I haven't drank in weeks: the use of past tense forms as past participles in English corpora

Kristina Geeraert and John Newman

University of Alberta

Abstract

We investigate a relatively understudied phenomenon, the use of the (standard) past tense verb form as a (non-standard) past participle in English, as in I haven't drank in weeks *and refer to this phenomenon as "past tense spreading". We explore this phenomenon in some familiar, large corpora of English, as well as utilizing the World Wide Web as a corpus through the Google search engine. The corpus-based approach allows us to examine details in the behaviors of many verbs across genres and to identify degrees of spreading among verbs. The web searches reveal differential behaviors for high-frequency and low-frequency verbs with respect to past tense spreading, an example, we claim, of Bybee's (2006) Conserving Effect. Past tense spreading also occurs more than expected with modal auxiliaries, a pattern which would not be predicted based solely on the non-standard character of the phenomenon.*

1. Introduction[1]

This paper is a corpus-based study of a phenomenon in non-standard usage of English which we call here "past tense spreading" (PTS).[2] PTS refers to the use of the past tense form of a verb in place of a distinct, prescribed past participial form in perfect tenses, as in *I haven't drank in weeks*, rather than *I haven't drunk in weeks*. PTS is by no means restricted to the verb *drink* and is found with a number of verbs in contemporary English.

We identify the verbs which show evidence of PTS and the frequency of the phenomenon in two readily available large corpora: the British National Corpus (BNC, see Aston and Burnard 1998: 28-41) and the Corpus of Contemporary American English (COCA, see Davies 2010a). Since the number of instances of PTS is relatively small in both of these corpora, we find it necessary to use the World Wide Web (WWW or web) as a "corpus", using the Google search engine. We also investigate the distribution of PTS across genres and its frequency with modal auxiliaries.

2. Background

PTS, though a relatively peripheral phenomenon in English, has existed for centuries. The phenomenon appears to have begun towards the end of the Middle English period with some increase in use in Early Modern English (Lass 1994:

89, Lass 2008: 170). The Oxford English Dictionary Online (2010) records, for example, *broke* as having been used as a past participle since the end of the 14th century, *took* since the 16th century, and *drank* since approximately the 18th century. In contemporary usage, PTS is documented in numerous non-standard varieties worldwide, leading Wolfram (2003: 146) to comment “past for perfect, as in *they had went there*, occurs in socially subordinate varieties of English wherever they are found throughout the world”. PTS has been documented in the linguistics literature for a number of varieties of English, including the USA (Atwood 1953, Wolfram 2003), UK (Cheshire 1982: 46-49, Wright 1981: 118-120), and Australia (Eisikovits 1987). We believe it is very widespread, beyond just those cases described in the literature, though whether it is quite as widespread as Wolfram states remains an open question.

Research on PTS has received rather less attention than spreading in the opposite direction, namely the use of the past participle for the (prescriptive) past tense, as in *I drunk*, *I rung*, *I swum*, etc. (cf. Bybee and Slobin 1982; Bybee 1985, 1995). Research on these verbs and others like them has tended to focus on the phonetic structure of verbs which pattern in this way, in particular the presence of the vowel [ʌ], followed by a velar or a nasal consonant, in the past participial form. Anderwald (2007, 2009) further explores these verbs, which she fittingly calls “Bybee verbs”, using some corpus-based methods (mining the Freiburg English Dialect Corpus and the Survey of English Dialects, containing dialectal data from across Great Britain – see Anderwald 2009 for details). As part of this study, Anderwald (2009: 8-11) calculated the approximate percentage of the number of verb types belonging to an inflectional type, using for this purpose Quirk et al.’s (1985: 115-120) list of strong verbs. The possible inflectional types she uses are based on identity (or not) of the infinitival, past tense, and past participial forms within the strong verb paradigm. Figure 1 is an adaptation of Anderwald’s figure (Figure 1.1 in Anderwald 2009: 8) in which she summarizes her findings, with an example verb provided for each class. As can be seen in Figure 1, class (b) – precisely the group in which past tense and past participial forms are identical – contains the largest percentage of such verbs, where the percentage is calculated with respect to the number of verb types rather than the number of tokens in a corpus. Class (b), it will be noticed, is the class to which Bybee verbs are attracted (identical forms of *rung* in *I rung/I have rung*). It is also the class which many PTS verbs join (identical forms of *drank* in *I drank/I have drank*).[3] Class (b), therefore, is not only the largest one of the five, but it is also the class currently undergoing the largest increase in varieties of English where spreading between past tense and past participial forms occurs. Anderwald incorporates this observation in her account of the historical processes at work in the English verb paradigm, proposing that verbs from the other classes are in part motivated to undergo analogical leveling due to the comparatively large size of class (b), though other factors are also considered relevant. Another pertinent fact about the classes in Figure 1 is that the (e) class contains a greater complexity of forms (a three-way distinction, as in *drink*, *drank*, *drunk*) compared with the other classes. A change from class (e) to class (b), therefore, results in reduction of the complexity of the system in some sense (cf. the idea of "paradigm economy" as a factor motivating the evolution paradigms in Carstairs 1983). Simplification within the paradigm is also claimed to be indicative of non-standard usage (Cheshire 1994: 126).

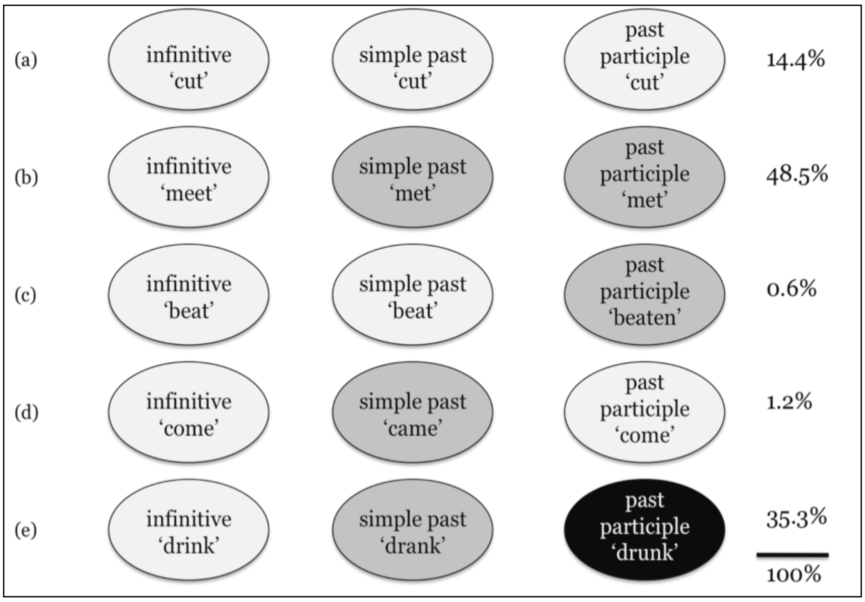

Figure 1: Identity of forms in the strong verb paradigm (adapted from Anderwald 2009)

It is clear that there are multiple forces which have shaped, and continue to shape, the variation in the form used as a past participle in the English perfect, in addition to any system-internal considerations such as economy within the paradigm. Miller (1987) illustrates the multi-layered complexity of the variation to be found with these verbs in his detailed account of the leveling at work in the English verb paradigm through three case studies of the verbs *bite*, *ride*, and *shrink*. Historical factors spanning England and the USA (all three verb paradigms have been leveling for centuries), geographical factors, social and ethnic factors, and education systems are all part of the larger story about each of these verb paradigms. A full account of PTS, in the spirit of Miller (1987), would necessarily involve many research methods (archival research, interviews targeting various populations, etc.).

The corpus-based approach which we have adopted in this paper can not possibly do justice to understanding PTS in its entirety or even do justice to all the distributional facts surrounding any one verb. We have demographic information for speakers in just one corpus (the speakers for the spoken part of the BNC). For the other corpora, we do not even have reliable information about exactly where the writers/speakers represented in the corpora are located, whether the writers/speakers are native-speakers, etc. It should be clear that we are working with a minimum of demographic information in the case of COCA and the web. However, a strength of the corpus-based approach lies in the fact that the corpora under consideration offer us a maximum amount of synchronic linguistic data from “ordinary” language use in the English-speaking world, with each corpus containing millions of words. Exploring PTS by means of these large corpora can reveal mainly “intra-linguistic” facts, i.e., facts from within the external products of language use, and can not reveal “extra-linguistic” facts (the “who, when, where, and why” behind the use). We readily acknowledge such limitations of the corpus-based approach to the study of PTS. We believe, nevertheless, that large corpora present a unique opportunity to explore PTS with reference to a large amount of naturally occurring data and for this reason a corpus-based approach should at least be tested in the way we proceed to do here (cf. also the informative overview of the use of corpora in the study of language variation in Bauer 2002).

3. Data and methodology

For the purposes of this study, we utilized three sources of data:

1. BNC. The BNC is a 100-million-word corpus of written and transcribed spoken British English. We accessed the BNC through Mark Davies’ website (http://corpus.byu.edu/bnc/) and made use of the five major genre categories of the BNC, as provided for in the Mark Davies’ interface at the time of carrying out this research in 2009-2010 (since changed): SPOKEN, FICTION, NEWSPAPER, ACADEMIC, MISCELLANEOUS.
2. COCA. The Corpus of Contemporary American (COCA) English contains over 400 million words (and growing). COCA was accessed in 2009-2010 through Mark Davies’ website (http://www.americancorpus.org). COCA allows for easy searches across five major (and relatively equally represented) genres: SPOKEN, FICTION, MAGAZINE, NEWSPAPER, ACADEMIC.
3. World Wide Web. We utilized Google’s search engine searching on the World Wide Web (WWW or web) in 2009-2010. Using the advanced search settings in Google, we restricted our searches to English websites.

Basically, we relied on one method of exploring PTS in these corpora which was by using individual search terms made up of fixed expressions, such as the sequences *have rode*, *has rode*, *had rode*, etc. for the non-standard forms and *have ridden, has ridden, had ridden*, etc. for the standard forms. This method had the advantage of directly targeting the patterns we were interested in, but it meant that we did not retrieve instances of the perfect where the auxiliary is separated by one or more intervening words, as in *Artists have in recent times dabbled in new media like acrylic*. We sacrificed the precision of our method in favor of the simplicity of searching for contiguous forms and reproducibility of results in all corpora.

Mark Davies' interfaces to the BNC and COCA allow for wildcard searches, as well as searches on parts of speech. For example, one can search for "vvd" (the part of speech tag for past tense of lexical verbs, i.e., verbs other than *be, have, do*, in both BNC and COCA), "vvn" for the past participle of the same verbs, as well as combinations of wildcard syntax with part of speech tags, e.g., "vv*" for all inflected forms of lexical verbs. We explored the possibility of utilizing such tags in the BNC and COCA, but we found the method too unreliable for our purposes. One needs to be aware, too, that the BNC and COCA have apparently been tagged using different algorithms and the difference in algorithms is directly relevant to the problem at hand. In the BNC, *took* in the sequence *we could have took over* (recorded as being spoken by an 80 year old retired miner) is tagged as a past tense. A comparable use of *took* in COCA *I would have took my two kids out of the house* (from the newspaper genre) is tagged as a past participle. Presumably, the tagging algorithm for the BNC relies primarily on the actual form of *took* to assign past tense status, whereas the algorithm for COCA relies rather on the occurrence in the frame *have X* to assign past participial status to a verbal *X*.

The Appendix lists all the verbs, their past tense and past participial forms considered as "standard" in this paper, selected from the same list used by Anderwald (2009), namely, Quirk et al. (1985: 115-120). Needless to say, we recognize that there is variation in usage (this is the starting point of our study, after all) and that some of these forms are local standards, so "standard" here simply means the reference forms assumed in this study. The (PTS) past tense forms *got*, *proved*, and *struck*, are well entrenched as past participles in the corpora used in this study, as opposed to *gotten*, *proven*, and *stricken*, respectively. In the BNC, for example, we found extraordinarily high percentages of use for the PTS forms: *got* (99%), *proved* (95%), and *struck* (99%). In other words, with these three verbs, the erstwhile past tense forms *got*, *proved*, and *struck*, represent the current dominant usage and we have not included these three verbs forms in the original list in Quirk et al. (1985) in our analysis.

There is currently much debate about using the web as a corpus (cf. Hundt, Nesselhauf, and Biewer 2007), with both pros and cons being highlighted. Some of the main criticisms leveled against relying on the web as a corpus are:

1. The web is "dirty" with numerous erroneous forms (Kilgarriff and Grefenstette 2003: 342).
2. Counts of the number of hits can be distorted due to the large amount of duplication on the web (Lüdeling, Evert & Baroni 2007: 14, Fletcher 2007: 31).

3. Search engines can be unreliable, returning substantially different counts for the same query on the same day (Kilgarriff 2007: 147-148), or returning hits that are not actually on the page itself, but rather contained in a link to the page (Keller & Lapata 2003: 469), or within titles or headings to these pages (Kilgarriff & Grefenstette 2003: 345).
4. Results are not returned in the format of easily readable concordance lines, a format much favored by corpus linguists.

Despite these issues, there is much to be said in favor of using the web as a corpus. Most obvious of all, the web provides data on a scale which is simply not matched by corpora such as BNC and COCA. Keller and Lapata (2003: 470) found evidence for claiming that counts obtained from the web were comparable to those obtained from standard corpora (a point we return to later). Importantly for our purposes, the magnitude of texts on the web makes the web particularly relevant when searching for relatively rare usage. Indeed, the web can be the only option to obtain data for a particularly rare phenomenon. Furthermore, some of the "dirty" forms referred to by Kilgarriff and Grefenstette (2003: 342) could be evidence of language change or evidence of a particular dialectal or regional use (Rosenbach 2007: 168-169), hence are potentially forms of some linguistic interest.

4. Results

4.1 Genre differences

Given the non-standard nature of the phenomenon under investigation, one would expect the more informal, spoken parts of the corpus to be where the phenomenon is most evident. We studied the occurrence of PTS forms in various genres in both the BNC and COCA. Figure 2 presents the results as mosaic plots, a simple but effective way to convey the relative proportions across categories. For these plots, the "expected" proportions are obtained by calculating the relative size of a genre compared to the whole corpus and applying that ratio to the total number of occurrences of PTS in the whole corpus. For example, in the BNC, the spoken part of the corpus consists of 10 million words = 1/10 of the total size of the BNC. The total number of PTS forms in the BNC is 323. Therefore, if PTS is proportionately distributed across all genres, we would expect 1/10 x 323 = 32.3 PTS forms in the spoken part of the corpus. In fact, 184 forms occur in the spoken part, so the "observed" size of PTS in the spoken part of the BNC (and COCA) is many times larger than the "expected" size, as shown in the darkest shaded portion in the top boxes of Figure 2.

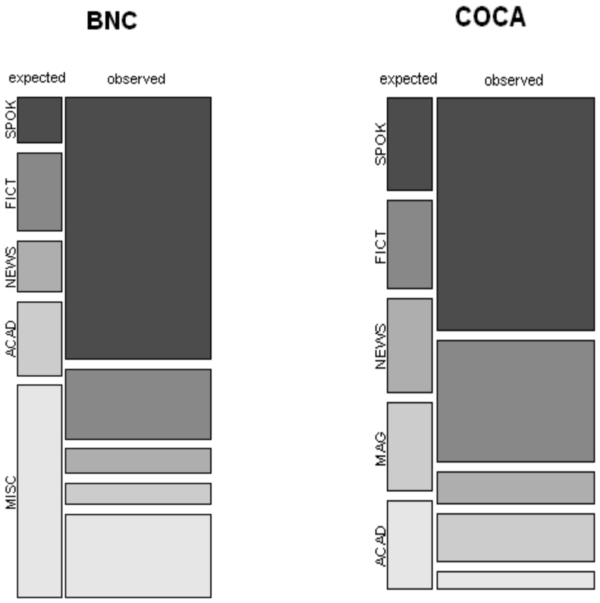

Figure 2: Observed and expected proportions of PTS use in BNC and COCA

What we see in Figure 2 is a snapshot of the corpus behavior of a particular non-standard usage: overrepresentation in the spoken part and variable degrees of representation in the other parts. In COCA, the fiction genre also has an overrepresentation of the non-standard, pointing perhaps to a greater tolerance and acceptance of non-standard forms in American usage, although one would need to follow up this observation with a closer study of genre preferences for individual instances of PTS. It may be surprising to see any occurrence of non-standard usage in the academic genre, but occurrences are found, in both BNC and COCA, and there can be various reasons for this. Sometimes, the non-standard form in the academic genre is reported speech, hence belongs more correctly to a conversational genre, as with *have wrote sic* in (1a), which occurs as part of a class drama where pupils are talking informally among themselves as part of the drama. The *sic* annotation here is part of the corpus itself. In (1b), *had wrote* also occurs as part of a direct quote, from the eighteenth century, a reminder as to how a corpus of contemporary Modern American English is not as clearly demarcated in time as one might expect. In (1c), *has showed* is part of serious academic writing in a journal.

1. (a) But a tribe member from the Onandage sic... said that we were taking their land and its sic realy sic ours and I have wrote sic

down 3 suggestions for it. [COCA/ACADEMIC, *Social Studies*, Sep/Oct 1991, *82*(5), p. 179, reporting on a class drama]

(b) He claimed Jekyll was “much incensed at the insolence of the Spanish protest which is to be considered in the Privy Council tomorrow August 18, 1737 by his Majesty. That Sir Jseph Jekyl sic had wrote a forcible letter to my Lord Chancellor Phillip Yorke, the first Earl Hardwicke on the occasion.” [COCA/ ACADEMIC, Thomas H. Wilkins, “Sir Joseph Jekyll and his Impact on Oglethorpe’s Georgia”, *Georgia Historical Quarterly* 2007, 91(2), 119-134]

(c) Although some observers have elaborated on the semi-democratic aspects of Palestinian rule, rather than a dictatorship, the Palestinian Authority has showed a certain degree of diversity in its rule. [COCA/ACADEMIC, Helena L. Schulz, “The ‘al-aqsa intifada’ as a result of politics of a transition”, *Arab Studies Quarterly* 2002, 24(4), p. 21]

4.2 PTS verbs

In a larger study on PTS verbs (Geeraert 2010), the first author examined around 50 PTS forms not just in constructions with the auxiliary HAVE as a lemma (subsuming all inflected forms), but in constructions with each inflected form of the auxiliary (*have drank*, *has drank*, *had drank*, and *’ve drank*). In the larger study, patterns with negatives were examined for each inflected form of the auxiliary (*hasn’t drank*, *hadn’t drank*, etc.), as well as for morphologically complex forms (*have underwent*, *have misspoke*, etc.). For the purposes of the present paper, however, we content ourselves with discussing a selection of these results just for the simplex form of the verb (e.g., *went*, not *underwent*) from the three corpora mentioned, with frequency numbers aggregated for inflected forms of the auxiliary, ignoring negative constructions (e.g., *haven’t went*).

In order to identify some main patterns in the fairly substantial amount of data obtained from the corpora, we begin by considering overall frequencies of the verbs and the extent to which they show evidence of PTS. In Figure 3, we plot the percentage use of PTS in the perfect construction against the frequency of the past participle, for all three corpora. The frequency count shown on the *X* axis is the ($\log_{10}$) frequency of the total number of past participial forms (standard + non-standard). The percentage use shown on the *Y* axis is based on the ratio of the frequency of the non-standard past participial forms relative to the total number of past participial forms (standard + non-standard). Each point represents one of the verbs, ordered from left to right in terms of increasing frequency on the *X* axis. This means that the verb represented by the nth point from left to right in one of these plots should not be equated with the verb represented by the nth point in either of the other plots, even though there can be quite a lot of overlap. Note that the scales for the *X* and *Y* axes in the three plots differ. This practice is usually avoided in presenting plots, but in this case, the main point to be made about these plots concerns the overall shape of the distribution of points and the axes have been adjusted to best reveal that shape. Extreme outliers have been removed from the BNC and COCA plots in order to reveal this distribution. These outliers include *bid*, *trod*, and *bade* in the BNC (with >6% PTS, see Table 1), and *bid*, *trod*, and *beat* in COCA (with >10% PTS, see Table 1).

Figure 3 presents an intriguing series of plots. In the BNC, there is a hint of higher PTS use in the lower frequencies (<2.5 on the log scale, about 316 in absolute frequency). In COCA, this trend seems a little more pronounced in the lower frequencies (<3, or 1,000 in absolute frequency). Using the Google search engine on the web, the trend appears quite marked, with PTS use noticeably higher, on the whole, in the lower frequencies (<6, or 1 million in absolute frequency). Of course, as one proceeds from the smaller corpus size to the larger corpus size, the absolute frequencies increase, quite dramatically in the case of the web. And the range of percentages of PTS use also varies noticeably between the BNC and COCA on the one hand and the Google searches on the other. However, a similarity in the shape of the distribution in each plot nevertheless emerges, with the logarithmic curve showing more PTS with lower verb frequencies.

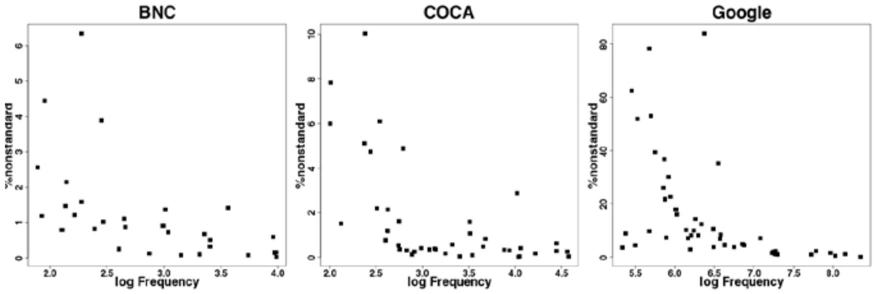

Figure 3: Scatterplots of % PTS x (log10) frequency of the perfect construction in the BNC, COCA, and the WWW (using the Google search engine). Note the different scales for the axes.

To appreciate better just what verbs occur where, we show the results from the web search in Figure 4. There are some verbs which appear as outliers, e.g., *bit*, *bid*, and *beat*. These three verbs have even higher PTS than verbs with similar frequencies and seem fairly well established as past participles. There is also a small group of verbs, *strode*, *wove*, *slew*, and *bade*, which appear as outliers in the lower left of the plot. These verbs have lower percentages compared with other verbs on the left side in this plot. Three of these verbs, *strode*, *wove*, and *slew*, would appear to be undergoing a different type of leveling to what we are exploring here, with *strided*, *weaved*, and *slayed* used more commonly as past participles. In the methodology we adopted for this study, only the past tense forms noted in the Appendix were searched and counted as possible past participles (*strode*, *wove*, and *slew* for these three verbs). So, in fact, a higher percentage of PTS use with these verbs would have been found if we had searched for *have strided*, *have weaved*, etc. *Bade*, however, appears to have a lower percentage of PTS because of the alternative verb form *bid*, which shows an 80% use with PTS. Putting aside these outliers, we see even more clearly the overall trend of higher PTS use in the lower frequencies and lower PTS use in the higher frequencies.

Even if each individual verb has its own unique history and set of sociolinguistic circumstances, the overall trend in Figure 3 calls for discussion. One might turn, most immediately, to Bybee's (2006: 715, 2007: 10-11) Conserving Effect as a motivating principle behind the trend: "high frequency sequences become more entrenched in their morphosyntactic structure and resist restructuring on the basis of productive patterns that might otherwise occur" (Bybee 2006: 715). One example of this principle from Bybee's work concerns regularization in the paradigm of the irregular verbs. For these verbs, Bybee observes how the regularization of the past tense suffixation of *-ed* to the stem of the base/present form, evident in the occurrence of *weeped*, *creeped*, *leaped*, etc., is resisted most in the case of high-frequency verbs where the past tense maintains a shortened vowel (*kept*, *slept*, etc.). Similarly, in our study, a high-frequency verb like *go* might be expected to "resist" a change to the structure of its perfect participial form more than a low-frequency verb like *ring* would. Hence, the argument would go, *have gone* remains in a high percentage of cases, whereas *have rung* has "succumbed" to the new formation *have rang*.

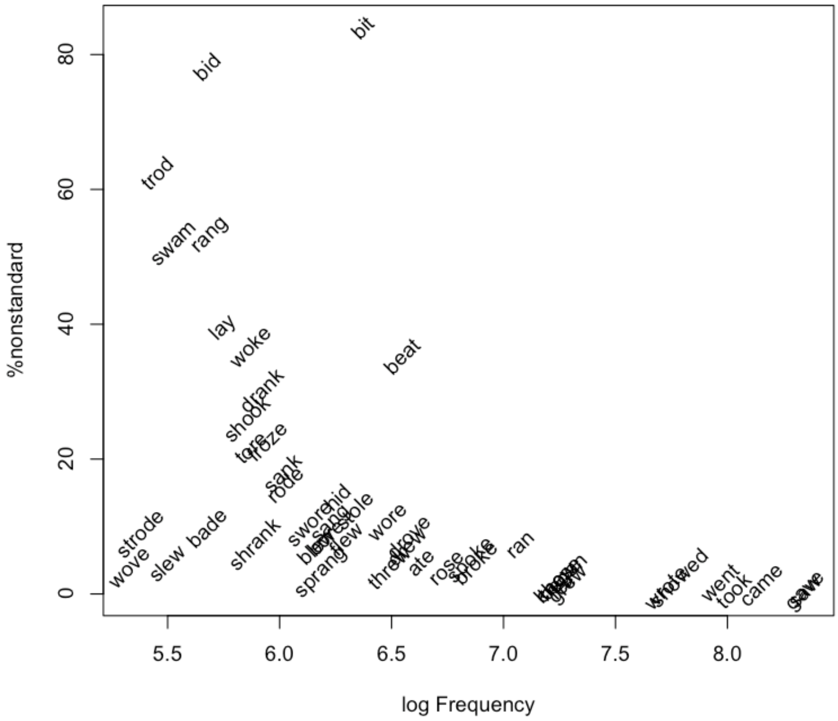

Figure 4: Detailed plot of PTS x Log$_{10}$ Frequency of verbs, based on WWW hits using Google search engine.

While a Conserving Effect, or some principle like it, is undoubtedly real, it does not explain away exactly the distribution we see in Figure 3. By itself, the Conserving Effect says nothing about the stages which typify the historical course of a change like PTS, e.g., whether PTS spreads through the verb system in a linear way by, say, increments of 10% per century or whether PTS spreads in an S-curve manner with slower early and final stages and rapid middle stages (cf. Denison 2003 for a critical review of the S-curve hypothesis). The pattern in the plots in Figure 3 represent a snapshot in time and can hardly be used as a basis for arguing for any particular diachronic sequence of stages of implementation. Figure 3 would be consistent, for example, with a linear increase in PTS or an S-curve type of increase, both broadly consistent with the Conserving Effect. But the pattern in Figure 3 is also consistent with a "stable, low-frequency" phenomenon, as opposed to a change in progress. The phenomenon is then not a change making its way gradually through the lexicon. By this account, the change is contained within certain boundaries defined by genre, domain of "standard" usage, and the effects of stigmatized usage.

Let us turn to a more detailed presentation of some key results, beginning with Table 1, showing the verbs most prone to PTS in the BNC and COCA (with a PTS rate higher than 1%). A number of observations may be made about the two lists in Table 1. Notice, to begin with, that there is considerable overlap in the two lists – something which could not be inferred from Figure 3. All the verbs listed under BNC, except *swore*, *ate*, and *sank*, also occur under COCA. These three verbs do occur with PTS in COCA, just not with a percentage higher than 1%. The raw frequency of occurrence of the non-standard form can be small, e.g., 1 in the case of past participial *tore* in the BNC, but there are 85 occurrences altogether of the perfect construction with *tore/torn*, so there are a robust number of corpus occurrences of the construction. Clearly, though, the non-standard usage is very infrequent for most of these verbs. There are two verbs, however, that stand out as having significantly higher percentages of PTS than the rest, namely *bid* and *trod*, in both the BNC and COCA. These verbs were not considered "standard", like *got*, *proved*, and *struck* mentioned in Section 3, but appear close to achieving that "standard" status, especially within COCA.

Among the verbs showing PTS in the BNC and COCA, we have some basis here for identifying the lists in Table 1 as representing the verbs which are most advanced as far as PTS is concerned. It is precisely this kind of quantification of trends as seen in a broad spectrum of language use (even across continents!) where the corpus-based approach shows its strength. It is tempting to speculate on the phonetic and orthographic similarity of some past participles in Table 1. Two main phonetic patterns emerge. The first pattern is illustrated by *bid*, *beat*, *bit*, and *hid*. This pattern has a high front vowel followed by an alveolar stop. The second pattern is characterized by the verbs *drank*, *rang*, *sang*, *sank*, and *swam*, all of which have a low front vowel followed by a velar or a nasal (similar to the Bybee verbs mentioned earlier).[4] These phonetic patterns might be seen as phonetic templates functioning as prototypes within the PTS verbs (cf. Bybee & Thompson 2000: 384, Bybee 2010: 79), with the *bid–bit–beat–hid* group representing the most central exemplars.

Table 1: Non-standard past participial use of lexical verbs in BNC and COCA >1% of combined standard and non-standard use. Frequencies are based on co-occurrence with all inflected forms of the auxiliary HAVE.

BNC	*freq*	*% non-standard*	*COCA*	*freq*	*% non-standard*
bid	15	60.00	*bid*	89	89.00
trod	12	32.43	*trod*	36	76.60
bade	3	12.00	*beat*	150	14.07
rang	12	6.35	*bit*	26	10.04
bit	4	4.44	*swam*	8	7.84
beat	11	3.79	*drank*	21	6.10
rode	3	2.14	*bade*	6	6.00
froze	2	2.56	*woke*	12	5.11
drank	3	1.58	*hid*	30	4.88
swore	2	1.46	*sang*	13	4.73
showed	52	1.42	*showed*	301	2.87
broke	14	1.36	*froze*	7	2.18
sank	2	1.22	*shook*	9	2.13
tore	1	1.18	*rode*	9	1.61
ate	5	1.11	*broke*	51	1.58
			rang	2	1.52
			lay	5	1.20
			tore	5	1.18
			spoke	35	1.07

We decided to probe more deeply into the web behaviors of selected low-frequency and high-frequency verbs by examining the percentage of PTS use in three Internet domains (.com, .ca, .uk). We are aware that .com covers more than just American-based websites. Still, we relied on this web domain to represent American English usage in preference to domains such as .org, .edu, .gov, etc., since we felt that the latter are less likely to yield the non-standard usage being targeted in this paper. We chose five representative verbs from each of the low-frequency and high-frequency groups for this purpose, with results summarized in Tables 2 and 3, respectively. Percent PTS use in COCA is also included for the sake of comparison. As can be seen from these results, there is a surprisingly high degree of consistency across the three Internet domains in the behavior of these verbs, for both frequency groups.

Table 2: Five low-frequency verbs and the frequencies/percentages of their use as past participles using the WWW (searched by Google), compared with percentages in COCA. Frequencies are based on co-occurrence with all inflected forms of the auxiliary HAVE.

	.com		*.ca*		*.uk*		*COCA*
	Freq.	*% PTS*	*Freq.*	*% PTS*	*Freq.*	*% PTS*	*% PTS*
bid	330,600	**77.29**	5,388	**89.50**	37,235	**92.06**	**89.00**
bit	1,806,700	**83.99**	39,821	**89.35**	125,410	**81.62**	**10.04**
rang	217,400	**52.08**	1,399	**34.28**	43,280	**59.04**	**1.52**
swam	163,900	**54.07**	1,685	**43.46**	9,757	**31.06**	**7.84**
trod	164,900	**63.42**	2,269	**71.15**	9,970	**47.41**	**76.60**

Table 3: Five high-frequency verbs and the frequencies/percentages of their use as past participles using the WWW (searched by Google), compared with percentages in COCA. Frequencies are based on co-occurrence with all inflected forms of the auxiliary have.

	.com		*.ca*		*.uk*		*COCA*
	Freq.	*% PTS*	*Freq.*	*% PTS*	*Freq.*	*% PTS*	*% PTS*
came	1,513,000	**1.12**	26,866	**0.91**	46,640	**0.95**	**0.26**
saw	360,300	**0.17**	7,289	**0.23**	13,510	**0.21**	**0.04**
took	421,400	**0.43**	10,222	**0.37**	25,560	**0.49**	**0.27**
went	1,391,000	**1.61**	38,070	**1.97**	29,470	**0.67**	**0.62**
wrote	432,500	**0.88**	13,577	**1.06**	17,790	**0.77**	**0.31**

One difference between the BNC and COCA that is worth a special mention concerns the use of PTS for the verb *do*, omitted from Table 1, which reported only on full “lexical” verbs. In the BNC, the sequence HAVE *did* rarely occurs as a past participle. Instead, such sequences, to the extent they occur, usually are part of different constructions and must be categorized accordingly, as shown in (2).

(2) a. cleft constructions: *The job they had, didn’t interest me.*
b. repetitions: *Did that have, did that have quite an impact.*
c. self-corrections: *So they had – did have a variety of different things.*

d. run-on sentences: *That was the first inkling we ever had – Did she have her cloth in hand?*
e. tag-questions: *Yeah we have, didn't you?*
f. interruptions: *At the moment Lincoln have [Did you know there's someone waiting for you at Radio].*

There are just three instances in the BNC that show *did* occurring in a past participle: *I've did er night shift once*; *...because the ones that we've did cost like I think it was eight hundred pounds*; and *you've did your undergraduate work at York*. The majority of instances of HAVE *did* in the BNC are not instances of PTS.

In COCA, however, the sequence HAVE *did* occurs quite frequently as a genuine past participle (89 occurrences in COCA), as in the examples in (3).

(3) a. he s got to protect his client, and he *should have did* it right then and then to start the clock ticking. (COCA: SPOKEN, 2008, CNN Nancy Grace)
b. if they wanted to express that way, they *should have did* it in another form (COCA: SPOKEN, 2008, NPR TalkNation)
c. There were a lot of things that night we *could have did* differently, and we should have (COCA: SPOKEN, 1998, ABC 20/20)

4.3 Modals and PTS verbs

One particularly intriguing usage context associated with a higher-than-expected incidence of PTS involves the presence of modal auxiliaries, as in *it must have took about three years, my father could have went to jail, it should not have came to this*, and *could one person have ran a massive scheme* Eisikovits (1987: 23) had already noticed in her analysis of Inner-Sydney English that PTS occurred in this environment more often than one would expect. We found a similar pattern in our data. For example, out of the 81 instances of *have went* in COCA, 67 (82.7%) occur in constructions with a modal auxiliary. On the other hand, 3,623 (just 40.9%) out of the 8,860 instances of *have gone* occur with a modal. This amounts to a *percent difference* of 41.8% in favor of co-occurrence of PTS with a modal auxiliary. We rely on this simple statistical measurement of percent difference to report on the use of PTS and modal auxiliaries in COCA, though the same patterns were evident in all our corpora. This statistical measurement seemed most appropriate for this analysis as some verbs had extremely large frequencies which would lead too easily to statistical significance using, say, chi-square tests. We searched for all modal auxiliaries within five word positions to the left of *have*, (the "L1-L5" range), individually inspecting forms with lower frequency but relying on automatic retrieval of forms with larger frequencies. Only verbs which showed a minimum of three instances were used in these calculations of percent differences. The results are summarized in (4a-b):

4. (a) Preferred use of non-standard past participles in the context of a modal, with percent difference compared with standard past participial form in a modal context:
froze (64%), *shook* (61%), *saw* (61%), *drove* (60%), *fell* (60%), *swam* (59%), *ran* (50%), *hid* (47%), *gave* (46%), *came* (43%), *went* (42%), *wrote* (40%), *chose* (39%), *took* (39%), *showed* (33%), *broke* (31%), *sang* (31%), *beat* (31%), *began* (19%), *drank* (15%)

(b) Preferred use of standard past participles in the context of a modal, with percent difference compared with non-standard past participial form in a modal context:
trodden (40%), *bitten* (33%), *woken* (24%), *worn* (11%), *spoken* (7%)

The preference for non-standard past participial use in the modal auxiliary context, as seen in (4), is striking. It is not a matter of a subtle point or two in percent difference which is the basis for this conclusion – the percent differences are large, staggeringly so in some cases. Four modals appear most frequent with PTS: *would*, *could*, *should*, and *must* (which have variable overall frequencies in COCA, as reported in Davies 2010b).

We leave unanswered the question of *why* the non-standard form is so strongly associated with the modal auxiliary context. One fact that would seem relevant to answering this question in future research is the phonetic reduction that occurs, often but not always, in the infinitival *have* in these constructions, reflected in the orthographic representations *could've*, *would'v*e, etc., or the reanalyzed variants *could of*, *would of*, etc. Or perhaps the presence of the modal auxiliary introduces an extra degree of processing complexity which goes hand-in-hand with the selection of a more colloquial, more easily accessible variant of the verb. Whatever the explanation might be, it seems that we must acknowledge a recurring constructional schema in these cases with PTS firmly entrenched for some speakers.

5. Conclusions

The advent of corpora and the potential of using the web as a corpus have created the opportunity to use data from these corpora to explore non-standard English usage. Corpora such as the BNC and COCA and the web have, of course, not been constructed for the explicit purpose of studying non-standard usage – usually we would collect data in different ways if we were documenting and analyzing non-standard usage, e.g., interviews which target non-standard usage, specialized corpora based on local usage, etc. It is, therefore, of interest to learn what can be discovered about non-standard usage using these kinds of general, all-purpose resources.

As would be expected of non-standard usage, PTS is most common in the spoken genre of corpora which distinguish genres, such as BNC and COCA. But, interestingly, PTS occurs in all the genres of these corpora, reflecting in part the complexity and subtlety associated with some “genres” and their labels in corpus linguistics. The ACADEMIC genre of COCA, for example, can include quoted speech (including very colloquial speech) embedded in a piece of academic writing, giving rise to informal, colloquial variants which would not normally be found in highly self-conscious, planned, and edited written academic discourse. The presence of PTS in an academic sub-corpus of English, therefore, must be interpreted cautiously: some occurrences may reflect a new or emerging standard usage, but other occurrences may be “posing” as academic writing when they really belong to a conversational genre. It is a reminder of just how mixed the catch can be when we trawl in an ocean.

Relying on a corpus-based approach, we were able to identify the verbs in both BNC and COCA which are most prone to PTS. Despite some variation between the British and American varieties of English represented in these two corpora, there is a high degree of agreement in the results from the two corpora. We took a percentage of >1% PTS use as a basis for identifying verbs most prone to PTS (cf. Table 1) and found that there were 15 such verbs in the BNC and 19 in COCA. Two verbs had an exceptionally high percentage of PTS, *bid* and *trod*, while all other verbs in Table 1 had below 15%, suggesting that PTS remains a relatively peripheral phenomenon in these corpora. Of particular interest, though, was the comparison of PTS usage in these corpora with Google-based searches. On the web, we found a much greater range in percentage of PTS use and differential behaviors for high-frequency and low-frequency verbs. Specifically, we found that high-frequency verbs evidence PTS considerably less often than low-frequency verbs, a pattern which is discernible, though to a lesser degree, in the BNC and COCA results. For example, a high-frequency verb like *go* manifests relatively less usage of the non-standard perfect construction *have went* when compared with a low-frequency verb like *ring* and its non-standard perfect *have rang*. We believe this pattern is evidence of Bybee’s Conserving Effect coupled, most likely, with the stable but marginal character of PTS. Additionally, one should explore potential correlations of PTS use with other frequency measures, such as frequencies of the past tense forms or the lemma. Recall that we used only the frequency of the combined standard plus non-standard past participial usage in the present study and this may not be the only relevant or optimal measure. In any case, other factors, beyond what we can measure using corpus-based methodologies, are most certainly relevant, as mentioned in Section 2. The intriguing strong association between a modal auxiliary context and PTS also invites further analysis, especially experimental analyses which investigate speakers’ mental processing of modal vs. non-modal contexts of usage.

Notes

[1] The authors would like to thank participants at the AACL 2009 conference and anonymous reviewers for helpful suggestions on earlier versions of this paper.

[2] Past tense spreading is referred to as “preterite shift” in Lass (1994: 89).

[3] Note that past tense spreading in some verbs results in one single form in infinitival, past tense, and past participial forms, such as *have beat*. In such cases, the forms migrate to class (a) in Figure 1.

[4] There could potentially be a third pattern, all containing the vowel phoneme /ow/ (*rode*, *froze*, *showed*, *broke*, *spoke*). The preceding and following consonants around the vowel are quite varied in these cases, to the point that this pattern is less distinctive, phonetically, than the other two patterns.

References

Anderwald, L. (2007), ‘‘He rung the bell’ and ‘she drunk ale’ – non-standard past tense forms in traditional British dialects and on the internet’, in: M. Hundt, N. Nesselhauf & C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam: Rodopi. 271-286.

Anderwald, L. (2009), *The morphology of English dialects: verb formation in non-standard English*. Cambridge: Cambridge University Press.

Aston, G. & L. Burnard (1998), *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atwood, E. B. (1953), *A survey of verb forms in the eastern United States*. Ann Arbor: University of Michigan Press.

Bauer, L. (2002), ‘Inferring variation and change from public corpora’, in: J. K. Chambers, P. Trudgill, & N. Schilling-Estes (eds.) *The handbook of language variation and change*. Malden and Oxford: Blackwell. 97-114.

Bybee, J. (1985), *Morphology: a study of the relation between meaning and form*. Amsterdam: Benjamins.

Bybee, J. (1995), ‘Regular morphology and the lexicon’, *Language and cognitive processes*, 10: 425-455.

Bybee, J. (2006), ‘From usage to grammar: the mind’s response to repetition’, *Language*, 82(4): 711-733.

Bybee, J. (2007), *Frequency of use and the organization of language*. Oxford: Oxford University Press.

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

Bybee, J. L. & D. I. Slobin (1982), ‘Rules and schemas in the development and use of the English past tense’, *Language*, 58(2): 265-289.

Bybee, J. & S. Thompson (2000), ‘Three frequency effects in syntax’, *Berkeley linguistics society*, 23: 65-85.

Carstairs, A. (1983), ‘Paradigm economy’, *Journal of linguistics*, 19(1): 115-128.

Cheshire, J. (1982), *Variation in an English dialect: a sociolinguistic study*. Cambridge: Cambridge University Press.

Cheshire, J. (1994), ‘Standardization and the English irregular verbs’, in: D. Stein & I. Tieken-Boon van Ostade (eds.) *Towards a standard English: 1600-1800*. Berlin: Mouton. 115-133.

Davies, M. (2010a), ‘The Corpus of Contemporary American English as the first reliable monitor corpus of English’, *Literary and linguistic computing,* 25(4): 447-464.

Davies, M. (2010b), *Word frequency lists and dictionary from the Corpus of Contemporary American English*. Available on line at: http://www.wordfrequency.info.

Denison, D. (2003), ‘Log(ist)ic and simplistic S-curves’, in: R. Hickey (ed.) *Motives for language change*. Cambridge: Cambridge University Press. 54-70.

Eisikovits, E. (1987), ‘Variations in the lexical verb in inner-Sydney English’, *Australian journal of linguistics*, 7: 1-24.

Fletcher, W. (2007), ‘Concordancing the web: promise and problems, tools and techniques’, in: M. Hundt, N. Nesselhauf & C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam: Rodopi. 25-45.

Geeraert, K. (2010), I haven’t drank in weeks: *preterite shift in English*, MSc thesis, University of Alberta.

Hundt, M., N. Nesselhauf & C. Biewer (eds.) (2007), *Corpus linguistics and the web*. Amsterdam: Rodopi.

Keller, F. & M. Lapata (2003), ‘Using the web to obtain frequencies for unseen bigrams’, *Computational linguistics*, 29(3): 459-484.

Kilgarriff, A. (2007), ‘Googleology is bad science’, *Computational linguistics,* 33(1): 147-151.

Kilgarriff, A. & G. Grefenstette (2003), ‘Introduction to the special issue on the web as corpus’, *Computational linguistics*, 29(3): 333-347.

Lass, R. (1994), ‘Proliferation and option-cutting: the strong verb in the fifteenth to eighteenth centuries’, in: D. Stein & I. Tieken-Boon van Ostade (eds.) *Towards a standard English: 1600-1800*. Berlin: Mouton. 81-113.

Lass, R. (2008), ‘Phonology and morphology’, in: R. Lass (ed.) *The Cambridge history of the English language volume 3: 1476-1776*. Cambridge: Cambridge University Press. 56-186.

Lüdeling, A., S. Evert & M. Baroni (2007), ‘Using web data for linguistic purposes’, in: M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam: Rodopi. 7-24.

Miller, M. I. (1987), ‘Three changing verbs: bite, ride, and shrink’, *Journal of English linguistics*, 20(3): 3-12.

Oxford English Dictionary online, 2nd ed. (OED) (2010). Available on line at: http://www.oed.com/.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik (1985). *A comprehensive grammar of the English language*. London: Longman.

Rosenbach, A. (2007), ‘Exploring constructions on the web: a case study’, in: M. Hundt, N. Nesselhauf & C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam: Rodopi. 167-190.

Wolfram, W. (2003), ‘Enclave dialect communities in the south’, in: S. Nagle & S. Sanders (eds.) *English in the southern United States*. Cambridge: Cambridge University Press. 141-158.

Wright, P. (1981), *Cockney dialect and slang*. London: B. T. Batsford.

Appendix

Verbs and their past tense and past participial forms, as assumed in this study

bear bore borne

beat beat beaten

begin began begun

bid bid/bade bidden

bite bit bitten

blow blew blown

break broke broken

choose chose chosen

come came come

draw drew drawn

drink drank drunk

drive drove driven

eat ate eaten

fall fell fallen

fly flew flown

freeze froze frozen

give gave given

go went gone

grow grew grown

hide hid hidden

know knew known

lie lay lain

ride rode ridden

ring rang rung

rise rose risen

run ran run

see saw seen

shake shook shaken

show showed shown

shrink shrank shrunk

sing sang sung

sink sank sunk

slay slew slain

speak spoke spoken

spring sprang sprung

steal stole stolen

stride strode stridden

swear swore sworn

swim swam swum

take took taken

tear tore torn

throw threw thrown

tread trod trodden

wake woke woken

wear wore worn

weave wove woven

write wrote written