He sang and *she danced.*
English gendered pronouns through the lens of Google Ngram Viewer

Katja Jasinskaja – University of Cologne
katja.jasinskaja@uni-koeln.de

1 Introduction

When I started to work at the University of Cologne in 2014 and became part of Klaus von Heusinger's team, it was not only exciting linguistic research that I got to do, but I also took on some administrative duties. One of my first tasks of the latter kind was to organize a series soft skills workshops for postdocs like myself at the time, and the topic of one of those workshops had to be *Gender and cultural diversity* (*in linguistics*). I must admit that until that point I had never given serious thought to the role of gender in society and remained comfortably oblivious of the extent of its impact on my own life. My attitude changed dramatically after the experience of both preparing and attending that workshop, and I continued to be curious about gender-related issues both in society and in language. In this short essay, I relate some (admittedly superficial) observations concerning the use of the English pronouns *he* and *she* that I made along the way.

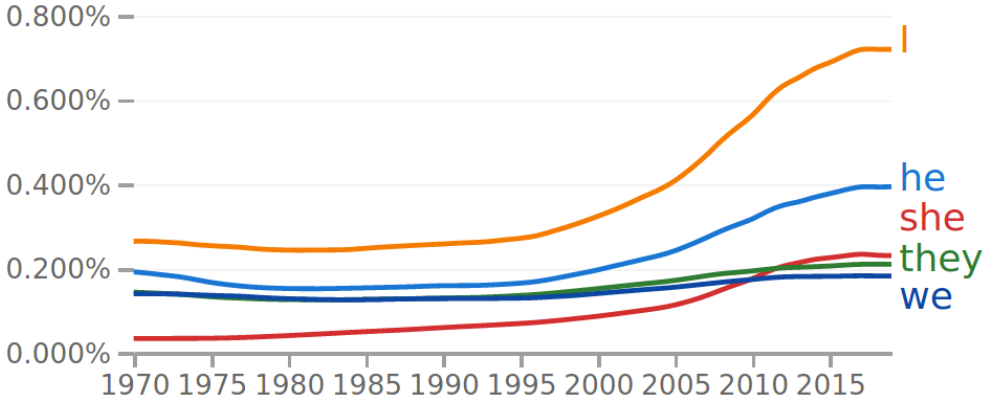
One convenient feature of the English pronouns is that the choice between *he* and *she* is largely determined by their reference, where *he* is used for male and *she* for female human individuals (if we disregard relatively rare uses of *she* for countries, vessels, and vehicles, and the purportedly gender-neutral generic uses of *he*, cf. Huddleston & Pullum 2002: 484–495). Secondly, *he* and *she* are unambiguously nominative, in contrast to the oblique *him* and *her* (same holds for the pronouns *they*/*them*, *I*/*me*, *we*/*us*). Therefore, by looking at the relative frequency of the pronouns *he* and *she* we can get a rough picture of how often a 3d person pronominal grammatical subject of a clause refers to a male or a female individual.

In what follows I report frequencies of the pronouns *he* and *she* over the past fifty years (1970–2019) in different contexts in the *Google Books Corpus*, using the search engine of *Google Ngram Viewer* (https://books.google.com/ngrams). Google Books is a corpus of millions of digitized books, and 361 billion words in its English section only, whose contents is split into case-sensitive n-grams, sequences of blocks of text separated by whitespace (e.g. 'I am' is a bigram, 'I am surprised' is a trigram, etc.). Despite its undeniably impressive size and ease of use, the corpus has a number of limitations, especially for diachronic study of language, due to unbalanced representation of text genres across centuries (an increasing skew towards scientific texts since 1900), poor optical character recognition (especially for older texts), and errors in metadata (Pechenick, Danforth & Dodds 2015). Therefore, all observations reported in this essay should be taken with a grain of salt, as questions for future research rather than ripe conclusions.

2 *he* and *she*

A simple frequency check on unambiguously nominative pronouns (Figure 1) shows that *he* has been persistently the second most frequent pronoun, losing the first place only to the first person pronoun *I* (see Dahl 2001 on egocentricity in discourse). In contrast, *she* used to be the least frequent pronoun of them all until very recently. (*You* and *it* are not included in the graphic, because they are ambiguous between nominative and oblique case, but yes, *she* is also less frequent than those pronouns.) The fact that *she* overtook the plural pronouns *we* and *they* around 2010 instills hope that things might be changing. Even so, in 2015 *she* only has about two thirds (0.22541%) of the frequency of *he* (0.33741%) in Google Books.

Figure 1: Frequency of unambiguously nominative pronouns *I*, *we*, *he*, *she*, and *they*



This, of course, can indicate a number of things. It could be that women are more often referred to by full name and title, than by a pronoun. It could be that women appear overwhelmingly in grammatical roles, other than the subject. But there is also a certain chance that women are not mentioned that much at all.

Seeing this general predominance of male gender among singular human pronominal subjects in English clauses, I asked myself if there were contexts where the pattern would be reversed. More specifically: Are there verbs that have female subjects more frequently than male subjects? To answer this question I performed searches for bigrams consisting of a pronoun *he* or *she* followed by a verb in simple past tense, e.g. *went*, *took*, *said*, *wanted*, etc. The selection of verbs used was largely opportunistic and no statistical analysis was performed. The cases I report here are where the difference to the general pattern was categorical, i.e. *she*+verb bigrams either tying in with or winning against the respective *he*+verb bigrams in frequency.

It is hardly surprising that for most general use verbs like *went*, *took*, *said*, *wanted*, the distribution of *he* and *she* subjects did not differ strongly from the general frequency pattern of the pronouns, the variant with *he* being substantially more frequent than the variant with *she*. On the other hand, it is to be expected that verbs describing activities conforming to traditionally female gender role would be more frequent with female subjects. This indeed is the case for the verb *cooked*, as shown in Figure 2. Interestingly, *he* and *she* are equally frequent with the verb *washed*, although the parity in washing was only established in the 1980s. Before that, disconcertingly, he washed more frequently than she did.

This could be a specific manifestation of the fact that work in general was done more frequently by men (Figure 3). Against this background, it is remarkable that both genders got roughly the same amount of relaxation in the last fifty years (Figure 3), which made me wonder whether verbs with a preference for female gender subjects could be found in the domain of free time activities. Indeed, the gender gap for the verb *sang* is much smaller, although *he* sang still more frequently than *she* did. On the other hand, more recently she beat him in dancing (Figure 4). Moving further into the domain of partner activities: While he was undeniably the more prolific kisser, she took the lead in hugging (Figure 5).

The partner activity that linguists know most about is, of course, that between the speaker and the listener in speech communication. Interestingly, it appears that most of both speaking and listening is done by men. However, I wonder if this impression could simply be the result of differences in decibels, women's voices being overheard due to lower speech volume. As it turns out, she whispered much more frequently than he did in the past five decades, while he shouted more than she did (Figure 6). Moreover, quite often she did not put her thoughts to words at all: While he and she wondered roughly the same amount of time, she actually asked

her question less frequently (Figure 7). Incidentally, Klaus von Heusinger criticized this behavioral pattern in me more than once.

Figure 2: he/she washed/cooked

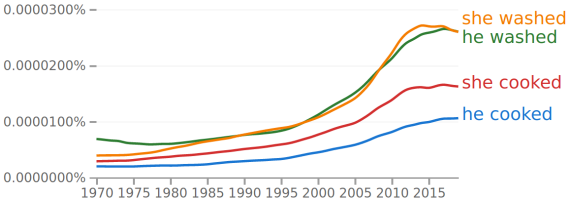


Figure 3: he/she worked/relaxed

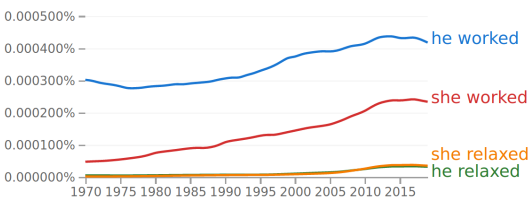


Figure 4: he/she sang/danced

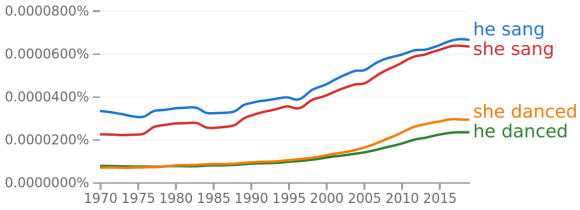


Figure 5: he/she kissed/hugged

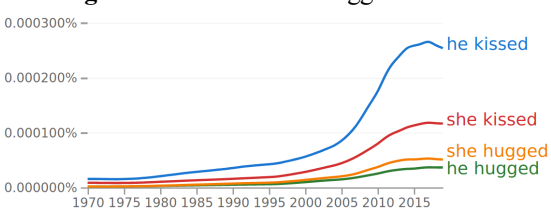


Figure 6: he/she whispered/shouted

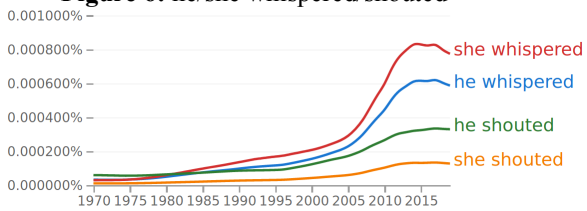


Figure 7: he/she wondered/asked

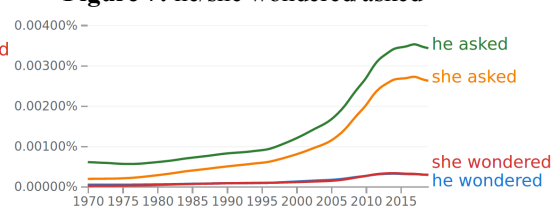


Figure 8: he/she ordered/begged

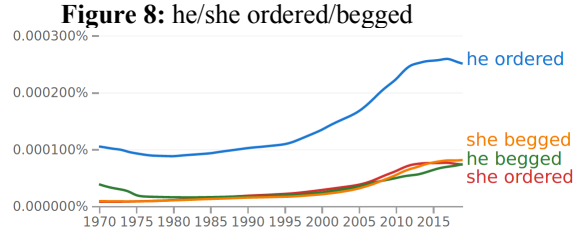


Figure 9: he/she said yes/no

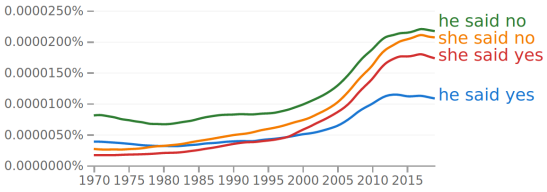


Figure 10: he/she blinked/blushed

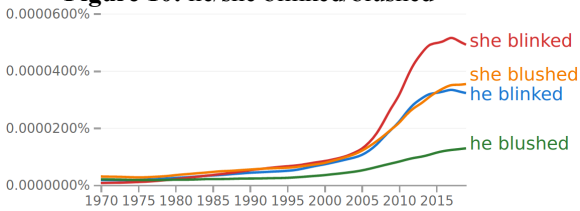
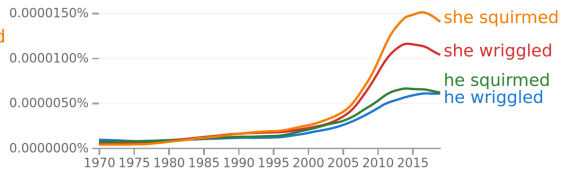


Figure 11: he/she wriggled/squirmed



Of course, it is easier to perform directive speech acts when you are in a position of authority. Differences in authority between male and female subjects can be seen in the distribution of pronouns with verbs like *ordered* and *begged* (Figure 8): While male and female subjects were roughly equal in the amount of begging they did, unsurprisingly, male subjects ordered more frequently. Finally, it is all a question of *what* you say – how you respond to questions, orders, requests, and pleas. Here the picture for female gender subjects is much more positive: While he said *no* more often than she did, she said *yes* more (Figure 9).

However, of all the verbs I have looked at, I was most surprised to find a clear preference for female subjects in the verbs *blinked*, *blushed*, and *wriggled*. I had never thought that women blinked more than men (Figure 10), but apparently they do (see Sforza et al. 2008). I could

have guessed that women might blush more than men, or at least, that blushing in women might have greater social significance (see e.g. Crozier 2016), but I did not expect to find the greatest relative gender gap to the advantage of the female with the verb *blushed* among all the verbs I considered (Figure 10). Finally, the verb *wriggled*, as well as its near-synonyms *squirmed* and *writhed*, turned out to have a large majority of female gender among human pronominal subjects (Figure 11), which was to me personally the most fascinating finding. I wonder (and ask) whether this could be because blinking, blushing, and wriggling is something we see other people do, but rarely admit to doing ourselves. A point of view character in a story rarely blinks, blushes or wriggles. Interestingly, the verbs *blushed*, *wriggled*, and *squirmed* also show a relatively low co-occurrence with first person subjects, while otherwise the pronoun I is the most frequent (cf. Figure 1). Could the relatively low frequency of male gender pronouns as subjects of these verbs be the consequence of the narrator taking his perspective?

3 Conclusion

These observations made me worry: How will Google Books affect the vulnerable artificial mind of a robot trying to navigate the complexity of human life? What will it learn about the human female from all those books? As far as I can see, there is nothing to prevent it from getting the impression that a woman is someone who cooks and hugs, probably wonders why, wriggles and squirms (not to be confused with dancing), but ultimately says yes, in a whisper. Or is there? I cannot express enough gratitude to Klaus von Heusinger to bringing these questions to my attention, wondering about them was a life-changing experience for me. However, the questions remain questions. There is no conclusion, except, perhaps, that those of us endowed with natural intelligence should write more books, with more verbs, and more pronouns.

References

Crozier, W. Ray. 2016. The blush: Literary and psychological perspectives. *Journal for the Theory of Social Behaviour* 46(4). 502–516.

Dahl, Östen. 2001. Egocentricity in discourse and syntax. *Functions of Language* 7(1). 37–77.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10(10). e0137041.

Sforza, Chiarella, Mario Rango, Domenico Galante, Nereo Bresolin & Virgillio F Ferrario. 2008. Spontaneous blinking in healthy persons: An optoelectronic study of eyelid motion. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)* 28(4). 345–353.