

Assignment 2 - Experimentation in Psychology, Linguistics and AI

Alexander Kern (5711088) & Benedetta Ghedi (7005571)

Hypothesis and research question

In our first experiment we investigated the influence of edit distance on perceived word-pair semantic similarity. We conducted a survey and collected data from 10 participants. The analysis of the data collected confirmed the null hypothesis: word-pairs with a similar cosine distance do not vary in perceived semantic similarity based on the edit distance. The power analysis suggested that a considerably larger number of participants is needed to yield significant results. The t-test conducted to investigate our hypothesis yielded a value of .968: there is no significant difference between the groups of word pairs. We believe that, beside the limited number of participants, the experiment was designed with an insufficient number of stimuli and, in particular, of fillers: some of the participants sent us questions regarding the syntactic similarity of many word pairs, suggesting that they were aware of the target of the investigation. We also believe that the phenomenon under investigation, if existent at all, is very subtle and would require a large scale experiment. We therefore decided to change the variable to investigate, using a similar design to our first experiment: instead of investigating the influence of the edit distance on perceived similarity, the experiment will test for the influence of concreteness on perceived similarity. We believe that the design of the experiment is better suited for the new hypothesis because the level of concreteness of words is more closely related to meaning and is processed more implicitly: we do not expect participants to actively notice the different groups of word pairs, unlike the previous experiment.

Our new hypothesis can be formulated as follows:

“Word pairs with a similar cosine distance vary perceived semantic similarity based on their average concreteness score.”

The null-hypothesis of our experiment goes as follows:

“Word-pairs with a similar cosine distance do not vary in perceived semantic similarity based on their average concrete score.”

We used the Simlex-999 database and calculated the average concreteness score for each word-pair. We selected three groups of word-pairs: one group of fillers and two groups with a similar cosine score, one rating high on average concreteness and one rating low. The groups consist of 20 word pairs each. We expect to collect data from at least 20 participants. We increased both the number of word pairs and participants in order to yield more significant results.

Methods

To test the effect of word pairs that vary in their level of concreteness we created three groups of word pairs, which were integrated in the survey. The first two groups represented the word pairs of interest, the third group consisted of filler questions for the survey. For all the words in the dataset we computed the mean concreteness score, which is the average of the concreteness score of both words and ranges between 1 and 5. The word pairs of interest were required to have a similar cosine distance, with either high or low mean concreteness score. The first 20 word pairs had a cosine distance between .4 and .6, and a mean concreteness score below or equal to 2.2. The second group consisted of 20 word pairs with a cosine distance also between .4 and .6, however, they were chosen to have a mean concreteness score above 4. We expect the group consisting of word pairs with a low mean concreteness score to be perceived more semantically similar than the group consisting of word pairs with a high mean concreteness score.

We used the same design as our last experiment, but we made some changes to our groups. Instead of having three groups of each 10 word pairs, we have enlarged the groups to 20 word pairs each, getting the total of the survey up to 60 word pairs. There were two reasons for this. First, we noticed that with only 10 word

pairs per group it was difficult to find a general effect. Secondly, participants in the first experiment replied that during the experiment they knew what the experiment was about and noticed which items were the fillers. That's why we paid more attention to the selection of word pairs for the filler group and also added more word pairs to the filler group. Our results also showed that we needed a higher amount of participants so we envisioned at least 20 participants partaking in this experiment.

Participants were first asked to consent to the study and were notified that they were allowed to stop at any point before the word pairs were shown. Word pairs of all groups were randomized in order, and were each presented to the participants on a different page accompanied by a ten-point scale. Participants were instructed to give a high score to word pairs which they thought were similar in meaning and lower scores to pairs they found semantically different. The survey was created and distributed using qualtrics.

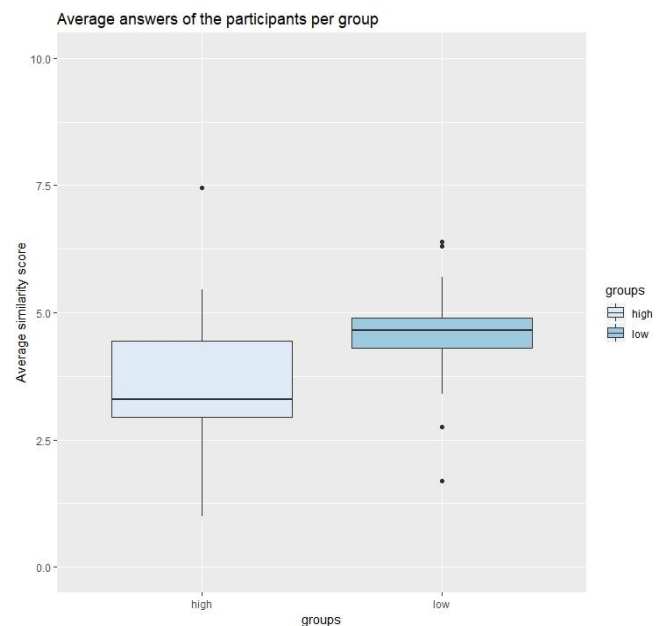
Results

An unpaired two-samples T-test was used to evaluate if the means of the two different conditions are different, according to the next hypothesis:

“Word-pairs with a similar cosine distance vary in perceived semantic similarity based on their average concrete score.”

Before a T-test could be used, the data was checked on the next three assumptions. The two samples were independent, were both normally distributed and had an equal amount of variance. We used the Shapiro-Wilk test to test for normality and a F-test to test for homogeneity in variances.

The T-test had a p-value lower than .05 which means that we can reject the null-hypothesis. The test showed that the group consisting of word pairs with a high mean concreteness score ($M = 3.579$, $SD = 1.421$) compared to the group of word pairs with a low mean concreteness score ($M = 4.495$, $SD = 1.063$) were seen as more dissimilar, $t(40) = -2.366$, $p = .0229$. The boxplot shows the distribution of the average answer of the respondents on the different groups. A Cohen's d was also computed post hoc, which found an effect size of .886.



4. Discussion and Conclusions

The analysis of the data collected from the survey confirmed our hypothesis: the concreteness score of word pairs influences their perceived semantic similarity. In particular, amongst word pairs with cosine similarity score between .4 and .6, those with low average concreteness score were perceived as more similar than those with high average concreteness score.

We did not make sure that all participants are native English speakers: our results could derive from the fact that non-native speakers might have a harder time grasping the meaning of abstract terms compared to concrete ones, skewing our results. Further studies would be taking this into account and test the hypothesis on a larger number of participants.

The results may bring forward new investigations on the reasons behind this phenomenon. From a cognitive point of view, for example, it can be argued that since the meaning of high concrete words is grounded in the real world, it is well defined and they hold a specific semantic content which is easier to compare and therefore differentiate. We can stipulate that sensory experience is strongly linked to word meanings, supporting the embodied cognition theory.