented in Figure 3, the most extreme deviations from expectation are observed for the main *have* clauses, followed by the preposed and postposed adverbial *have* clauses – these residuals point to the *have* perfect as an outlier in the three-way comparison. Also noteworthy is the similar degree of deviation observed between the *be* and *be-with* samples with respect to main clauses.

clause type	be	be-with	have
main	437	147	408
nominal	259	141	86
postposed adv	454	141	41
preposed adv	511	254	31
relative	56	18	36

Table 4: Clause type distribution across full samples. $\chi^2(8, N = 3020) = 521.72, p < 0.001$

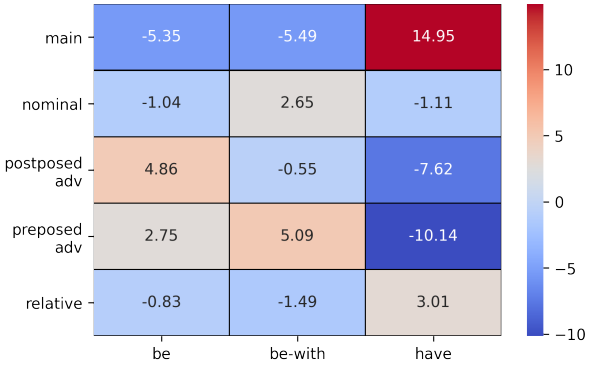

Figure 3: Standardized residuals from χ^2 test of clause type distribution

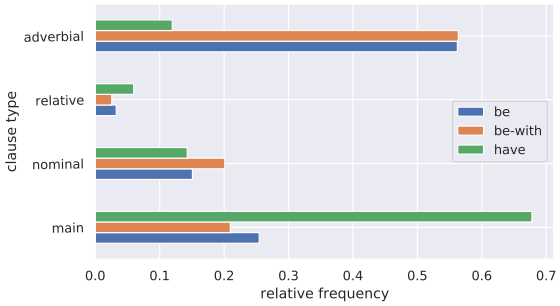

Figure 4: Distribution of constructions by clause type

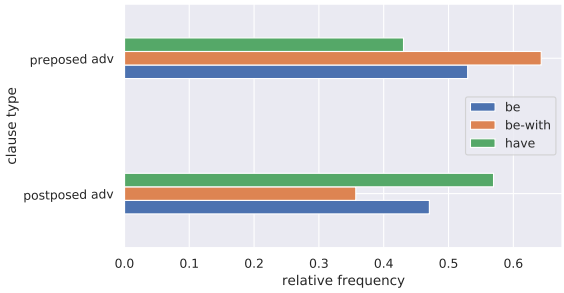

Figure 5: Distribution of constructions by adverbial clause type

To aid in the examination of the data in Table 4, the exemplar counts are normalized to relative frequencies and for better visualization presented in Figure 4 where pre- and postposed adverbials are collapsed into one class. We observe that adverbial clauses prevail in the *be* perfect ($f/n = 0.56$) and predicate adjective ($f/n = 0.56$) samples, exhibiting near identical relative frequencies, while main clauses dominate almost two-thirds of the *have* perfect sample ($f/n = 0.68$). When adverbial clauses are further isolated into separate subsamples and their counts are similarly normalized to relative frequencies (Figure 5), we find that (i) preposed adverbial clauses prevail within the *be* ($f/n = 0.53$) and *be-with* ($f/n = 0.64$) samples, and (ii) post-posed adverbial clauses prevail within the *have* sample ($f/n = 0.56$).

These quantitative findings are consistent with the pragmatic function of *have* perfects in English. It has been suggested that *have* perfects typically tend to code new information (Fenn, 1987; Michaelis, 1994; Portner, 2003); it is unusual – although not impossible – to use *have* perfects to elaborate on old information. English perfects tend to be reserved for new information, while simple pasts – for elaborations of that information (i.e. old information). The canonical – but not only – function of the *have* perfect in English is the introduction of new information. Because topic shifting is understandably less frequent in discourse than topic persistence, it is unsurprising that 68% ($n = 408$) of *have* perfect clauses in this study occur in matrix clauses rather than in embedded ones where the likelihood of presupposition and givenness is higher. And when *have* perfects do occur in adverbials they tend towards postposition. The higher incidence of postposed vis-à-vis preposed adverbial clauses among *have* perfects, revealed in this study, is in agreement with Ford (1993, 23-25), who found that the majority of perfect adverbial clauses is postposed in American English (Table 5).

While adverbial postposition is attested for both *be* and *have* perfects in this study, *be* perfect adverbials substantially outnumber *have* perfect ad-

perfect clause	count	relative freq.
postposed adv	135	0.69
preposed adv	48	0.25
stand-alone adv	11	0.06

Table 5: Distribution of present perfect by adverbial clause type in American English (adapted from Ford (1993, 24)

verbials. This outperformance is important for the present analysis because adverbial clauses are typically known to convey topical and backgrounded information (Thompson, 2011), rather than introduce new information, as would be expected from canonical perfects. Consider the exemplar in (31) from the study, in which the italicized postposed adverbial codes information of local significance, acting as a time adverbial with little anaphoric or cataphoric anchoring.

(31) To all those wonderful men who let me offer comments -- of course I looked -- and who contributed more than a loonie for the questionable privilege of me making your gift look like you'd just wrapped it yourself, thank you. [¶] The reason I was so happy -- even bursting into an off-key carol *after the elementary school kids were finished their concert*, was that each year I am heartened by the good nature of complete strangers who understand the spirit of the season is contagious. [¶] To all of those people who recognized me in the mall, bless you for reading this newspaper and helping pay my salary.

The backgrounding tendency of *be* perfect adverbials is even more evident when they are preposed; in such cases, they tend to perform global, discourse-organizing (Ramsey, 2011) functions. By way of illustration, consider two exemplars from the study. In (32), the italicized preposed adverbial follows a series of culinary descriptions and shifts topics from food to a depiction of the surrounding environment.

(32) [¶] I am a dessert lover at heart and decided to sample Ken's baklava ($4) with no regrets. This delicious dessert was crafted with several layers of phyllo pastry and walnuts. The taste of cinnamon and nutmeg were not overpowering. A clear, buttery and sweet- tasting sauce covered the entire piece. Its heat gently warmed the pastry. I could have chosen a variety of pies or muffins for dessert. [¶] *By the time I was finished my meal*, I was still quite comfortable sitting in the wooden sturdy chair at the matching table. Plenty of natural light flooded through the only large window along the front wall. A unique wall border separated the light-coloured upper wall and the lower sea-foam green coloured wall.

A similar pattern pattern can be found in (33) where the italicized preposed adverbial shifts topics from ideation to action.

(33) Before the Anti Wal-Mart War began, I had my own ideas about what the Chandler Park School could be put to use for. [¶] For years I have wanted to start a Youth Recreation Centre in Smithers, and for the past two years, I went to school to learn about business management. [¶] *Once I was finished school*, I was excited to get my plans into action - I went to Nadina and got a business plan form, and asked about small business grants. [¶] People at Community Futures Development Corporation of Nadina told me that they couldn't help me in the grant department, and gave me a form for arranging financing.

5.3 Definiteness in the direct object slot

The pragmatic specialization of the *be* perfect is evident not only in its syntactic tendencies but also in the morphology of its direct object slot. The tendency of this slot toward definiteness was already demonstrated in experimental work with native speakers of Canadian English (Yerastov, 2012, 442-443). But this study found additional, corpus-based evidence to reinforce the definiteness claim – in the context of the topicalization argument.

The direct objects of the three constructions were parsed with spaCy's (Honnibal et al., 2020) part of speech and dependency models, and slot-initial material was aggregated by category. Three cohesive categories emerged: 1) definites (definite, demonstrative and possessive determiners; demonstrative, personal and reciprocal pronouns; null anaphora); 2) indefinites (indefinite determiners and pronouns, quantifiers, wh- complementizers and relativizers); 3) undetermined nouns. Table 6 summarizes counts for each of these types across the three constructions. A chi-square test of independence revealed significant differences in these count distributions ($p < 0.001$), with a moderate effect size ($V = 0.3995$). The analysis of standardized residuals, presented in Figure 6, shows that the most pronounced deviations from expectation pertain to the indefiniteness marking of *have*

definiteness marking	be	be-with	have
definite	1290	479	230
indefinite	0	31	254
undetermined	429	192	119

Table 6: Distribution of definiteness marking in the direct object slot of the *be* perfect.
$\chi^2(4, N = 3024) = 965.41, p < 0.001$.

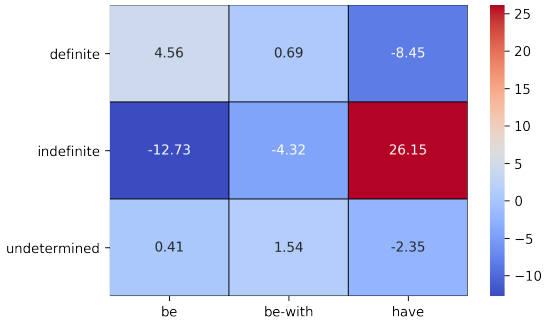

Figure 6: Standardized residuals for definiteness marking in direct object slots.

and *be* perfects.

Figure 7 visually reinforces the findings in Table 6 but in terms of relative frequency. We observe that definiteness marking of the direct object slot is strongest in the *be* perfect ($f/n = 0.75$) followed by the predicate adjective construction. Conversely, indefiniteness marking is strongest in the direct objects of *have* perfects ($f/n = 0.42$). With respect to undetermined nouns, we observe that *be-with* ($f/n = 0.27$) has a slight edge over both *be* ($f/n = 0.24$) and *have* ($f/n = 0.19$). It should be noted that undetermined noun phrases occurring in *be* perfects are either bare plurals (e.g. *chores*) or mass singulars (e.g. *school*). As such, they frequently code specific and culturally salient entities, which already carry some degree of definiteness signal in them. The prevalence of definiteness in the direct object slot is counter-expectational to the canonical tendency of direct objects in English to contain new information; it can be interpreted from a holistic perspective which takes into account the pragmatic function of the *be* perfect to background information and mark topics.

6 Conclusion

Taken together, the analyses of semantic density, clause distribution, and direct object marking converge on a unified characterization of the *be* perfect. The construction systematically patterns with predicate adjectives along the following dimensions: (i) greater semantic homogeneity, (ii) preference for embedded – and particularly adverbial – clauses, (iii) preference for pre-position among adverbial clauses, (iv) preference for definiteness marking in the direct object slot. Most importantly, the construction as a whole shows evidence of a specialized pragmatic function that consists in coding topical and backgrounded information, in clear contrast to *have* perfects, which introduce new information in matrix clauses. The topic marking function of the *be* perfect adds to the inventory of distinguishing characteristics of the construction already present in the literature (Yerastov, 2012, 2015; Hinnell, 2012; Fruehwald and Myler, 2015).

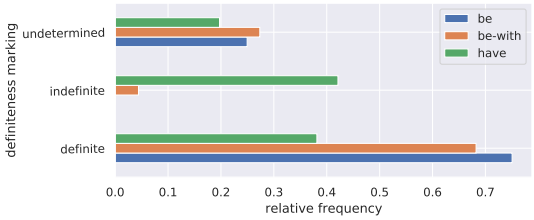

Figure 7: Relative frequency of definiteness making across the three constructions.

The distributional contrasts point to the conclusion that a constructional blend has taken place – much along the lines proposed by (Barlow, 2000), wherein syntactic, semantic, and pragmatic properties are shared across the three constructions. On the one hand, the *be* perfect inherits resultative semantics and transitive complementation from its *have* perfect relative; on the other hand, the *be* perfect inherits topicalization tendencies from its predicate adjective relative.

Methodologically, this paper demonstrates that SBERT embeddings can be used to construct prototypical representations of constructions, offering a scalable and interpretable complement to traditional quantitative analysis. The relationships observed in the constructional similarity matrix and in the centroid-based subsamples were replicated by a quantitative analysis of the entire clause type distribution. The cross-validation of these results suggests that embeddings-based methods can reliably capture distributional tendencies within a usage-based framework.

Future work should investigate the relationship between sentence-wide pragmatic signals and signals originating specifically from constructional slots.

References

E Bagby Atwood. 1953. *A Survey of Verb Forms in the Eastern United States.* University of Michigan Press.

Michael Barlow. 2000. Usage, blends and grammar. In Michael Barlow and Suzanne Kemmer, editors, *Usage-based models of language*, pages 315–345. CSLI Publications, Stanford, CA.

Joan L Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Peter Fenn. 1987. *A semantic and pragmatic examination of the English perfect*, volume 312. Gunter Narr Verlag.

Cecilia E Ford. 1993. *Grammar in interaction: Adverbial clauses in American English conversations*. Cambridge University Press.

Josef Fruehwald and Neil Myler. 2015. I'm done my homework—case assignment in a stative passive. *Linguistic Variation*, 15(2):141–168.

Elaine Gold. 2007. Aspect in bungi: Expanded progressives and be perfects. In *Proceedings of the 2007 annual conference of the Canadian Linguistic Association*, pages 1–11.

Adele E. Goldberg. 2005. *Constructions at Work: The nature of generalization in language*. Oxford University Presss.

Jennifer A. J. Hinnell. 2012. A construction analysis of [be done X] in Canadian English. Master's thesis, Simon Fraser University, Burnaby, BC.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adrienne Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Paul J. Hopper and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language*, 56(2):251–299.

Laura A Michaelis. 1994. The ambiguity of the english present perfect. *Journal of linguistics*, 30(1):111–157.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Patrick Murphy. 2018. I'm done my homework: Complement coercion and aspectual adjectives in Canadian English. *Oslo Studies in Language*, 10(2).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Paul Portner. 2003. The (temporal) semantics and (modal) pragmatics of the perfect. *Linguistics and philosophy*, 26:459–510.

Violeta Ramsey. 2011. The functional distribution of preposed and postposed 'if'and 'when'clauses in written discourse. In *Coherence and grounding in discourse: Outcome of a symposium, Eugene, Oregon, June 1984*, pages 383–408. John Benjamins Publishing Company.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Sandra A Thompson. 2011. "Subordination" and narrative event structure. In *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984*, pages 435–454. John Benjamins Publishing Company.

Walt Wolfram. 1996. Delineation and description in dialectology: The case of perfective I'm in lumbee english. *American Speech*, 71(1):5–26.

Yuri Yerastov. 2012. Transitive be perfect in Canadian English: An experimental study. *Journal of Canadian Linguistics*, 57(3):1001–1031.

Yuri Yerastov. 2015. A construction grammar analysis of the transitive be perfect in present-day Canadian English. *English Language & Linguistics*, 19(1):157–178.

Yuri Yerastov. 2017. The kids are finished school: A corpus study of geographical distribution. *Aorists and Perfects: Synchronic and diachronic perspectives*, 29:179.