3 Recent shifts with three nonfinite verbal complements in English: data from the 100-million-word *Time* corpus (1920s–2000s)

MARK DAVIES

3.1 Introduction

The goal of this chapter is to show how robust data from a large corpus of English from the 1900s to the 2000s can shed light on shifts in verbal syntax, in ways that might not be possible with smaller corpora.

By way of introduction to this topic, we note that some languages have large historical corpora with robust data that allow researchers to look at a wide range of linguistic changes. For example, the *Corpus del Español* (www.corpusdelespanol.org) contains 100 million words from the 1200s to the 1900s, and the *Corpus do Português* (www.corpusdoportugues.org) contains 45 million words from the 1300s to the 1900s. Large corpora like these have been used to look at a wide range of changes in the language – lexical, morphological, syntactic, and semantic. (For a few examples dealing with infinitival complements, see Davies 2000, 2002, 2003, 2004a, 2004b, 2005, 2008, 2010.)

English corpus linguists, on the other hand, have tended to create much smaller corpora, in the belief that only 'small, carefully-constructed' corpora can be textually accurate enough to provide useful data. Examples of such corpora (from among many) are the 1.6 million word *Helsinki Corpus*, the 1.8 million word ARCHER corpus, and the *Diachronic Corpus of Present-Day Spoken English* (DCPSE), among others.[1]

The BROWN[2] family of corpora are representative of this tendency to use smaller corpora. As is well-known and as is discussed in other chapters in this volume, the 2 million words from the 1960s in the Brown (US) and LOB (UK) corpora have been supplemented by the 2 million words from the 1990s in the Frown (US) and FLOB (UK) corpora, and this allows users to carry out comparisons of the two decades.[3] This approach has been very useful for looking at high-frequency constructions, such as modals and auxiliaries – where even with the small corpora, researchers are able to find enough tokens to support their analyses (see, among others, Leech 2003; Leech and Smith 2009; Leech, Hundt, Mair and Smith 2009; Mair 1997a; Mair and Hundt 1997; Mair and Leech 2006; Smith 2002). Similarly, the

Diachronic Corpus of Present-Day Spoken English (DCPSE), which is composed of two small corpora of spoken British English from the 1950s and 1990s (about 1 million words), has been used for a number of insightful investigations of recent changes in English (see e.g. Close and Aarts 2010; Aarts, Close and Wallis 2010).

However, as most linguists are aware, there are many mid- to low-frequency constructions – such as verbal complementation – that cannot be studied (easily – or at all) with these smaller corpora. When the corpora yield only 10–20 tokens for a particular construction (such as [*to*-V] or [V-*ing*] with a particular verb, e.g. *proposed* [*to leave/leaving*]), that is often not sufficient to accurately map out the linguistic shift. For such constructions, we need much larger corpora.

The awareness that small corpora are not adequate for mid- to low-frequency constructions has led many corpus linguists who work on syntactic changes in the 1800s–1900s to create and then use their own proprietary corpora. This has been the case with nearly all linguists who have studied verbal complements in the 1800s–1900s. For example, Vosberg (2003a, 2003b) and Rohdenburg (2006b, 2007, 2009b) have created a corpus comprised of newspapers, *Project Gutenberg*, *Literature Online* (ProQuest), as well as using the BROWN and the BNC. Cuyckens and De Smet (2007) and De Smet (2008) have created corpora (CLMET, CLMETEV, CNN, and others) that has novels from *Project Gutenberg*, *Literature Online*, and they also use other modern corpora like BROWN, the *Bank of English* (BoE), and the BNC. Rudanko (e.g. 2000, 2003, 2005, 2006) uses the *'Century of English'* corpus from the 1700s (cf. Milik 1995) and the CONCE corpus from the 1800s (Kytö, Rudan Ro and Smitterberg, 2000), as well as modern corpora like the BNC, the BoE, and BROWN. Finally, Mair (e.g. 2006a) – who is the creator of the Frown and FLOB portions of BROWN – has, in addition to BROWN, used text archives of newspapers from the 1900s, the *Oxford English Dictionary* (*OED*), and modern corpora like the BNC and the BoE. While proprietary corpora based on large text archives are useful for obtaining data for individual studies, unlike publicly available corpora like BROWN+ and the DCPSE, these text archive-based corpora have the downside of not existing in a form that can easily be reused by others to check results and carry out follow-up studies.

In addition to using text archives like *Project Gutenberg*, *Literature Online*, or archives of newspapers and magazines, within the past two or three years, another potentially useful set of "corpora" (in the broad sense) have become available. These are the *Google Books* (books.google.com) and *Google News* (http://news.google.com/archivesearch) archives. Each of these contains millions of books or articles – and thus tens or hundreds of billions of words of text – from throughout the history of English (and other languages as well). For linguists who are interested in finding the first occurrence of a

word or phrase, or searches that involve a small, finite set of search strings, these text archives can be a powerful research tool.

In spite of their potential, however, *Google Books* and *Google News* archives are quite limited for diachronic syntax. The main problem, of course, is that the texts are not lemmatized or tagged for part of speech, which makes syntactically oriented searches quite difficult. For example, in order to look at the construction [Verb NP into V-*ing*] (*talk someone into making dinner*, *spooked them into cancelling the show*), one would have to look – one by one – at hundreds of matrix verbs (*talk*, *spook*, etc.) followed by different types of noun phrases (*her*, *Bill*, *the people*, etc) followed by any and all [V-*ing*] forms (*making*, *cancelling*, etc.) – all of which would involve hundreds of thousands or millions of individual search strings.

3.2 Using text archives: the *Time Magazine Corpus of American English*

As was mentioned previously, most researchers who work on low-frequency constructions like verbal complements end up creating their own proprietary corpora. One of the best sources for these corpora are text archives, composed of tens or hundreds of millions of words of text. In this chapter, I will focus on the use of a text archive that was converted into a structured linguistic corpus in 2007 – the *Time Magazine Corpus of American English* (hereafter the '*Time* corpus'), which is freely available online at http://corpus.byu.edu/time.

This corpus is composed of more than 275,000 articles from *Time* magazine archive (www.time.com/time/archive) – which includes more than 100,000,000 words of text from 1923 to 2007. The texts themselves are extremely accurate – having used the corpus for more than three years, I have encountered very few typographical errors. All of the texts were downloaded, along with the metadata (title, author, year, etc.) and imported into a relational database. The texts were then lemmatized and tagged for part of speech, using the same CLAWS tagger that has been used to tag the *British National Corpus* (BNC) and other corpora. The corpus was then integrated into the same corpus architecture and interface that have been used for other corpora from http://corpus.byu.edu.

The corpus allows users to carry out many different types of syntactically oriented searches. At the most basic level, users can input a string such as [end] up [v?g*] (*ended up paying*, *ends up saying*, etc), where [v?g*] matches tags like [vvg] *paying*, *going* or [vhg] *having*. Within about one second they will see the frequency of the construction in the 100 million words of text from the 1920s to the 2000s, as shown in Figure 3.1.

Users can also see the frequency of any word, phrase, syntactic construction, or collocates of any word in a particular year, and they can see the data displayed in tabular format – one entry for each matching string. They can

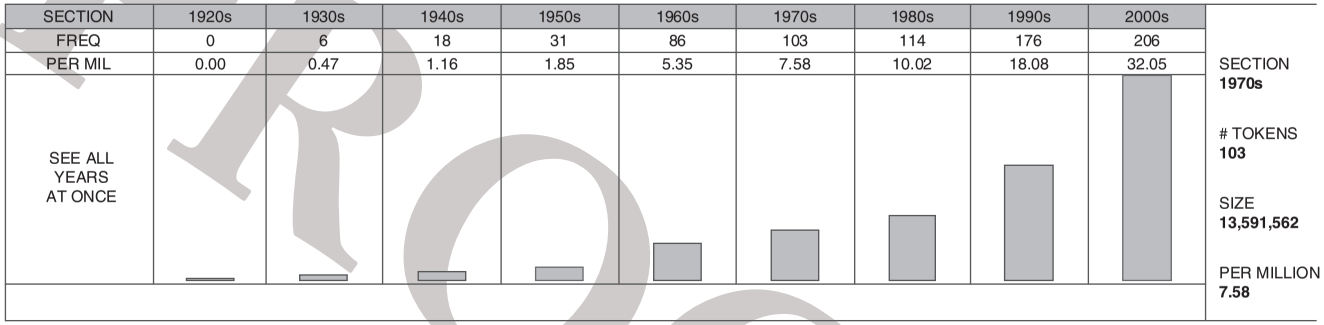

Figure 3.1. Decade-by-decade frequency for *[end] up [V-ing]* from the *Time* corpus, 1920s–2000s

also compare one section against the other, such as the search for phrasal verbs presented in Table 3.1, which in less than one second finds all instances with *up* that are much more common in the 1970s–2000s (left) compared to the 1920s–1950s (right) (note that for clarity in this display, here I have limited it just to infinitival forms of the verb).

Although the *Time* corpus is large, annotated, and allows for a wide range of searches, it has one obvious limitation. Unlike BROWN – which contains texts from several different genres – the *Time* corpus contains texts from just one genre (magazines) and just one magazine within that genre.

However, the concern about corpora composed of just one genre (or even just one source) is not unique to the *Time* corpus. In English historical linguistics it is quite common to use corpora that are composed of just one genre – such as the *Old Bailey Corpus*, the *Corpus of Early English Correspondence* (CEEC), the *Lampeter Corpus of Early Modern English Tracts*, the *Zürich Corpus of English Newspapers* (ZEN), and many others.[4] In all of these cases, each corpus is just part of an overall mosaic for the particular time period. To completely and totally flesh out changes from the 1900s, one would probably want data from corpora containing other genres as well.

However, in spite of the fact that the *Time* corpus contains data from just one genre, it is possible to compare its data against those of a more balanced corpus. Using the *Time* corpus, we can simply search for phenomena where – based on data from BROWN – we already have a relatively good idea of what changes have occurred in the 1900s. As can be seen in the many examples at http://corpus.byu.edu/coha/ compare-smallCorpora.asp, the data from the *Time* corpus are surprisingly similar to those of the more genre-balanced BROWN family of corpora.

So while the *Time* corpus does not paint a complete picture of changes in the 1900s, its data do agree quite well with smaller, more balanced corpora – for those constructions where the smaller corpora are able to provide sufficient data. The difference – as we will see – is that the 100-million-word *Time* corpus adds to this in a very important way, by providing data on some constructions that cannot be studied with smaller corpora.

3.3 Three shifts in verbal complementation during the 1900s

The goal of this chapter is to show how robust data from a large corpus of English from the 1900s can shed light on shifts in verbal syntax, in ways that would not be possible with smaller corpora. To do so, I will focus on three different aspects of verbal complementation during the 1900s–2000s. These constructions are:

(1) V NP *into* [V-*ing*]: e.g. *we talked Bill into staying*

(2) V [*to* V/V-*ing*]: e.g. *he started* [*to walk/walking*] *down the street*

(3) V (*for*) NP [*to* V]: e.g. *I'd really like* (*for*) *them to leave now*

Table 3.1. *Phrasal verbs with* up, *1970s–2000s and 1920s–1950s*[a]

	Word/Phrase	Tokens 1	Tokens 2	PM 1	PM 2	Ratio		Word/Phrase	Tokens 2	Tokens 1	PM 2	PM 1	Ratio
1	FREE UP	15	0	0.36	0.00	36.47	1	BOLSTER UP	39	0	0.74	0.00	74.24
2	BULK UP	13	0	0.32	0.00	31.61	2	MUSTER UP	9	0	0.17	0.00	17.13
3	BUCKLE UP	11	0	0.27	0.00	26.75	3	CLUTTER UP	8	0	0.15	0.00	15.23
4	FESS UP	8	0	0.19	0.00	19.45	4	SLOW UP	45	3	0.86	0.07	11.74
5	SUIT UP	14	1	0.34	0.02	17.88	5	CAST UP	6	0	0.11	0.00	11.42
6	SCROUNGE UP	7	0	0.17	0.00	17.02	6	PLOW UP	5	0	0.10	0.00	9.52
7	BOOT UP	6	0	0.15	0.00	14.59	7	PRICK UP	5	0	0.10	0.00	9.52
8	BUMP UP	5	0	0.12	0.00	12.16	8	WHISTLE UP	5	0	0.10	0.00	9.52
9	DIAL UP	5	0	0.12	0.00	12.16	9	BEAR UP	11	1	0.21	0.02	8.61
10	RATCHET UP	5	0	0.12	0.00	12.16	10	RIG UP	7	1	0.13	0.02	5.48

[a] In this table, PM1 and PM2 show the normalized frequency (per million words) in the two sections (PM1 = 1970s–2000s and PM2 = 1920s–1950s). For example, *free up* occurs 15 times in the 1970s–2000s, but does not occur in the 1920s–1950s. The 'ratio' column is the ratio of these two normalized figures, and the results have been smoothed (by adding .1 to the denominator) to avoid division by zero.

There certainly are other aspects of verbal complementation that we could consider in this chapter. For example, there is variation with the verb *help*:

(4) V *(to)* V *she helped John (to) clean the room*

This is a construction that has already been studied by many others, such as Kjellmer (1985), Řeřicha (1987), Mair (1995, 2002), Rohdenburg (2009b), and Callies (this volume). Because *help* is a high-frequency verb, there are enough data in even a small corpus like BROWN to answer most of the important questions. As a result, I will not discuss that construction here. In this chapter, I will focus on the three constructions (1)–(3) above, where only a large, robust corpus can provide us with the needed data.

I should note that our goal in this chapter is *not* to provide in-depth and complete analyses of each of these three constructions. That would not be possible in this limited amount of space. Rather, the goal is to show how the *Time* corpus can provide the needed data, and suggest some possible areas of study for future researchers.

3.4 V NP *into* [V-*ing*]: e.g. *we talked Bill into staying*

This is a construction that has received a fair amount of attention over the past decade or so, as evidenced by studies like Rudanko (2000: ch. 5, 2003, 2005, 2006), Rudanko and Luodes (2005: ch. 2), Gries and Stefanowitsch (2003), Wulff, Stefanowitsch, and Gries (2007), and Hunston and Francis (2000: 102–3). In spite of the insights of all of these studies, however, they are all essentially synchronic in nature, and deal with data from the 1990s and 2000s – with little or no attention to the historical development of the construction. The two exceptions are Rudanko (2005, 2006), which as an aside very briefly look at the construction in the Brown family of corpora, and suggest that the construction is expanding its scope in English. But other than a short table with frequencies in the four corpora in the Brown family, we have little sense of what has happened with the construction throughout the rest of the 1900s.

I have replicated the search in the BROWN corpora (using a version of these corpora that was annotated with the CLAWS tagger), using the following three searches:

(5) a. [vv*] * [nn*] *into* [v?g*] (e.g. *talked the people into leaving*)
b. [vv*] [np*] *into* [v?g*] (e.g. *coerces Bill into doing*)
c. [vv*] [p*] *into* [v?g*] (e.g. *forced them into buying*)

The following table represents the overall frequency of the construction in the four corpora:

	1960s	1990s
American	6	11
British	3	9

In terms of the 21 specific matrix verbs that take the construction, we find the following:

Brown: *talk* (2), *fool, goad, gull, terrify, entice*
LOB: *nag, shame, harass*
Frown: *talk* (4), *coax, fool, seduce, force, charm, coerce, deceive, hound, persuade, pressure*
FLOB: *talk* (2), *fool* (2), *bluff, shame, spur, deceive, cow*

The sparse BROWN data leave us with a number of questions:

1. Is there really an increase over time, or is the number of tokens too small to draw this conclusion?
2. If there is an increase, has it taken place recently (e.g. from the 1960s to the 1990s – the period of the BROWN corpora), or before that time?
3. In terms of the semantics of the construction, how has the class of matrix verbs that take the construction changed over time?

With more than twenty-five times as much data as the BROWN corpora (100 million vs. 4 million words), the *Time* corpus yields more than 1,101 tokens for this construction (compared to the 29 from the BROWN corpora), and is thus able to answer most of these questions quite well. First, the *Time* data confirm what the BROWN corpora suggest in question 1 above – the construction is clearly increasing over time. As Table 3.2 shows, it increased in frequency (per million words) nearly fourfold from the 1920s to the 1970s, although it has stayed relatively constant since.

To answer question 2 above, Table 3.2 shows that the construction did increase from the 1960s to the 1990s (as the BROWN corpora suggest it did), but the increase was very slight (11.2 to 11.6 tokens per million words), and this is probably not statistically significant. However, the data clearly show that the greatest increase was before the 1960s.

Table 3.2. *Overall frequency of* V NP into V-ing*, 1920s–2000s*

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	TOTAL
Tokens	25	84	154	171	180	174	119	113	81	**1,101**
Size (millions)	7.635	12.658	15.454	16.788	16.081	13.592	11.372	9.735	6.427	
Tokens per million	**3.3**	**6.6**	**10.0**	**10.2**	**11.2**	**12.8**	**10.5**	**11.6**	**12.6**	

Table 3.3. *Frequency of* V NP into V-ing *by verb, 1920s–2000s*

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	Total
TALK	0	5	32	41	33	37	15	11	9	183
TRICK	0	7	7	4	8	6	6	8	18	64
PRESSURE	0	0	3	3	15	18	10	8	7	64
COAX	0	2	0	5	6	8	3	4	4	32
GOAD	0	1	7	6	8	2	3	1	3	31
LURE	1	3	5	4	6	4	1	4	1	29
FOOL	2	5	2	3	6	2	2	3	1	26
FORCE	0	1	2	5	5	5	5	3	0	26
SCARE	0	6	5	4	1	2	4	2	1	25
PROD	1	1	3	5	6	1	2	3	1	23
PUSH	0	0	3	1	5	7	3	2	1	22
BULLY	1	2	4	5	1	1	1	5	1	21
CON	0	0	0	5	6	7	1	0	1	20
DELUDE	0	1	0	0	7	4	1	2	2	17
COERCE	1	2	2	0	2	3	4	2	1	17
BLACKMAIL	0	3	3	4	1	2	1	1	1	16
MANEUVER	0	1	3	4	2	2	3	0	0	15
SHAME	0	0	2	3	3	0	1	3	3	15

Regarding question question 3 above (the semantics of the construction), Table 3.3 shows the frequency (by decade) of all verbs that occur at least fifteen times total in the corpus.

Few of these are surprises; they certainly fit in with the general meaning of this construction, which is that 'X causes Y to do something via emotional or physical force'. And such a list is typically what we would find with a small 4-million-word corpus like BROWN, where there are few tokens and of course even fewer types. With a more robust corpus, however, we can see interesting ways in which the construction has extended its use. For example, Table 3.4 presents a full listing of all 168 verbs in the corpus, with the decades in which they first occur.

Some of these verbs are strange enough that it might be difficult to believe that they actually all participate in the construction, but in fact they do. Table 3.5 presents a handful of the more interesting ones.

In this chapter I will not examine the details of the semantic extension of the matrix verbs, such as:

- whether certain types of control have become more common (compare the idea in Wulff *et al.* 2007 that matrix verbs in British English tend to represent physical force more, while those in American English relate more to persuasion);
- when the romance-related uses arose (*charm/smooch someone into doing something*), or
- whether the metaphorical extension of physical force has increased or decreased in American English (e.g. *drive, push, pound, elbow, drill, move, nudge, budge, jar*).

Table 3.4. *New matrix verbs with* V NP into V-ing *construction, 1920s–2000s*

Decade	Verbs occurring for the first time in this decade
1920s	fool (2 tokens), seduce 2, tease 2, beguile, browbeat, bully, cajole, coerce, frighten, heckle, heffle, hornswoggle, hypnotize, lure, prod, rouse, spirit, stampede, terrify, terrorize, worry
1930s	trick 7, scare 6, talk 5, argue 4, blackmail 3, wheedle 3, bulldoze 2, coax 2, drive 2, startle 2, anger, badger, bamboozle, bludgeon, bluff, catspaw, charm, coddle, convert, deceive, delude, egg, flatter, force, goad, gull, harangue, impress, intimidate, inveigle, jockey, josh, kid, maneuver, nag, sting, trap
1940s	mislead 5, pressure 3, push 3, blitz 2, hoodwink 2, irritate 2, shame 2, stir 2, bewitch, bother, challenge, chivie, deflect, draw, enthuse, entice, harass, hex, hound, humiliate, hurry, hustle, lull, mesmerize, persuade, pound, preach, rush, shock, sidetrack, smooch, taunt, torture, urge, woo
1950s	con 5, pour 3, throw 3, bribe 2, jolt 2, activate, beat, beg, cramp, dope, dupe, ease, elbow, encourage, feint, jar, lead, needle, panic, politic, provoke, salestalk, threaten, work
1960s	tempt 2, confound, dump, gig, graft, harden, incite, insinuate, jolly, manipulate, propagandize, shill, spur, steer, stiffen
1970s	embarrass 3, bore, brainwash, divert, drill, filibuster, hook, hurl, move, shake, svengali
1980s	blow 2, galvanize 2, press 2, blarney, drag, guide, nudge, stimulate
1990s	spook 2, bargain, bomb, budge, cross, euchre, lock, muscle, numb, pummel, snooker, sober
2000s	catapult, chase, jawbone, poison, wrangle

But with 168 types and 1,101 tokens over an 85-year period (compared to 29 tokens with 29 types in the 1960s and 1990s with the BROWN corpora), it would of course be quite possible to answer questions like these, and this would give us valuable insight into the role that prototypes have played in the development of the construction (cf. Gries and Stefanowitsch 2003).

3.5 [*to*-V] vs. [V-*ing*]: e.g. *he started* [*to walk/walking*] *down the street*

The alternation between these two constructions has been discussed at length in a number of articles and books during the past decade or two. The two constructions have been compared in terms of dialect differences, syntactic constraints, and semantic and pragmatic contrasts.[5]

What has been studied somewhat less is the historical development of these two constructions, although there have been some very good studies here as well. Fanego (1996, 2007) does a very good job looking at the contrast in Early Modern English, but since her focus is on the 1700s–1800s, there is obviously no data from the 1900s–2000s. Rohdenburg (2006b, 2007, 2009b) and Vosberg (2003a,b) do look at the 1800s–1900s, and use an eclectic collection of newspapers, magazines, and books from Project Gutenberg from this time period. Mair (2001, 2002, 2006b) likewise

Table 3.5. *Selected examples of* V NP into V-ing, *1920s–2000s*

1920s	HEFFLE	Last week Senator Heflin tried, at length, to *heffle* the Senate into adopting a resolution condemning the nameless bottle-thrower [note the play on the senator's name]
1920s	HORNSWOGGLE	Hilda invited to Rackham, with the idea of *hornswoggling* them into buying the place, gouty Lord Mere de Beaurivage.
1930s	EGG	its Soviet and Chinese members tried to *egg* the League into cracking down as hard as possible on Tokyo.
1930s	STING	This was supposed to have *stung* Dictator Stalin into assuming a defiant attitude.
1930s	CATSPAW	Then his boss *catspawed* him into marrying a European mistress who was getting troublesome.
1940s	PREACH	a week's visit to the Chemurgic Institute, which he *preached* mid-Trinity farmers into founding at Romayor
1940s	POUND	with a completely wrong belief that you can *pound* your neighbors into loving you as an Apache pounds his woman into dazed rapture
1940s	SMOOCH	they had encouraged his wife to *smooch* the customers into buying more drinks
1950s	CRAMP	enabling Rattigan to dramatize incidents that the stage *cramped* him into reporting at secondhand
1950s	DOPE	pearl fishermen made plans to *dope* stubborn oysters into yielding up their precious pearls
1950s	JAR	I want to *jar* the observer into thinking, to make him uncomfortable.
1960s	JOLLY	all trying to *jolly* the reader into putting up once more with that old boudoir Bolshevik
1960s	GIG	He can and has *gigged* the Administration into paying closer heed to the Vietnamese refugee problem
1960s	STIFFEN	We hope this will *stiffen* them into resisting the automobile, and preserving the amenities
1970s	DRILL	and to *drill* the people into becoming active participants in public health campaigns
1970s	HOOK	The building tension undoubtedly *hooked* first-night viewers into sticking with the series
1970s	SVENGALI	he *Svengalied* willing authors into writing potboilers and racy romans clef
1980s	BLARNEY	Moynihan, who … *blarneyed* Nixon into endorsing the idea [note: Moynihan was Irish]
1980s	BLOW	a providential "wayward wind" will *blow* him into drawing fewer minority conclusions
1990s	EUCHRE	trying to portray the Democrats as the high-tax party, by *euchring* them into proposing an increase
1990s	SOBER	such a frightening specter will *sober* both countries into backing off their nuclear one-upmanship
2000s	POISON	We have to get rid of the states that *poison* their people into believing that terrorism is the only means of improving their lives

uses an eclectic corpus of (primarily) Late Modern English, composed of quotations from the *Oxford English Dictionary*, newspapers, the BROWN corpora, and a handful of American novels. Rudanko uses the *'Century of English'* corpus from the 1700s (cf. Milik 1995) and the CONCE corpus from the 1800s (cf. Kÿto *et al.* 2000).

The bottom line from all of these studies and all of these corpora is that there has been an overall shift from [*to*-V] to [V-*ing*] over time, relating to what Rohdenburg (2006b, 2007, 2009a) has called the Great Complement Shift in English, by analogy with the Great Vowel Shift from Early Modern English. Let us consider this analogy a little more closely, because it provides us with insight into what type of data from our corpora might be the most useful.

It is widely believed that the Great Vowel Shift involved a simple 'push chain' or 'drag chain', which was responsible for all of the vowel raising. However, Görlach (1991) suggests the picture was much more complex. The Great Vowel Shift was a combination of push and drag changes, operating at different times. As Görlach shows, the only way to understand the relationship between the different competing shifts is to have fine-grained data from a nearly continuous corpus of texts. It is not sufficient to sample the texts from one year and then come back thirty to fifty years later and sample more texts. In this case, important intermediate, cumulative shifts in the vowel system would have been missed.

It is the same with the [*to*-V] to [V-*ing*] shift. Just as the Great Vowel Shift had raising with one set of vowels and then another and then another, we need to look at the Great Complement Shift in the same way – by a series of shifts from [*to*-V] to [V-*ing*] with different matrix verbs, one after another. And just as with the vowel shifts, the sequencing of these micro-level shifts in the overall Great Complement Shift can provide us with important clues about what may have been driving the overall shift, as we consider why some verbs changed before others (e.g. *start*, *begin*, *continue*, *try*, *love*, *prefer*, *bother*). Did higher-frequency verbs shift before lower-frequency verbs? Did certain semantic classes (e.g. aspectual verbs, or verbs of emotion) lead the way?

While the studies listed at the beginning of this section are all valuable, the problem with them is that we have data on different verbs using different corpora from different time periods. As a result, there is really no way to answer questions like those posed in the previous paragraph. In order to answer these questions, we need a corpus robust enough to look at all relevant verbs in one consistent corpus across a fairly large time period. While the data are far too sparse with a small, discontinuous corpus like BROWN, they are readily available with the 100-million-word *Time* corpus.

In terms of the BROWN data, consider Table 3.6. For each of the four corpora (Brown, LOB, Frown, FLOB) it shows the number of tokens of [*to*-V]:[V-*ing*]; for example, there are 50 tokens of [*start* to V] in the Brown corpus, and 52 tokens of [*start* V-*ing*]. The column labeled [American] shows the overall percentage of [V-*ing*] with that verb over time, e.g. 51 percent of the tokens with *start* are [V-*ing*] in the 1960s (Brown), and this increases to .61 in the 1990s (Frown), and the data in the [British] column

Table 3.6. to- V vs. V-ing *by verb in the BROWN corpora*

	Brown	Frown	American	χ^2 (p)	LOB	FLOB	British
Increase							
START	50:52	59:94	.51 >.61	.23	36:48	47:51	.57 >.52
BEGIN	252:47	203:85	.16 >.30	**<.001**	249:22	212:23	.08 >.10
LIKE	125:43	115:53	.26 >.32	.23	126:37	109:55	.23 >.34
LOVE	10:2	19:5	.17 >.21	.77	7:0	15:7	.00 >.32
HATE	8:2	6:6	.20 >.50	.15	1:2	4:7	.67 >.64
BOTHER	13:0	8:1	.00 >.11	.22	9:1	18:0	.10 >.00
PROPOSE	12:1	12:3	.08 >.20	.35	34:1	15:1	.03 >.06
CEASE	18:0	10:1	.00 >.09	.19	36:2	20:0	.05 >.00
FEAR	4:0	0:2	.00 > 1.00	**.014**	1:0	1:2	.00 >.67
AFFORD	16:0	18:1	.00 >.05	.35	20:0	14:0	.00 >.00
Little change							
TRY	344:6	371:6	.02 >.02	.90	350:6	322:14	.02 >.04
CONTINUE	117:5	169:7	.04 >.04	.96	97:9	108:6	.08 >.05
ATTEMPT	129:0	140:0	.00 >.00	.96	111:0	161:0	.00 >.00
INTEND	41:0	55:0	.00 >.00	.71	65:1	81:5	.02 >.06
MANAGE	31:0	55:0	.00 >.00	.69	57:0	72:0	.00 >.00
DESERVE	4:0	7:0	.00 >.00	.71	7:0	9:2	.00 >.18
HESITATE	8:0	6:0	.00 >.00	.70	11:0	5:0	.00 >.00
Decrease							
PREFER	21:2	19:0	.09 >.00	.19	33:3	19:1	.08 >.05
STAND	3:19	8:10	.86 >.56	**.03**	3:20	5:8	.87 >.62

work the same way. Finally, the [c^2 (p)] column gives the p value from the chi-square test for the American shift.[6] For example, for the shift with *start* in Brown and Frown, the p value is .23, which is not statistically significant. (I have only calculated the p value and chi-square for the two American corpora, since that is what I will compare to the American *Time* corpus).

Two important things stand out in the BROWN data. First, the data are too sparse to provide statistically significant values. For example, it does seem that there is an increase in [V-*ing*] with *hate* (20% > 50% [V-*ing*] from the 1960s > 1990s), but since there are just 22 tokens, the p value is .15 – greater than the statistically significant value of $p < .05$. In fact, there are only 3 out of 16 verbs where there is a statistical significance (the verbs *begin*, *fear*, and *stand*). The second important fact with the BROWN data is that there is really no way to show how these shifts are related, to see the sequencing in terms of the Great Complement Shift. Because we only have two time periods (1960s and 1990s), even if one shift occurred mainly between 1950 and 1970 and the other was between 1980 and 1990, there would be no way to know this.

The data from the *Time* corpus are much more robust. As we will see, nearly all of the shifts are statistically significant, and we can also sequence the shifts with the different matrix verbs. Table 3.7 shows the data for the same 19 verbs as in the Table 3.6.

Table 3.7. to-V vs. V-ing *by verb in the* Time *corpus*

Verb	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	tokens	χ^2 (p)
Increase											
START	228	885	1314	1809	1487	1248	1172	1423	1513	11079	<.0001
	0.434	0.553	0.727	0.726	0.792	0.705	0.627	0.651	0.683		
BEGIN	1393	3990	5336	6136	5563	4982	4064	2923	1597	35984	<.0001
	0.14	0.30	0.33	0.40	0.42	0.47	0.48	0.46	0.50		
LIKE	462	1072	1777	1883	1687	1278	958	923	790	10830	<.0001
	0.028	0.034	0.024	0.027	0.027	0.036	0.056	0.052	0.096		
LOVE	60	108	162	148	174	145	137	218	227	1379	<.0001
	0.100	0.046	0.043	0.108	0.098	0.103	0.153	0.197	0.260		
HATE	29	61	137	135	90	51	47	67	43	660	<.0001
	0.241	0.082	0.080	0.089	0.156	0.118	0.234	0.254	0.419		
TRY	1519	3842	6035	6922	6265	5718	4901	4379	3223	42804	<.0001
	0.009	0.016	0.020	0.012	0.019	0.013	0.009	0.021	0.029		
BOTHER	26	126	168	274	229	177	82	123	57	1262	<.0001
	0.038	0.032	0.042	0.029	0.031	0.079	0.098	0.138	0.263		
PROPOSE	296	408	322	252	204	174	190	116	57	2019	<.0001
	0.034	0.061	0.078	0.111	0.245	0.414	0.463	0.534	0.526		
FEAR	160	238	189	145	107	105	83	71	40	1138	<.0001
	0.038	0.034	0.048	0.021	0.140	0.133	0.181	0.296	0.450		
Little change											
CONTINUE	689	1189	1154	1302	1567	1878	1630	1162	656	11227	0.048
	0.086	0.083	0.094	0.076	0.091	0.086	0.111	0.096	0.108		
PREFER	145	259	322	302	413	318	245	188	126	2318	0.007
	0.090	0.062	0.056	0.053	0.051	0.075	0.102	0.122	0.119		
ATTEMPT	541	720	379	409	572	690	580	328	166	4385	0.55
	0.006	0.003	0.005	0.002	0.002	0.001	0.000	0.000	0.000		

Table 3.7. *(cont.)*

Verb	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	tokens	χ^2 (p)
MANAGE	153	552	846	1362	1388	1193	827	695	350	7366	0.01
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003		
AFFORD	100	202	261	281	272	227	177	155	119	1794	0.67
	0.000	0.005	0.004	0.004	0.000	0.009	0.000	0.000	0.000		
DESERVE	18	33	44	54	44	54	62	87	58	454	0.63
	0.000	0.000	0.023	0.056	0.023	0.056	0.016	0.011	0.034		
HESITATE	73	108	108	152	119	86	66	31	12	755	0.45
	0.000	0.000	0.000	0.000	0.000	0.012	0.000	0.000	0.000		
Decrease											
STAND	61	113	165	164	223	208	112	96	57	1199	<.0001
	0.738	0.327	0.412	0.274	0.166	0.091	0.071	0.156	0.123		
INTEND	445	608	835	926	1191	879	756	507	205	6352	<.0001
	0.045	0.018	0.007	0.005	0.007	0.002	0.001	0.002	0.000		
CEASE	203	339	217	168	128	109	85	66	26	1341	<.001
	0.241	0.322	0.171	0.155	0.281	0.165	0.200	0.242	0.154		

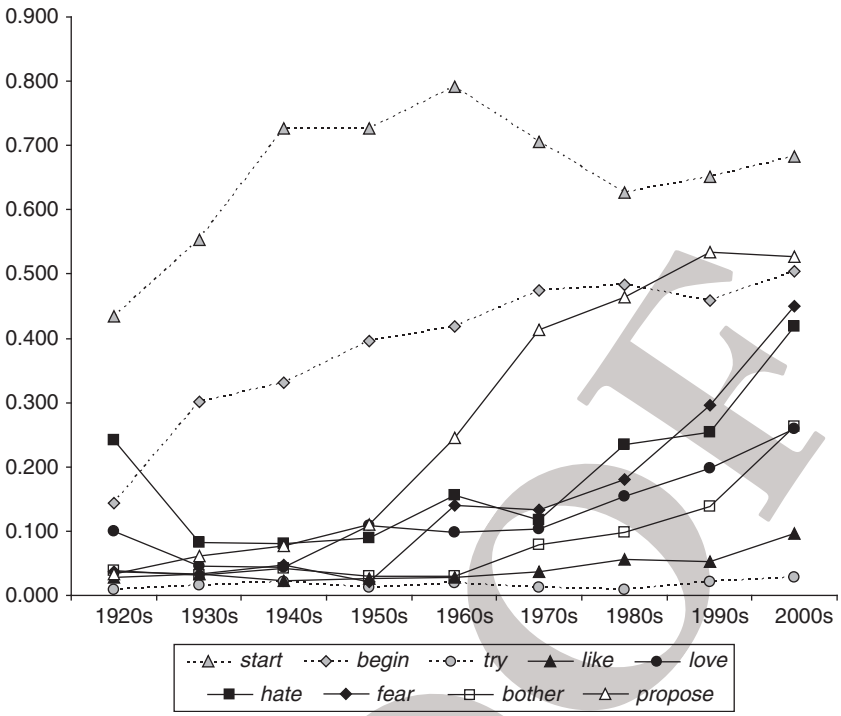

Figure 3.2. % of [*V-ing*] (vs [*to-V*]) by verb, 1920s–2000s

In general terms, the *Time* data and the BROWN data support each other. In both cases, the following verbs have clear increase in [V-*ing*] over time: *start*, *begin*, *like*, *love*, *hate*, *propose*, *bother*, and *fear*. In BROWN, *cease* decreases in British English, but it increases in American English (based on a one token increase from the 1960s to the 1990s). Based on 1,541 tokens, the *Time* corpus shows a moderate decrease with *cease*. In BROWN, *afford* shows a slight increase in [V-*ing*] (based again on a one token increase), but based on 1,794 tokens, *Time* shows it staying relatively unchanged. BROWN shows a slight decrease in [V-*ing*] with *prefer* (but this is due to just two [V-*ing*] tokens in the 1960s and none in the 1990s), while the 2,318 tokens in *Time* show a slight increase.

Notice, however, that whereas there are almost no shifts in the BROWN corpora that are statistically significant, 15/19 verbs in *Time* show a statistically significant shift. Whereas the data in the BROWN corpora are *suggestive* of change, the data in *Time* confirm this, and they show that the change in fact has occurred.

Finally, consider Figure 3.2, which perhaps shows more clearly the relative chronology of the shift from [*to*-V] to [V-*ing*] with most of the

verbs in the [increase] section of Table 3.7. As this chart shows, the high-frequency verbs *start* and *begin* increased more than almost any other verbs from the 1920s to the 1940s. In the last thirty or forty years, however, their increase has leveled out, while *propose*, *bother*, and the 'emotion' verbs *love*, *like*, *fear*, and *hate* have experienced a significant increase since the 1950s/1960s, and the largest increase was with the emotion verbs, especially with the strongly negative verbs *hate* and *fear*.[7]

Again, the *Time* corpus provides us with enough data for most verbs to draw statistically significant conclusions, whereas smaller corpora like BROWN do not. In addition, it is perhaps the only one that is 'continuous' enough (with millions of tokens each decade) to allow us to see relationships between shifts with different matrix verbs, and thus to perhaps begin to investigate some of the semantic factors that may be contributing to the overall syntactic shift.

3.6 V (*for*) NP [*to* V]: e.g. *I'd really like (for) them to leave now*

In the Government and Binding version of generative grammar, this construction was known as the Exceptional Case Marking construction, and the discussion revolved around how the subject (*them*, in the example above) could receive case (see Radford 2004: 128–31 for a concise discussion of ways in which this construction has been analyzed in different models of syntax). In this section, I will focus only on those verbs that allow both [+*for*] and [−*for*], such as *like*: *I'd like (for) them to leave now*. I will not consider verbs like *wait*, *care*, *arrange*, or *yearn*, which do not allow [−*for*], e.g. * *I wait/cared/yearned/arranged him to do it*.

As seen in Table 3.8, there are a number of verbs that do allow variation with the use of *for* – it can optionally appear before the subject of the embedded clause. Note that most of these are from the *Time* corpus. In cases where there were no tokens with a particular verb in the *Time* corpus, the examples come from the *Corpus of Contemporary American English* (COCA).[8]

There has been relatively little study of the verbs that allow both [+*for*] and [−*for*], especially their historical development. There are of course many studies that deal with [*for* + NP + to + V] generally (including diachronic studies like Fischer 2000, Pak 2006, Fanego 2007, Cuyckens and De Smet 2007, and McFadden 2008), but there has been little discussion of the construction in the context of verbal complements (*I want* (*for*) *him to do it*). In two synchronic studies, Erdmann (1993) compares contemporary American and British English, and Wagner (2000) discusses their use in the *British National Corpus*.

The only study to look at the historical development of [*for* + NP + to + V] with verbal complements is Cuyckens and De Smet 2007, which was released later as chapter 6 in De Smet (2008). De Smet shows that the

Table 3.8. *[±*for*] with selected verbs in the* Time *and COCA corpora*

	[−*for*]	[+ *for*]
LIKE	I would *like you to meet* the man I am going to marry (1950)	I'd *like for you to give* this to the kids at the New York Foundling Hospital (1949)
LOVE	Beleaguered Japanese chief would have *loved Clinton to drop* in (1998)	The state's Republicans would *love for him to run* for political office (1990)
HATE	we *hate you to convert* the people to democracy (1995)	Dad's cookouts had always been such fun that we all *hated for them to end* (COCA, 2008)
WISH	John Le Mesurier *wishes it to be* known that he conked out on Nov. 15 (1984)	but you desperately *wish for it to be* leavened with a little humor (COCA, 1991)
WANT	Heine's placid father *wanted him to be* a comfortable merchant (1937)	I haven't thought about it. I just *want for him to be* all right (1960)
PREFER	Helen … would *prefer her to marry* Pyrrhus (1925)	I'd much *prefer for nobody to know* I've been in it (1979)
ASK	I *ask you to appoint* a doctor to take my life (1935)	have wired us specifically *asking for you to broadcast* to the men in the Philippines (1942)
BEG	He writhed under the scorching heat, *begged someone to shoot* him (1949)	Treglia's tearful widow Tilda, as she *begged for someone to identify* the killer (1977)
EXPECT	and don't *expect him to be* an Angel from Heaven! (1925)	how could you *expect for me to want* to see you (COCA, 1992)
INTEND	Gandhi himself said: "God *intends me to live*" (1943)	God never *intended for me to work* hard (2001)
MEAN	Harold Stassen did not *mean it to be* in any way final (1942)	God never *meant for it to be* so difficult (1973)
NEED	You don't *need me to be* a nursemaid of any sort (1978)	You *need for me to show* you what time it is, baby (COCA, 1994)
ALLOW	"Old Mike" … would only *allow two people to pet* him (1928)	to *allow for reasonable people to spend* a week last winter buying plastic sheeting (2003)

overall frequency of verbs taking [*for* + NP + *to* + V] has increased during the last three hundred years (2008: 175 *et passim*). But of course this includes verbs like *wait*, *arrange*, and *yearn*, which do not alternate with [−*for*]. The question, then, is how much [+*for*] has been increasing in those cases of verbal complements where they do alternate with [−*for*], and this is something that has not been discussed in previous studies.

The problem with looking at this construction is that with most of these verbs, the [+*for*] construction is still quite rare. For example, I searched for *[vv*] for [p*] to [v*]* (*I want for you to leave*, *they needed for someone to do something*) in the four BROWN corpora. There were only four tokens in the 4 million words of text: none from Brown, one from the LOB, three from Frown, and none from FLOB:

(6) a. *The other one, Diablo, does not* ***like for me to get*** *on the back*. (LOB)
b. *Joe* ***liked for me to eat*** *it all up and want more*. (Frown)
c. *Well, I don't* ***intend for them to be*** *mine*. (Frown)
d. *A voice* ***asking for someone to put*** *another spoon of sugar in*. (Frown)

Table 3.9. *Overall percentage of [*±for*] in the* Time *corpus with* ask

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
[− *for*]	181	264	331	418	284	273	201	227	146
[+ *for*]	0	0	2	1	0	0	0	1	0
+ *for* %	0	0	0.006	0.002	0	0	0	0.004	0

Table 3.10. *Overall frequency of [*+for*] in the* Time *corpus, with verbs allowing alternation*

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
EXPECT	0	0	0	0	0	0	0	0	0
NEED	0	0	0	0	0	0	0	0	0
WANT	0	0	0	0	1	1	0	0	0
WISH	0	0	0	1	1	1	1	0	0
MEAN	0	0	0	0	0	1	3	3	1
INTEND	0	0	0	3	4	4	2	0	7
ASK	0	0	2	1	0	0	0	1	0
BEG	0	0	2	1	0	1	0	1	0
ALLOW	0	1	0	0	0	1	0	1	2
LOVE	0	0	0	0	0	0	0	1	2
HATE	0	0	0	0	0	0	0	0	0
Tokens	0	1	3	8	6	13	9	6	14
No. words (millions)	7.6	12.6	15.5	16.8	16.1	13.6	11.4	9.7	6.4
Tokens per million	0.00	0.08	0.19	0.48	0.37	0.96	0.79	0.62	2.19

Although the increase from one token in the 1960s to three tokens in the 1990s is suggestive of a shift towards [+*for*], the data are of course rather sparse.

There are at least two ways to measure the increase in [+*for*]. First, we could measure it as a percentage of all [±*for*] with a given verb. For example, in the *Time* corpus, Table 3.9 shows the figures for [±*for*] as verbal complements with *ask* and a pronominal subject (e.g. *she asked for* ***him*** *to keep quiet*).

The percentage of tokens with [+*for*] is so small that they are not overly insightful. Rather than look at the percentage, I instead calculated the normalized frequency of [+*for*] with the verbs that allow both [+*for*] and [−*for*]. In other words, the writers had the option of choosing either complement, and they chose [+*for*]. Table 3.10 shows the frequencies by verb and decade for the search *[verb] [for]* * *{1,3} to [v**],[9] meaning that there are between 1 and 3 elements in the noun phrases between *for* and *to*.

Before summarizing these data in Figure 3.3, I should note that one problem with the data was the existence of full noun phrases with *ask* and *beg*, with examples like the following:

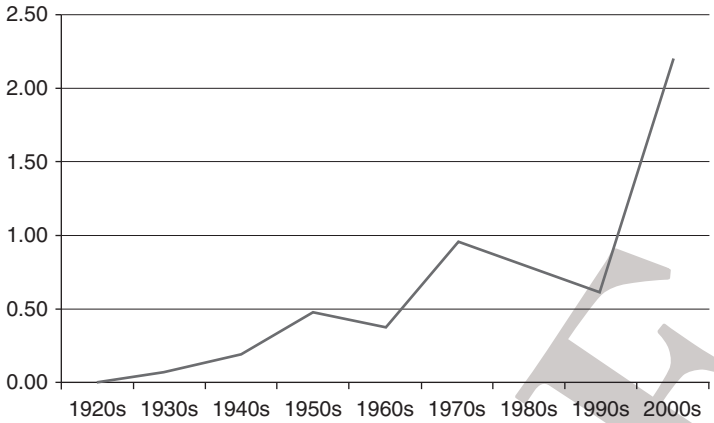

Figure 3.3. Frequency of [+*for*] in the *Time* corpus, with verbs allowing alternation

(7) a. Bevin *asked for a commission to investigate* the catering business from top to bottom. (1943)
b. Sir Cecil's second letter … *asked for international courts to try* arch war criminals. (1945)
c. Asia is a field that is almost *asking for an enemy to come* by night and sow tares in it. (1951)
d. I am *asking for the chairman to rule* that my question is pertinent. (1951)
e. A housewife was politely turned down when she *asked for a policeman to baby-sit* at her home while she went to the police station. (1957)

A sentence like *She asked for a policeman to babysit* is ambiguous between the following readings:

(8) a. 'she asked [$_{PP}$ for [$_{NP}$ a policeman]] (so as to get him) [$_{CLAUSE}$ to babysit]'
b. 'she asked [$_{CLAUSE}$ for a policeman to baby-sit]'

It would appear that interpretations like (8a) were much more common through the 1940s (see (7a–b) above), whereas starting in about the 1950s, interpretations like (8b) were more common (see (7c–d) above). And yet there are still cases like (7e), which can be interpreted in both ways. Due to the ambiguity, I have not included in the frequency count cases of *ask* and *beg* with full NPs.

The data from Table 3.10 can be summarized in Figure 3.3. This shows that (even with a small decrease in the 1990s), there clearly has been a

general increase in [+*for*] with infinitival verbal complements in American English (this chart shows the figures per million words).

3.7 Conclusion

As many previous studies have shown, small corpora such as the BROWN family of corpora, ARCHER, or the DCPSE can provide useful data for high- and medium-frequency constructions, such as modals and auxiliaries, and certainly much valuable work has been done in this area. For low-frequency constructions like verbal complementation, however, much larger corpora – such as the 100 million word *Time* corpus – are often necessary, as I have shown throughout this chapter.

As I mentioned previously, my goal in this chapter was *not* to provide in-depth and complete analyses of each of these constructions. Rather, the goal was to show how the *Time* corpus can provide the needed data, and suggest some possible areas of study for future researchers – such as possible motivations and mechanisms for syntactic shifts with these constructions.

With the [V NP *into* V-*ing*] construction (*we talked Bill into staying*), the BROWN corpora only yield 29 tokens with 21 matrix verbs, while the *Time* corpus yields 1,101 tokens with 173 matrix verbs. The data from *Time* show that the construction is in fact increasing over time (only hinted at in BROWN), and that the majority of this increase occurred from the 1920s to the 1940s, with more modest increases since then. With the four BROWN corpora that are widely available, of course, there are no data from before the 1960s, and because the corpora only sample the data every thirty years, there is no way to be more precise about the exact decades in which the increase was most pronounced. The *Time* data also show the extreme lexical creativity with this construction, involving matrix verbs like *heffle*, *hornswaggle*, *egg*, *sting*, *catspaw*, *preach*, *pound*, *smooch*, *cramp*, *dope*, *jar*, *jolly*, *gig*, *stiffen*, *drill*, *hook*, *Svengali*, *blarney*, *blow*, *euchre*, *sober*, and *poison*. With 173 different matrix verbs and 1,101 tokens, we can begin to answer questions about semantic differences with this construction between American and British English, raised by Wulff *et al.* (2007). With the sparse BROWN data, we can barely begin to look at semantic issues.

With the [V (*to* V/V-*ing*)] construction (*he started* [*to walk/walking*] *down the street*), I looked at 19 different matrix verbs (*begin*, *like*, *try*, *prefer*, etc.). Because of the small number of tokens, there is statistically significant data for only 3/19 verbs in the BROWN corpora. With *Time*, however, the data provides statistically significant data for 15/19 verbs. Perhaps more importantly, with the *Time* data, we can begin to see the chronology of [*to*-V] towards [V-*ing*] with the different matrix verbs. High-frequency aspectual verbs like *start* and *begin* underwent the greatest shift between the 1920s and the 1940s, but by the late 1900s the major shifts were with *propose*, *bother*, and the 'emotion' verbs *love*, *like*, *fear*, and *hate*. As with the

Great Vowel Shift – where the actual shift was more complex than simple vowel raising with all vowels at one time – the shifts with the different verbs in the Great Complement Shift in the 1800s–2000s have most likely occurred in more complex patterns as well, and only robust data from a corpus like *Time* allow us to see how (and perhaps why) the shift spread from one verb to another over time.

With the [V (*for*) NP to V] construction (*I'd really like (for) them to leave now*), we saw that there were only 4 tokens in BROWN, compared to 60 in *Time*. While 60 tokens is still not overly robust, it does allow us to see that the use of [+*for*] with verbs (where there is alternation) is definitely on the increase. In addition, with a much larger database, we can begin to look at when pragmatic shifts may have occurred, in which the complement clause shifted from being more adverbial (e.g. *Bevin asked for a commission to investigate the catering business*; 1943) to being more integrated as an object-like complement of the verb (e.g. *I am asking for the chairman to rule that my question is pertinent*; 1951), which in turn may provide important clues about the overall semantic and pragmatic motivations for the shift towards [+*for*] with these verbal complements.

In summary, with the increasing use of corpora such as the *Time* corpus, we can obtain robust data and begin to map out (and hopefully find motivations for) syntactic changes in a way that has not been possible before this time.