Follow-up to Language Log guest post 24 March 2018

This is a follow-up to my guest post on Language Log on 24 March 2018 (http://languagelog.ldc.upenn.edu/nll/?p=37417), on the German expression *ich gehe davon aus* ('I assume') and the apparent fact that it attracted the attention of purists and peevers only after it started to become relatively more frequent in the 1970s. This follow-up addresses the discrepancy between what I said in the post and the Google n-gram graph that Mark Liberman included in his comments.

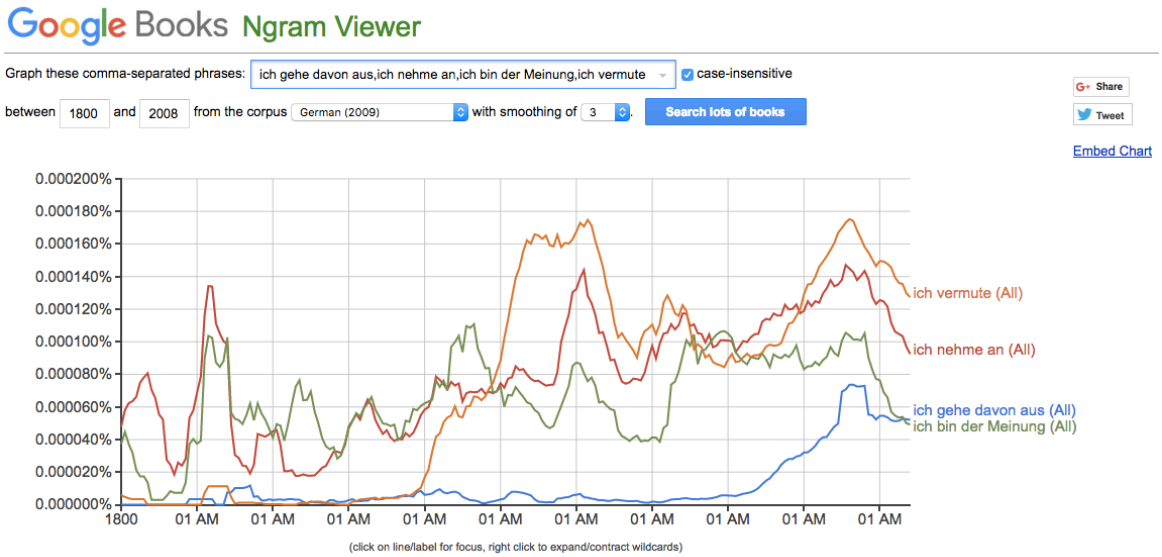

Fig. 1: The graph on which I based my statements in the Language Log guest post

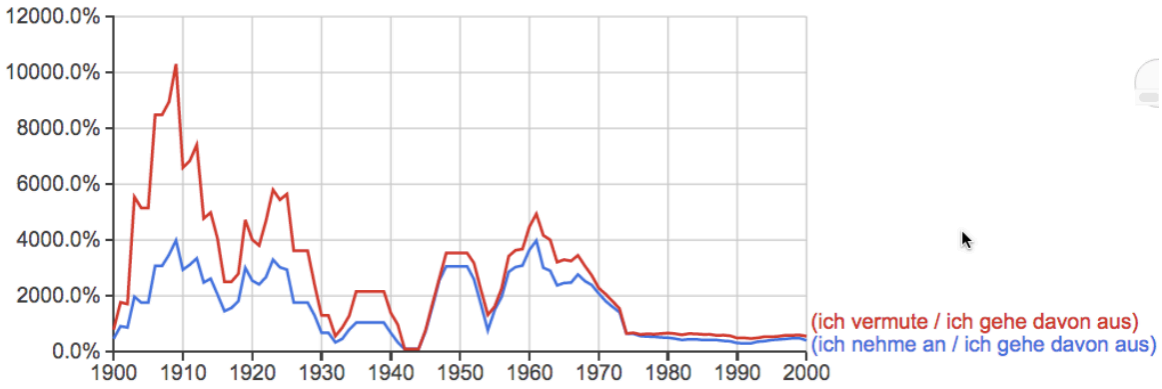

Fig. 2: The graph Mark included in his comments on my post

Figure 1 shows the original Google n-gram data on which I based my statement that frequency of *ich gehe davon aus* started to increase substantially in the mid-1960s (note that the labels on the x-axis are affected by a glitch that was discussed on Language Log a few years ago (http://languagelog.ldc.upenn.edu/nll/?p=24717). The interval between the x-axis labels is 20 years.) Figure 2 is a copy of Mark's graph; in addition to showing that he's not subject to the x-axis glitch, it seems incompatible with what I said in the post, though part of the difference is that Fig. 1 shows (roughly) absolute frequency and Fig. 2 shows the frequency of two expressions relative to each other.

The explanation of the apparent discrepancy involves three separate factors. First, and fairly trivially, it seems likely that Mark and I have accessed slightly different Google corpora or have used slightly different settings. Figure 3 shows what I get when I run the same comparison as in Figure 2. There are certainly differences, but the overall pattern seems broadly similar, and there's still no sign of any increase in *ich gehe davon aus.*

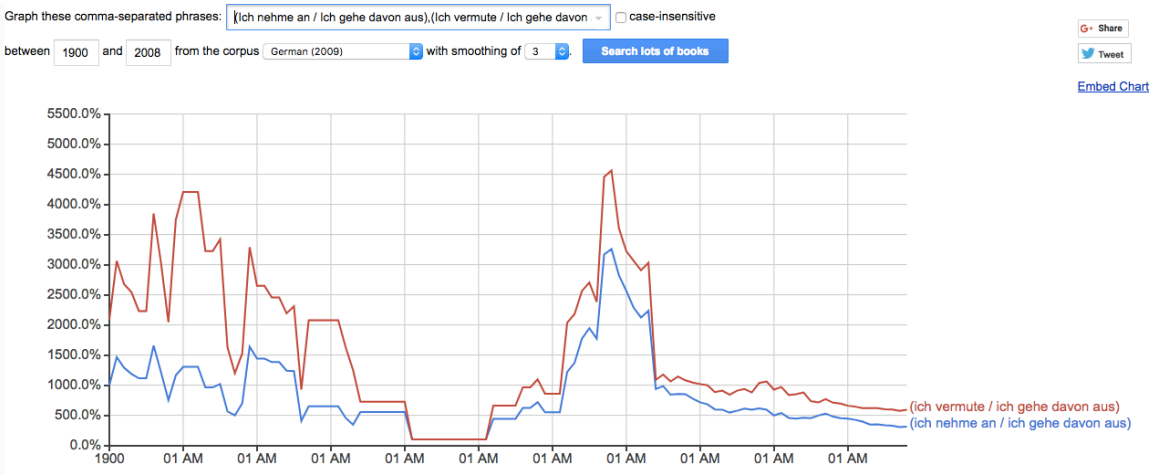

Fig. 3: The graph that I get from Google n-grams for the same comparisons in Fig. 2

Second and somewhat more significant is a possible effect of our pattern recognition skills: by putting *ich gehe davon aus* in the **denominator** of the quantitative comparisons, as is done in Figures 2 and 3, the graph needs to take the y-axis up to 5000% or even 10000% to allow for the early days of the 20th century when *ich gehe davon aus* was extremely rare, so that after 1970 the vertical scale is too compressed to see any change. If instead we invert the proportions, with *ich gehe davon aus* in the **numerator**, the pattern I discussed shows up clearly, as can be seen in Figure 4. (I have no idea where the upward blip in both curves in the early 1990s comes from.)

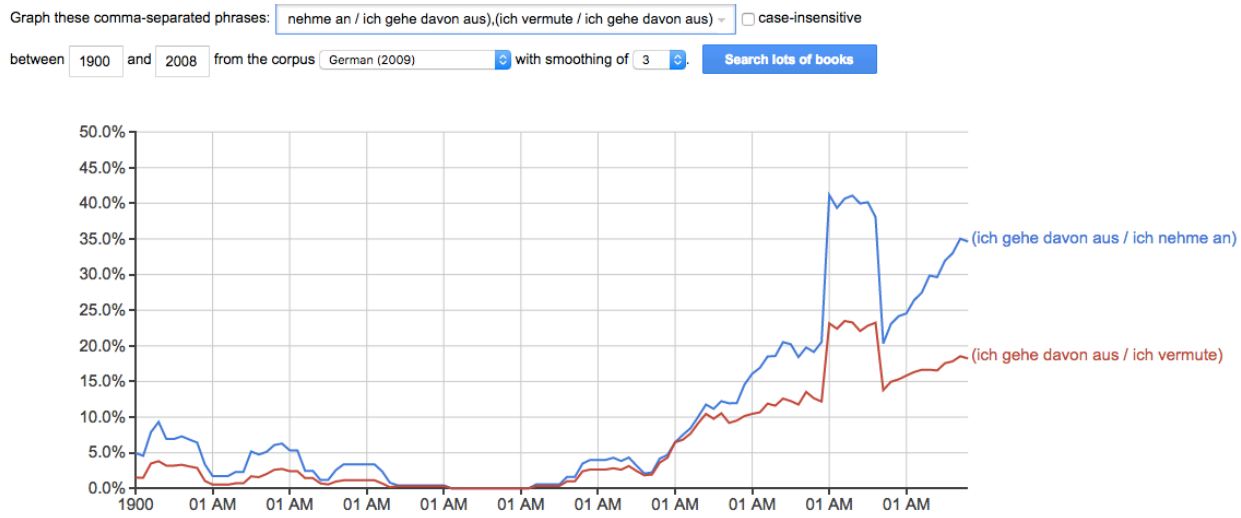

Fig. 4: The data from Fig. 3 plotted with the proportions inverted

Third and perhaps most important, my original data in Figure 1 were based on a **case-insensitive** search, which means that it will include both sentence-initial and sentence-internal instances of the phrases in question. But the n-gram viewer tells me that "case-insensitive searches and compositions cannot be combined", which means that the

graphs in Figures 2-4 are based only on sentence-internal instances. Since all of these phrases are frequently found sentence-initially, this is undesirable. If we redo Figure 4 asking for the same expressions beginning with upper-case *Ich*, we get the picture in Figure 5, which shows that the relative increase in sentence-initial *Ich gehe davon aus* is even greater. (It also, for some reason, makes the early-1990s blip less prominent.)

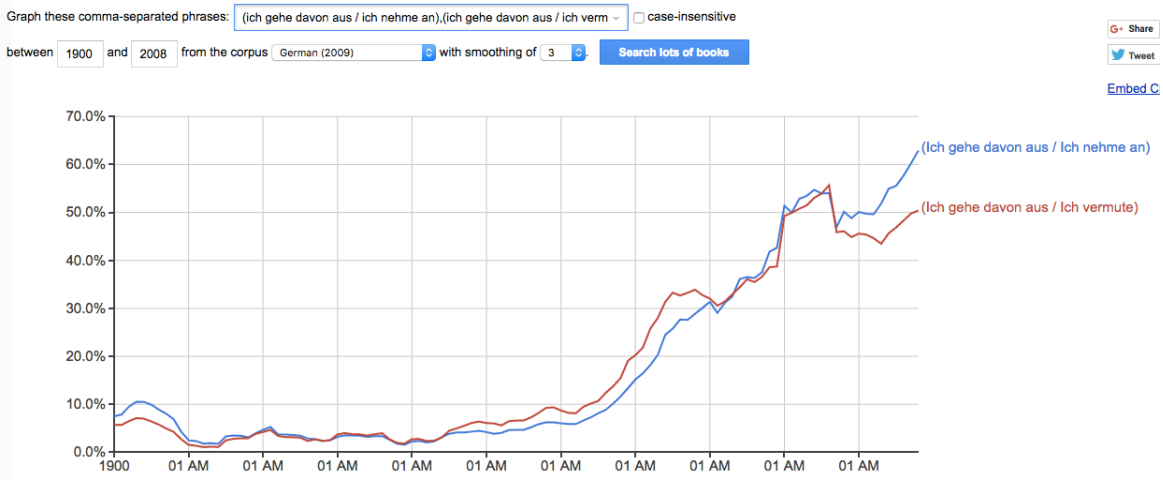

Fig. 5: Data for the same phrases as in Fig. 4, but only sentence-initial.

All of this shows that some caution is required in drawing conclusions based on Google n-grams, but it does suggest that there really has been a change in the relative frequency of *ich gehe davon aus.*