Tracing the evolution of gender bias in large diachronic corpora

Nikolaus Ritt, Irene Böhm, Magdalena Schwarz & Vanja Vukovic
University of Vienna

Keywords: gender bias, distributional semantics, corpus linguistics, Google Books

We report on a project that attempts to (a) describe the evolution of gender bias in the use of English during the last 200 years, as well as to (b) understand the dynamics of that evolution.

That language use in most contemporary western societies is gender biased has long been recognized as a socio-linguistic issue of salient concern (Lakoff 1973; Henley 1987; Cameron 1998; Romaine 1999). However, although the phenomenon is increasingly approached with corpus linguistic methods, (Baker 2010; Baker 2013; Baker 2014; Baker 2015; Caldas-Coulthard, Moon 2010; Konnelly 2020; Norberg 2012; Norberg 2016; Pearce 2008), diachronic studies of gender bias are still rare (but see Baker 2014). Our project attempts to fill that gap.

We operationalize gender bias in terms of the collocational strength obtaining between any English noun, adjective, or verb on the one hand, and items taken to refer to 'female' and 'male' humans, on the other, i.e., the pronouns *he* and *she*, and nouns such as *woman*, *daughter*, *mother*, *man*, *son*, *father*, etc. We regard the use of a word to be biased if it collocates more strongly with items referring to one gender than with items referring to the other. To assess collocational strength, we use the odds ratio (i.e. the ratio of the odds of a specific word (e.g. *colleagues*) to appear in a specific construction with a 'male' item (e.g. *his* NOUN) and its odds to appear in the 'female' counterpart of that construction (e.g. *her* NOUN).

Our data base are the 2012 versions of the American and British Google Books Corpora, which we used to trace bias evolution for a (random) sample of – so far – more than 500 nouns, adjectives and verbs in constructions like (*his*/*her*) + NOUN, (*he*/*she*) + VERB, (*he*/*she*) + (*is*/*was*) +ADJ, ADJ + (male or female NOUN). For each word in a construction, we calculated and visualised a bias trajectory. To illustrate this by way of example: in the first half of the 19th century, the noun *colleagues* was about 76 times more likely to appear after *his* than after *her* but is now only more than twice as likely (see figure 1; note that this example of radical bias decrease is not representative of the whole set).

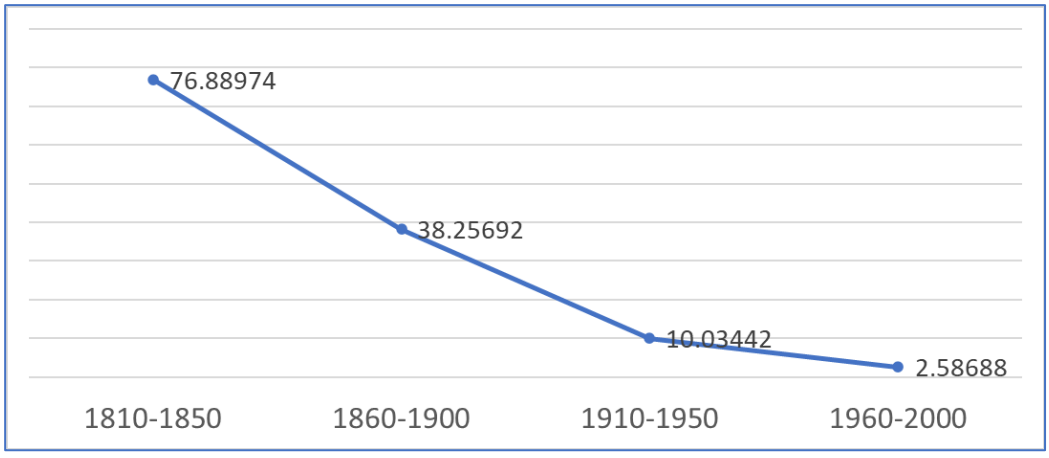

Figure 1: Odds-ratio trajectory of *his/her colleagues* in Google Books (American)

Going beyond the mere description of individual bias trajectories, our talk reports some exciting findings derived from our sample. For instance, our data show that (a) overall gender bias declined consistently but moderately during the 19th century, and remained rather stable during the first half of the twentieth century before declining steeply in the wake of the 1960ies; (b) throughout the observation period verbs seem to have attracted more male subjects but more female objects; (c) bias seems to be more pronounced and more stable among attributive adjectives than among predicative ones; or (d) adjectives indexing stable states

(such as *pretty*, or *strong*) are more resistant to bias loss than adjectives indexing changeable states (such as *tired*, or *sad*).

Overall, our data show that even though gender bias has decreased, it is still strong. At the same time, we think our results provide a promising basis for studying the interaction between deliberate and conscious attempts at bringing about social change and the inertia that characterizes the evolution of conventional language use.

References

Baker, Paul (2010), Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English, *Gender and Language* 4 (1).

Baker, Paul (2013), Introduction: Virtual Special Issue of Gender and Language on corpus approaches, *Gender and Language* 1 (1).

Baker, Paul (2014), *Using Corpora to Analyze Gender*, London: Bloomsbury Academic.

Baker, Paul (2015), *Language and masculinities*: *Performances, intersections, dislocations*, London: Routledge.

Caldas-Coulthard, Carmen Rosa, Rosamund Moon (2010), 'Curvy, hunky, kinky': Using corpora as tools for critical analysis, *Discourse & Society* 21 (2), 99-133.

Cameron, Deborah (1998), *Feminism and linguistic theory*, Basingstoke [u.a.]: Macmillan.

Davies, Mark. (2011-) *Google Books Corpus*. (Based on Google Books n-grams). Available online at http://www.english-corpora.org/googlebooks/.Based on: Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser 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*. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331 (2011) [Published online ahead of print 12/16/2010].

Henley, Nany M. (1987), This new species that seeks a new language: on sexism in language and language change, in: Joyce Penfield (ed.), *Women and Language in Transition*, SUNY Press, 3-27.

Konnelly, Lex (2020), "The woman in the background": Gendered Nouns in CNN and FOX Media Discourse, *Journal of English Linguistics* 48 (3), 233-257.

Lakoff, Robin (1973), Language and woman's place, *Language in Society* 2 (1), 45-79.

Norberg, Cathrine (2012), Male and female shame: A corpus-based study of emotion, *Corpora* 7 (2), 159-185.

Norberg, Cathrine (2016), Naughty Boys and Sexy Girls, *Journal of English Linguistics* 44 (4), 291-317.

Pearce, Michael (2008), Investigating the collocational behaviour of *man* and *woman* in the BNC using Sketch Engine, *Corpora* 3 (1), 1-29.

Romaine, Suzanne (1999), *Communicating gender*, Mahwah, NJ: Erlbaum.