From opportunistic to systematic use of the Web as corpus: *Do*-support with *got (to)* in contemporary American English

Christian Mair

The Oxford Handbook of the History of English
Edited by Terttu Nevalainen and Elizabeth Closs Traugott

Print Publication Date: Nov 2012 Subject: Linguistics, Sociolinguistics, Historical Linguistics
Online Publication Date: Nov 2012 DOI: 10.1093/oxfordhb/9780199922765.013.0023

Abstract and Keywords

The chapter argues that the best way to profit from the rich corpus-linguistic working environment available to the student of the history of English is to use traditional (and sometimes small) linguistic corpora together with larger textual databases and digital archives, including the World-Wide Web, in a coordinated way. Linguistic corpora (ARCHER, Brown family, BNC, COCA, COHA) are sufficient to document the successive waves of grammaticalisation which have added have to, have got to and, more recently, want to or need to to the older form must, producing the complex layered system of present-day English modal markers of obligation and necessity. Using do-support with modal *got* (*to*)/*gotta* as an illustration, the paper shows that, in spite of its known deficiencies as a linguistic corpus, the World-Wide Web can help fill in the language-historical picture in useful ways where even the biggest available corpora fail to produce sufficient evidence.

Keywords: Web, corpora, English, language change, Representative Corpus of Historical English Registers, Brown family, modal auxiliary function, syntax

1. “Caveat Googlator?”

When Brian Joseph (2004: 382) coined this memorable phrase to draw attention to linguists’ increasing readiness to retrieve data from the Web, he at once acknowledged a major recent trend in the profession and expressed skepticism with regard to the often opportunistic ways in which the new resource was being used. Such opportunism is encouraged by three factors. First, the Web contains a huge amount of data and is hence often the only source for examples of very rare usages and constructions. Second, the Web is a self-updating corpus and hence an excellent source for neologisms and recent grammatical innovations. Third, it is a deceptively convenient source of data—with a “corpus architecture” that is taken for granted and (p. 246) retrieval software that is considered a global default standard and therefore not worth thinking about.

Measured by the traditional philological and linguistic standards of data quality, the Web is inferior to corpora designed, compiled, and annotated by linguists for linguistic research. Documentation of sources is frequently incomplete; the authenticity of the material may be in doubt; conventions of representation, including the transcription conventions vital in the case of orthographically represented spoken language, are (a) not standardized across sources and (b) not usually made explicit in many cases. On the retrieval side, current search engines for the Web are seriously deficient by the standards of customized linguists’ software, for example, because insufficient options are provided for truncation or the search for discontinuous constituents. While all this makes it easy to join in with Brian Joseph's exhortation “Caveat googlator!”, I am firmly convinced that the Web is indispensable as a source of linguistic data in the study of ongoing language change. Here the question is not whether we should use the Web, but how we can domesticate its protean richness. For all the dubious quality of much data it contains, the Web has three advantages over traditional corpora. Two—the sheer quantity of material and its recency—are largely uncontroversial. A third advantage frequently goes unnoticed, namely its diversity. Standard American English, the default language of the early Web, has been losing ground, both to other languages (Danet and Herring 2007), and—more important to the student of the history of English—other standard and nonstandard varieties of English itself.

How can we best domesticate the Web as corpus then? This short chapter is not the place to provide an exhaustive survey of all the varied research that is going on in this field at present (for surveys, see e.g. Hundt, Nesselhauf, and Biewer 2007 or Bergh and Zanchetta 2008; Gatto 2011). But two major current trends can be identified.

(1) Direct and systematic use of the Web as corpus
Developing mostly from the type of opportunistic use described above, direct and systematic analysis of Web language is characterized by attention to issues such as the size and composition of the particular portion of the Web that represents the focus of attention and to the vagaries of data retrieval from the Web. Much of this direct analysis is not quantitative but qualitative and discourse-analytical, for example, studies of multilingual usage in community Web forums (e.g. Androutsopoulos 2006). However, studies such as Mair (2007a, 2007b) have been able to show that most national varieties of English (except standard American English, which—as the default language of the Web—is pervasively distributed across sub-domains) can be researched on the Web given a few simple methodological precautions, such as checking the validity of Web findings against results from more traditional corpora. In a study on quotative usage, Buchstaller et al. (2010) have even proposed statistical normalization procedures that have made it possible to integrate Web data into a variationist analysis. Finally, Web-based quantitative comparative cultural (p. 247) studies ("culturomics"—Michel et al. 2010) holds great potential, particularly for the study of lexical change, which should be exploited in historical linguistics. The two major drawbacks of accessing data directly on the Web remain that (a) the database is unstable, which precludes replicability of the results, and (b) piggybacking on existing search engines severely constrains a linguist's options for more sophisticated searches.
(2) Compiling offline corpora from Web sources
A large number of researchers (e.g. Davies 2001; Mukherjee and Hoffmann 2006; Altmann, Pierrehumbert, and Motter 2009; Mair 2011) have used the Web as a source for the compilation of large offline corpora (which on completion of compilation are frequently made available online again). The Corpus of Contemporary American English (COCA; Davies 2010) is probably the most widely used resource compiled in this way. This procedure allows compilers full independence with regard to corpus architecture and retrieval. The database can be held constant, or expansion can be controlled (e.g. through regular and documented annual updates), which ensures replicability of results, makes possible statistical modeling of ongoing trends, and cumulative progress through cooperation of several research groups on the analysis of the same material. Another interesting example of domesticating the Web as corpus in this way is the WebCorp project, based at Birmingham City University.[1] Here, one of the corpora downloaded and annotated for analysis is supposed to be a representative "mini Web" rather than a collection of texts documenting a specific regional or functional variety of English. The technological challenges that present themselves in compiling corpora from the Web

are the focus of the activities of SIGWAC, the Association of Computational Linguistics' "special interest group on the web as corpus".[2]

2. Integrating the Web, Web-Derived Corpora, and Traditional Corpora in a Corpus-Linguistic Working Environment

Both working directions described above are worth pursuing and even complementary in their strengths and weaknesses. Regardless of which one is adopted for a particular study, however, work will only be successful if the traditional corpora, with (p. 248) their order and structure, and the rapidly expanding new data sources are systematically integrated into a state-of-the-art corpus-linguistic working environment.

This will be demonstrated with research on current changes in the English modal system. Using small corpora such as ARCHER (A Representative Corpus of Historical English Registers) and the BROWN family, Krug (2000) and Leech (2003), among others, have shown that the expression of obligation and logical necessity has been undergoing rapid change in the recent history of English. *Have to, have got to, want to/wanna*, and *need to* have been added to *must* to produce a complex layered system, and for some of these forms epistemic uses have become more prominent. The history and current development of *(have) got to/gotta* is an important part of the story, because—together with *going to, want to*, and *need to*—it shows that modal auxiliary function is increasingly being performed without traditional modal-verb syntax. However, given the spoken-informal stylistic profile of *got to/gotta*, its history tends to be recorded insufficiently in traditional historical corpora—a gap that we can attempt to fill on the basis of much larger Web-derived corpora or even the Web itself. One trait that is particularly unexpected in the grammaticalization of a modal operator is the emergence of *do*-support, which is impressionistically familiar with possessive *got* and modal *got to* from informal and nonstandard American usage.

3. *Do we got Change? do*-Support with Possessive/Modal *Got (to)*

The variable usage serving as a study example here is well illustrated by the following instance from contemporary US media usage, retrieved from the COCA corpus:

(1)

A: Are you going to keep Louis Freeh?
B: Well, he's appointed by the president. …
A: … does he got a chance to stay. (COCA, CNN)

Got is used as a main verb here. *Got (to)* certainly lacks many defining features of main-verb syntax, for example, an infinitive (**to got*) or an inflected past and participle (**gotted*), but is certainly part of a group of modal and aspectual expressions that vacillate between auxiliary and main verb syntax (*dare, need, ought, used to*). Plausible assumptions, which need to be tested systematically on the basis of corpora, are that:

(a) regionally, this use is mainly North American;
(b) diachronically, it is increasing in frequency; and
(c) sociolinguistically, it is not part of the standard (yet?).

(p. 249) The *Dictionary of American Regional English* (*DARE*) has the first attestation from a 1938 source and refers to the form as "[used by] younger people, colloquial" (s.v. *get*). Pullum claims that in "standard English, necessity GOT only occurs as a complement to perfect HAVE. In colloquial American English, it occurs without HAVE as in *I gotta go*, with negative *I don't gotta go*, polar interrogative *Do you gotta go?*, etc. Some speakers have zero inflection in the 3rd singular present (*She gotta find another place*)" (1997: 89). Trudgill suggests that the construction is beginning to spread into standard usage and claims that "American Standard English currently admits a new verb *to got* in *You haven't got any money, do you?* but not (or not yet) in *You don't got any money, do you?*" (2002: 169).

Hardly any relevant examples can be found in traditional small corpora. There are no attestations in corpora of edited written American English such as Brown and Frown. The 2 million word Corpus of Spoken Professional American English contains one. Corpus size being the limiting factor, we thus have a good case for consulting the Web. Using the Google Advanced Search mode, regional concentrations of the phrase *do I got* were polled in two ways: by restricting searches to the top-level national domains, and by restricting them by region. The general expectation is that the search by region will produce more hits than the search by top-level national domain (most drastically so in the case of us/United States). The figures for *looking* were added in order to poll for the occurrence of a regionally and stylistically unmarked single word (Table 1). The frequency of *looking* (n_2) was divided by the frequency of *do I got* (n_1) in each case (see rightmost column). Assuming an (p. 250) average frequency quotient (n_2: n_1) of 4,313 for the whole English-language Web, very high values (such as the 12,381 obtained for the .uk top-level national domain) were assumed to signal relative underrepresentation of *do I got* (a calculation which, as we shall see below, is fraught with considerable uncertainties).

Table 1. *do I got* on the Web (24 January 2011)

	do I got **(n_1)**	***looking*** **(n_2)**	**n_2: n_1**
English Web	619,000	2,670,000,000	4,313
.uk [top-level domain]	10,500	130,000,000	12,381
“United Kingdom” [region]	29,300	207,000,000	7,065
.ie [top-level domain]	1,850	9,140,000	4,941
.au [top-level domain]	3,680	28,200,000	7,663
“Australia” [region]	6,000	62,200,000	10,367
.ca [top-level domain]	2,610	20,600,000	7,893
“Canada” [region]	16,100	119,000,000	7,391
.nz [top-level domain]	1,370	9,870,000	7,204
“New Zealand” [region]	2,300	11,000,000	4,783
.us [top-level domain]	1,150	7,000,000	6,087
“United States” [region]	496,000	2,190,000,000	4,415

As usual in Google-based Web linguistics, we are overwhelmed by sheer numbers and lack of transparency.[3] Two obvious sources of noise in the data are that many of these hits are not instances of the construction looked for and that phrase-frequency counts in Google are based on guesswork and extrapolation rather than fact (see Introduction). Comparison across domains could still be useful, though, if noise levels were shown to be constant. Table 1, however, does not give cause for optimism. The only thing it shows beyond doubt is what we have good evidence for anyway, namely that *do I got* is in regular use. A laborious manual post-edit of the Google results would thus be sure to yield many interesting examples. Any further interpretation of the figures, on the other hand, is, quite frankly, pointless. Not even the North American territorial base of the construction—established beyond doubt, for example, by its extreme scarcity in the British National Corpus and its regular occurrence (ca. 100 instances) in the Longman Corpus of Spoken American English—shows up clearly in Table 1. Spot checks reveal some possible reasons: most cases in New Zealand, for example, are accidental co-occurrences (of the type *Do I? Got no time …*), typing errors (*Do I got to the doctor?*), or

the expected quotations from American sources, including multiple occurrences of the transcription of the lyrics of American pop songs.

The way forward in this situation is not to access the Web directly (the first strategy listed in section 1 above), but to do so indirectly, on the basis of two large Web-derived[4] corpora (the second strategy outlined above). For exemplification, I will choose WebCorp, because of its claim to offer a representative mini-web, and COCA, a corpus targeting the regional variety in which we can hope to find the most examples.[5] In both corpora, the number of relevant hits is small enough for human inspection of all the attestations, which allows a few interesting conclusions as to the quality of Web-derived data. Thus, in addition to trivial typing errors (⟨got⟩ ← ⟨get⟩) the returns of a search for *do I got* in WebCorp yield borderline examples of the following kind, showing the plural form *informations*, a typical feature of nonnative varieties of English:

> **(2)** where do you got your informations from? wikipedia (WebCorp)

(p. 251) Example (3), from a role-play,[6] is probably genuine in the sense of having been produced by a native speaker of English, but not authentic, as the person is not speaking in his or her own voice but impersonating a character.

> **(3)** “What do they got over there we ain't got here?”

In spite of having to discard several false returns, WebCorp leaves us with rich and interpretable material: 54 “clean” cases of *do I got* and 93 of *do I gotta*. They show that *do-got* is chiefly used in writing either in passages of direct speech (4) or to serve as a marker of informality, as in the sports-related Internet newsletter in (5):

> **(4)** Mr. D'Ambrosi—who testified yesterday that he had paid $16,500 to Mr. Scopo on lesser projects after the union leader told him ‘everybody else does it[’] asked on the tape: ‘Who do I **got to go see**? Tell me who I got to go see[’] (WebCorp)
> **(5)** They feel that Brown is the best one out there, but what does that mean for the defense. Take away the injury-riddled players, and **what do you got left**? Not much. (WebCorp)

COCA yields somewhat fewer, namely 23, clean instances of *do I got(ta)*, from a total of 24 hits. This shows that WebCorp's “mini-web” is a bigger but also a much messier resource than COCA. The material includes interesting cases of variation in online speech production such as the following:

> **(6)** and what **do we have to** do to—**do I got to** hug you like—what—what **do we need to** do (COCA, NBC)

> **(7)** I'll tell you—how long **have we got**, Harry? **Do I got** time enough to get in this window? (COCA, CBS)

In (6), the contemporary statistical default option *do we have to do* is used to start off a sequence; the second question *do I got to* introduces a distinctly informal note, whereas the third expression of obligation *do we need to* uses a form that has boomed in this function only very recently and hence might be used for emphasis. Variable grammar is also in evidence in (7). Most importantly, however, the tidy diachronic layering of COCA reveals an increase of *[do] [p*] got* of around 50 percent between 1990–94 and 2005–10 in real time.

The middle ground between insufficient attestation (traditional corpora) and total confusion (whole Web) also proves the most fruitful area to explore when it comes to studying the history of the construction. The *Oxford English Dictionary* has almost (p. 252) nothing to offer.[7] *DARE*'s first attestation from 1938 is implausibly late. A trawl of Google Books and Project Gutenberg (a site already used for corpus compilation with promising results—see De Smet 2005)[8] yields several examples from the second half of the nineteenth century, with most attesting the form in representations of immigrant English and African American or Creole-influenced varieties of English. The following example is from William Henry Drummond's (1854–1907) popular collection *The Habitant and Other French-Canadian Poems* (1897). Irish-born Drummond migrated to Canada at the age of 10, where he lived among French speakers in Lower Canada. In the poem "Pelang", the persona looks back on her younger days:

> **(8)** Ah me! I was foolish young girl den
> It's only ma own plaisir I care,
> …
> **Don't got** too moche sense at all dat tam
> …
> (Drummond 1897: 75)

At around the same time, the *Brotherhood of Locomotive Engineers' Journal* (1869: 262) shows a similar example of foreigner talk:

> **(9)** An intelligent gentleman from Germany, on his first visit to an American church, had a contribution box with a hole in the top presented to him, and whispered to the collector: "I **don't got** mein bapers unt can't vote … "

Captain Clutterbuck's Champagne—A West Indian Reminiscence, first published in *Blackwood's Magazine* in 1861 and then as a book in the following year (Anon. [Hamley] 1862: 154), has Daddy, a Jamaican sorcerer, asking:

> **(10)** You **don't got** lilly bit of nice baccy for the old man?

While simplification through second-language acquisition or creolization is one plausible scenario to account for the establishment of *do*-support for *got to* in North America, the sole early attestation of the construction that is from formal writing presents somewhat of a puzzle. It is from a letter dated 20 July 1865 by an American missionary posted in the Ottoman Empire, which was published in *Evangelical Christendom* (1867: 499). The missionary comments on corruption in the Ottoman legal system:

> **(11)** An unprincipled fellow has a quiet neighbour whom he wishes to spite. He has only to demand 20000 piastres of him and carry the demand into court to succeed. If he gets his claim, he ruins the man. If he **don't got** it, the man has to pay the fee due to a 20000 piastres (p. 253) decision, or, most likely, the Cadi, Mufti, and Medjliss will tell the poor fellow: “He has no claim at all against you, but you had better settle the matter. … “

The rest of the publication, which runs to considerable length, contains numerous verbatim or pseudo-verbatim quotations, which, however, never attempt representation of colloquial or nonstandard speech.

4. Provisional Conclusion and Further Challenges for Web-Based Corpus Linguistics

As these preliminary findings have shown, use of existing corpora, the Web, and intelligently selected Web-derived resources produces enough material to put the study of a previously neglected instance of recent morphosyntactic change in English on a new level. Diachronically, we have been able to antedate *do*-support for *got (to)* by almost a century, and suggest an origin in language contact in multilingual environments. Sociolinguistically, we have been able to confirm the new form's regional base in North America (where it still has very informal, borderline nonstandard status). Examples showing variable use of standard/formal and nonstandard/informal forms by the same speaker in close proximity provide yet another potentially interesting research focus.

As a variable, *do*-support for *got* is rather typical of the type of morphosyntactic innovation that is commonly investigated in corpus-based studies of ongoing change. Web-based corpus-linguistics of the future, however, will have to break new ground in at least two further directions:

(1) As an increasingly **multilingual** communicative domain, the Web confronts linguists with multilingual texts, including texts displaying phenomena such as code-mixing and code-switching, which have traditionally been outside the remit of corpus-based studies of language change in progress.

> **(2)** The Web is also developing into an increasingly **multimodal** resource, combining text, sound, and visual images in diverse formats.

Currently, corpus linguists operate in a textual universe. At least for the world's more widely used languages, we now have immense amounts of written text to investigate. But even for English, Spanish or German we still have a shortage of spoken language, especially the baseline register of spontaneous dialogue. What little we have we usually access through orthographic transcriptions rather than the recorded sound (or the filmed interaction). Web-based research needs to join current efforts to bring the spoken word and the visual gesture back into the remit (p. 254) of corpus study, for only then will (corpus) linguistics become an equal partner in the multidisciplinary science of the Web envisaged by Berners-Lee et al.:

> Web science is about more than modeling the current Web. It is about engineering new infrastructure protocols and understanding the society that uses them, and it is about the creation of beneficial new systems. It has its own ethos: decentralization to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways, and fairness. It uses powerful scientific and mathematical techniques from many disciplines to consider at once microscopic Web properties, macroscopic Web phenomena, and the relationships between them. Web science is about making powerful new tools for humanity, and doing it with our eyes open. (Berners-Lee et al. 2006: 770–71)

In this spirit, I hope to have identified some immediately realistic and some more remote steps that need to be taken in Web-assisted corpus linguistics if we wish to move on from Joseph's cautionary slogan to a more optimistic *Non caveat googlator*!

References

Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter. 2009. ‘Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words’. PLoS ONE 4: e7678. DOI: 10.1371/journal.pone.0007678.

Androutsopoulos, Jannis. 2006. ‘Multilingualism, Diaspora, and the Internet: Codes and Identities on German-Based Diasporic Websites’. *Journal of Sociolinguistics* **10**: 520–47.

Anon. [Hamley, William G.] 1862. *Captain Clutterbuck's Champagne—A West Indian Reminiscence*. Edinburgh: William Blackwood and Sons.

Bergh, Gunnar, and Eros Zanchetta. 2008. ‘Web Linguistics’. In *Corpus Linguistics: An International Handbook*. Vol. I, ed. Anke Lüdeling and Merja Kytö, 309–27. Berlin: Mouton de Gruyter.

Berners-Lee, Tim, Wendy Hall, James Hendler, Nigel Shadbolt, and Daniel J. Weitzner. 2006. ‘Creating a Science of the Web’. *Science* 313 (11 August): 770–71.

Brotherhood of Locomotive Engineers' Monthly Journal 3. 1869. Fort Wayne, IN.

Buchstaller, Isabelle, John R. Rickford, Elizabeth Closs Traugott, Thomas Wasow, and Arnold Zwicky. 2010. 'The Sociolinguistics of a Short-Lived Innovation: Tracing the Development of Quotative *All* Across Spoken and Internet Newsgroup Data'. *Language Variation and Change* 22: 1–29.

Danet, Brenda, and Susan C. Herring, eds. 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford: Oxford University Press.

DARE= Cassidy, Frederic G., and Joan Houston Hall, comps. 1985–2012. *Dictionary of American Regional English*. 5 vols. Cambridge, MA: Belknap Press.

Davies, Mark. 2001. 'Creating and Using Multi-Million Word Corpora from Web-Based Newspapers'. In *Corpus Linguistics in North America*, ed. Rita C. Simpson and John M. Swales, 58–75. Ann Arbor, MI: University of Michigan Press.

———. 2010. 'The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English'. *Literary and Linguistic Computing* 25: 447–64.

(p. 255) De Smet, Hendrik. 2005. 'A Corpus of Late Modern English'. *ICAME Journal* 29: 69–82.

Drummond, William H. 1897. *The Habitant and Other French-Canadian Poems*. New York: G. P. Putnam's Sons.

Evangelical Christendom: A Monthly Chronicle of the Churches. Vol. 8, New Series. 1867. London.

Gatto, Maristella. 2011. 'The "Body" and the "Web": The Web as Corpus Ten Years on'. *ICAME Journal* 35: 35–58.

Hundt, Marianne, Nadja Nesselhauf, and Carolin Biewer, eds. 2007. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Joseph, Brian D. 2004. 'The Editor's Department: On Change in *Language* and Change in Language'. *Language* 80: 381–83.

Krug, Manfred. 2000. *Emerging English Modals: A Corpus-Based Study of Grammaticalization*. Berlin: Mouton de Gruyter.

Leech, Geoffrey. 2003. 'Modality on the Move: The English Modal Auxiliaries 1961–1992'. In *Modality in Contemporary English*, ed. Roberta Facchinetti, Manfred Krug, and Frank R. Palmer, 223–40. Berlin: Mouton de Gruyter.

Mair, Christian. 2007a. 'Change and Variation in Present-Day English: Integrating the Analysis of Closed Corpora and Web-Based Monitoring'. In *Corpus Linguistics and the*

Web, ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, 233–47. Amsterdam: Rodopi.

———. 2007b. 'Varieties of English around the World: Collocational and Cultural Profiles'. In *Phraseology and Culture in English*, ed. Paul Skandera, 437–68. Berlin: Mouton de Gruyter.

———. 2011. 'Corpora and the New Englishes: Using the "Corpus of Cyber-Jamaican" (CCJ) to Explore Research Perspectives for the Future'. In *A Taste for Corpora. In Honour of Sylviane Granger*, ed. Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot, 209–36. Amsterdam: Benjamins.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2010. 'Quantitative Analysis of Culture Using Millions of Digitized Books'. *Science* 1199644. 16 December 2010. DOI: 10.1126/science.1199644.

Mukherjee, Joybrato, and Sebastian Hoffmann. 2006. 'Describing Verb-Complementational Profiles of New Englishes: A Pilot Study of Indian English'. *English World-Wide* 27: 147–73.

Pullum, Geoffrey K. 1997. 'The Morpholexical Nature of *to*-Contraction'. *Language* 73: 79–102.

Trudgill, Peter. 2002. *Sociolinguistic Variation and Change*. Washington, DC: Georgetown University Press. (p. 256)

Notes:

(1) See http://www.webcorp.org.uk/.

(2) See http://www.sigwac.org.uk/.

(3) A striking illustration is that—on the strength of these figures—*looking* is far more common in the "English" subset of the Web than in the whole Web.

(4) This term is obviously ambiguous. The identification of suitable material for download can take place in the public domain through the use of standard search engines and retrieval software piggybacking on them, but material can also be derived from password protected sites and the "deep" Web through targeted searches.

(5) Problems in the current architecture of or access to WebCorp meant that retrieval took an uncomfortable two and a half hours. No such problems occurred in the use of COCA.

(6) From http://www.angelfire.com/az3/twohourwargames/batreps/CA/LGrognard.pdf, accessed 12 May 2010. This link was correct at the time of access given, but the URL and text have been removed since then. This is a minor illustration of a point made above, that the results of Web-based corpus research are not fully replicable unless the data analyzed are stored offline.

(7) http://www.oed.com/. There is a single instance of *do I got*, from the transcripts of the American television series *The Sopranos* (entry for *moolie*).

(8) See http://books.google.com/ and http://www.gutenberg.org/.

Christian Mair

Christian Mair holds a chair in English linguistics at the University of Freiburg in Germany. He has been involved in the compilation of several linguistic corpora (among them F-LOB and Frown, updates of the classic LOB and Brown corpora, and the Jamaican component of the International Corpus of English). His research over the past two decades has focused on the corpus-based description of modern English grammar and variation and change in standard Englishes worldwide. His most recent books are Twentieth-Century English: History, Variation, and Standardization (2006) and Change in Contemporary English: A Grammatical Study (with G. Leech, M. Hundt, and N. Smith, 2009). christian.mair@anglistik.uni-freiburg.de