The BE06 Corpus of British English

- 1 million-word corpus of written, published British English
- 500 2000-word texts first published in paper form and later archived on the World Wide Web
- Part of the Brown 'family' of corpora (including BLOB-1931, Brown, LOB, Frown, FLOB, AmE06) in that it uses the same sampling frame (500 files of 2000 words taken from 15 subgenres of writing)
- Available at the Corpus Query Processor system (CQP) at Lancaster University (https://cqpweb.lancs.ac.uk/)
- Produced by Paul Baker of Lancaster University

Creation

- Text collection began in 2007 with the aim of collecting texts from the mid – 2000s, and completed in May 2008
- 82% of texts collected were published between 2005-2007, making the median point of the corpus 2006 (hence BE06)
- These texts were drawn from 4 major categories (Press, General Prose, Learned Writing and Fiction) and 15 subcategories
- As far as possible, BE06 text samples were matched to texts in the LOB, resulting in 37% of the corpus being drawn from beginnings of texts, 26% middles of texts, 5% ends of texts, and 32% full texts
- Where texts were not long enough to make a 2000 word sample, several short texts were combined into one sample
- Problem: What constitutes a ‘British text’?

Annotation/Markup

- Logs were made of the sample's title, author, date published, word count, website address, and whether the sample was taken from the beginning, middle or end of the text
- Lemmatised
- Grammatically annotated with the CLAWS7 tagset and the Oxford Simplified Tagset
- Semantically annotated with the USAS tagset

Usage

- The frequency of pronouns was compared across the 4 British corpora (BLOB, LOB, FLOB and BE06)
- There are higher numbers of all first person pronouns (except *our/ours*) for 2006 than for any other year
- Second person pronouns also increased between 1991 and 2006
- This suggests that colloquialisation (associated with first and second person pronouns) was higher in 2006 than in previous sampling periods
- Male third person pronouns were lower in 2006 than in previous years
- In contrast, there has been an overall increase in female pronouns
- This suggests a move away from male bias in language use. However, female pronouns are still collectively less frequent than male pronouns, suggesting that the male bias still exists

References

- Baker, Paul. 2009. The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14 (3), 312–337
- Hardie, Andrew. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17 (3). 380–409.
- The corpus can be accessed at https://cqpweb.lancs.ac.uk/

ISLRN

Title:	BE06 Corpus
Full Official Name:	The BE06 Corpus of British English
Resource Type:	Primary text
Source/Url:	https://cqpweb.lancs.ac.uk
Format/MIME Type:	Text/xml
Size/Duration:	1010996 words
Access Medium:	Corpus Query Processor System at Lancaster University
Description:	500 2000-word British English texts drawn from the internet which were previously published in paper form, collected from the period 2003-2008.
Version Type:	?
Media Type:	Text
Language(s):	English
Resource Creator:	Paul Baker
Distributor:	Paul Baker
Rights Holder:	Paul Baker