STUDYING RARE LINGUISTIC PHENOMENA WITH CORPORA

Nadina CEHAN, PhD Candidate, "Babeş-Bolyai" University of Cluj-Napoca

Abstract: The study sets out to investigate the 'plupluperfect' structure, which is a relatively rare nonstandard feature found in counterfactual conditionals. It looks into this construction's geographical spread using available corpora. As the study unfolds, the strengths and weaknesses of the method come to the fore, as well as the limits of this kind of empirical linguistic analyses.

Keywords: corpus linguistics, plupluperfect, nonstandard English

Introduction

Due to relatively recent technological developments which have allowed for large amounts of data to be stored and easily searched, corpora can now be used to collect information about both common and rare linguistic phenomena. One such rarity is the so-called plupluperfect (*cf.* Christophersen 1986, Declerck 2006) construction which appears in past counterfactual conditionals, as exemplified here:

(1) The thought crept into him that it would have been better *if* there *had have been* trees outside. (BNC, ADA, W_fict_prose)

(2) *If* it *had have been*, *if* we*'d have known* we'd have pulled the tables further, pulled the tables further across that way. (BNC, HYD, S_meeting)

(3) So *if* you *hadn't have been* prepared, you wouldn't have been lucky. (COCA, 2003, SPOK, Ind_Oprah)

(4) But hey, you can't buy this kind of publicity. *If* I *hadn't have gotten* it, maybe somebody else might've. (COCA, 2011, FIC, FantasySciFi)

(5) I would have worked harder *if* I *had have known* that life could be so soon over. (Strathy, 1945, NF, StreamRunsFast)

Apart from the third conditional, no other environments are known to trigger the doubling of the perfect. This nonstandard lives alongside the intrusive *would* after *if*, in contexts such as:

(6) *If* more *would have gone* back then the strike would have come to an end quicker. (BNC, HMM, S_interview_oral_history)

Whereas *would* is also expected after *if* where *if* can be replaced by *whether*, as in:

(7) I don't know *if* it *would have been* there in Flaubert's day. (BNC, G1A, W_fict_prose)

no such possibility exists for the plupluperfect. In other words, it will never follow in sentences introduced by *I wonder if*, *I don't know if* and others alike.

When contracted, *had* is identical in form with *would*, and both auxiliaries are possible after conditional *if* in nonstandard English. Therefore, the contracted form *'d* was ignored and only the full form of the plupluperfect was introduced as a search term in the available corpora. The aim was to reveal the frequency and spread of the plupluperfect in the different extant varieties of English, both at a national and international level. Four easily accessible, online corpora were used: the BNC (http://corpus.byu.edu/bnc/), COCA

(http://corpus2.byu.edu/coca/), Strathy Corpus of Canadian English (http://corpus2.byu.edu/can/) and GloWbE (http://corpus2.byu.edu/glowbe/), all hosted by Brigham Young University. The same platform is used to access all of them, meaning that they can be inquired in an identical manner, introducing the same search phrases every time. This is highly advantageous, as once the search syntax is learnt and decided upon, the researcher can confidently expect similar results to be returned across the corpora.

Accessing the corpora

In order to enhance search ease, the corpora have been annotated with CLAWS 7 (Constituent Likelihood Automatic Word-tagging System), a descendant of the first part-of-speech tagger used to annotate the LOB (Lancaster-Oslo-Bergen) corpus in the 1980s. There are many benefits in having access to an annotated corpus, as the following brief discussion will show, yet certain limitations are also worth pointing out.

A relevant particularity of the CLAWS 7 tagset is that it differentiates between verbs which can be auxiliaries and lexical verbs. *Be*, *do* and *have* are each treated separately and given unique tags for each of their forms. Consequently, when looking for the past participle, the search term *[vvn]* will only return the past participles of so-called lexical verbs, such as *studied* and *come*. *Been*, *done* and *had* are excluded. This is somewhat surprising, since they too can be lexical verbs in certain contexts. The instances where they behave as auxiliaries are well-defined within larger grammatical patterns and in addition an end-user may always input the exact form, rather than the tag, as a search term, which makes the *[vbn]*, *[vdn]* and *[vhn]* seem useless (for a detailed explanation of the tags, visit *CLAWS part-of-speech tagger for English* and *UCREL CLAWS 7 Tagset* websites; for an in-depth discussion on tagging, see Leech, Garside and Bryant 1994 and Garside 1996). In practice, what needs to be done is to include a wildcard, that is to type *[v*n]* so as to retrieve all past participles.

Complex systemic structures are not easy to analyse in corpora, since they typically involve quite a few variables and are subject to syntactic transformations, such as inversion. Nevertheless, they would be nearly impossible to identify and analyse if the corpus were not annotated. For the structure in question, it does not make much of a difference if as search terms we introduce *if * had have* or *if * had have [v*n]*, since the *had have* group is already abnormal in an *if*-clause, but some possible irrelevant cases are nevertheless avoided. However, one could not properly identify standard third conditional clauses without the *[v*n]* tag, since *if * had* would inevitably include *if*-clauses of the second type where *had* can be followed by any noun phrase. Moreover, the possibility to search for any past participles is useful when looking for inversion of the plupluperfect structure, by keying in *had * have [v*n]*, which rids final results of hundreds of irrelevant entries.

Because the tagger couples each word with a part-of-speech category, certain higher-level analyses cannot be run in a straightforward manner, but rather require the researcher to be creative and pattern-aware. The plupluperfect, as a construction, can be minimally coded *as if [Subject] had have [past participle]*, *if [Subject] had not/n't have [past participle]*, and for inversion *had [Subject] have [past participle]*, *had [Subject] not have [past participle]*. While for the past participle there is a tag, the syntactic function of Subject is not signaled in any way. Nevertheless, it can be expressed by nouns and personal pronouns, as well as by more complex noun phrases. *If* does not usually allow anything to come between it and the

Subject, with the exception of words such as *only*, *by chance* and *perhaps*. Taking these and the plupluperfect structure together is not necessary, since both *if by chance* and *if perhaps* are rare enough to allow for a quick check of all the entries where they appear. Subjects expressed by a single word can be searched for with only a wildcard, while Subjects expressed by noun phrases can be searched for with * *[nn*]*, where the wildcard is expected to retrieve any article, determiner, possessive pronoun or adjective that may come before any noun, as given by *[nn*]*. In order to check for even more complex, three-element noun phrases, all is needed is the addition of a wildcard before the noun tag, due to the robustness of the *if [Subject]* pattern. Personal pronouns could also form part of a multi-word Subject (*e.g. If pitiful me had known…*), which requires the use of the *[pp*]*, instead of *[nn*]*. Finally, entering *if * [nn*] / [pp*] had have [v*n]* and their variants will also show instances with *if only*.

It may also be worth pointing out that punctuation is not treated in the same way in the considered corpora. In some cases it seems to be ignored completely, as in the BNC, in others it is not. Usually, the wildcard (*) by itself does not return any punctuation marks, but in GloWbE it does. This has no serious effect on the analysis, but it does mean that even more useless results need to be ignored in the final tally.

The British National Corpus

The BNC, compiled in the 1990s, contains a wide variety of texts which together amount to around 100 million words, 10 million of which are spoken language. Of these, approximately half come from naturally occurring conversations (*i.e.* less than 5% of the entire corpus). The search for the plupluperfect construction (with *had* full form, not contracted to *'d*), yielded a total of 40 instances, 10 of which came from written sources and 30 from spoken situations. This result reinforces the idea that nonstandard forms are a lot less likely to be encountered in writing.

The BNC provides the opportunity to see who the speakers were in terms of age, gender, occupation and dialect. Unfortunately, the records are not complete for all and the dialect could not be identified for 13 of them. For the rest, the regional distribution is the following (the list is in alphabetical order):

Dialect / Accent	**Number of speakers**
Central Midlands	2
Central Northern England	2
East Anglian	1
Home Counties	1
Irish	4
Lancashire	2
London	1
Lower south-west England	1
Merseyside	1
South Midlands	1
Welsh	1

Table 1. Regional varieties of British English where the plupluperfect is found.

The list documents the fact that the nonstandard plupluperfect is widely spread in the British Isles. Although it does not seem to be able to provide a clear picture, the strength of the analysis can be verified in other ways. For instance, since Scottish is not on the list, one can use the SCOTS (http://www.scottishcorpus.ac.uk/) corpus to verify whether the plupluperfect is found in the region, an inquiry which returned no results.

If the above returns of the plupluperfect should appear disappointing in any way, it is worth remembering that standard conditionals are in themselves rare structures, as compared to others (*cf.* Biber 1993), and the nonstandard form is expected to be even rarer, since most corpora, the BNC included, rely heavily on published, edited material. Moreover, dialect can be ascertained only when interviewing native speakers, and in this particular case, there was only a fraction of the corpus available for study and incomplete records.

The Corpus of Contemporary American English

More than 450 million words in length, COCA is a monitor corpus which was released in 2008. The number of plupluperfects found was quite low: a total of 42 instances of the construction were identified, 33 of which belong to the spoken part. This is very close to what the BNC returned, which is quite unexpected given the difference in size. Unfortunately, the BNC and COCA cannot be compared, since the spoken component of COCA consists of transcripted naturally occurring language from TV and radio shows. In such situations, people tend to be more careful with their language and their contributions are influenced by the format of the show they appear on. Very little personal information is ever shared (*cf.* Lindquist 2009). Nevertheless, the low frequency of plupluperfects may in fact be characteristic of American English.

It is impossible to find out the backgrounds of the speakers. Hence, not even a tentative picture of the nonstandard's geographical spread in the USA can be given. It may be the case that it has diffused throughout the country, or that there are a few pockets of tight communities using it. It may also be worth mentioning that of the entries which appear in the written news section of the corpus, most plupluperfects appear within quoted discourse, suggesting that the reporters either did not edit the speech, or that they did not notice the nonstandard so as to correct it (*cf.* Ishihara 2003 on the nonstandard third conditional with *would* which is typically overlooked by native speakers).

The Strathy Corpus of Canadian English

Geographical spread can be considered on a smaller or a larger scale, at a national or international level. Should the nonstandard plupluperfect be found in other native-English speaking countries, such as Canada, it would contribute to the evidence that this construction is old, as it travelled with British-born immigrants within the growing Empire, and has spread the world over. Strathy is a 50 million word corpus of Canadian English in which a total of 11 instances of the construction were found, 5 of which belong to the spoken component. The smaller number of returns is unremarkable, given the size of this corpus.

As with COCA, there is no information available about the speakers' backgrounds, which impedes further analysis. Another similarity is that three of the results in the news section evince the nonstandard within quotation marks, meaning that they could be accepted

as instances of spoken language. Furthermore, the searches indicated that there are some glitches in the corpus' construction, as the search for *if * [pp] had have [v*n]* returned

(8) they would be far fewer than -- than would have been present *if --if they had have been* present as it were in -- in November. (Strathy, WalkertonInquiry, 2001)

This instance of the nonstandard did not show up when the phrase *if * had have [v*n]* was typed in because the *--if* element is mistakenly analysed as a whole by the tagger.

Corpus of Global Web-based English

GloWbE is a 1.9 billion word corpus based on content taken from websites from twenty English-speaking countries. It allows researchers to look at data for each country at the same time, which makes it a very powerful tool. For the plupluperfect construction, only data from the US, Canada, Great Britain, Ireland, Australia and New Zealand were compiled, with the following results:

Country	*United States*	*Canada*	*Great Britain*	*Ireland*	*Australia*	*New Zealand*
Totals	62	22	192	45	106	24

Table 2. Plupluperfect total instances from GloWbE for each country.

It seems, then, that the nonstandard construction is alive and well in the former empire and the Commonwealth. However, nothing more can be said about it, for a number of reasons.

First, it seems that each country component is different in size. For instance, the US sample is larger than the Australian one. Moreover, there is no way of telling whether the web-texts that make up the components are balanced in terms of style and genre. This makes any kind of comparison irrelevant, as the frequency of the construction in any one country may be accounted for by a number of unknown factors. Thus, to say that the plupluperfect is more frequent in the British Isles than in Australia would be completely hazardous at this point. In addition, since the backgrounds of the writers are not known, with web material it may be the case that the authors are non-natives and they might not even live in that country to which the web-domain belongs.

Further difficulties in relying on Web-data, although in this case careful selection and compiling has already taken place, arise when multiple entries are encountered and when the results evince questionable language. This entry, for example, makes no sense and had to be disregarded when the total figures were put together:

(9) shouted out not where they expect the good word. Most mr car insurance online *had spaceship builders have been* Scott said in a like the KDL and forestall the others (GloWbE, cesl.arizona.edu, US G)

Results returned by GloWbE require careful inspection in order to ensure validity. (For an in-depth discussion on working with Web material for linguistic research, *cf*. Hundt 2013).

Conclusions

Corpora cannot answer all questions researchers might have. Each corpus has a number of particularities which resulted from the conditions in which it was compiled. Thus, while it cannot be contended that the BNC, COCA and Strathy are each representative of their national varieties, in their own right, they do not seem to be comparable. Their size, overall

makeup and genre balance are different. Moreover, although most plupluperfects were found in the spoken sections, these corpora's respective spoken components are very different in terms of sources. In addition, the corpora cover different time spans: Strathy covers 1920s to 2000s, but the BNC has material only from the early 1990s. Although corpora compiled from top level domains are certainly representative of national varieties of English (*cf.* Cook and Hirst 2012), GloWbE's national components are also different in size, which makes comparison difficult, again.

Putting together all the instances of the plupluperfect found in three of the major national varieties of English, the results are the following:

Variety	*British English*			*American English*			*Canadian English*		
Corpora	BNC W	BNC S	GloWbE	COCA W	COCA S	GloWbE	Strathy W	Strathy S	GloWbE
Instances	10	30	192	9	33	62	6	5	22
Totals	232			104			33		

Table 3. The number of plupluperfect instances found for three major English varieties. W and S stand for the written and spoken components of a corpus.

Unfortunately, the question remains whether these numbers are actually indicative of the frequency with which the nonstandard is met, although one may suggest that the plupluperfect is most frequently encountered in British English. More information is needed in order to statistically confirm such a claim.

The data presented here also seems to suggest that Australian English is closer to British English (see the results for GloWbE), while Canadian English is a lot closer to American English. It would be interesting to correlate the results for the plupluperfect with other nonstandard constructions or language features. Furthermore, given the dispersion of the plupluperfect on the British Isles (see the results for the BNC), it would be worthwhile to see whether such data can be correlated with the historical demographics of the former Empire.

Bibliography

Biber, Douglas. 1993. 'Representativeness in Corpus Design' in *Literary and Linguistic Computing* 8(4). pp 243-257.

Christophersen, Paul. 1986. 'A history of the plupluperfect' in *English Today* 8(2). p 36.

CLAWS part-of-speech tagger for English, available online at http://ucrel.lancs.ac.uk/claws/.

Cook, Paul and Graeme Hirst. 2012. 'Do Web Corpora from Top-Level Domains Represent National Varieties of English?', available online at http://www.cs.toronto.edu/~pcook/CookHirst2012.pdf

Declerck, Renaat. 2006. *The Grammar of the English Verb Phrase. Volume 1: The Grammar of the English Tense System*. Berlin: Mouton de Gruyter.

Garside, R. 1996. 'The robust tagging of unrestricted text: the BNC experience' in J. Thomas and M. Short (eds.). *Using corpora for language research: Studies in the Honour of Geoffrey Leech*. Longman, London. pp 167-180.

Hundt, Marianne. 2013. 'Using web-based data for the study of English' in Krug, Manfred and Julia Schlüter (eds.). *Research Methods in Language Variation and Change*. Cambridge: CUP.
Ishihara, Noriko. 2003. '"I Wish I Would Have Known!": The Usage of *Would Have* in Past Counterfactual *If-* and *Wish-* Clauses' in *Issues in Applied Linguistics* 14(1). pp 21-48, available online at http://escholarship.org/uc/item/5wd0w3sz.
Leech, G., R. Garside and M. Bryant. 1994. 'CLAWS4: The tagging of the British National Corpus' in *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)* Kyoto, Japan. pp 622-628.
Lindquist, Hans. 2009. *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.
UCREL CLAWS7 Tagset, available online at http://ucrel.lancs.ac.uk/claws7tags.html.