ALTERNATIVE COMPLEMENTATION IN PARTIALLY SCHEMATIC CONSTRUCTIONS: A QUANTITATIVE CORPUS-BASED EXAMINATION OF COME *TO* V2 AND GET *TO* V2

Nicholas A. Lester, B.A.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2012

APPROVED:

Stefanie Wulff, Major Professor
Nancy J. Caplow, Committee Member
Willem de Reuse, Committee Member
Brenda Sims, Chair, Department of Linguistics and Technical Communication
James D. Meernik, Acting Dean of the Toulouse Graduate School

Lester, Nicholas A. Alternative complementation in partially schematic constructions: A quantitative corpus-based examination of COME *to* V_2 and GET *to* V_2. Master of Arts (Linguistics), May 2012, 81 pp., 14 tables, 3 figures, references, 39 titles.

This paper examines two English polyverbal constructions, COME *to* V_2 and GET *to* V_2, as exemplified in Examples 1 and 2, respectively.

(1) The senator *came to know* thousands of his constituents

(2) Little Johnny *got to eat* ice cream after every little league game

Previous studies considered these types of constructions (though *come* and *get* as used here have not been sufficiently studied) as belonging to a special class of complement constructions, in which the infinitive is regarded as instantiating a separate, subordinate predication from that of the "matrix" or leftward finite verb. These constructions, however, exhibit systematic deviation from the various criteria proposed in previous research. This study uses the American National Corpus to investigate the statistical propensities of the target phenomena via lexico-syntactic (collostructional analysis) and morpho-syntactic (binary logistic regression) features, as captured through the lens of construction grammar.

ACKNOWLEDGEMENTS

Above all, I would like to thank Dr. Stefanie Wulff for her guidance, patience, and dedication to this project, and for introducing me to the principles and framework structuring this research. I would also like to thank Dr. Nancy Caplow, Dr. Willem de Reuse, and Haj Ross for their insights and suggestions, as well as Dr. Stefan Th. Gries for lending his technical expertise.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The present paper investigates a particular pair of constructions in English, COME *to* V_2 and GET *to* V_2[1], as exemplified in Examples 1 and 2 below.

(1) You don't *get to see* the actual lethal injection

(2) Your in-laws will *come to appreciate* your thoughtful choices

In Example 1, GET seems to provide some aspect of its centric lexical semantics (meaning roughly *receive*) tinctured with some special modality (expressing the desire of the subject to achieve the activity represented in V_2 and the difficulty or unconventionality of doing so). In Example 2, a similar situation arises in which COME seems to offer aspects of its centric lexical semantics (deictic progress towards reference point – here an end-state) while providing aspectual modification by designating that the action be durative or iterative and require some time to culminate. Formally, these constructions are structurally analogous to a narrow subdivision of the well-researched infinitival complementation pattern (i.e. V_1 *to* V_2, e.g., *Mary* expects to win *the race*)[2]; however, these constructions are particularly idiosyncratic and conform but little to the conventional classificatory criteria ascribed them by their superficial structure (see the following sections)[3]. Furthermore, both COME and GET are extremely frequent verbs in regular usage, whose lexical-semantic contribution is often significantly altered by the slightest structural or collocational variance in the broader clausal context in

---

[1] Elements in all capital letters indicate lemmas, or uninflected forms of the lexeme in question. Subscripts refer to the iconic ordering of elements (in this case, exclusively verbs) in relation to one another. Variable elements are represented by single capital letters; *V* is equivalent to "(any) verb"; all others represent indefinitely long arguments (e.g., *X, Y, Z*).

[2] Elements in italics indicate lexically specified material (i.e. elements attested as they appear without variation, morphological or otherwise).

[3] One such discrepancy is that COME/GET *to* V_2 do not seem to include two propositions or two events.

which they appear. Above all, though, these constructions are extremely under-researched. For these reasons, a fine-grained corpus-based approach was chosen to identify specifically those features of the target phenomena which differentiate them from their superficial analogues so as to develop not only a more accurate typology of English non-finite constructions, but to bolster further the claims of cognitive linguistics (especially those of construction grammar) regarding the nature of these constructions. Through the use of two distinct statistical analyses – one semantic, the other syntactic in its orientation – this paper demonstrates empirically the integral relatedness of morpho- and lexico-syntactic features within constructions and provides new evidence for the effects of specified material on the overall meaning of partially schematic constructions.

CHAPTER 2

LITERATURE REVIEW

This section begins with a review of complementation as handled variously by the generativist, cognitive, and typological literature. Next comes an account of the motivations for and general framework of construction grammar, followed finally by an introduction to the theoretical underpinnings of quantitative corpus linguistics.

2.1 Complementation

The term complementation refers in its broadest sense to the combination of multiple propositions within a single sentence, this being accomplished with a conspicuous lack of overt coordination or non-embedded subordination (i.e. conjunctive morphemes/lexemes such as *because* or *while*). Further excluded from this definition are purposive clauses (identified as a sort of subordinate structure; cf. Dixon, 2006) and relative clauses (which act as adjectival complements to NPs, rather than propositional complements to VPs). As mentioned previously, this study focuses on V_1 *to* V_2 structures (traditionally known as "reduced infinitival complements"). This class of English complements has undergone the scrutiny of most schools of syntax and cognitive linguistics, though common threads run throughout the respective frameworks. Taking this into consideration, the following sections outline the most common interpretations of the three largest of these camps: generative/derivational, cognitive/functional, and typological linguistics.

2.1.1 Generativist/derivationalist Approaches

Early generativist approaches handled English complement clauses as representing an "NP-over-S" construction, meaning that some sentence-like element S behaves like a noun (cf.

McCawley, 1988). This ultimately amounts to a statement about the argument structure of complement taking predicates (CTPs); namely, CTPs may either have valency structures similar to transitive verbs (in that they require the presence of two or more core arguments) or to intransitive verbs (in the special case of SUBJECT-SUBJECT RAISING; see 2.1.1.1 below). The evidence for this claim comes from many different phenomena, including (a) the occurrence of complement clauses with otherwise intransitive verbs (e.g., *appear*), in which case the complement clause shares a subject with the matrix verb but expresses a proposition which is, as a whole, conceptually equivalent to the subject of the matrix clause (Example 3 – Surface; Example 4 -- Deep), and (b) the possibility of certain variations, for example pronominalization of the complement clause (Examples 5 and 6) and passivization of the complement clause across the matrix verb (Examples 7 and 8).

(3) He appears to be tired

(4) [he is tired] appears

(5) He wants *to drink a beer*

(6) He wants *that*

(7) He seems to like her

(8) *She* seems *to be* like*d by him*

The surface appearance of non-finite verb forms in the complement clause and the separation of subject from the nuclear proposition are said to be the result of certain transformations, more specifically rules which impose deletion or reorganization of arguments (for instance, the *to*-V structure is argued to derive in some accounts from an underlying "*for* X *to* Y" complement pattern, e.g., *John*$_i$ *wants* [for John$_i$ to go to the store] => *John wants to go to the store*).

Chomsky (1965) investigated more closely two subtypes of reduced sentential complementation, which pattern differently in their various surface realizations. Using *persuade* and *expect*, as exemplified in Examples 9–12, Chomsky distinguished between transformations involving what would come to be known as RAISING (*expect*) and EQUI-NP DELETION (*persuade*; Langacker, 1995; Rosenbaum, 1967). The primary distinction lies in the "logical argument" structure of each construction, reflected in deep structure but nullified in surface structure.

RAISING

(9) He expected John to go to the store for milk.

(10) He expected that John would go to the store for milk.

EQUI-NP

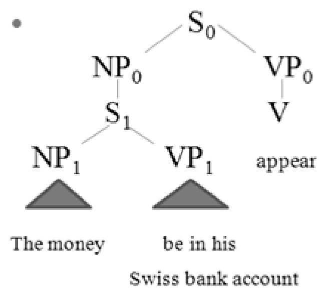
(11) Harry persuaded John to go to the store for milk

(12) *Harry persuaded that John (would) go to the store for milk

2.1.1.1 RAISING

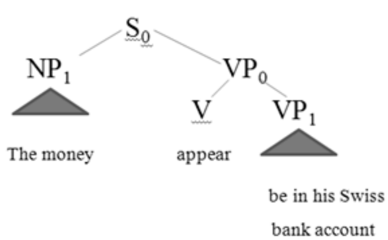
There are two major sub-types of RAISING verbs – A-type and B-type. A-type RAISING verbs pattern like *seem* and *appear* (as in *The money appears to be in his Swiss bank account* – exemplified below in Figure 1). The logical argument structure of such verbs construes the complement clause (e.g., [*the money*] BE *in his Swiss bank account* in Figure 1) as a fully realized S which shares no arguments with the matrix clause. The matrix verb is then, in deep structure, an intransitive (valence-1) verb, superordinate to the complement clause. McCawley (1988) illustrates this relationship in a tree-diagram, in which the complement S is represented as a subject NP-over-S as shown in Figure 1.

Figure 1: Tree-diagram of A-type RAISING, deep structure[4]



The process of RAISING[5], then, breaks the subject NP out of the complement S ("raising" it into the subject-NP slot of the matrix clause). The remainder of the embedded predicate is "kicked" to the end of the matrix clause and morphologically reduced to become a non-finite (*to*-infinitive) form. Thus, the logical relations of the embedded predicate are maintained while the valence requirements of the matrix verb are satisfied by the raised NP. The resulting structure is shown in Figure 2.

Figure 2: Tree-diagram of A-type RAISING, surface structures

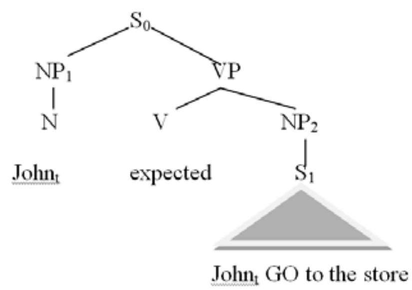


B-type instances of RAISING verbs, on the other hand, pattern like *expect* in Chomsky's and

[4] From McCawley (1988)
[5] For our purposes, only SUBJECT-SUBJECT RAISING (outlined above), as opposed to other documented cases such as SUBJECT-OBJECT RAISING, will be examined, as both COME *to* V_2 and GET *to* V_2 can only be interpreted as sharing notional *subjects* – not objects – between "matrix" and "reduced complement" clauses (though this is by no means the contention of the author).

later analyses. These differ from A-type primarily in that the NP-over-S complement is sister to V (i.e. a direct object of the matrix verb). The deep structure for *John expected to go to the store* is illustrated below in Figure 3:

Figure 3: Tree-diagram of B-type RAISING, deep structure



In true generativist fashion, a series of tests were catalogued whereby one could identify RAISING structures (cf. Postal, 1974; Chrubala & Genabith, 2007). Among these were Dummy-Subjects Examples 13 and 17), WH-Cleft (Examples 14 and 18), VP-Drop (Examples 15 and 19), and Pronominalization (Examples 16 and 20). A-type RAISING structures typically allow Dummy-Subjects[6] (e.g., existential *it*) but disallow WH-Cleft, VP-DROP, and Pronominalization (exemplified in Examples 13–16). B-type RAISING verbs allow WH-Cleft and Pronominalization but disallow Dummy-Subjects and VP-DROP (exemplified in Examples 17–20).

A-type

(13) *It* seemed to be raining outside.

(14) ?**What* it seemed was to be raining.

(15) Did it seem to rain? *It seemed <Ø>.

(16) *It seemed *that* [*to be raining*]

[6] This is only the case when the verb's semantics do not place specific requirements on the subject-type

B-type

(17) *It* expects to rain tomorrow.

(18) *What* he expected was to win the election.

(19) Did he expect that? *Yes, he expected <Ø>.

(20) He expected *that* [*to win the election*]

Given these criteria, if either COME *to* V_2 or GET *to* V_2 were part of either of these classes of verbs, they should pattern like Examples 13–16 or 17–20. Sure enough, the A-type pattern seems to match with COME (Examples 21 –24).

(21) *It* came to exist.

(22) **What* it came was to exist.

(23) Did he come to know her well? *Yes, he came <Ø>.

(24) *He came *that* [*to know her well*]

However, despite the formal similarity, there are some strong conceptual irregularities between the two that must be addressed in more detail in Section 5.1 below. Suffice it to say for now that the requirement that the subjects of RAISING verbs have discontinuous (i.e. indirect) semantic relationships with the matrix verb does not hold with COME. One consequence of this is the discrepancy between the sense of COME in a purposive intransitive motion construction (e.g., *The Russians came to aid the* Greeks) and COME *to* V_2 (e.g., *The Russians came to respect the Greeks*); that is, in the supposed deep structure of each of these constructions, COME occupies the same V slot but exhibits different syntactic and semantic relationships (for instance, the COME of purposive intransitive motion selects exclusively for motile subjects, the same not being true for the target construction, e.g., 'Disinterested' *has* *come to mean 'uninterested'*). This superficial polysemy is not a characteristic of RAISING verbs but can easily be accounted for using Construction grammar.

Applying the above tests to GET produces similar results (Examples 25-28), with the addition of the incongruity in Example 25, which is associated with B-type raising verbs:

(25) *It* got to be raining

(26) *What he got was to go to the movies

(27) Did he get to go? *Yes, he got <Ø>.

(28) *He got that.

GET *to* V_2 is then less likely than COME to belong to the A-type class of verbs. A further test strengthens the validity of this claim: passivization across the matrix verb (Examples 29-36):

A-type

(29) Her friend seemed to loathe him.

(30) He seemed to be loathed by her friend.

B-type

(31) He expected to escape the officer.

(32) *The officer expected to be escaped by him

COME

(33) He came to appreciate her family.

(34) Her family came to be appreciated by him.

GET

(35) You'll also get to see Puerto Ricans at work.

(36) *?Puerto Ricans will get to be seen by you at work.

Again, GET *to* V_2 behaves more like a B-type RAISING verb (compare Examples 35–36 with 31–32). This skewing of tendencies is possibly a reflex of the similarity of the prototypical GET (i.e., a valence-2 verb meaning *receive*) to the simple transitive character of B-type verbs (just as COME skews to the simple intransitives, the A-types). These are truly surface-level similarities, however, and a more thorough-going investigation is necessary to tease out the intuitive dissimilarity between the target phenomena on the one hand, and RAISING verbs on the other. The strangeness in (Example 36) can most readily be attributed to higher level lexico-syntactic constraints governing the felicitous interaction of GET and V_2 (see Section 4). Notice also that the same argument foreshadowed in the discussion of the "polysemy" of COME holds for GET as well. As a B-type construction, GET *to* V_2 would necessarily allow for a V_2 subject non-co-referential with that of the matrix verb – a trait which the target construction explicitly disallows, as its inclusion generates a separate construction, the causative (e.g., *Bill got* Jim *to paint his house*).

2.1.1.2 EQUI-NP DELETION (CONTROL)[7]

Instances of EQUI-NP DELETION (in later studies, CONTROL) pattern like *persuade* in Examples 11-12. In these cases, the matrix and complement clauses have at least one identical argument. In stricter terms, the matrix clause shares an argument semantically with the complement cause (Chrupala & Genabith, 2007). Important to this analysis are those instances in which the identical argument is the notional subject of both predicates, as in Example 37.

(37) John wants [*for John*] to earn the merit badge

[7] As with the treatment of RAISING, only the variety of EQUI amenable to the target constructions is described here – namely, "subject-controlled forward EQUI;" see (37).

Postal (1974) specifies that NP_2 (i.e. the subject of the complement clause) is a full lexical NP in deep structure, which is deleted due to its redundancy in relation to the matrix argument, leaving behind a now-reduced non-finite predicate. This process is illustrated in Example 38.

(38) $John_i$ wants $[John]_i$ to earn the merit badge =>

John wants Ø to earn the merit badge

Jackendoff (1972), however, counted EQUI as a "rule of [semantic] interpretation," which takes the subject of the deep structure to be phonologically null just as it is on the surface. Again, the two NPs are considered to be co-referential; the crucial difference lies in the positing of an empty node (symbolized "Δ") which is interpreted according to semantic – not syntactic – rules of co-referentiality. A sample representation of this structure is presented in Example 39.

(39) John wants Δ_{NP} to earn the merit badge

In government and binding theory, Chomsky (1981) introduced the notion of "control" along with that of an argument PRO -- a phonologically empty [+anaphor, +pronominal] subject argument, which, however, maintains features of person, number, and gender as represented by "null-case."[8] PRO Is *controlled* by the matrix NP in that its features are determined by those of the matrix subject to which it is bound via its antecedent/anaphoric relation. Therefore, this approach seeks to parameterize syntactically the binding of co-referential arguments, without

[8] PRO must not be governed because of its pronominal characteristics, but it must take case-marking because it is functionally the subject of the complement clause (an impossibility for an ungoverned entity); null-case was posited to reconcile these conflicting requirements (cf. Martin, 2001).

need of special semantic processes essentially corollary to established syntactic derivations (as with Jackendoff's proposal). An example follows in Example 40:

(40) $[[\text{He}]_{NPi}$ $[\text{tried}]_{VP}$ $[[\text{PRO}]_i$ $[\text{to catch}]_{Vi}$ $[\text{the wild hog}]_{NP}]_{S2}]_{S1}$

As with RAISING, diagnostic transformations may be undertaken to discover whether a verb ought to be classified as belonging to EQUI/CONTROL. In fact, EQUI/CONTROL verbs are for the most part operationally opposite vis-à-vis RAISING verbs, as demonstrated in Examples 41–44.

(41) **There* tried to be fumes in the air.

(42) ?*What* he tried was to win the race.

(43) Did he try to make it on time? Yes, he tried <Ø>.

(44) He tried *that* [*to win the race*]

Of these four tests, the most telling (i.e. maximally contrastive) are Examples 41 and 44 in terms of distinguishing this complement structure from RAISING. Notice, also, that neither COME *to* V_2 nor GET *to* V_2 constructions fit this mold, the former conforming in a shallow sense to the behavior of Type-A Raising verbs (see Examples 21-24), the latter torn between RAISING and EQUI/CONTROL (see Examples 25-28).

This causes some problems, it turns out, because in every instantiation of the target phenomena, the grammatical and semantic subject of the complement clause is necessarily co-referential with the matrix subject – a fact that suggests conceptual congruity with the semantic properties of CONTROL verbs (verbs which *must* share an argument semantically) as opposed to RAISING verbs of either type. A conflict arises with COME, then, which patterns ostensibly like Type-A RAISING verbs. Though GET is not as jarring given its semi-conformability to the standards of CONTROL, it still exhibits transformational incongruities worthy of more than a cursory glance.[9]

2.1.2 Cognitive/Functional Approaches

This section briefly outlines two different approaches to complementation: that exemplified by Cognitive grammar (henceforth CG; e.g., Langacker, 1987) and that of the various studies of iconicity/isomorphism (e.g., Haiman, 1980; Verspoor, 2000; Givón, 1991).

Cognitive linguistics in general diverges from the structuralist schools in several ways. Croft and Cruse (2004) list three major hypotheses adopted by most or all cognition-based linguistic frameworks which contrast sharply with the views of other systems[10] (p. 1; listed in Examples 44–46):

(44) "[L]anguage is not an autonomous cognitive faculty"

(45) "[G]rammar is conceptualization"

(46) "[K]nowledge of language emerges through language use"

The statements in Examples 44–46 combine to form a view in which (1) language is the result of the interaction of cognitive processes through which primary conceptual meaning is symbolized via conventional networks of formal-semantic pairings; (2) these cognitive processes are not restricted to language, but apply also in the organization and operationalization of other mental activities (e.g., sensory-motor, perception, relation/comparison, etc.); and (3) language acquisition is an iterative inductive process through which these particular patterns of activation are etched out and entrenched.

[9] This incongruity is intensified when taken with the fact that GET does not instantiate an event all-together separate from that of V_2.

[10] Sometimes added to this list in recent research is the requirement that all cognitive systems (including language) are "embodied," or determined by the physiological qualities of the body, its environment, and the relations between the two.

2.1.2.1 Cognitive Grammar

A general assumption of the compositional structure of complex elements in CG is that of the modifier-head-complement arrangement (Achard, 1998). In this system, “modifiers” express relationships between “reified” (nominal) concepts and their relative positioning along a valuative scale and are “instantiated” (elaborated or grounded) by “heads” (the schematic type of the concept, e.g., the noun-class schema [THING/…]; Langacker, 1995). “Complements” differ from modifiers in that they instantiate heads, highlighting some salient sub-part thereof. In this sense, reduced complement clauses are taken to instantiate some specific-level construal of the broader (in the sense of “semantically less specified”) matrix clause (Achard, 1998).

According to this interpretation, V_1 (the head in the above discussion, V_2 being the complement) should represent a conceptually more abstract event-type, a subpart of which is then narrowly focused to include material salient to the original conception. This is perhaps an attractive account, in that the very category of actions or states seems to inhere within that of temporal deictics like COME (in the target construction) for the simple reason that actions necessarily endure, if only for a moment; however, other things endure as well, like objects as we perceive them. Therefore, although it may not be reasonable to assert that in all cases temporal deictics represent a broader category of actions, it is acceptable to locate these V_1 types (i.e., like COME) among the tools of event construal. The problem arises when accounting for why these verbs should occur with the *to* + infinitive form, which has been associated with future-oriented *atemporal* event-types (due to considerations of the prepositional weight of *to*, both in the sense of an “away” vector -- towards future -- and the representation of the following verb form as “ungrounded,” or conceptually outside of time and space – consider *He wants to eat ice cream*, in which *to eat* makes no reference to, and is therefore not grounded in, a specific conceptual scene; Verspoor, 2000). If COME references continuity of development through duration or iteration in this construction, then some other compositional constraints must be at play.

2.1.2.1.1 Critique of RAISING/CONTROL

Langacker (1995) argues strictly that “'[r]aising' sentences and their 'nonraising' counterparts are not derived from the same underlying structure, nor one from the other (p. 36).” He points out (at least) three criteria by which these two seemingly related or synonymous utterances can be disentangled from one another: (1) the semantic roles of complementizers, (2) role-relations of participant NPs, and (3) the differences in the “highlighting effects of focal prominence.” These subtle aspects facilitate the various modes of construing a conceptual event available to a speaker. Any alternation in the symbolic representation of a concept, then, reflects a difference in the core concept, affecting both its interior (“immediate scope” – e.g., valuations and characteristics of participants integrally related to an action-event) and external (“overall scope” – e.g., valuations of perceiver not integrally related to an action-event) features. Therefore, the comparative methods of the generativists were misguided in that they sought to categorize syntactic types based on the possibility or impossibility of variation across truly distinct constructions (Langacker, 1995).

Langacker (1995) did note, however, the relatedness of RAISING and CONTROL structures. Arguing against the Chomskyan dichotomy, which was born from internal considerations of a theory rather than from empirical considerations of the structures themselves, he recognized that, ultimately, both constructions profile events and participants in the same way, though the possibility of complex variation in the construal of each component within them allows for some freedom in overall representation.[11]

Of the three criteria mentioned above, only the first does not apply to the classification of COME/GET *to* V_2. The other two, however, do offer insights into the incompatibility of these and other traditionally defined complement structures in English. In the first place, the agent/experiencer role of the subject in these constructions holds a special semantic relationship with the verb of the main clause, perhaps through a metaphorical linking (i.e. *He* COME as an intransitive clause stands as the source context for a construal of future-deictic development from a reference point prior to that of V_2; *He* GET *X* exhibits a similarly metaphorical sense of reception in the construal of unconventional achievement/permission). The second point raises an empirical question which would require the analysis of arguments in relation to V_1 and V_2 semantics, which is unfortunately beyond the scope of this study. However, the previous documentation of the unusual behavior of the target phenomena will tentatively be accepted as evidence for their conceptual otherness.

2.1.2.2 Grammatical Iconicity

After years of dormancy during the reign of structuralism, which stressed the necessary arbitrariness of linguistic representation, and spurred on by advances in neurobiological and cognitive-theoretical modeling, studies in iconicity have resurfaced and are gaining empirical validation. The primary assertions of this theory are based in the views, espoused above, of cognitive linguistics in general; they take as their foundation the argument that, as linguistic

[11] For a full treatment, see Langacker (1995).

processes are not substantially different from (and indeed are comprised of a combination of) other, more general cognitive processes, these should have a non-arbitrary effect on their symbolization. The iconic relationship between schemas and representations is known as isomorphic, or constituting a point-to-point mapping of one domain onto another (Haiman, 1980).

Among the principles of iconicity put forth in Givón (1991), two apply to an analysis of the complement type exemplified in COME/GET *to* V_2: the "proximity principle" and the "coding device" of the finite verb form in temporal integration. The former states basically that entities that appear closer together in their syntactic organization are closer together "functionally, conceptually, or cognitively" (p. 89). Thus, elements construed to be related to one another will appear nearer each other than elements construed as unrelated. A corollary of this statement is that the proximity of "functional operators" to their relevant "conceptual unit[s]" is determined by the relevance of their relationship. Thus, modificatory units will tend to be closest to the elements they modify (consider verbal morphology and verb stems). This seems to fit the constructional semantics of GET/COME *to* V_2 rather well, in that the lack of any intervening NP or lexical subordinator (aside from *to*, which will be discussed shortly) suggests the conceptual relatedness intuitively drawn from the meaning of these constructions (e.g., aspectual-type modification in COME, modality-type modification in GET).

Givón (1991) explains the second principle listed above as follows:

> The more integrated the two events are, the less main-clause like – finite – will the morphology of the complement verb be, with the scale of finiteness of prototype main-clauses being: FINITE>SUBJUNCTIVE>INFINITIVE>NOMINAL>BARE STEM (p. 96).

According to this proposition, infinitives are medially integrative. The underlying principle structuring this hierarchy is that a higher degree of integration is realized in a more nominal complement element. A revised model is presented in Givón (1993), adapted here in Table 1:

Table 1: Finiteness scale (adapted from Givón, 1993: 27)

least finite	**example**
lexical-nominal	remov*al*
bare-stem	remove
-ing infinitive	remov*ing*
to infinitive	*to* remove
modal	*may* remove
aspectual	*having* remov*ed*
most finite	

This scale was designed with issues of “implicativity, co-temporality, and control” in mind and developed through an investigation of deontic and epistemic verbs (Verspoor, 2000). Verspoor (2000) argues, however, that this system not only has broader application (e.g., to emotion, assessment, and judgment verbs), but multiple dimensions (i.e. finiteness interacts with other conceptual effects, such as “grounding” and direct/indirect perception). Grounded concepts are those which schematically reference a concrete entity/experience. Thus, finite verbal morphology indicates groundedness, and following another of Givón’s principles of iconicity (1991) – the principle of quantity – the greater the amount of morphology, the more complex (or grounded) the core concept. The cline in Table 1 can then be combined with a parallel model in which the least finite forms are equivalent to iconically simpler units, the most finite to iconically larger and more complex units. Directness of perception is encoded in a similar way, with the less intricate (lower level) cognitive function of direct experience being represented with less complex units, the more complex (higher level) function of indirect experience with more complex units.

These analyses are problematic when applied to COME/GET *to* V_2 in at least two ways. First, both COME and GET represent highly grounded scenarios in which the referents are clearly instantiated and well defined conceptually. With COME, it seems the event captured in V_2 represents a grounded activity – a specific event instantiation with direct participants (which, however, agree with the matrix verb) whose tense is determined by V_1 morphology – with a special aspectual construal based on the deictic nature of COME (*He has come to know her* is conceptually equivalent to *He knows her* in terms of the participant relationships, etc.). GET functions similarly, offering what appears to be a modal (i.e. assessment/permission) augmentation of the nuclear, grounded event instantiated in V_2. In terms of the directness of experience, within the immediate scope, the subject of V_1 in both constructions is necessarily a direct participant in the nuclear event. When this subject form corresponds to the speaker in overall scope (i.e. use of the first person pronoun, as in *I have come to fear her wrath*), there can be no question that the experience is construed as direct (a clear violation of the predictions outlined above).

2.1.3 Typological Approaches

Noonan (1985), in his comprehensive review of complementation across languages, distinguishes between two possible reduced complement scenarios: clause union and simple clause reduction. Clause union describes a situation in which the complement and matrix predicates share a set of grammatical relations. In these cases, the matrix clause typically retains its subject role whereas the complement clause retains its object roles, though in the case of SUBJECT-SUBJECT RAISING, the shared argument is indeed the subject. Simple clause reduction, on the other hand, refers to complement-matrix relationships not characterized by any sharing of arguments (i.e. all grammatical relationships are retained by their respective notional predicates).

The target phenomena might be interpreted as instances of simple clause reduction, though the fact that the grammatical/semantic subject of V_1 is necessarily identical to that of V_2 (if COME and GET are interpreted to be full lexical verbs and not auxiliaries) goes against the general trend of this complementation strategy. However, the description of clause union outlined above also seems inappropriate as it is clearly skewed toward <X Verb Y *to* Z> configurations (e.g., *He wants John to clean the kitchen*), which allow for the sharing of an argument beside Subject-V_1 and Object-V_2. This is not a corollary form of the target constructions, and in the case of GET, would result in an entirely different interpretation (not surprisingly, that of the causative, as in <X CAUSE Y *to* Z>; see 2.1.1.1 for an example), which is the predominant construction embodied in this form). However, as hinted above (i.e. argument sharing across clause boundaries), COME and GET might, in a broader sense, plausibly instantiate a more-or-less marked realization of clause union.

Shifting to a diachronic perspective on this issue, Givón (1995) discusses embedded complementation structures in terms of a process of grammaticalization. He proposes that the well-attested route along which main verbs become reinterpreted as tense-aspect-modality (TAM) auxiliaries, reducing finally to clitics, has its source in the extremely tight-knit bundling of main and complement predicates. As morphological marking migrates from the "semantic main verb” (in the complement clause) to the newly grammaticalized auxiliary, the main verb comes to rely on the auxiliary for all of its TAM designation, and as such is “governed” in the strictest sense of the word. Givón (1991), mentioned earlier, identifies iconic proximity as a fairly reliable signifier of close syntactic (tracing backwards, psycho-structural/conceptual) bonds. Therefore, the main-clause types most likely to undergo this process are those whose verbal elements are closest together. COME and GET seem an excellent fit for this model, given their V_1 *to* V_2 structure and intuitive aspectual/modal semantics (durative/iterative and drawn-out for COME, unconventional permission or achievement for GET – see Section 1 and Section 5 for elaboration). One issue plagues this account, however – Givón (1995) equates the iconic proximity of the elements in these types of constructions with a structure similar to that of Type-B RAISING verbs (see Figure 3). Concerning COME in particular, this analysis does not hold (if COME is, in these instances, truly considered to have some kind of TAM effect -- see Section 5). This leaves us with two options – (a) consider COME/GET to belong to a category so poorly suited to them that they share only a few characteristics with their fellow members; or (b) account for the intuitive similarity of our target phenomena and the process in question by redesigning the structural interpretation thereof. The second option is implicitly adopted in the remainder of this research[12,13].

In a more recent typological description, Dixon (2006) argues that complement clauses must (a) have the internal structure of a main clause in terms of the presence and appropriate inflection of all core arguments unless a higher order grammatical rule alters this structure

[12] That is, construction grammar will be used to interpret the target phenomena, which will, however, continue to be equated with the process of grammaticalization outlined here.
[13] A full diachronic treatment of complement constructions in flux (using construction grammar) is far beyond the scope of this paper, though the possibility of this is assumed based on similar work (e.g., Rudanko, 1998; 2002).

(e.g., EQUI in English; see 2.1.2.2 above); (b) function as a core argument of a superordinate matrix predicate; and (c) comprise a distinct proposition[14]. In addition to these requirements, Dixon posits cross-linguistic semantic categories for which verbs generally allow for the implementation of a "complementation strategy."[15] According to this distinction, there are two overarching sets of complement taking predicates: primary and secondary. Of these, the primary types are split into two subcategories – types A and B – and the secondary types into three subcategories – types A, B, and C (p. 9). Those which are of interest to this study, however, are only Primary-B, Secondary-A, and Secondary B.

Primary-B verbs are those which take either NPs or clausal complements as their O argument (e.g., *She recognized him*/*that he was a jerk*). Already we see a minor issue, in that the target phenomena never take an NP. Dixon (2006) outlines four semantic classes of these verbs – ATTENTION (e.g., *see/recognize*), THINKING (e.g., *think/assume/remember/believe*), LIKING (e.g., *love/fear/enjoy*), and SPEAKING (e.g., *say/report/describe/promise/order*) (p. 10). Clearly, none of these is an appropriate fit for either COME or GET, nor are these verbs included in the lists offered by Dixon. Therefore, this category must be discarded in pursuit of an accurate description of the target constructions.

Secondary-A verbs, if in the form of a lexical verb (as opposed to an affix or modifier), can be transitive (with an O complement, e.g., *He began his work*) or intransitive (with an S

[14] Dixon (2006) offers a fourth criterion which has to do with the ability of the complement clause to occur with certain verbs of reporting, emotion, etc, but this typological consideration lies outside the scope of the present investigation.

[15] This term refers not only to complementation as laid out thus far, but also to other forms, e.g., serialization or morphological augmentation of the verb stem, exploited by languages without overt complementizers (including rigid structural constraints, e.g., the presence of an uninflected verb form to the right of an inflected form, as in English V_1 *to* V_2).

complement,[16] e.g., *He began to work*). Of these, only the latter (in English) take reduced complement clauses with the requirement that it have the same subject as the matrix clause. This class consists of four types: negators (e.g., *no/not*), modals, “beginning-type” (e.g., *begin/continue*) and “trying-type” (e.g., *try/attempt*) (Dixon, 2006, pp. 12-13). We can therefore easily discard this as a possible category, as the closest match (i.e. the category of Modal for GET, to the exclusion of COME), at least in English, shares nothing in the way of morpho-syntactic features with GET as defined here.

Secondary-B verbs, like Secondary-A verbs, take an O argument reduced complement, the subject of which is usually omitted when identical to the subject of the matrix clause (e.g., *I plan* [I]$_{\emptyset}$ *to go to work on time today*). The primary difference is that this class does not require the subject of the complement clause to match the subject of the matrix clause. It includes verbs like *want*, *hope (for)*, *plan (for)*, etc (p. 13). This class is a possible fit for the target phenomena, if only for its somewhat vague characterization, a reflex of the typological nature of the assertion. However, it still presupposes a dipropositional scenario in which the matrix and complement verbs reference two distinct events. For this reason, this category, too, must be set aside in favor of a different theoretical approach.

2.2 Construction Grammar(s)

Based on early investigative work, such as Lakoff (1987)’s exhaustive case study of *there,* along with more formal (lexical) semantic theories like Fillmore’s Frame Semantics (e.g., Fillmore (1982)), Construction grammar is a cognitive/functional approach to language which asserts the explanatory potential of “constructions,” or “form-meaning pairings,” over that of

[16] S refers here to the subject of a transitive or extended transitive clause (Dixon, 2006, p. 7).

traditional generativist approaches which hold semantic and syntactic, as well as discourse-pragmatic, effects as belonging to functionally discrete systems (Goldberg, 2003). Goldberg (2006) defines constructions as follows:

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency. (5)

Implicit in this statement is the notion that constructions apply not only to individual lexical items, but indeed to all coordinates defined within the gradient axes of structural complexity (from morphemes to sentential argument structure) and internal specification (from substantive or fully specified constructions – idioms – such as *going great guns*,[17] to schematic or fully abstract constructions – argument structure constructions – such as the double-object construction <SUB V Obj_1 Obj_2>). Furthermore, this statement makes it clear that frequency of use affects our cognitive storage and the granularity of our analysis of component-whole relationships within constructions. Therefore, elements occurring frequently as collocates may not always be amenable to dissection simply because they conform to common combinatory restrictions,[18] but may rather be activated as wholes due to their highly entrenched cognitive routines.

2.2.1 The Problem of Polysemy

One of the most critical aspects of such an approach is its solution to the problem of polysemy. In general, lexemes of sufficient overall frequency (especially verbs) will have a

[17] This example comes from Goldberg (2003).
[18] This has implications most importantly for language acquisition, but it also impacts any theory of autonomous syntax in that (a) general structural rules are not responsible for the generation of all well-formed utterances of a language and (b) inductive approaches to the categorization of linguistic information are not only possible but necessary.

variety of uses, their meaning depending on the context of the larger utterance. Goldberg (1995)'s classic example of KICK shows that what is most readily classifiable as a monotransitive verb (*he kicked the* ball) can, in fact, occur in the ditransitive (*he kicked her the ball*), caused-motion (*he kicked the ball onto the roof*), and various other constructions. The predominant view in generativist frameworks is that such formally distinct uses of the same verb must be accounted for in the lexicon (the source of all semantic meaning) and therefore listed as separate entries[19]. This view has two disadvantages: (1) it leads to a bloated, partially redundant lexicon, and (2) it blatantly ignores the interconnectedness of the meanings of KICK in these different instances (Goldberg, 1995; Ch. 1).

Similarly, GET and COME are attested in many different constructions, sometimes displaying *prima facie* all-together different semantics (compare *get* [*receive*] *a package* to *get* [*convince*] *him to apply*). Rather than referring to each variation as a distinct form, construction grammar suggests looking past the mere lexicon to broader structures such as argument type and ordering. Thus, one should compare a special instance like *get* [*convince*] against other similar verbs, such as *force* or *persuade*, which, as it turns out, occur in the exact same construction: <SUB V Obj *to* V>. We can thus attribute the shared meanings to their belonging to the same argument structure construction (CAUSATIVE construction – something like "X causes Y to Z")[20].

---

[19] Since verbs are considered the governing entity in terms of their projecting of argument requirements, and since these different realizations of KICK cannot simply be attributed to inclusion of optional material (esp. with the indirect object of the ditransitive), each separate set of core-argument structures would require its own entry in a lexicosemantic framework (see Goldberg (1995, p. 8) for a review of this literature).

[20] For more specific evidence of the influence of argument structure constructions on the syntax-lexis interface and their contribution to overall semantics, see e.g., the case studies in Goldberg (1995).

2.2.2 GET/COME to V_2 as Constructions

Following the above criteria, this research takes the unpredictable meanings of each target V_1 (and of the construction as a whole) as evidence that some form of internal construction interaction is responsible. In that sense, each interior element is considered as holistically and through various component relationships a meaningful contributor to the overall interpretation. Moreover, the semi-idiomatic nature of these constructions are treated as an indication of their being conventionalized, or put another way, cognitively entrenched, based on their frequency of use and, again, their non-compositional semantics.

2.3 A Note on Quantitative Methods in Corpus Linguistics

In the past fifteen years or so, many technological and theoretical developments have augmented the capabilities of corpus linguists in investigating a progressively expanding scope of phenomena (Gries, 2007). As Gries and Stefanowitsch (2004) point out, early approaches to corpus studies were interested primarily in linear phenomena such as immediate collocations, often interpreting their results on the basis of frequency alone (see e.g., Kennedy, 1991) and geared toward reinforcing “either/or” statements about grammaticality (Gries, 2007). Regarding analyses based on raw frequency, the conclusions drawn in these studies were often misguided and unable to withstand statistical evaluations, which sometimes uncovered drastically different outcomes (see an example in Gries, 2009: Chapter1). Regarding grammaticality as comprising “either/or” statements, empirical research in cognitive psychology and other fields has consistently indicated a prototype-based structuring to cognitive processes, including language use (see Lakoff, 1987 for a review of this literature). This means that the categories we activate in the use of language are measured according to grammatical prototypes and an “ungrammatical” periphery; there exists no solid boundary beyond which everything is unacceptable (Langacker, 1987). Therefore, simple binary statements about grammaticality are extremely suspect. New applications of statistical tests for corpus data have proven to overcome this issue by providing “frequencies, percentages/probabilities, [and] statistical methods” which “cover the middle ground between what is possible/grammatical and what is not” (Gries, 2007, p. 4). Therefore, we can now generate data commensurable in their description of the organization of linguistic phenomena with the prototype theory of cognition. Furthermore, under the guiding assumptions of construction grammar, Stefanowitsch and Gries (2003) developed a method (collostructional analysis – See Section 3.4.1) by which not only mere linear collocations, but also the broader, schematic patterns of which they are a part, could be investigated and their relationships in the domain of the syntax-lexis interface espoused. This statistical approach to corpus linguistics is the one adopted here, with the addition of a multi-factorial analysis (binary logistic regression – see Section 3.4.2) following Wulff et al. (*to appear*). In so doing, this study adopts a somewhat naïve perspective, separating out the analyses into theoretically unrelated semantic and morpho-syntactic statistical cross-cuttings of the data. Construction grammar predicts that these tests should produce related results, but this is by no means a necessary outcome. Therefore, any corroboration of the results of these analyses can be taken as independently and empirically reinforcing the claims of this framework.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This section begins with a description of the American National Corpus (henceforth ANC) and presentation of the reasoning behind its being chosen for this particular study. This is followed by an account of the software employed in the tasks of data retrieval, refinement, and coding, as well as the procedures involved in each of these tasks. This section concludes with a discussion of the statistical models used in the analysis of the syntactic and semantic features of the target constructions, including a partial demonstration thereof.

3.2 American National Corpus

All data in this study was culled from the ANC (second release), a general corpus consisting of over 22 million words of American English from the mid-1990s to early 2000s, unpacked using the Biber part-of-speech tagset (one of many available options, applied via a computer parser) (Reppen et al., 2005). Some terminological clarification is provided below.

3.2.1 General Corpora

The term “general corpus” (as opposed to the various specialized corpora) refers to a corpus collected without selectional restrictions of genre or register. More simply, a general corpus contains data drawn from both written and transcribed-spoken sources, spanning formal and informal modes of communication. Therefore, this style of corpus is the most broadly representative of a linguistic community and thus the most desirable corpus type for exploratory research. Most importantly, it avoids distributional biases which might skew frequency – overall or in terms of representativeness or extent of variation in internal variable elements. Additionally, it is not only more likely to contain the greatest number of instances of a given target construction (general corpora are, in general, the largest, containing tens or hundreds of millions of words), but it also provides a broad-level perspective on the patterns of distribution of the target phenomenon across genres and registers, thereby suggesting appropriate directions for further, more narrowly defined inquiry.

3.2.2 Biber Tagset

The Biber tagset is a system of annotation which codes only for the lexical category of each word. The tag immediately follows the textual item and consists of an underscore and a capitalized two- or three-letter designation, as illustrated in Example 47 below.

(47) We_PRP got_VBD to_TO do_VB whatever_WDT we_PRP wanted_VBD

There are many advantages to the choice of this tagset over others, such as those based on syntactic parse. In the first place, it must be noted that the ANC has not been hand-edited for accuracy (meaning that all tags remain as the admittedly imperfect parsing program labeled them). Intuitively, then, the choice of tagset should in some sense be guided by the following maxim: the less complexity required on the part of the parsing program, the lower the chance of problematic or erroneous tags being introduced. Furthermore, the Biber tagset is more likely to over-predict instances of V_2 (i.e. to interpret nominal and other lexemes following *to* as verbs, as in, for example, come_VB to_TO work_VB, in which *work*, labeled as an uninflected verb form, is actually a locative noun) than the converse. Finally, part-of-speech annotation imposes the fewest – possibly antagonistic – presuppositions on the internal structure of the target constructions. Its maximally surface-oriented character provides the researcher with the most immediate access to the data and strengthens the validity of subsequent decisions of inclusion or exclusion (now the object of manual inspection).

3.2.3 Motivation for the Choice of American English

The decision to investigate American English (as opposed to British English) was made in consideration of one primary concern: the researcher is a native speaker of a dialect of American English. Thus, in order not to encounter any unanticipated differences in use or construal of the target phenomena found possibly in other varieties of the language, this research has been restricted to American English.

3.3 Data Retrieval

In order to extract all and only instances of the constructions under investigation from so large a corpus, this research employed a concordance generator, AntConc (see 3.3.1). Concordance generators are applications which allow for corpus files to be loaded and explored using various search types (as well as more broadly exploratory functions). Generally, the output is returned in the form of a 'keyword in context' (KWiC) table, which contains the search term along with the preceding and following context. The advantage of such a table (especially for this project) is that it allows the researcher to identify any discourse-pragmatic effects relevant to the proper construal of the search term.

3.3.1 AntConc 3.2.1w

All files of the ANC, after having been extracted into Biber-tagset text files, were loaded into the concordance generator AntConc 3.2.1w. As GET/COME *to* V_2 are partially filled constructions containing not only variable argument slots, but also variable verbal morphology, a batch search was conducted in order to capture without risk of error all verbal inflection, the search terms of which contained regular expressions (to capture all variable arguments efficiently; see 3.3.1.2). The results were then saved into a text file to be loaded into *R* for clean-up (see 3.3.2).

3.3.1.1 Batch Searches

AntConc allows the user to search for multiple items simultaneously to be returned in a single KWiC table (thus eliminating the tedious step of integrating multiple tables after the fact). A batch search was deemed appropriate for this study because of the somewhat cumbersome task of using regular expressions to code for five word-internal (morphological) alternations, and moreover, for the sake of maximal accuracy. This search contained for each of the target constructions, in addition to a statistical sample of all V_1 *to* V_2 constructions, a term for the morpho-syntactic variations shown in Table 2:

Table 2: Biber annotation of verb forms

Tag	**Meaning**
_VB	uninflected verb forms, i.e. objects of auxiliaries or modals and *to*-V forms
_VBZ	third person singular forms
_VBD	simple past tense forms
_VBN	past participles, or *–ed*/*-en* forms
_VBG	Progressive participles, or *-ing* forms

Thus, each batch search contained five search terms designed to capture all morphological variants of V_1, fifteen being entered in all.

3.3.1.2 Regular Expressions

Regular expressions (regex) are essentially special character strings that can be used as variables to search out and capture a target string, such as a morpheme, word, or broader, more schematic construction. In this case, regular expressions were only used in defining the

V_2 (except for the statistical sample, in which V_1 is also unspecified) as a variable slot containing any number of word characters and spanning from the white space after the Biber tag of *to* to the underscore in the next available _VB tag. An example of the code retrieving *come to* V_2 is provided below (Example 48):

(48) \bcome_VB to_TO \w+?_VB\b

A breakdown of this code can be found below in Table 3. The first column contains the code itself, generally broken according to relevant functional distinctions; the second column contains the formal designation of each element; and the third column contains a description of the element's function

Table 3: Code and explanation for search term retrieving come to V_2

Element	Characterization	Function
\b	Word boundary	Makes sure that words ending in "come" are not retrieved, e.g., "become"
come_VB to_TO	Specified character strings	All results will have this and the empty space after it[21]
\w	word character (alphanumeric and "_")	This finds the letters of V_2.
+	1 or more of leftward element	V_2 can be of any length > 1
?	Non-greedy – stops previous code (\w+) at first instance of next element (_VB).	V_2 should be only one word
_VB	Biber Tag	V_2 should be uninflected
\b	Word boundary	Makes sure only uninflected verb forms are returned (e.g., no returns of _VBZ)

The "\w" regex, in conjunction with the "+" ("one or more of what precedes me") and non-greedy quantifier "?" ("only so many of what precede me as occur before the first instance of

[21] This could be portrayed by the regular expression \s; however, this symbol was not employed in the actual code nor is its function an essential one in the retrieval of these data.

what follows me”), disallows the presence of more than one word following the to_TO. Therefore, only instances of the constructions which were strictly of the form V_1 *to* V_2 were returned (excluding, e.g., adverbial or other intervening material). This search term is preferable for many reasons: it allows for the direct comparison of formally regular structures, whereas any other approach would necessarily lead to asymmetries in the statistical analyses (of those instances with and without intervening material); its form is the most structurally compact and therefore the most likely to exhibit strong, perhaps prototypical iconic-semantic relationships; and only a negligible amount of instances containing extra material were returned from each of the two target constructions.

3.3.2 R 2.12.1.[22]

R is a programming language and statistical environment which allows one to search through, manipulate, and analyze data sets by loading them as text files and applying various functions. In this stage of the research, *R* was used to clean up and arrange the KWiC table returned from AntConc. The new table was saved into a text file and exported to Microsoft Excel.

3.3.2.1 Data Clean-up and Arrangement

For ease of further inspection, as the data returned from the AntConc search was still annotated (i.e. still had Biber tags on each word) and arranged in three columns where four are preferable (see below) the KWiC table was loaded into *R* to remove these tags from V_1 and V_2. Although a code could have been written to remove *all* annotation in the document, this would have unnecessarily increased the possibility of errors in the programming syntax and/or results.

[22] NB: some regular expressions in *R* require an extra back-slash (e.g., \\b rather than \b for word boundaries). This has no effect on the interpretation of the code.

In addition to creating an easy-to-read format by eliminating tags from the target phenomena, some adjustments to the arrangement of the data were necessary. Because we are interested in the frequency of each V_2 lemma (see Section 3.4.1), and for ease of manipulation of the data along this criterion, a table consisting of four basic columns (preceding context, V_1, V_2, and following context) is the most suitable. These tasks were carried out simultaneously with the function *gsub*, which is composed crucially of two core arguments. The first argument of *gsub* isolates a target character string, while the second designates a new string to replace the first. In order to remove only those tags mentioned earlier, while concurrently rearranging the data, two more regular expressions were necessary – tab-stops and back-referencing. Tab-stops, represented in code as \t, serve as column separators. Back-referencing allows one to refer back to previous elements in the code (in this case, elements in the first argument of *gsub*, specially demarcated by parentheses). The numerals indicate which parenthetical character string is being referenced, starting from the left with \\1. Therefore, by placing the V_1 lexeme and the regular expression designating any final non-finite V_2 inside of parentheses (without their tags), we allow for them to be referenced by \\1, and \\2, respectively, while excluding their tags and *to*. This is not enough, however, because the issue of the tri-columnar, as opposed to the preferred quad-columnar formatting still remains. Therefore, tab-stops were inserted around each of the back-references. An example of the code used to clean-up and arrange the expression *come to* V_2 is provided below in Example 49:

(49) gsub("(\\bcome)_VB to_TO (\\w+?)_VB\\b", "\t\\1\t\\2\t"…)

Table 4 breaks down the two primary *gsub* arguments using the terms introduced earlier. The back-references appear beneath their antecedent expressions and the tab stops are situated according to their place of insertion.

Table 4: Structure of the gsub function and outcome of its application to come to V_2

Target Code	come_VB to_TO (\\w+?_VB)_VB				
Argument 1	(\\bcome)_VB		to_TO	(\\w+?)_VB	
Argument 2	\t	\\1	\t	\\2	\t
Resulting Code	\t\\bcome\t\\w+\t				

In saving this output to text file (for easy conversion to Excel), the appropriate column names were appended to the original file in a header line.

3.3.3 Microsoft Excel

The spreadsheet software Excel was used in the final stages of data retrieval – the weeding-out of false hits and the systematic coding of entries. As a working knowledge of this program is assumed, a preliminary description of its functions will not be offered. It should be noted, however, that tab stops are realized in Excel as columnar breaks. Thus, the table generated by the above process consisted of 4 columns (again, PRECEDING, V_1, V_2, and FOLLOWING).

3.3.3.1 Weeding-out False Hits

The reliability of the Biber annotation, as mentioned previously, has not been hand-checked in this release of the ANC. Therefore, it was necessary to identify and exclude any false hits. One of the most frequent issues was the tagging of nominal locatives as _VB following *to* (especially present in the overall V_1 *to* V_2 table), for example, go_VB to_TO school_VB/work_VB. Another typical issue was the return of idiomatic or phrasal constructions, also including mis-tagged nominals, as in came_VBD to_TO power_VB. However, inaccurate annotation was not the only issue; there were also instances of purposive V_2's (he_PRP came_VBD to_TO see_VB her_PPN), which, though formally identical to the target COME *to* V_2, are subject to different semantics (i.e. *come* in these instances is the lexical-semantically full, prototypical intransitive deictic of physical motion towards a reference point). In the case of GET, some instances reflected the sense of 'become' associated with the phrasal construction <GET *to be* X>, as in *They're getting to be real stinkers*. The entirety of the results was manually investigated, and each false hit was coded as such and set aside for possible future inquiry. Out of 975 total instances of COME *to* V_2, 478 were deemed true hits; out of 1,126 total instances of GET *to* V_2, 597 were deemed true hits; out of 82,428 total instances of V_1 *to* V_2, 68,983 were deemed true hits.

3.3.3.2 Coding

A new table was generated in Excel to handle the parsing of morpho-syntactic and semantic elements of the constructions. This table comprised 22 columns covering 19 variables, as shown in Table 5:

Table 5: Coded variables and their contents

#	Column Header	Description of Contents
1	R1	clause-initial adverbial material
2	A1	grammatical subject NP
3	R2	post-subject/pre-verb adverbial material
4	Negation?	presence or absence of negation, coded for type – not, never, etc
5	V1 LEMMA	lemmatized form of V_1 verb
6	V2 LEMMA	lemmatized form of V_2 verb
7	A2/3	any remaining arguments
8	R3	post-verbal adverbial material
9	TENSE	tense of V_1
10	V1 ASPECT	aspectual morphology of V_1
11	V2 ASPECT	aspectual morphology of V_2
12	MOOD	modality type (e.g., indicative)
13	V1 VOICE	voice-related morphology of V_1

14	V2 VOICE	voice-related morphology of V_2
15	PERSON	personal agreement of V_1
16	NUMBER	numerical agreement of V_1
17	COMP	complementation type of construction, if applicable (e.g., subordinate clause)
18	MODAL	presence or absence of modal verb, coded with appropriate lexeme
19	INF	whether V_1 is itself non-finite

The first eight of these were ordered as an iconic representation of the entries, the left-most column containing the left-most element in the data file. The remaining three non-variable columns were reserved for organizational purposes: NUMBER (within the coding table -- continuous), ORIGINAL NUMBER (from the weeding-out table – discontinuous), and FILE (the original ANC file designation). Each of the true hits was manually coded into this table according to the relevant criteria. An entry of "None" was provided for non-applicable variables. All lexical items were entered without their annotation.

3.4 Statistics

A two-fold analysis was conducted, with one approach focusing on the semantic parameters and selectional preferences of V_2 relative to V_1, the other exploring the main effect interactions of the morpho-syntactic variables in determining the target construction relative to the V_1 *to* V_2 sample. The former involved a collostructional analysis, the latter a logistic regression.

3.4.1 Collostructions

The term "collostruction," as it appears in, for instance, Stefanowitsch and Gries (2003), refers to the combination of two opposite approaches to the investigation of a construction – namely, the top-down variety, which identifies those lexical items exhibiting significant distributional preference for filling a particular slot in a particular target construction (collexemes) and the bottom-up variety, which reveals those constructions strongly associated with a particular lexical item (collostructs). The ultimate goal of such a combination is to account for a single relationship – that between a schematic construction and a lexical item contained therein – as accurately as possible by relativizing their level of attractedness to one another given the possibility that one may have a high frequency of occurrences out of all instances of the other but be proportionally more prevalent elsewhere in the corpus (that is, a lexeme might make up a huge percentage of all instances of a slot in one construction, but, due to high overall frequency within the corpus, it may exhibit stronger ties with other constructions, thus producing asymmetric values for mere collexeme- and collostruct-oriented analyses; Stefanowitsch and Gries, 2003)[23].

3.4.1.1 Fisher Exact Test

Turning away from frequency-based expressions of attractedness, Stefanowitsch and Gries (2003) suggest the Fisher exact test as the statistical model for collostructional analysis. The advantages of this model are that it makes no distributional assumptions (these are rather specified in the tabulation, solving the above problem of asymmetries between collexeme and

---

[23] Certain criticisms have been raised against Collostructional analysis, especially by proponents of frequency-based interpretations (cf. Bybee, 2010). It has been argued, for instance, that CA is ineffective due to its alleged lack of semantic input (i.e. it only looks at comparative frequencies of occurrence without addressing funtional,high frequency semantic cores in relation to which the other attested lexemes are radially proximal/peripheral). However, these claims fail to recognize that CA simply intends to determine *what* items appear more or less frequently than expected in a variable slot in a Cx. Further judgments can then be made on the part of the researcher, e.g., classifying the data into semantic classes. The advantage lies in the attested strength of the CA to eliminate distributional skewing effects among high frequency lexemes while distilling overwhelming type counts (the majority exhibiting low token counts, following Zipf's law) into a concentrated set of elements occurring more frequently than chance would predict. Therefore, any attempt at classification of the results of CA into semantic classes will be speculative but crucially applied to more reliable results than approaches based merely on frequency can provide. Furthermore, as Bybee (2010) counts high frequency elements as being prototypic, a speculative interpretation of CA could take the greatest collostructional strength to indicate prototype status. Indeed, this study's collostructional strengths of V_2 do reflect both constructionally feasible (i.e. conformable to the higher-level constraints of the construction) and highly frequent lexemes. For these reasons, the Collostructional analysis and its results will be considered here accurate and relaiable.

collostruct strength) and that it has no difficulty handling small sampling sizes and low-frequency instances (which, as stated in Zipf's Law, make up the majority of collexemes occurring in a given construction) (p. 218). This test does, however, typically consist of thousands of component calculations. Luckily, this burden is diminished significantly by the use of Gries' collostructional analysis script for *R*.

In the collostructional analysis, the Fisher exact test is applied to a four-by-four table containing the frequency of a lexeme L in construction C, not-L in C, L in not-C, and not-L in not-C, along with the totals for L, not-L, C, and not-C. This structure is reflected in Table 6.

Table 6: Template for tabulation in Collostructional analysis

	Target Collexeme (L)	Corpus (V2) - Target Collexeme (¬L)	Total
Target Collostruct (C)	L in C	¬L in C	(L in C) + (¬L in C)
Corpus (V1 *to* V2) - Target Collostruct (¬C)	L in ¬C	¬L in ¬C	(L in ¬C) + (¬L in ¬C)
Total	(L in C) + (L in ¬C)	(¬L in C) + (¬L in ¬C)	Corpus [V_1 *to* V_2]

The Fisher test, then, computes the probability that a specified slot in a specified construction will be filled by a given collexeme, relative to its frequency and the frequency of the construction, based on a comparison against the frequencies of all lexemes and constructions in the sample.

For this study, it was necessary to conduct two collostructional analyses as both the COME and GET constructions needed to be compared individually and separately to the fully schematic statistical sample (V_1 *to* V_2). For each of these tabulations, the V_2 lemma frequencies were collected from the fully parsed tables for COME, GET, and V_1 *to* V_2 using the pivot table function in Excel. The totals for the constructions were calculated by subtracting the total frequency of COME and GET from V_1 *to* V_2, respectively (once for each table). Example tabulations for the collostructional strength of GET *to see* and COME *to see* are shown in Table 7.

Table 7: Sample tabulations of COME *to see* and GET *to see*

	see	*¬see*	Totals
COME *to* V_2	23	455	478
¬COME *to* V_2	1115	67390	68505
Totals	1138	67845	68983

	see	*¬see*	Totals
GET *to* V_2	114	483	597
¬GET *to* V_2	1024	67362	68386
Totals	1138	67845	68983

As shown in the Table 7, the total sample size is equivalent to all instances of V_1 *to* V_2, including the COME and GET constructions, rather than to the entirety of the ANC. Due to the relatively low frequencies of the target constructions with respect to the enormous 22-million-word corpus, an ultra-fine-grained approach was selected. In order to increase the granularity of focus, the domain of distribution was restricted to a structurally homogenous sample. This way, the results will indicate specifically the effect of the presence of COME or GET (as opposed to other available collexemes) on the behavior of the broader collostruct.

3.4.1.2 Collostructional Analysis 3.2 (R-script)

Rather than calculating all collostructional strengths manually, this study used the application Collostructional Analysis 3.2 (Gries, 2007).

3.4.1.2.1 Data Preparation

First, the data needed to be prepared in the appropriate format to be read properly into the application. Accordingly, two new text files were created. This step included generating frequency lists for all instances of any V_2 lemma in the target constructions as well as in the V_1 to V_2 data set. Within the input text files, an alphabetically-sorted list was created of all V_2 lemmas[24]. Next to this were placed two columns: one for the frequencies of each verb in the overall sample and another for those in the relevant construction, left to right. Whenever a gap appeared (i.e. a zero-instance of a given lexeme), a blank was inserted. An excerpted sample of the input table can be found in Appendix A.

3.4.1.2.2 Execution

This script opens an interface that allows one to select the type and parameters of a small group of analyses (including, of course, the collostructional analysis). Having successfully arranged the input text files, the parameters were specified as follows: corpus size = 68983 (frequency of V1 to V2 in ANC); frequency of COME = 478/GET = 597; index of association strength = 2 (log-likelihood); sorting = 4 (values will be sorted by collostructional strength); and decimals = 6 (answers will contain up to six decimal places).

The results are automatically output into a text file, which contains the following information: frequency of the words in the corpus, observed frequency of the words within target construction, expected frequency of the words within the target construction, percentages of how many instances of the word occur within the target construction, the relation of the word to the target construction (attraction or repulsion), how much the word helps to guess the construction, how much the construction helps to guess the word, and the

[24] This list was simply extracted from the V_1 *to* V_2 parsed table, as it encompasses both COME and GET.

collostructional strength as represented in a p_{log}-value (significance>1.30103).

3.4.2 Logistic Regression

Binary logistic regression tests for the probability of either realization of a two-level dependent variable given various predictor variables (and possibly, their interactions). More simply, the dependent variable represents two possibilities, X and Y. The independent variables represent the (coded-for) properties attending each instance of X and Y. The logistic model determines the significance of the variables in predicting the outcome of the target realization of the dependent variable (i.e. that level which is of interest to the researcher). Once these variables are tested for significance in the logistic model, the results are submitted to a series of iterative comparative statistics (in this case, ANOVAs), after each of which the least significant determinant is removed and a new model generated. The final outcome of this process is known as the minimal adequate model, which contains only the significant main effects[25] (cf. Gries & Wulff, under revision). Following Wulff et al. (to appear), this process was carried out in *R*. The function *glm* was used to calculate the logistic models, ANOVA(model.glm, type="III", test.statistic="Wald") for comparisons, and lrm(formula = formula(model.glm), x = T, y = T, linear.predictors = T) to test the strength and accuracy of the minimal adequate model.

In this study, the dependent variable consisted of a *Yes* and *No* level. *Yes* indicated the presence of the target construction; *No* indicated the presence of any other V_1 *to* V_2 construction. The predictor variables and their sub-levels are listed in Table 8:

[25] That is, the minimal adequate model will only contain significant main effects without the inclusion of interactions. If interactions are included, non-significant main effects are retained if they participate in any significant interactions.

Table 8: Predictors and levels for logistic regression

#	Predictor	Variable Levels
1	Tense	None, Future, Past, Present
2	AspectV1	Simple, Progressive, Perfect
3	AspectV2	Simple, Progressive, Perfect
4	Mood	None, Indicative, ConditionalSub
5	VoiceV1	Active, Passive
6	VoiceV2	Active, Passive
7	Person	None, First, Second, Third
8	Number	None, Plural, Singular
9	COMP	None, RelativeCl, SubordConjunction, ThatCl, IndirQuestion
10	MODAL	None, Modal
11	INF	None, Inf
12	Participle	None, ING, EN/ED

3.4.2.1 Data Preparation

Two new text files were created to accommodate the input requirements of the *glm* function, one for COME and the other for GET. The V_1 *to* V_2 coding for the above-mentioned variables was appended to that of each target collostruct, and a new column was generated for the dependent variable (containing the levels *Yes* and *No*, as mentioned earlier).

3.4.2.2 COME to V_2

An initial modeling of the COME data set yielded an "aliased" variable for Number, meaning that one of the levels of this variable was not computable with any reliability as contributing to the determination of the dependent variable. As such, it was removed from the *glm* function and a new model was created. Altogether, seven iterations were performed before the minimal adequate model was reached.

3.4.2.3 GET to V_2

The initial modeling of GET *to* V_2 produced two aliased variables, one for Number and the other for Person. After removing these and re-computing the model, six iterations were performed to reach the minimal adequate model.

CHAPTER 4

RESULTS

4.1 Collostructional Analysis

The following results come from the application of the collostructional analysis (Gries, 2007), which measures the level of attractedness of V_2 lemmas to the target collostructs (COME *to* V_2 and GET *to* V_2). As such, the values offered here reflect broadly the significance of correlation of semantic and syntactic features in the collexeme/collostruct interface. The collostructional strength values reported below are interpreted for significance according to the following criteria: p_{log}-value > 3 = p-value < .001 (extremely significant); p_{log} -value > 2 = p-value < .01 (highly significant); p_{log} -value > 1.30103 = p-value < .05 (significant). Note that, as this analysis employs a logarithmic adjustment of infinitesimal p-values, the figures representing collostructional strength are all necessarily positive (for both those lexemes attracted and those repulsed). For this reason, the output of Collostructional Analysis 3.2 (Gries, 2007) includes an extra column indicating repulsion or attraction (see Appendices B and C).

4.1.1 COME to V_2

Table 9 shows the V_2 lemma and the p_{log}-value indicating collostructional strength for the lemmas attracted to the target collostruct. As mentioned previously, only lemmas with an observed frequency of 4 or greater in the target construction were included. Furthermore, this table only reflects results of significance p<.05 or greater. In this sample, only one V_2 lemma of frequency 4 or higher was significantly repulsed from the collostruct: BE (p_{log}=52.47; shown under the bolded line). The full table of results can be found in Appendix B.

Table 9: Significant collexemes of COME *to* V_2

V_2 Lemma	P_{log}-value	V_2 Lemma	P_{log}-value
mean	187.38	love	24.67
pass	166.44	view	21.19
know	140.63	regard	20.93
realize	130.92	recognize	20.71
expect	119.93	exist	20.18
call	96.2	see	19.71
believe	69.03	rely (*on*)	19.62
dominate	66.2	accept	19.25
seem	40.09	refer (*to*)	11.07
appreciate	35.497	include	9.17
regret	34.82	stand (*for*)	9.01
understand	30.06	represent	7.895
rest	25.21	think	6.93
be	52.47		

The values for MEAN, PASS, KNOW, REALIZE, and EXPECT demonstrate an extreme level of attractedness between these collexemes and COME *to* V_2. Moreover, all of the values in this table are of a *p*-value<.001 and therefore very highly attracted.

A closer inspection of the lexemes themselves yields evidence of emerging semantic classes – that is, evidence that this is not a random assortment of verbs, but rather a structured set of elements. These categories are defined and their membership laid out in Table 10 and the following six sections[26].

[26] These classes (as well as those outlined below for GET *to* V_2) are categorized according to the intuitions of the researcher and in no way reflect any output of the Collostructional analysis.

Table 10: Semantic classes of V_2 lexemes in COME *to* V_2

Class	**Members**
Valuation	view, regard, see
Emotion	appreciate, love, regret, accept
epistemic/factive	know, realize, understand, recognize, think, believe, respect, recognize
Existential	pass, exist, (be)
Designation	call, refer (to), mean, seem, represent, stand (for)
organization/ hierarchy	dominate, rely (on), include

4.1.1.1 Class 1: Valuatives

This class is composed of VIEW, REGARD, and SEE as realized in the valuative construction <X VERB$_{valuative}$ Y *as* Z>. These all express confident and developed opinions or perceptions about someone or something (e.g., *He has come to view aid and charity as industry*)

4.1.1.2 Class 2: Emotionals

This class is composed of APPRECIATE, REGRET, LOVE, and ACCEPT, which all appear both as complement taking predicates (<X VERB$_{Emotion}$ *that* Y>) and monotransitives (<X VERB$_{Emotion}$ Y>). These express dynamic judgments, the assertion of which usually entails a process of contact or experience with their objects (e.g., *I came to appreciate her*).

4.1.1.3 Class 3: Epistemics/factives

This class is composed of KNOW, REALIZE, UNDERSTAND, and RECOGNIZE. Of these, KNOW is of interest as it appears in both complement-taking and monotransitive constructions. All share the quality of being factive, or implying the truth of the complement proposition regardless of negation of the main proposition. A more peripheral subset of this category includes THINK and BELIEVE, which are only factive in a weak sense. An even more peripheral member is EXPECT, which seems to behave formally as the other epistemics, but which connotes a habituality left somewhat obscure in the other members (except for, perhaps, RECOGNIZE) (e.g., *I came to realize why leopard is the most desirable of all the fur prints)*.

4.1.1.4 Class 4: Existentials

This class has two central members, PASS and EXIST, as well as the strongly repulsed BE. Of the first two, PASS is a gradual dynamic existential which emphasizes the activity, while EXIST is a concrete static existential which emphasizes the accomplishment of a developmental process. There are, in fact, instances of BE which mirror those of EXIST (e.g., *How does a specification of direction come to be?*); however, the predominance of its instances being tied to passive V_2 structures in COME *to* V_2 and its disproportionately high frequency in V_1 *to* V_2 probably account for its repulsion. A third attracted member is REST, which refers more narrowly to a specific stative quality of an object (namely, that of the terminus of a gradual cessation of motion) (e.g., *Flying squirrels came to exist in the world*).

4.1.1.5 Class 5: Designators

This class is actually split according to two types of conceptual focus – CALL and REFER (*to*) focus on the entity doing the designating, while MEAN, SEEM, REPRESENT and STAND (*for)* focus on the thing being designated (e.g., *Feminism came to mean anti-pornography statutes*).

4.1.1.6 Class 6: Organizationals/hierarchicals

This class contains DOMINATE, RELY, and INCLUDE. All refer to hierarchical relationships which are defined rather inflexibly, and which are typically not instantaneous in their inception (e.g., *This self came to dominate our reception of input*).

4.1.2 GET to V_2

Table 11 shows all and only the significant V_2 collexemes of the GET *to* V_2 construction. Those above the bolded line are attracted to this construction, those beneath it repulsed (full results can be found in Appendix C).

Table 11: Significant collexemes of GET to V_2

V_2 Lemma	**P_{log}-value**	**V_2 Lemma**	**P_{log}-value**
see	379.8	pick	12.87
play	79.94	act	12.47
meet	64.8	travel	9.14
watch	53.4	read	8.79
choose	40.6	talk	7.02
hang (*out*)	39.97	do	6.76
spend	35.94	sleep	6.57
hear	16.52	decide	6.36
go	15.36	keep	4.76
dance	14.65	eat	4.61
live	14.07	write	1.94
be	63.2	make	3.16
get	10.24	have	3.01

The collostructional strength of SEE is exceptional, and all verbs other than WRITE (even those exhibiting repulsion) are extremely significant (p-value<.001).

As with the results in Table 9, these comprise a series of cohesive semantic sub-groups[27], reflected here in Table 12 and the following six sections.

[27] Note that these classes do not overlap with those of COME *to* V_2, indicating that, despite their formal similarity, these constructions are cognitively distinct.

Table 12: Semantic classes of V_2 lexemes in GET to V_2

Class	**Members**
Perception	see, watch, hear
Socializing	meet, hang (out), spend (time with), talk (to)
Selection	choose, decide, pick
Recreation	play, dance, travel, read
vital processes	sleep, eat, live
"bleached"	go, do, keep,(be, get, make, have)

4.1.2.1 Class 1: Perception/interaction

This class contains SEE, WATCH, and HEAR. The situations expressed by these propositions conform to the broad-level unconventionality constraint in that the events or entities being perceived are rarely encountered first-hand by most people (i.e. special concerts, presidential addresses, celebrity panel discussions, etc.). Formally, they are all monotransitive and literally and simply perception-based (that is, neither SEE nor HEAR appears as a complement taking predicate in any of its instantiations) (e.g., *We don't get to see the fun part*).

4.1.2.2 Class 2: Social

This class contains MEET, HANG (*out*), SPEND (*time with*), and TALK (*to*). Similar to Class 1, these verbs refer to special social events of an extraordinary character (e.g., *We'll probably get to meet Serena Altshul*).

4.1.2.3 Class 3: Selection

This class contains CHOOSE, DECIDE, and PICK. These are all characterized by authority based on group consensus, such as law-based claims about the powers of judiciaries, or a superior's assent/assignation, such as decisions made in work-place projects (e.g., *A man gets to choose a potential bride from among two dozen women*) .

4.1.2.4 Class 4: Recreation

This class contains PLAY, DANCE, TRAVEL, and READ, which again emphasize actions not easily engaged in every day. They therefore focus on the satisfying of an ongoing desire normally hindered by quotidian concerns and responsibilities (e.g., *Maybe I'll get to travel this year*).

4.1.2.5 Class 5: Vital Processes

This class contains SLEEP, EAT, and LIVE. In line with the broader trends presented in the other classes, these vital processes are highlighted as unconventional or licensed by an exterior, super-ordinate entity or force (e.g., [*Pets get*] *to sleep on the bed for warmth*).

4.1.2.6 Class 6: "Bleached"

This class contains GO, DO, and KEEP (attracted) and BE, GET, MAKE, and HAVE (repulsed). The term "bleached" is meant to reflect the diverse functions of these verbs, all of which participate in special syntactic structures (auxiliaries, phrasal constructions, etc.) aside from their centric, lexical use as main verbs. The interesting thing to note here is that the majority of these are repulsed, and strongly so, indicating perhaps a semantic discontinuity between the GET to V_2 collostruct and lexically bland elements (e.g., *I don't get to go skiing*).

4.2 Logistic Regression

The following results reflect the application of a binary logistic regression as outlined in Section 3.4.2. As mentioned previously, this test uncovers those morpho-syntactic variables contributing significantly to the determination of the target constructions.

4.2.1 COME *to* V_2

The minimum adequate model (reflected in Table 13) exhibits a highly significant and strong correlation of the predictor variables and the selection of the target construction (log–likelihood ratio $\chi 2$= 453.66; df =12; p = 0). The accuracy of the minimum adequate model in predicting the choice of COME *to* V_2 given various arrangements of the predictor variables was also good (R^2=.369; C=.814; Dxy=.629).[28] Negative coefficient values indicate the independent variable's repulsion of the target construction; positive values indicate attraction (the full results can be found in Appendix D).

Table 13: Significant predictor variables for COME *to* V_2

Predictors	**Coefficient**	***p*-value**
TENSE=None	-3.8883	<.00001
TENSE=Past	-0.2685	<.00001
AspectV1=PerfectProgressive	-9.0227	<.00001
AspectV1=Progressive	-2.4654	<.00001
AspectV1=Simple	-2.7332	<.00001
VoiceV2=Passive	1.7198	<.00001
COMP=None	-1.7695	<.00001
COMP=RelativeCl	-1.2531	.0005
COMP=SubordConjunction	-1.8117	<.00001
COMP=ThatCl	-1.5322	<.00001
INF=None	-2.5194	.0008

The first striking aspect of the results in Table 13 is the clustering of strong aspectual repulsions, which leaves one level (AspectV1=Perfect) as the only statistically felicitous formulation. Combining this feature with the values of negative correlation for TENSE=None (-3.8883) and TENSE=Past (-0.2685), we find a tendency for COME to prefer present perfect

[28] R^2 is considered stronger as it approximates 1; C-values are generally considered strong > .8.

construction (e.g., *He has come to know her*).[29]

The second aspect of interest is the clustering of strong repulsions in the variable COMP (complementation type of the construction). In a fashion similar to that of AspectV1, only one possible variable level remains: COMP=IndirQuestion (e.g., *I know how he came to occupy that post*). We can also see that COME *to* V_2 is by default attracted to non-lexical complementation strategies (INF=None, -2.5194).

Table 13 also illustrates a possible explanation for the singularly (and quite intensely) repulsed V_2 lemma, BE (collostructional strength = 52.47). The highly significant attractedness of COME to passivized V_2s (coef.= 1.7198; *p*-value<.00001), which also require BE (in an auxiliary rather than copular capacity), is a competing and superior constraint on the overall construction. Thus, in order to facilitate processing, BE will occur only within the passive construction, itself conformable to the construction's higher level event-construal type.

4.2.2 GET to V_2

The minimal adequate model for GET *to* V_2 (reflected in Table 14) exhibits a highly significant and moderately strong correlation between the predictor variables and the selection of the target construction (log–likelihood ratio $\chi 2$= 182.25; df =11; p = 0). The predictive accuracy of the model is near-optimal (R^2= .147; C=.664, Dxy=.329). The full results can be found in Appendix E.

Table 14: Significant predictor variables for GET to V_2

Predictors	Coefficient	*p*-value
TENSE=Past	- 0.8501	0.0002
TENSE=Present	-0.9975	<0.00001

[29] There were no attestations of TENSE=Future/AspectV1=Perfect in the data set.

AspectV1=Progressive	1.8973	0.0005
AspectV1=Simple	3.4303	<0.00001
Mood=Indicative	0.6673	0.0009
VoiveV2=Passive	-1.5495	0.0005
INF=None	1.6836	0.0022

As we can see in Table 14, the GET construction prefers to occur with progressive or simple aspect marking, in the indicative mood, and in full clauses (i.e. GET is not attracted to reduced infinitival clauses such as *to get to go*[30]). Taken alone, these values do not serve adequately to differentiate this construction from <X GET *to be* Y> (meaning 'become'). However, a comparative inspection between these and the collostructional results (Table 11) reveals an alternative mode of disambiguation – the repulsion of BE as V_2. Therefore, GET *to* V_2 will retain its core constructional meaning given that BE does not instantiate V_2, the occurrence of which would most likely hinder comprehension by activating two competing constructions simultaneously. Further supporting this mode of disambiguation is the negative correlation of GET *to* V_2 and a passivized V_2, which would naturally require the presence of BE as an auxiliary.

[30] This might be due to the principle of *horror aequi*, which states that immediate clustering of the same or similar structures is disprefered by language users (cf. Rohdenburg, 2003).

CHAPTER 5

DISCUSSION

5.1 Theoretical Findings

The results supported the claim that GET/COME *to* V_2 need to be considered discrete constructions, not only in relation to their closest, fully schematic structural analogue (i.e. V_1 *to* V_2), but also in relation to patterns of complementation in general, whose minimally "dipropositional" character (cf. Dixon, 2006) is not intuitively nor empirically discernible in the target constructions. Regarding COME, the collostructional analysis yielded not a random distribution of V_2 lemmas, but rather a highly structured set of semantic classes. Taken as a whole, these classes exhibit commonalities conformable to the presupposed lexical-semantic contribution of COME, as well as its broader constructional semantics.

Among the qualities shared by the V_2 classes (see Tables 10 and 12) that seem explicitly tied to the lexical semantics of COME, we find that the participants of the event must undergo a change in state (a metaphorical extension of COME's deictic motion involving change of location). "Change of state" is here meant to capture all together valuative judgments (from "undecided" to "decided"; e.g., *come to regard as*), emotional vicissitudes (e.g., *come to love/hate*), epistemic transitions (from "not knowing" to "knowing", e.g., *come to realize*), existential realization (*come to exist*), changes in or designation of names (e.g., *come to be called*), and changes in hierarchical orientation/situatedness (e.g., *come to dominate/rely on*).

Furthermore, the event expressed by V_2 matches the construal strategy of COME's intransitive deictic motion construction; namely, the event necessarily focuses the end state (grounded temporally through an inheritance of the tense placed, perhaps by analogy, on

COME) with only optional profiling of the source (e.g., the use of *from* to indicate a source location – deictic motion – or state – the target usage.

Finally, COME undergoes another, related metaphorical extension, in which a normally "physical" path is mapped onto a temporal one. Thus, the end-state is always construed as following the implicit source-state (past → future orientation), and through tendencies identified in the logistic model, the end-state is nearer the speaker in time than the onset of its becoming (as attested in the construction's preference for present perfect morphology).

Among the construction-level constraints on V_2 (i.e. those unable to be predicted by a summing of the purely lexical semantics of each word), we find that the end-state must be amenable to a developmental (iterative or gradient) process of transition. A corollary constraint is that the process is unmarkedly slow. This accounts for infelicitious V_2 instantiations such as **The ghost came to vanish*, in which the V_2 carries a semelfactive aktionsart – this change of state implies no intervening experience between the source and end states (here, the perception of presence and absence of the ghost). COME *to* V_2 does allow some verbs traditionally interpreted as immediate, such as a large portion of Class 3 (4.1.1.3; e.g., *realize*); however, in this construction, the end-state refers to the culmination of an extensive series of compatible and corroborating (e.g., similarly oriented) experiences, such that the most natural interpretation of *come to realize* is not one involving an unmotivated, unexpected flash of insight, but rather a more or less seriously considered examination of experiential evidence leading ultimately to a well-reasoned conclusion.

Viewed from another perspective, this construction-level semantic constraint also accounts for the repulsion of progressive morphology evidenced in the logistic model (see section 3.2.1). Progressive aspect typically marks the action as unfinished, in process, and thus conceptually focuses the immediate and on-going sub-processes of an action (i.e. chewing, swallowing, etc. in *is eating*). However, COME *to* V_2 has a tendency to obscure the antecedent events into a vectored, linear blur (a consequence of the focusing of the end-state), thereby clashing with the progressive aspect's principal function.

Much like the results of the collostructional analysis, those of the logistic regression, given the predictive strength of the model and high significance of correlation between predictor variables and choice of COME *to* V_2 (see section 3.2.1), represent a highly structured, internally coherent system. Furthermore, this system is integrated and commensurable with both the higher-level constructional semantics and the selectional restrictions of the V_2 slot captured in the collostructional analysis. We have seen the correlation of morpho-syntax with (1) lexis (disambiguation of passive and copular BE V_2 through the repulsion of the latter; above, section 3.2.1), (2) prototypical semantic features of COME (the preference of the present perfect in relation to deictic motion towards a referent space or goal; above, this section), and (3) higher-level, non-compositional semantic features of the construction (the repulsion of progressive aspect due to obscuring of antecedent events; above, this section). For these reasons, COME *to* V_2 must be considered a construction[31], but more than that, it must be recognized as functionally discrete from other types of complementation, as evidenced by its

[31] Keep in mind that constructions are characterized overall by a pairing of form-meaning, but that in partially filled constructions, as evidenced here and elsewhere, e.g., Wulff (2008), the "filled" unit imposes, according to salient aspects of its prototypical lexical semantics, strong restrictions on the acceptability of possible lexico-syntactic and morpho-syntactic choices. Put another way, "filled" slots in otherwise schematic constructions, as fully realized constructions, may very well contribute more substantially to the meaning of the utterance as a whole than variable instantiations. The latter, though they typically (and in this research) comprise structured classes, do not so much govern their meaning relative to broader construction as they are governed by it.

significant divergence from the random sample.

A similar set of arguments can be made for GET *to* V_2, although it remains unclear whether or to what extent this construction is related to COME *to* V_2 (e.g., whether they are subordinate to the same argument structure constructions, composed of similar sub-constructions, etc.). First of all, the collostructional analysis revealed a structured set of classes of verbs occupying the V_2 slot. Just as with COME, these classes are all in some way compatible with both the prototypical lexical semantics of GET and the higher-level semantics of the construction as a whole. In this construction, the most basic transitive meaning of GET, akin to *receive*, selects for and blends with V_2s that profile the same kind of participant – namely, a volitional or rational subject. This requirement is attested in all of the classes outlined in section 4.1.2 (see the examples in each subsection). This is not only evidence for the necessary and essential relatedness of the lexically specified material in a partially-substantive construction (here, GET) to the overall construction, but also for the level of event-integration in this construction. The two verbal elements, GET and V_2, are so fully converged that they share a subject, both in an abstract-schematic (identical participant constraints) and empirical (the subject of GET is indeed always a participant in the V_2 event) sense.[32,33] In terms of the higher-level constructional semantics, both GET and the semantic classes attested in V_2 contribute to the overarching sense of the subject's achieving the event represented by V_2 despite hindrances of social convention or opportunity (consider, e.g., *Dennis got to hang out with Cordy while she was bathing*). Again, the prototypical lexical meaning of GET is applied in

[32] In fact, the inclusion of a V_2 subject other than that of GET (e.g., *He got John to realize his mistake*) enacts a separate construction – the causative.

[33] This goes against the purely syntactic approach (see section 1.1.2.2), which argues for coreference of subjects as opposed to recognizing the overt lack of an independent subject for V_2 as signifying its conceptual absence.

that the subject somehow *receives* the right or chance to engage in the nuclear activity.

Though the logistic model generated for the morpho-syntax of GET *to* V_2 was not as strong or accurate in its predictive force as that of COME *to* V_2, it did illustrate highly significant and structured interactions of the independent variables in the selection of GET. Perhaps the most interesting aspect of the results was the relationship between voice (V_2) and the collostructional repulsion of BE (see Tables 11, 14; section 4.2.2). These data produce a rather clear picture of the construction's internal structure relative to formally related constructions with, however, radically different semantics (i.e. <X GET *to be* Y>, meaning 'become'). We see here through the integration of syntactic and semantic constraints a previously unaccounted for mode of disambiguation between these two structures, which, moreover, fits nicely the contention made by Construction grammar that "polysemy" is a product of complex interaction between individual lexemes and broader-level schematic constructions (cf. Goldberg, 1995). In fact, these results take this claim one step further in accounting for "polysemous" verb senses in structures of varying idiomaticity.[34]

5.2 Methodological Findings

This research served essentially two functions in furthering corpus methodology. In the first place, it provides further empirical support to methods attested elsewhere (Stefanowitsch and Gries, 2003; Gries and Stefanowitsch, 2004; Wulff et al, *to appear*). The feasibility of the results in the face of native-speaker intuitions of the constructions under investigation is perhaps only mildly interesting (see e.g., Gries, 2003 for an illustration of the inadequacy of

[34] It is assumed here, though not investigated empirically, that the construction in which GET means roughly *become* contains more lexically specified material, i.e. BE *getting to be*, than does the more schematic GET *to* V_2, and therefore "reserves" these substantive slots as contributing to the generation of the broader constructional meaning.

native-speaker judgments), but the identification of features which disambiguate what are superficially the same sentence, yet easily differentiated by speakers, and that at the level of granularity afforded by the integration of collostructional analysis and logistic regression, indeed provides strong evidence for the validity of multi-faceted statistical manipulation of large-scale corpus data.

Secondly, this research bolsters the claims of cognitive linguistics generally, and construction grammar in particular, while further demonstrating their usefulness in the interpretation of corpus data. The approach to language which readily acknowledges the inextricable interrelatedness of form and meaning most easily accounts for data such as these and in fact facilitates the process of analysis. Especially in the case of logistic regression, a tool not designed with a construction-based approach to language in mind (as, indeed, the collostructional analysis was), we find clear traces of the multiple levels of the syntax-semantic interface – evidence from the one test points inevitably to evidence from the other. Therefore, the methodology employed here testifies to an emerging symbiosis between quantitative semantic and syntactic investigations of corpora and cognitive approaches to linguistics.

5.3 Limitations of Current Research

As mentioned earlier (section 3.2.2), one limitation of this research was the corpus itself. The ANC has not been hand-edited for accuracy, resulting in the possibility of the exclusion of pertinent data due to flawed tagging or spacing issues between words. Another issue with the ANC is its size. Were these methods applied to a larger corpus (say, the British National Corpus – over 100 million words), there would likely be a proportional increase in the frequency of COME/GET *to* V_2. This is useful not only in terms of its greater representativity of the target phenomena vis-à-vis more instances to be compared, but also with respect to the predictive accuracy of the logistic regression (which tends to over-predict the strength and significance of interactions between variables with sampling sizes < 1000; Gries, 2009).

Another issue at stake was the choice of including only those instances of the construction with absolute verbal continuity (i.e. those instances without any intercessory material dividing up the chunk V_1 *to* V_2). The solution to this issue is, however, somewhat unattractive, as it seems the only way to capture fully any length greater than one word of, say, an adverbial phrase separating any of the elements in the construction would be to search for V_1, code out any entries lacking a subsequent V_2, and manually investigate each entry for true hits. As the number of instances of GET/COME overall is much higher than the amount immediately abutting a "*to* V_2" structure, and as this process would have to be mirrored in the statistical sample (most likely an astronomical amount, perhaps in the hundreds of thousands), this is an unfeasible tactic.

Finally, the fact that this was an exploratory study impacted the clarity of the results. Parameterizing the statistical sample to include only V_1 *to* V_2 structures collapses a number of different phenomena into one, including structures of interest in direct comparison to the target structures (especially GET, for reasons outlined in section 5.1). Furthermore, a more detailed investigation might contrast COME/GET *to* V_2 against known instances of so-called RAISING and CONTROL V_1s to produce more attenuated distinctions from each class.

5.4 Directions for Future Research

Besides the suggestions offered in the preceding section, there are three distinct avenues of future endeavor which might proceed from the current research. Firstly, the arguments of the constructions, here identified only in COME/GET and even then ignored, could be coded along a series of parameters to capture more clearly any effects the construction might have in their selection. We have earlier alluded to some trends along these lines (section 5.1), but without a systematic inquiry yielding substantial, quantitative measures, these remain hypothetical descriptions.

Another, more divergent track would be a large scale comparative study of GET. This research has already demonstrated (one-sidedly) some commonalities and differences between two of these (<X GET *to* V_2> and <X GET *to be* Y>). A more thorough-going experiment would clarify these issues, providing new insight into both "polysemy" and idiomaticity effects horizontally (the combinatorics of fully specified material in the same construction, i.e. the effects that, for example, *getting* and *to be* have on each other when combined in an idiom chunk) and vertically (the effects of varying sizes of idiom chunks in constraining higher-level semantic features of the construction).

Finally, and perhaps most relevant to the current investigation, a diachronic study of the development of COME and GET might be performed to test theories like that laid out in Givón (1995), which claim that these types of structures are the result of a process of grammaticalization in which V_1s bleach semantically and become functional entities like TAM auxiliaries or clitics (see section 2.3). Such a study could determine why COME/GET cling so peculiarly to their lexical semantics and simultaneously project broader organizational and argument type constraints on the rest of the construction.

The success of this research provides an optimistic perspective on the convergence of empirical methodology and abstract-theory building, while concurrently illustrating the descriptive potential of multi-modal approaches to linguistic phenomena. The more closely we as researchers can approximate the multifactorial nature of cognitive processing in our collection and analysis of real, usage-based data, the more definite and internally cohesive our picture of language will become.

APPENDIX A

EXCERPTED SAMPLE OF COLLOSTRUCTIONAL INPUT FOR COME TO V2

WORD	WHOLE_CORPUS [V_1 *to* V_2]	COME
abandon_vb	26	2
accept_vb	108	7
adopt_vb	32	2
appear_vb	90	2
apply_vb	59	1
appreciate_vb	20	6
arrest_vb	11	1
assert_vb	13	1
assume_vb	34	1
be_vb	11588	28
bear_vb	24	1
believe_vb	160	18
bomb_vb	16	1
call_vb	173	23
care_vb	42	2
characterize_vb	30	1
comprise_vb	4	1
condemn_vb	3	1
consider_vb	87	1
control_vb	86	1
corrupt_vb	1	1
define_vb	77	3
demand_vb	12	1
depend_vb	21	1
describe_vb	108	1
designate_vb	12	1
despise_vb	5	1
disdain_vb	4	1
dissolve_vb	12	1

APPENDIX B

FULL RESULTS FOR COLLOSTRUCTIONAL ANALYSIS OF COME TO V_2

words	word.freq	obs.freq	exp.freq	relation	faith	coll.strength
mean_vb	100	31	0.692924	attraction	0.31	187.383596
pass_vb	120	30	0.831509	attraction	0.25	166.444392
know_vb	590	45	4.088254	attraction	0.076271	140.633847
realize_vb	62	21	0.429613	attraction	0.33871	130.917771
expect_vb	29	16	0.200948	attraction	0.551724	119.925217
call_vb	173	23	1.198759	attraction	0.132948	96.201997
believe_vb	160	18	1.108679	attraction	0.1125	69.029219
dominate_vb	25	10	0.173231	attraction	0.4	66.203013
seem_vb	40	8	0.27717	attraction	0.2	40.091292
appreciate_vb	20	6	0.138585	attraction	0.3	35.496871
regret_vb	5	4	0.034646	attraction	0.8	34.819138
understand_vb	172	11	1.19183	attraction	0.063953	30.055451
rest_vb	44	6	0.304887	attraction	0.136364	25.210128
love_vb	72	7	0.498906	attraction	0.097222	24.674701
view_vb	36	5	0.249453	attraction	0.138889	21.187101
regard_vb	18	4	0.124726	attraction	0.222222	20.933039
recognize_vb	64	6	0.443472	attraction	0.09375	20.711469
hope_vb	7	3	0.048505	attraction	0.428571	20.345332
exist_vb	67	6	0.464259	attraction	0.089552	20.176149
perceive_vb	2	2	0.013858	attraction	1	19.89634
retell_vb	2	2	0.013858	attraction	1	19.89634
see_vb	1138	23	7.885479	attraction	0.020211	19.710316
rely_vb	42	5	0.291028	attraction	0.119048	19.619547
accept_vb	108	7	0.748358	attraction	0.064815	19.252209
doubt_vb	12	3	0.083151	attraction	0.25	16.479108
refer_vb	61	4	0.422684	attraction	0.065574	11.067359
resemble_vb	32	3	0.221736	attraction	0.09375	10.339357
corrupt_vb	1	1	0.006929	attraction	1	9.946088
ignite_vb	1	1	0.006929	attraction	1	9.946088
mingle_vb	1	1	0.006929	attraction	1	9.946088
partition_vb	1	1	0.006929	attraction	1	9.946088
reassemble_vb	1	1	0.006929	attraction	1	9.946088
revere_vb	1	1	0.006929	attraction	1	9.946088
revolve_vb	1	1	0.006929	attraction	1	9.946088
soul_vb	1	1	0.006929	attraction	1	9.946088
include_vb	129	5	0.893872	attraction	0.03876	9.172774
stand_vb	81	4	0.561269	attraction	0.049383	9.007482
represent_vb	95	4	0.658278	attraction	0.042105	7.895337
question_vb	18	2	0.124726	attraction	0.111111	7.559991
think_vb	471	9	3.263674	attraction	0.019108	6.926745
enjoy_vb	67	3	0.464259	attraction	0.044776	6.235538
condemn_vb	3	1	0.020788	attraction	0.333333	6.154759

abandon_vb	26	2	0.18016	attraction	0.076923	6.126921
extract_vb	28	2	0.194019	attraction	0.071429	5.846632
own_vb	28	2	0.194019	attraction	0.071429	5.846632
define_vb	77	3	0.533552	attraction	0.038961	5.52125
comprise_vb	4	1	0.027717	attraction	0.25	5.489041
disdain_vb	4	1	0.027717	attraction	0.25	5.489041
savor_vb	4	1	0.027717	attraction	0.25	5.489041
adopt_vb	32	2	0.221736	attraction	0.0625	5.349227
despise_vb	5	1	0.034646	attraction	0.2	4.997577
eclipse_vb	5	1	0.034646	attraction	0.2	4.997577
haunt_vb	5	1	0.034646	attraction	0.2	4.997577
name_vb	38	2	0.263311	attraction	0.052632	4.72443
rival_vb	6	1	0.041575	attraction	0.166667	4.608745
care_vb	42	2	0.291028	attraction	0.047619	4.369148
encompass_vb	8	1	0.055434	attraction	0.125	4.014915
portray_vb	9	1	0.062363	attraction	0.111111	3.778139
trigger_vb	10	1	0.069292	attraction	0.1	3.569337
arrest_vb	11	1	0.076222	attraction	0.090909	3.382882
prefer_vb	11	1	0.076222	attraction	0.090909	3.382882
demand_vb	12	1	0.083151	attraction	0.083333	3.214692
designate_vb	12	1	0.083151	attraction	0.083333	3.214692
dissolve_vb	12	1	0.083151	attraction	0.083333	3.214692
signal_vb	12	1	0.083151	attraction	0.083333	3.214692
assert_vb	13	1	0.09008	attraction	0.076923	3.061712
occupy_vb	13	1	0.09008	attraction	0.076923	3.061712
suspect_vb	13	1	0.09008	attraction	0.076923	3.061712
feel_vb	134	3	0.928519	attraction	0.022388	2.935185
exert_vb	14	1	0.097009	attraction	0.071429	2.921594
donate_vb	15	1	0.103939	attraction	0.066667	2.792496
fear_vb	15	1	0.103939	attraction	0.066667	2.792496
bomb_vb	16	1	0.110868	attraction	0.0625	2.672944
embrace_vb	18	1	0.124726	attraction	0.055556	2.457909
exceed_vb	18	1	0.124726	attraction	0.055556	2.457909
seize_vb	18	1	0.124726	attraction	0.055556	2.457909
guide_vb	20	1	0.138585	attraction	0.05	2.269192
honor_vb	20	1	0.138585	attraction	0.05	2.269192
permit_vb	20	1	0.138585	attraction	0.05	2.269192
depend_vb	21	1	0.145514	attraction	0.047619	2.183032
paint_vb	21	1	0.145514	attraction	0.047619	2.183032
favor_vb	23	1	0.159373	attraction	0.043478	2.024581
possess_vb	23	1	0.159373	attraction	0.043478	2.024581
perform_vb	88	2	0.609773	attraction	0.022727	1.997122
bear_vb	24	1	0.166302	attraction	0.041667	1.95148

appear_vb	90	2	0.623632	attraction	0.022222	1.933938
forget_vb	26	1	0.18016	attraction	0.038462	1.815866
grant_vb	28	1	0.194019	attraction	0.035714	1.692599
rule_vb	28	1	0.194019	attraction	0.035714	1.692599
characterize_vb	30	1	0.207877	attraction	0.033333	1.579943
share_vb	103	2	0.713712	attraction	0.019417	1.568853
form_vb	108	2	0.748358	attraction	0.018519	1.446771
assume_vb	34	1	0.235594	attraction	0.029412	1.381147
distinguish_vb	34	1	0.235594	attraction	0.029412	1.381147
notice_vb	36	1	0.249453	attraction	0.027778	1.292933
learn_vb	122	2	0.845368	attraction	0.016393	1.149113
use_vb	450	5	3.11816	attraction	0.011111	0.973724
apply_vb	59	1	0.408825	attraction	0.016949	0.613307
wonder_vb	68	1	0.471189	attraction	0.014706	0.452113
reflect_vb	69	1	0.478118	attraction	0.014493	0.436591
sign_vb	69	1	0.478118	attraction	0.014493	0.436591
express_vb	73	1	0.505835	attraction	0.013699	0.378651
control_vb	86	1	0.595915	attraction	0.011628	0.229404
consider_vb	87	1	0.602844	attraction	0.011494	0.220043
end_vb	96	1	0.665207	attraction	0.010417	0.147141
look_vb	358	3	2.480669	attraction	0.00838	0.103172
describe_vb	108	1	0.748358	attraction	0.009259	0.077188
serve_vb	112	1	0.776075	attraction	0.008929	0.059719
speak_vb	112	1	0.776075	attraction	0.008929	0.059719
play_vb	397	3	2.75091	attraction	0.007557	0.022192
drive_vb	129	1	0.893872	attraction	0.007752	0.012241
be_vb	11588	28	80.296073	repulsion	0.002416	52.474168
get_vb	1620	1	11.225374	repulsion	0.000617	15.903197
find_vb	504	1	3.492339	repulsion	0.001984	2.509084
take_vb	906	4	6.277895	repulsion	0.004415	0.966688
live_vb	270	1	1.870896	repulsion	0.003704	0.493386
leave_vb	259	1	1.794674	repulsion	0.003861	0.423489
write_vb	212	1	1.469	repulsion	0.004717	0.170345
run_vb	196	1	1.358132	repulsion	0.005102	0.104973
spend_vb	193	1	1.337344	repulsion	0.005181	0.094151
turn_vb	174	1	1.205688	repulsion	0.005747	0.037609
support_vb	159	1	1.10175	repulsion	0.006289	0.009788

APPENDIX C

FULL RESULTS FOR COLLOSTRUCTIONAL ANALYSIS OF GET TO V_2

words	word.freq	obs.freq	exp.freq	relation	faith	coll.strength
see_vb	1138	114	9.8486	attraction	0.100176	379.802006
play_vb	397	30	3.43576	attraction	0.075567	79.941206
meet_vb	279	23	2.414551	attraction	0.082437	64.811112
watch_vb	231	19	1.999145	attraction	0.082251	53.352906
choose_vb	64	10	0.553876	attraction	0.15625	40.609346
hang_vb	48	9	0.415407	attraction	0.1875	39.9703
spend_vb	193	14	1.670281	attraction	0.072539	35.941956
spank_vb	2	2	0.017309	attraction	1	19.005443
hear_vb	232	10	2.007799	attraction	0.043103	16.516296
go_vb	1052	23	9.10433	attraction	0.021863	15.358791
stroll_vb	3	2	0.025963	attraction	0.666667	15.203684
dance_vb	31	4	0.268283	attraction	0.129032	14.648881
live_vb	270	10	2.336663	attraction	0.037037	14.072128
meddle_vb	4	2	0.034617	attraction	0.5	13.494918
pick_vb	104	6	0.900048	attraction	0.057692	12.865577
act_vb	108	6	0.934665	attraction	0.055556	12.468307
yell_vb	9	2	0.077889	attraction	0.222222	9.592019
immortalize_vb	1	1	0.008654	attraction	1	9.501058
pocket_vb	1	1	0.008654	attraction	1	9.501058
work	1	1	0.008654	attraction	1	9.501058
travel_vb	64	4	0.553876	attraction	0.0625	9.135414
read_vb	208	7	1.800096	attraction	0.033654	8.79138
dip_vb	11	2	0.095197	attraction	0.181818	8.730322
name_vb	38	3	0.328864	attraction	0.078947	8.128198
talk_vb	247	7	2.137614	attraction	0.02834	7.01971
do_vb	1655	25	14.32288	attraction	0.015106	6.764843
interview_vb	18	2	0.155778	attraction	0.111111	6.724735
sleep_vb	93	4	0.80485	attraction	0.043011	6.566184
cash_vb	19	2	0.164432	attraction	0.105263	6.513182
wrap_vb	19	2	0.164432	attraction	0.105263	6.513182
decide_vb	96	4	0.830813	attraction	0.041667	6.358533
command_vb	3	1	0.025963	attraction	0.333333	5.716683
mow_vb	3	1	0.025963	attraction	0.333333	5.716683
hit_vb	60	3	0.519258	attraction	0.05	5.677667
congratulate_vb	4	1	0.034617	attraction	0.25	5.054443
lead_vb	72	3	0.62311	attraction	0.041667	4.7657
keep_vb	535	10	4.630054	attraction	0.018692	4.764247
eat_vb	127	4	1.099097	attraction	0.031496	4.614266
ring_vb	5	1	0.043272	attraction	0.2	4.566456
tour_vb	5	1	0.043272	attraction	0.2	4.566456
touch_vb	32	2	0.276938	attraction	0.0625	4.562669
ride_vb	33	2	0.285592	attraction	0.060606	4.452965

inflict_vb	7	1	0.06058	attraction	0.142857	3.863564
shout_vb	7	1	0.06058	attraction	0.142857	3.863564
soar_vb	7	1	0.06058	attraction	0.142857	3.863564
articulate_vb	9	1	0.077889	attraction	0.111111	3.360928
participate_vb	100	3	0.865431	attraction	0.03	3.243718
redeem_vb	11	1	0.095197	attraction	0.090909	2.972626
anticipate_vb	12	1	0.103852	attraction	0.083333	2.807914
pitch_vb	12	1	0.103852	attraction	0.083333	2.807914
rub_vb	12	1	0.103852	attraction	0.083333	2.807914
behave_vb	13	1	0.112506	attraction	0.076923	2.658412
experience_vb	13	1	0.112506	attraction	0.076923	2.658412
fulfill_vb	15	1	0.129815	attraction	0.066667	2.396151
impress_vb	16	1	0.138469	attraction	0.0625	2.280078
enjoy_vb	67	2	0.579839	attraction	0.029851	2.146286
pretend_vb	19	1	0.164432	attraction	0.052632	1.97819
write_vb	212	4	1.834713	attraction	0.018868	1.935018
express_vb	73	2	0.631764	attraction	0.027397	1.902275
appreciate_vb	20	1	0.173086	attraction	0.05	1.890238
laugh_vb	20	1	0.173086	attraction	0.05	1.890238
hand_vb	21	1	0.18174	attraction	0.047619	1.807556
paint_vb	21	1	0.18174	attraction	0.047619	1.807556
define_vb	77	2	0.666382	attraction	0.025974	1.75537
grab_vb	22	1	0.190395	attraction	0.045455	1.729642
land_vb	22	1	0.190395	attraction	0.045455	1.729642
examine_vb	79	2	0.68369	attraction	0.025316	1.686153
claim_vb	23	1	0.199049	attraction	0.043478	1.656062
borrow_vb	24	1	0.207703	attraction	0.041667	1.586439
relax_vb	25	1	0.216358	attraction	0.04	1.520442
repeat_vb	25	1	0.216358	attraction	0.04	1.520442
sit_vb	85	2	0.735616	attraction	0.023529	1.493783
cross_vb	27	1	0.233666	attraction	0.037037	1.398201
win_vb	164	3	1.419306	attraction	0.018293	1.348931
celebrate_vb	28	1	0.242321	attraction	0.035714	1.341472
skip_vb	28	1	0.242321	attraction	0.035714	1.341472
characterize_vb	30	1	0.259629	attraction	0.033333	1.235773
practice_vb	31	1	0.268283	attraction	0.032258	1.186457
call_vb	173	3	1.497195	attraction	0.017341	1.181539
leave_vb	259	4	2.241465	attraction	0.015444	1.133569
join_vb	100	2	0.865431	attraction	0.02	1.096776
stick_vb	34	1	0.294246	attraction	0.029412	1.050892
jump_vb	35	1	0.302901	attraction	0.028571	1.009424
select_vb	35	1	0.302901	attraction	0.028571	1.009424
notice_vb	36	1	0.311555	attraction	0.027778	0.969637

steal_vb	36	1	0.311555	attraction	0.027778	0.969637
view_vb	36	1	0.311555	attraction	0.027778	0.969637
catch_vb	106	2	0.917356	attraction	0.018868	0.965516
tell_vb	281	4	2.43186	attraction	0.014235	0.857843
buy_vb	283	4	2.449169	attraction	0.014134	0.835376
operate_vb	40	1	0.346172	attraction	0.025	0.825545
run_vb	196	3	1.696244	attraction	0.015306	0.825313
let_vb	118	2	1.021208	attraction	0.016949	0.740895
visit_vb	125	2	1.081788	attraction	0.016	0.629952
shoot_vb	47	1	0.406752	attraction	0.021277	0.620787
finish_vb	50	1	0.432715	attraction	0.02	0.547843
raise_vb	134	2	1.159677	attraction	0.014925	0.5059
collect_vb	54	1	0.467333	attraction	0.018519	0.461891
attend_vb	55	1	0.475987	attraction	0.018182	0.442221
post_vb	56	1	0.484641	attraction	0.017857	0.423224
hold_vb	148	2	1.280837	attraction	0.013514	0.348614
compete_vb	61	1	0.527913	attraction	0.016393	0.337546
say_vb	648	7	5.60799	attraction	0.010802	0.326355
push_vb	62	1	0.536567	attraction	0.016129	0.322128
fly_vb	64	1	0.553876	attraction	0.015625	0.292863
close_vb	66	1	0.571184	attraction	0.015152	0.265583
vote_vb	66	1	0.571184	attraction	0.015152	0.265583
sign_vb	69	1	0.597147	attraction	0.014493	0.22813
investigate_vb	72	1	0.62311	attraction	0.013889	0.194518
walk_vb	76	1	0.657727	attraction	0.013158	0.15514
fall_vb	84	1	0.726962	attraction	0.011905	0.092709
consider_vb	87	1	0.752925	attraction	0.011494	0.074241
continue_vb	87	1	0.752925	attraction	0.011494	0.074241
expand_vb	88	1	0.761579	attraction	0.011364	0.068629
die_vb	89	1	0.770233	attraction	0.011236	0.063278
wear_vb	89	1	0.770233	attraction	0.011236	0.063278
come_vb	333	3	2.881884	attraction	0.009009	0.004842
take_vb	906	8	7.840801	attraction	0.00883	0.003282
serve_vb	112	1	0.969282	attraction	0.008929	0.000973
speak_vb	112	1	0.969282	attraction	0.008929	0.000973
test_vb	112	1	0.969282	attraction	0.008929	0.000973
be_vb	11588	36	100.2861	repulsion	0.003107	63.222069
get_vb	1620	4	14.01998	repulsion	0.002469	10.241572
know_vb	590	1	5.106041	repulsion	0.001695	5.008658
find_vb	504	1	4.36177	repulsion	0.001984	3.819554
think_vb	471	1	4.076178	repulsion	0.002123	3.37833
make_vb	1470	7	12.72183	repulsion	0.004762	3.158708
have_vb	2748	16	23.78203	repulsion	0.005822	3.009439

kill_vb	303	1	2.622255	repulsion	0.0033	1.329648
give_vb	470	2	4.067524	repulsion	0.004255	1.311924
provide_vb	293	1	2.535712	repulsion	0.003413	1.222578
help_vb	407	2	3.522303	repulsion	0.004914	0.790406
stop_vb	233	1	2.016453	repulsion	0.004292	0.636443
become_vb	221	1	1.912602	repulsion	0.004525	0.533482
create_vb	219	1	1.895293	repulsion	0.004566	0.516884
look_vb	358	2	3.098242	repulsion	0.005587	0.451175
stay_vb	207	1	1.791441	repulsion	0.004831	0.420951
sell_vb	205	1	1.774133	repulsion	0.004878	0.405602
put_vb	342	2	2.959773	repulsion	0.005848	0.355962
use_vb	450	3	3.894438	repulsion	0.006667	0.226405
show_vb	274	2	2.37128	repulsion	0.007299	0.06217
save_vb	141	1	1.220257	repulsion	0.007092	0.04282
grow_vb	132	1	1.142368	repulsion	0.007576	0.018719
drive_vb	129	1	1.116405	repulsion	0.007752	0.012712
learn_vb	122	1	1.055825	repulsion	0.008197	0.003036
set_vb	120	1	1.038517	repulsion	0.008333	0.001461
try_vb	119	1	1.029862	repulsion	0.008403	0.000883

APPENDIX D

FULL RESULTS OF LOGISTIC REGRESSION FOR COME TO V_2

	Coef	S.E.	Wald Z	P
Intercept	6.2707	0.9259	6.77	0.0000
Tense=None	-3.8883	0.5243	-7.42	0.0000
Tense=Past	-0.2685	0.3088	-0.87	0.3847
Tense=Present	-1.3910	0.3193	-4.36	0.0000
AspectV1=PerfectProgressive	-9.0227	29.5298	-0.31	0.7599
AspectV1=Progressive	-2.4654	0.3364	-7.33	0.0000
AspectV1=Simple	-2.7332	0.1859	-14.70	0.0000
VoiceV2=Passive	1.7198	0.2279	7.55	0.0000
COMP=None	-1.7695	0.3238	-5.46	0.0000
COMP=RelativeCL	-1.2531	0.3595	-3.49	0.0005
COMP=SubordConjunction	-1.8117	0.3841	-4.72	0.0000
COMP=ThatCL	-1.5322	0.3739	-4.10	0.0000
INF=None	-2.5194	0.7493	-3.36	0.0008

	2.5 %	97.5 %
(Intercept)	4.357474	8.0441826
TenseNone	-4.959376	-2.8893552
TensePast	-0.864773	0.3515865
TensePresent	-2.009594	-0.7521649
AspectV1PerfectProgressive	NA	25.7433838
AspectV1Progressive	-3.158867	-1.8328359
AspectV1Simple	-3.104726	-2.3753943
VoiceV2Passive	1.278443	2.1735452
COMPNone	-2.425063	-1.1494199
COMPRelativeCL	-1.975462	-0.5611506
COMPSubordConjunction	-2.583301	-1.0730787
COMPThatCL	-2.282836	-0.8125705
INFNone	-3.919689	-0.8934032

APPENDIX E

FULL RESULTS OF LOGISTIC REGRESSION FOR GET TO V_2

	Coef	S.E.	Wald Z	P
Intercept	-5.1015	0.9493	-5.37	0.0000
Tense=None	-0.9058	1.2347	-0.73	0.4632
Tense=Past	-0.8501	0.2295	-3.70	0.0002
Tense=Present	-0.9975	0.2296	-4.34	0.0000
AspectV1=PerfectProgressive	-4.5469	28.3103	-0.16	0.8724
AspectV1=Progressive	1.8973	0.5477	3.46	0.0005
AspectV1=Simple	3.4303	0.5615	6.11	0.0000
Mood=Imperative	-7.3099	31.6462	-0.23	0.8173
Mood=Indicative	0.6673	0.2005	3.33	0.0009
Mood=None	1.5553	1.1831	1.31	0.1886
VoiceV2=Passive	-1.5495	0.4454	-3.48	0.0005
INF=None	1.6836	0.5507	3.06	0.0022

	2.5 %	97.5 %
(Intercept)	-7.0744850	-3.3297756
TenseNone	-4.0063855	1.2791802
TensePast	-1.3091163	-0.4069103
TensePresent	-1.4567084	-0.5541503
AspectV1PerfectProgressive	NA	36.7947639
AspectV1Progressive	0.9309936	3.1339374
AspectV1Simple	2.4526114	4.6983118
MoodImperative	NA	36.9917561
MoodIndicative	0.2817545	1.0693222
MoodNone	-0.5194128	4.5954678
VoiceV2Passive	-2.5271319	-0.7503351
INFNone	0.6362723	2.8065547

REFERENCES

Achard, M. (1998). *Representation of cognitive structures: A syntax and semantics of French sentential complements*. Berlin: Mouton de Gruyter.

Bybee, J. (2010). *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT Press.

Chomsky, N. (1981). *Lectures on government and binding*. Cambridge: The MIT Press.

Chrupala, G. and van Genabith, J. (2007). Using very large corpora to detect raising and control verbs. In M. Butt and T. Holloway (Eds.), *Proceedings of the LFG07 conference* (148-162). Stanford: Stanford University Press.

Croft, W. and Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.

Dixon, R.M.W. (2006). Complement clauses and complementation strategies in typological perspective. In R. M. W. Dixon and A. Aikhenvald (Eds.), *Complementation: A cross-linguistic typology* (1-48). Oxford: Oxford University Press.

Fillmore, C. (1982). Frame semantics. In Linguistic Society of Korea (Ed.), *Linguistics in the morning calm* (111-137). Seoul: Hanshin Publishing Co.

Givón, T. (1991). Isomorphism in the grammatical code: Cognitive and biological perspectives. *Studies in Language, 15* (1), 85-114.

Givón, T. (1993). *English grammar*. Amsterdam: John Benjamins.

Givón, T. (1995). *Functionalism and grammar*. Amsterdam: John Benjamins.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Science, 7* (5), 219-224.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gries, S. Th. 2007. Coll. Analysis 3.2. A program for *R* for Windows 2.x

Gries, S. Th. (2007). Introduction. In S. Th. Gries and A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis* (1-17). New York: Mouton de Gruyter.

Gries, S. Th. (2009). *Statistics for linguistics with R: A practical introduction*. Berlin: Mouton de Gruyter.

Gries, S. Th. and Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics, 9* (1), 97-129.

Gries, S. Th. and Wulff, S. (*under revision*). Regression analysis in translation studies.

Haiman, J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language 56* (1), 515-540.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge: The MIT Press.

Kennedy, G. (1991). "Between and through: the company they keep and the function they serve." In K. Ajimer and B. Altenberg (Eds.). *English Corpus Linguistics* (95-110). New York: Longman.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Langacker, R. (1987). *Foundations of cognitive grammar, Volume I: Theoretical prerequisites*. Stanford: Stanford University Press.

Langacker, R. (1995). Raising and transparency. *Language, 71*, 1-62.

Martin, R. (2001). Null case and the distribution of PRO. *Linguistic Inquiry, 32* (1), 141-166.

McCawley, J. (1988). *The syntactic phenomena of English*. Chicago: University of Chicago Press.

Noonan, M. (1985). Complementation. In T. Shopen (Ed.), *Language typology and syntactic description, Volume II: Complex constructions* (42-138). Cambridge: Cambridge University Press.

Postal, P. (1974). *On raising*. Cambridge: The MIT Press.

Reppen, R., Ide, N., and Suderman, K. (2005). *American National Corpus*. LDC2005T35. CD-ROM. Philadelphia: Linguistic Data Consortium.

Rohdenburg, G. (2003). Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In G. Rohdenburg and B. Mondorf (Eds.), *Determinants of grammatical variation in English* (205-249). New York: Mouton de Gruyter.

Rosenbaum, P. (1967). *The grammar of English predicate constructions*. Cambridge: The MIT Press.

Rudanko, J. (1989). *Complementation in case grammar: A syntactic and semantic study of selected patterns of complementation in present-day English*. Albany: State University of New York Press.

Rudanko, J. (1998). *Change and continuity in the English language: Studies on complementation over the past three hundred years.* Lanham: University Press of America.

Rudanko, J. (2002). *Complements and constructions: Corpus-based studies on sentential complements in English in recent centuries*. Lanham: University Press of America.

Stefanowitsch, A. and Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8* (2), 209-243.

Verspoor, M. (2000). Iconicity in English complement constructions: Conceptual distance and cognitive processing levels. In K. Horie (Ed.), *Complementation: Cognitive and functional perspectives* (199-225). Amsterdam: John Benjamins.

Wulff, S. (2008). *Rethinking idiomaticity: A usage-based approach*. New York: Contiuum.

Wulff, S., Röme, U. and Swales, J. (*to appear*). Attended/unattended *this* in academic student writing: Quantitative and qualitative perspectives. In E. Csomay (Ed.), *Contemporary perspectives on discourse and corpora: New registers, analyses, texts, and tools*. Special issue of *Corpus Linguistics and Linguistic Theory*.