

1 Three-site attachment experiment series: The pretest

To verify the suitability of the experiment items, a pretest was done. First, we threw out items where, according to our intuition, it was plausible that the first and the third NP would belong together. This can cause a bias due to the structure of parse trees of such sentences, so it had to be avoided.

The main pretest was an offline plausibility rating experiment: 23 people filled in a questionnaire on the web, giving the plausibility for each of the three attachments for each potential experiment item. The goal was to remove items that showed a clear bias towards one or two of the attachment possibilities.

Given the sentence *De acteur in de film over de stuntman die populair was* (the actor in the movie over the stuntman that was popular), the three cases to rank would be:

- De acteur die populair was
- De film die populair was
- De stuntman die populair was

The ranking was done on a scale from one (bad, not plausible) to five (good, plausible) for each of the cases. The instructions asked to tell how good each of the “words” (we avoided the technical term NP) would fit the relative clause, and the subjects were encouraged to give the same score for more than one case if appropriate. The complete questionnaire can be found in the appendix and on the web. When gaps were found in the questionnaire, all three cases of the concerning sentences were removed from the data for that questionnaire.

After collecting enough results, a statistical measure was computed from the logged data of the questionnaires. For this, the data was first *normalized*: The mean for each subject was subtracted from all values, so that the new mean would be zero for each subject. Then, the values were scaled to give a standard deviation of one. In this way, individual differences in the general treating of the questionnaire were removed. If, for example, one subject considers ranking three as normal and one and five as bad and good, while another one thinks most cases are good for a ranking of four, and never worse than three, the further evaluation would have been biased without the normalization.

After the normalization, the *mean* and *standard deviation* were calculated for each case of each sentence, using the data of all questionnaires together. This data was used for additional checking of the results by hand: One sentence was removed by hand: Though all three attachment possibilities were similar in plausibility there, the overall plausibility of the cases (*De puree/De schaal/De taart waar over gemorst is*) was quite bad (average of -1.15).

The main check, however, was done automatically: For each sentence, an *F-value* telling how strongly the three cases differed was computed from the data. Then, the sentences were sorted using that value, giving a list with F-values from 0.02 to 6.17. All sentences with a value of more than 2.4 ($p = 0.1$) were removed, as well as the generally implausible sentence mentioned above. Our selection of $p = 0.1$ was done to reject even items with a low possibility of bias towards one or two of the possible attachments. Finally we had 48 experiment items remaining for the use in the two production experiments.