Statistical Analysis of Corpus Data with R

Exercise Sheet #2

In this exercise, you will familiarize yourself with the *zipfR* package, although, especially for the second part, some more general R skills will also come in handy.

1 Type richness of German NP and PP constructions

Although the notions of productivity and type richness have been used mostly in morphology, similar ideas can also be applied to the study of syntax. The *zipfR* package comes with data sets reporting frequency information about the "expansions" of the German NP and PP phrases in the German TIGER treebank (use `?TigerNP.spc` for more information).

Import the German NP and PP data sets. For the data set with more tokens, compute a binomially interpolated vocabulary growth curve (VGC). For the smaller one, estimate an LNRE model (pick your favorite model, or try them all) and use it to compute an expected VGC up to the size of the larger data set. Plot the interpolated and extrapolated VGCs, and determine which of the two constructions appears to be more productive (in terms of type growth).

2 Data lost with cut-off points

Before you tackle this part of the exercise, read the second case study in the *zipfR* tutorial.

It is common, in collocation studies and similar work, to discard bigrams (i.e., sequences of two words) below a certain occurrence threshold, typically $f < 2$ (discarding bigrams that occur once, the *hapax legomena*) and $f < 5$ (discarding bigrams that occur 4 times or less). In this exercise, we try to assess the effect that such cuts have on the proportion of types considered, on the basis of a sample from a relatively small corpus.

First, load the file `bigrams.100k.tfl` (containing bigrams extracted from the first 100,000 tokens of the Brown corpus) as a *zipfR* type frequency list, and generate a frequency spectrum from this (see the *zipfR* documentation).

What proportion of bigram types occur only once? What proportion of types occur 4 times or less?

Now, suppose that we want to use our data to estimate the proportion of bigram types that would be lost by using the same two frequency cut-offs if we had a 1 million word corpus. How will you go about it?