The British component of the International Corpus of English (ICE-GB), Release 2 , and: *Diachronic Corpus of Present-Day Spoken English (DCPSE)* , and: *The International Corpus of English Corpus Utility Program (ICECUP), version 3.1* (review)

Mark Davies

Language, Volume 85, Number 2, June 2009, pp. 443-445 (Review)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.0.0105*

The British component of the International Corpus of English (ICE-GB), Release 2. CD-ROM. Ed. by BAS AARTS and SEAN WALLIS. London: Survey of English Usage, University College London, 2006.

Diachronic Corpus of Present-Day Spoken English (DCPSE). CD-ROM. Ed. by BAS AARTS and SEAN WALLIS. London: Survey of English Usage, University College London, 2006.

The International Corpus of English Corpus Utility Program (ICECUP), version 3.1. CD-ROM. Ed. by SEAN WALLIS. London: Survey of English Usage, University College London, 2006.*

Reviewed by MARK DAVIES, *Brigham Young University*

In the movie *Back to the future* (1985), the main character is transported back thirty years into the 1950s, but the knowledge and experience that he took from the 1980s provide for a happy ending. This is somewhat analogous to the two recent corpora under discussion: The British component of the International Corpus of English (ICE-GB) and the Diachronic Corpus of Present-Day Spoken English (DCPSE). These corpora are reminiscent of corpora from thirty to forty years ago, when million-word corpora were the norm. In terms of relationships to older corpora, it is also interesting to note that half of the DCPSE actually ıs a corpus from the late 1960s to early 1980s. But as in the movie, these corpora have been (re)structured and annotated in ways that make them much more useful than other small corpora of bygone days, and I believe that both still fill an important niche in today's world of English corpora.

ICE-GB is the British component of the International Corpus of English, a project that will eventually contain components from approximately twenty English-speaking countries (see http://www.ucl.ac.uk/english-usage/ice/). Each component contains one million words—600,000 spoken and 400,000 written. IGE-GB is without a doubt the most advanced component of the overall ICE project in terms of its annotation and interface, which is something that serves as the focal point of this review. DCPSE, the second corpus under review, is composed of two parts. The first is the 600,000-word spoken part of ICE-GB and the second is the 400,000-word London-Lund Corpus, a corpus of spoken British English from the late 1960s through the early 1980s. In addition to the corpora users receive ICECUP ('ICE Corpus Utility Program') 3.1, the software and interface that are used to access the two corpora.

The ICECUP corpus interface allows users to interact with the corpus and corpus data in a number of different ways. Via expanding tree diagrams, users can browse through the corpora (based on several different criteria), and can limit the search to a particular 'node' of the corpus, or to texts identified by a given speaker or text variables. They can search through a lexicon of all forms in the corpus, or a 'grammaticon' of all syntactic tags and the associated words. In the 'Keyword in context' display, there are many options for customizing the display, increasing and decreasing context, and so on.

Users can carry out basic searches via the 'text fragment search', such as *end* up <V>* for *ended up leaving*, *ends up watching*, and so on. The heart of the search engine, however, is an extremely powerful (and yet relatively easy to use) interface that looks for 'fuzzy tree fragments'. Users create chart-like maps of the query by adding nodes and indicating part of speech, word forms, wildcards, and the like. The power of the fuzzy-tree-fragment searches comes from the fact that the corpora are not just tagged for part of speech, but are also parsed. Thus users can search for complex syntactic structures like 'notional direct objects', 'floating NP postmodifiers', 'cleft operators', and more than fifty other features. Users can also save query results, and can

*ICE-GB costs about $800 for an individual license ($50 for student license) or about $1,500 for an institutional network license (payment is in pounds sterling). The DCPSE costs about $700 for an individual license ($50 for student license) or about $1,400 for an institutional network license.

later combine queries at virtually any level of complexity. And lest users think that it is all too complex, the creators have written a 340+ page book (*Exploring natural language*) that guides them carefully and clearly through the full range of possibilities. In summary, I am not aware of any other interface or parsed corpora that allows users to perform such complex syntactic searches with such ease.

In spite of the power of the ICECUP software, there are a number of features that would make it even more useful. The entire interface seems to be oriented toward syntactically based searches, and there is much less attention given to lexically oriented queries. Perhaps the most obvious example is that it is impossible to search by lemma (e.g. [take] part in = *takes*/*taking*/*took part in*); rather one has to include all of the variant forms as word alternates. Beyond this, however, it would be nice to be able to see charts that compare the frequency of words and phrases across the different sections of the corpora. It would also be useful to use frequency information as part of the query, such as verbs that are more frequent in the 1960s vs. the 1990s, or adverbial phrases in conversations that are more frequent than in academic texts. Other lexically oriented features that are common in concordancing and text-retrieval programs are also absent, such as the ability to find the collocates of a given word or phrase, which would be useful for a wide range of semantically oriented investigations.

While there are many useful aspects of the corpora and the software, my major concern is the very limited size of the corpora—just one million words each. As Tony McEnery and Andrew Wilson (2001) and others have noted, small one-million-word corpora were the norm in the 1960s and 1970s (e.g. Brown and LOB), but corpora of this size were severely criticized by Noam Chomsky and others for not having enough data for insightful analyses. By the late 1980s, however, 'mega-corpora' of 100 million words and more were the norm, and this criticism was much less valid. Let us briefly consider whether one-million-word corpora such as IGE-GB and DCPSE are in fact overly small for certain types of investigations.

Consider first an example that Chomsky himself provides. As noted by McEnery and Wilson (2001), in an early critique of corpus linguistics he claimed that *perform* could not take direct objects that were mass nouns (e.g. *perform magic*), and he stated that he did not need a corpus to confirm his intuitions as a native speaker. Data from large corpora, however, show this 'intuition' to be false. There are hundreds of examples of *perform* + mass noun in large corpora such as the British National Corpus (BNC, 100 million words; http://corpus.byu.edu/bnc), the TIME Corpus of Historical American English (100+ million words; http://corpus.byu.edu/time), and the Corpus of Contemporary American English (COCA, 400+ million words; http://www.americancorpus.org). But because of their limited size, neither ICE-GB nor DCPSE contains a single example of the construction. In this particular case, then, one could therefore infer that the corpus data is no more enlightening than the faulty native-speaker intuition.

As a second example, let us briefly consider the construction *end up V-ing* (e.g. *ended up paying too much*), which is a construction that exhibits interesting diachronic and genre-based variation. The 550 tokens from the TIME Corpus show that the construction has increased steadily since the 1920s, with notable increases in every decade since that time. But the data from the DCPSE (four tokens from the late 1960s to early 1980s and seven from the 1990s) is too small to make any real claims. In terms of synchronic genre-based variation, data from the BNC (809 tokens) and COCA (6,170 tokens) show conclusively that the construction is much more common in 'informal' genres such as [spoken] than in formal genres such as [academic] journals. ICE-GB, however, provides just eleven tokens from spoken (600,000 words) and seven from written (400,000 words). This limited data suggests roughly equal frequency in each genre (per million words), which is very much at odds with the data from the much larger corpora.

The very limited size of the corpora is probably also the reason that the corpus interface and architecture is oriented almost entirely toward syntactic searches, rather than lexically oriented investigations. With just one million words, there just are not enough tokens to permit useful analyses of most lexical items. For example, in the DCPSE there are only twelve tokens of *crack* as a noun, and these yield only four noun collocates within a five-word span (and none of which occurs more than once). In other diachronic corpora such as the TIME Corpus, however, there

are nearly 2,300 tokens, and nearly 300 different noun collocates that occur two times or more. The same is true for IGE-GB, the synchronic corpus. There are only twenty-eight tokens of *crack* as a noun, and only two of the noun collocates occur two times or more. This is of course much less than in a corpus like COCA, where there are nearly 6,000 tokens and 1,275 of the noun collocates occur two times or more. This limitation in terms of lexically oriented searches is not a serious criticism of ICE-GB and the DCPSE, however, since they never really make the claim that the corpora can or should be used for anything besides grammatical research.

In terms of the big picture, we note that on the website for IGE-GB and the DCPSE, it suggests that the two corpora 'will permit research into synchronic and diachronic grammatical variation' and that the DCPSE 'will be a major new resource for linguists interested in ''current change'' '. In other words, the corpora can be used to examine changes from the 1960s to the 1990s (via the DCPSE) and to compare different genres from the 1990s (via IGE-GB). If one looks at high-frequency features such as tense, aspect, mood, voice, modals, and pronominal usage (features that have already been exhaustively studied with other small corpora of English), then this is probably true. In fact, major publications to this effect are beginning to appear (see Leech et al. 2009). But our sense is that such diachronic and genre-based investigations of medium- and low-frequency syntactic constructions (as well as most lexically oriented queries) would be very difficult with these small one-million-word corpora.

In summary, ICE-GB and DCPSE fulfill a very useful niche in the world of English corpora. For high-frequency constructions and structures, the quality of the tagged corpora and the well-designed corpus interface will allow researchers to carry out advanced, fine-grained analyses of English syntax that would be difficult or impossible with almost any other corpora.

REFERENCES

McEnery, Tony, and Andrew Wilson. 2001. *Corpus linguistics*. 2nd edn. Edinburgh: Edinburg University Press.

Leech, Geoffrey; Marianne Hundt; and Christian Mair. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press, to appear.

Nelson, Gerald; Sean Wallis; and Bas Aarts. 2002. *Exploring natural language*. Amsterdam: John Benjamins.

Department of Linguistics and English Language
Brigham Young University
Provo, UT 84602
[mark_davies@byu.edu]

Serial verb constructions: A cross-linguistic typology. Ed. by Alexandra Y. Aikhenvald and R. M. W. Dixon. Oxford: Oxford University Press, 2006. Pp. xxiv, 369. ISBN 019923342X. $55.

Reviewed by N. J. Enfield, *Max Planck Institute for Psycholinguistics**

This welcome volume offers a set of empirical studies of serial verb constructions (SVCs) framed in a consistent general typological-descriptive framework. Previous work on SVCs has focused on studies of single languages, language groups, or restricted geographical areas (e.g. Matisoff 1969, Thepkanjana 1986, Sebba 1987, Steever 1988, Jarkey 1991, Bisang 1991, Veenstra 1996, Crowley 2002), or has been theoretically and methodologically varied (e.g. Joseph & Zwicky 1990, Lefebvre 1991), more narrowly formal (e.g. Schiller 1991, Déchaine 1993, Stewart 2001), or historical in focus (Lord 1993).

*With thanks to Michael Cysouw, Martin Haspelmath, and Brian Joseph for comments on a draft. This work is supported by the Max Planck Society.