CHAPTER 13

Zooming out

Overall frequencies and Google Books

13.1 Normalised frequencies of the phenomena studied

Tanja Säily

In Part II of this volume we examined variation and change in several features from the perspective of historical sociolinguistics. This chapter widens the focus from historical sociolinguistics to frequency change in general: Section 13.1 compares the overall frequencies of the phenomena studied (excluding *thou*, which is too infrequent for quantitative analysis), while Section 13.2 extends the time period studied to the 21st century by looking at the phenomena in the Google Ngram Viewer and critically analysing the usefulness of the big and messy database of Google Books for studies of grammatical variability.

Biber (2010: 242) talks about the "pervasive linguistic characteristics" of registers, meaning their most frequent and typical features. How frequent are the linguistic features we have studied? Figure 13.1 shows that our most frequent feature is the variable *has/hath*, which constitutes between 0.33% and 0.43% of the words in each time period in the CEECE, undergoing fluctuations and an overall slight increase in frequency over time. The second most frequent feature is the suffix *-ity*, which stays within the range of 0.21–0.29%, also increasing over time (note that we are here looking at its token frequency, whereas Chapter 12 analysed its productivity or type frequency). The rest of the features are surprisingly similar in frequency, varying between 0.02–0.11%. We can see that affirmative DO decreases and the progressive aspect increases within this range over time, while the variables of indefinite pronouns and *its/of it/thereof* seem to fluctuate with less of a clear pattern of change.

We can also look at the changes in another way, by breaking down the variables into individual variants and analysing their normalised frequencies alongside those of the changes lacking a variable. Figure 13.2 shows increases in the normalised frequencies of the incoming variants *does*, *-body* and *its* as well as the progressive aspect (we have chosen *does* instead of *has* here because its frequency more closely matches that of the other changes). Interestingly, while *-body* consistently gained ground in terms of its proportion of the variable of indefinite pronouns

DOI 10.1075/ahs.8.c13

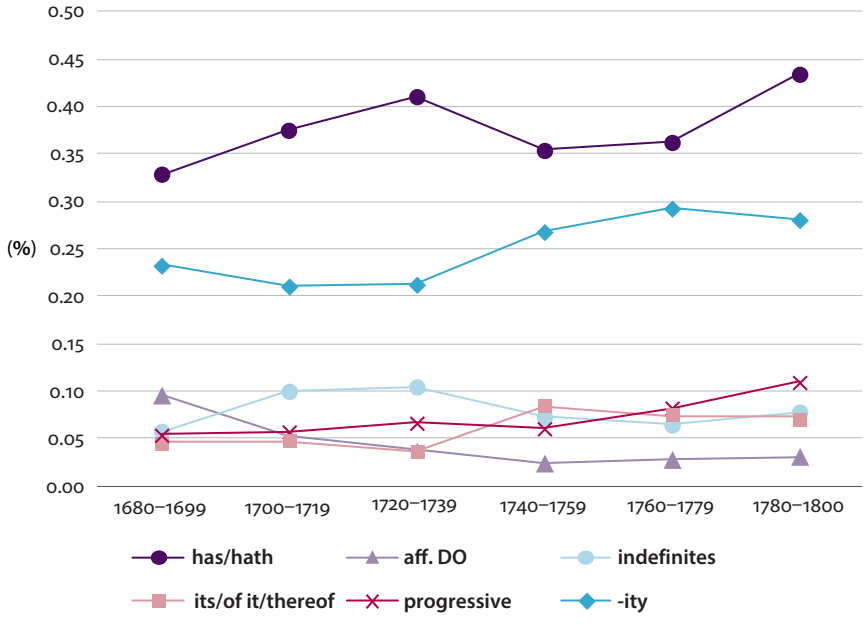

Figure 13.1 Normalised frequencies of the chief variables and other phenomena studied

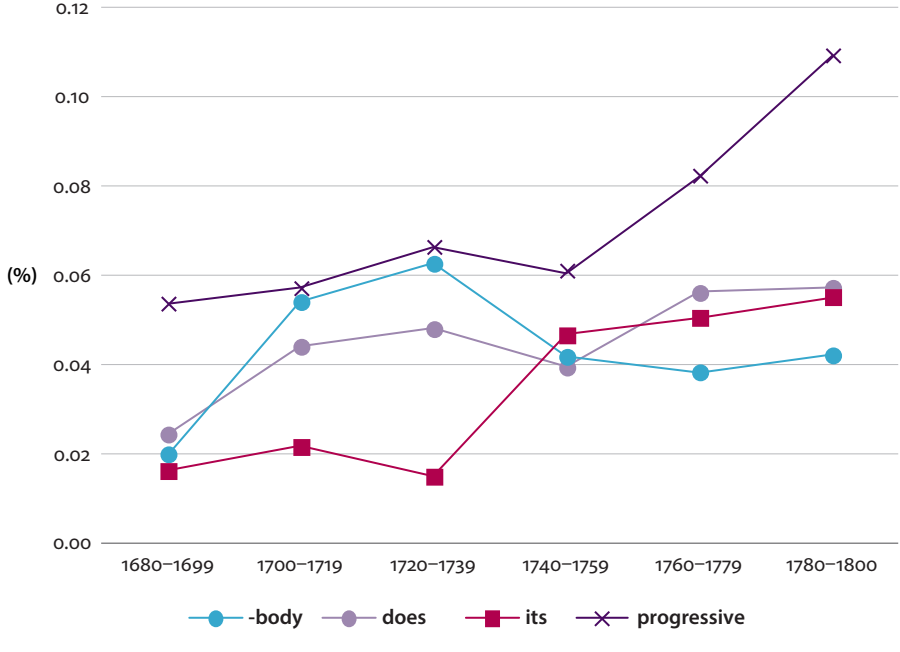

Figure 13.2 Normalised frequencies of three incoming variants and the progressive aspect

with singular human reference (Chapter 9), its normalised frequency does not show a consistent increase as there is a sharp drop between the periods 1720–39 and 1740–59.

We could also compare these frequencies with those found in other sources. The eighteenth century is covered by such materials as the *Old Bailey Corpus* and the *Corpus of Late Modern English Texts*, and it would certainly be useful to complement our data with these larger datasets in the future. As a first step, the next section compares the results we have gained using CEECE with those observable in the massive Google Books database.

13.2 Google Books: A shortcut to studying language variability?

Mikko Laitinen and Tanja Säily

This section looks into evidence provided by the big but messy database of Google Books (Google 2013a), which extends diachronically to Early Modern English, covering also the time period examined in this book. The purpose is to explore what types of broad diachronic evidence related to grammatical variability can be obtained from this massive freely available database with respect to the changes studied in this volume, to complement the insights into sociolinguistic variability offered by our specialized and tailor-made corpus in Part II.

The motivation for this type of comparison of big and messy databases with small and tidy corpora comes from two sources. The first is related to Google itself and the effects its search engines are said to have caused on information seeking in general. Google Books is, just like many but not all of the tools provided by Google Inc., an easily available free tool that enables access to a vast amount of information. The success of Google has led to the coining of a popular phrase, the Google generation, which in popular myth refers to the generation born after 1993 and is used to refer to a generation whose first port of call for knowledge is the Google search engine. In a recent report on information seeking patterns, commissioned by the British Library and JISC (Joint Information Systems Committee), the authors explore the myths around the Google generation and point out that "in a real sense, we are all Google generation now" (Rowlands et al. 2008: 301). There now exists convincing evidence from libraries' deep log statistics that people of all ages use the Internet and its various technologies in surprisingly simple ways. Indeed, the investigation shows that the digital information world is characterized by massive choice, easy access, and simple to use tools, but people "from undergraduates to professors" exhibit "a strong tendency towards shallow, horizontal, 'flicking' behavior in digital libraries" (Rowlands et al. 2008: 300).