*English Language and Linguistics*, 24.1: 243–248. © Cambridge University Press 2018

REVIEW

doi:10.1017/S1360674318000278

**Paul Baker**, *American and British English: Divided by a common language*. Cambridge: Cambridge University Press, 2017. Pp. xiii + 264. ISBN 1107088863 Hardback, ISBN 1107460881 Paperback.

Reviewed by Gunnel Tottie, University of Zurich[1]

The catchy subtitle of this book is an (undocumented) quotation from George Bernard Shaw; a more appropriate one would have been *A corpus-driven study of variation and change*. The work is clearly not intended for readers whose main concern is to learn about differences between American and British English, but for linguists interested in what can be achieved by using corpora with a *corpus-driven* approach, taking data as a tabula rasa, rather than a *corpus-based* one. As most readers will be aware, corpus-based work is based on hypotheses and prior research but corpus-driven research departs from raw data and uses methods and computer software 'unhampered' by biases inherent in pre-formed hypotheses (cf. Tognini-Bonelli 2001).

The author presents his goals in chapter 1, p. 3: 'to address the following questions: to what extent are British and American English different, and in what ways, and how have these differences altered over the last 100 years?' by using a corpus-driven approach. He is not a naïve practitioner; he admits that his approach also has disadvantages and that 'most research falls on a cline between the two' (p. 5) but aims to use corpus-driven methodology as far as possible. The reviewer therefore needs to ask whether this is useful and successful for his purposes, compared with corpus-based studies, as exemplified by e.g. Leech *et al.* (2009).

In chapter 1 Baker also introduces the 'Brown family' of eight comparable corpora of written English that he uses, four British and four American, comprising material from 1931, 1961, 1991/92 and 2006. He then introduces the methods and measures on which he mostly relies: Keyword search (including Key Clusters, Letter Sequences, and Tags), Coefficient of Variation, and Correlation as well as the computational tools employed for analysis. Good explanations of these measures are given in the text. A Keyword is 'a word [or other member of this category] which occurs more often in one corpus when compared against another' (p. 14); the extent of the difference is identified by means of a statistical test. The Coefficient of Variation is used for considering multiple measurements (p. 16) and 'is calculated by taking the standard deviation of a set of values and then dividing it by the mean of that set of values' (p. 17). The Correlation measure takes all eight corpora into account and indicates whether lines on a graph showing increase or decrease of an item in American or British English have a parallel, divergent, or convergent development. A combination

[1] I thank Sebastian Hoffmann for discussing an earlier version with me; all opinions and/or mistakes are mine.

of these three is used to calculate the range of differences between the eight corpora he used.

Chapter 2 discusses spelling differences between American and British English; it is the least interesting in the book, because of the problem with corpus-driven analysis mentioned in the introduction (p. 5): it 'can tell us what we already know or expect to find'. This applies especially in this age of word-processors tailored to American and British spelling, and Baker also admits that there is not much news here.[2]

Chapter 3, entitled 'Letter sequences and affixation' is much more innovative though not entirely successful.[3] The author's goal is to 'consider variation from the perspective or morphology, particularly change around affixation' (p. 57), but he is aware that 'morphemes and affixes are somewhat difficult to isolate and count' (p. 58). Baker makes ingenious use of WordSmith and Word to surmount some of the difficulties in isolating meaningful sequences of letters, and he succeeds in identifying a number of affixes, such as *-ology*, *trans-*, *techn-*, *-graph* and *-ism,* although no great changes or differences between varieties could be demonstrated. For these affixes he takes a diachronic approach and provides etymological information, but a problem arises with the five-gram affix *-cious* that he postulates. The most frequent word containing this sequence in British English is *conscious* and related forms. Clearly, *-cious* derives from both Latin *-scire* and *-osus*; one wonders whether it is a good idea to discuss only the form *-cious* without including words like *pretentious*, *licentious* or *facetious,* where the ending is pronounced in the same way but spelled with a *t*. The author's purpose is clearly to discuss written English, but the corpus-driven approach seems unsuitable here. Similar objections can be made to the author's treatment of the suffix *-sion* without mentioning *-tion,* as in *attention*, *objection*, *fruition*, *contrition*, etc. The problem of productivity is also treated in very inconsistent fashion – broached with *-ology* and *-ism* and mentioned in the context of *-fess* but left out with most of the affixes discussed. Although I know that this requires a corpus-based perspective, what I particularly miss in a chapter devoted to sequences of letters is at least a mention of the increasing use of alphabetisms and acronyms like *BMI*, *DNA*, *LGBT*, *IRS*, *JFK* or *AIDS*, *NATO*, *NASA*, *NAFTA* etc. (cf. Tottie 2002: 112ff.).[4] These sequences are extreme instances of densification that might have attracted the attention of the author. Couldn't they have been caught by using Baker's ingenious search methods devised for isolating letters and sequences?

The author is on firmer ground when he moves on to discuss words in the following chapters. Chapter 4 deals with higher-frequency words, defined as those that occur 1,000 times across four of the related corpora. Relevant items are first presented in tables and then discussed under the headings densification, democratization and informalization. The problem with this chapter is that, because of the smallness of

[2] Figure 2.1 on p. 31 has a confusing heading at the top: 'Preference for *a* (rather than *e*)'.

[3] The author lists affixes found in the American 2006 corpus as a source of inspiration, among them *-y*, as in *campy*, *buzzy*; the inclusion of *minstrelsy* puzzles me.

[4] Some mostly amused attention has been given to jocular texting abbreviations like *lol*, *imho* etc., but much less to the abbreviations used in newspapers and technical texts, where new concepts, governing bodies and institutions constantly appear and are quickly only referred to by their abbreviated forms.

https://doi.org/10.1017/S1360674318000278 Published online by Cambridge University Press

these corpora, the most key lexical items found were words like *big*, *city*, *several*… (American), and *be*, *have*, *very*… (British). When cut-offs were lowered, clusters like *at a time*, *of the state*, *it will be* were found in American English and *a bit of*, *when I was*, *it is true* in British English, interesting *per se*, but not impossible in the other variety. As Baker points out (p. 121), salient lexical differences discussed by 'other commentators on American and British English were too infrequent to meet the cut-offs … specified for this chapter'.

For chapter 5, Baker therefore lowers the threshold to 100 tokens across the corpora and supplements them with a few intuition-based observations. The chapter treats lower-frequency words that are usually present in standard lists of British–American differences, like *gas/petrol*, *elevator/lift* etc. and therefore offers no great surprises; it will be of particular interest and practical use to future compilers of such lists, however. Discussions and explanations of differences between the varieties are usually correct but there are a few slipups in the area of Americana, e.g. the somewhat ambiguous statements on p. 133 that the word *administration* 'refers to the governing political party' or on p. 134 that 'American English uses *jail* as a short term equivalent of *prison*'. And it is puzzling that the author does not discuss meaning differences between lower- and upper-case spellings of words like *d/Democrat* and *r/Republican* and that he spells *Medicare* and similar terms with a lower-case letter.[5]

Chapter 6 is a major departure from earlier chapters: it is based on corpora annotated for parts of speech, thus abandoning one of the tenets of the corpus-driven approach. The author bases his work on tags applied to individual words as well as fixed sequences and makes good use of the information obtained in this way; results are presented in mostly excellent tables and figures. However, in some areas Baker's discussion of results falls short, and although he has made the excuse that his chapter on grammatical features cannot be as detailed as that in Leech *et al.*'s volume (2009), parts of it are unnecessarily superficial or mistaken. Thus Leech *et al.* specify that their treatment of relativizers comprises only adnominal items (and therefore excludes sentential *what*), they mention the important distinction between restrictive and non-restrictive clauses as well as the *which*-hunt by American editors and spell-checkers. Those restrictions, which Baker omits, are important for the correct evaluation of results. He thus includes *what* among the items that 'regularly introduce relative clauses' (p. 153) and says nothing about the use of *which* in non-adnominal clauses or its survival in Academic English (of which this book is an excellent example).

Baker also links the strong increase in the frequency of *both* in American English to a decrease of *the two* (p. 170). He regards *both* only as a 'densified' equivalent of *the two* and seems unaware of the possible non-equivalence of the two expressions, or the fact that it is not always possible to substitute one for the other; cf. (1a) and (1b), which are at least propositionally equivalent, and (2a) and (2b), where (2b) is doubtful.

[5] On pp. 127–8 there is what is probably an editing mistake: the text reports that *transportation* is preferred to *transport* in British English, but table 5.5 gives the correct information that it is American English that prefers it.

https://doi.org/10.1017/S1360674318000278 Published online by Cambridge University Press

(If there is indeed an ongoing blurring of the two expressions in American English, it would have been interesting to know in how many instances this was the case.)

(1) (a) **The two** writers died in 1616
    (b) **Both** writers died in 1616
(2) (a) **The two** brothers became bitter enemies
    (b) ?**Both** brothers became bitter enemies

Baker concedes that there are no 'ground-breaking discoveries' in this chapter but observes that, by adding new corpora, his approach validates earlier corpus-driven work on grammatical change. Moreover, he points out the fact that the results of his corpus-driven approach agree with those obtained in the corpus-based paradigm; he stresses that this provides strong support for corpus-driven research – but he does not make the point that this also supports corpus-based research, and ultimately, a combination of the two.

Like the previous chapter, chapter 7 is based on annotated corpora, but the tagsets are semantic. Older tagsets used for the Brown Corpora (see e.g. Leech & Fallon 1992) are updated and applied to the newer corpora, and the coefficient of variation and correlational methodology are applied. The results are well presented in tables and graphs; the outcome is of great sociocultural interest but, as Baker himself notes (p. 204), they show no 'major and long-lasting differences between America and Britain, although the increased references to law and order and weapons and war in US English should not be overlooked'.

In chapter 8, Baker acknowledges that there are areas where corpus-driven methods are difficult to apply, such as the study of swearing and profanity, language and identity, and politeness phenomena. In this chapter he therefore definitively switches to the corpus-based methodology, looking for exponents of these categories. He also includes discourse markers, a comeback in written-language research, as most recent research on those has centred on spoken language. There are good reasons for including many of them in research on written language, where they are, as Baker points out, on the increase, and many typically spoken expressions are finding their way into writing.

The items that Baker considers in this chapter are those found in the 'Discourse Bin' from the semantic tagging used for the previous chapter. It is a motley crew that includes both words and phrases: *all right*, *anyway*, *as it is*, *bloody*, *etc*., *on the other hand*, *nevertheless*, *please*, *sorry*, *thank you* and *yeah,* some of them rising in frequency, some declining. The inclusion of some of these is puzzling to me. One example is *bloody*, which had already been treated in the section on Swearing; it is not quoted here in a function typical of discourse markers but as a modifier of nouns and adjectives (*bloody hell*, *bloody marvellous*) which are themselves not serving as discourse markers.[6]

Including *as it is* as a discourse marker meaning 'in the existing circumstances' (p. 226) is totally justified, and uses like the invented ones in (3) and (4) would be natural and idiomatic.

[6] As Baker mentions, *bloody* also has different meanings in the two varieties, British *bloody* 'fucking' and American English *bloody* 'blood-stained'. He points out (on pp. 250f.) that it is possible to be misled by such homonyms and gives the example of *drug*.

https://doi.org/10.1017/S1360674318000278 Published online by Cambridge University Press

(3) I have enough to do **as it is**
(4) **As it is**, I cannot be of help to you

However, none of the examples Baker quotes (on pp. 226–7; my numbering here) is relevant. As he points out, *as it is* is part of a causal construction in (5) and of a comparative construction in (6); neither these two or (7) can be paraphrased 'in the existing circumstances'.

(5) … I [had] better say sorry before I even start **as it is** now compulsory for anyone with an opinion to have to apologise for it a day or so later. (British English 2006)
(6) It is still a popular notion, but **as** false **as it is** popular. (British English 1931)
(7) This social value…is in little danger, important **as it is**, of being underrated. (American English 1931)

In (7) *as it is* is not a fixed expression but an instance of a fully variable syntactic construction, which is interesting in its own right as it can be either causal or concessive and is sometimes ambiguous between the two (Kjellmer 1992; Tottie 2001). See (8) and (9):

(8) Tired **as he was**, he fell asleep immediately. (Kjellmer 1992)
(9) Tired **as he was**, he felt obliged to finish the chapter. (Kjellmer 1992)

In chapter 9, Baker gives a thoughtful summary of processes that he has broached in earlier sections – Americanization, densification, democratization, informalization, colloquialization, grammaticalization and technologization. He discusses possible future developments but wisely refrains from hard and fast predictions.

Baker does not summarize the arguments for and against the corpus-driven approach here but they are discussed throughout the volume; his successive drift towards a corpus-based approach speaks for itself. One aspect of corpus-driven work that is sometimes ignored in this work is the necessity of careful post-editing – the lack of it will result in mistakes like those I have discussed in connection with the author's treatment of relativizers, *both/the two* and *as it is*.

The book is well written and has an abundance of clear tables and figures. Because of both its merits and its shortcomings, I would recommend it for a class or seminar on corpus linguistics, used together with e.g. Tognini-Bonelli (2001), Leech *et al.* (2009) and Andersen (2016), for a fruitful discussion of the pros and cons of corpus-driven vs corpus-based research. I agree with Andersen (2016: 39) that a combination of the two is necessary to produce fully accountable results.

*Reviewer's address:*
*English Department*
*University of Zurich*
*Plattenstrasse 47*
*CH-8032 Zurich*
*Switzerland*
*gtottie@mac.com*

https://doi.org/10.1017/S1360674318000278 Published online by Cambridge University Press

## References

Andersen, Gisle. 2016. Using the corpus-driven method to chart discourse-pragmatic change. In Heike Pichler (ed.), *Discourse-pragmatic variation and change in English: New methods and insights*, 21–40. Cambridge: Cambridge University Press.

Kjellmer, Göran. 1992. Old as he was: A note on concessiveness and causality. *English Studies* 73, 337–50.

Leech, Geoffrey & Roger Fallon. 1992. Computer corpora – what do they tell us about culture? *ICAME Journal* 16, 29–50.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nick Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam and Philadelphia: John Benjamins.

Tottie, Gunnel. 2001. *Tall as he was*: On the meaning of complement *as*-constructions. In Christiane Dalton-Puffer, Arthur Mettinger & Nikolaus Ritt (eds.), *Words: Structure, meaning, function*, 307–21. Berlin: Mouton de Gruyter.

Tottie, Gunnel. 2002. *An introduction to American English*. Malden, MA, and Oxford: Blackwell.

(Received 3 September 2018)