1 Further Explanation about User Test

1.1 Questionnaire Survey

We used the User Experience Questionnaire (UEQ) as developed by Schrepp et al. to measure the user experience (UX) of our product. UEQ is a well-known and popular method for such a UX measurement. Further, providing our results using UEQ allows other approaches to compare their results with ours. The UEQ measures three different qualities of the product:

- *Pragmatic quality*: provides a representation of the basic usability, e.g., is the product considered attractive, efficient and reliable?
- *Hedonic quality*: provides a representation of aspects that do not have a clear connection to the task-related goals, e.g., is the product considered stimulating and innovative?
- *Attractiveness*: presents a combined representation of the general appearance the product has on users.

To specify these three qualities further, they are divided into six scales as presented in table 1. To measure the scales, the UEQ contains a set of items as presented in Appendix X. Each item consists of two terms that have the opposite meaning. The order of the terms is random, which means that half of the items start with the positive term and the other half of the terms start with the negative term. A 7-point system is used to reduce central tendency bias. An example of an item is:

demotivating o o o o o o o *motivating*

The items are divided between -3 and +3. -3 represents the most negative answer and + 3 represents the most positive response. To validate the design of the UEQ, Schrepp et al. measured the validity of the items by conducting 11 user tests with a total of 144 participants along with an online questionnaire with 722 participants. The results of this study show for each scale reliability, as measured by Cronbach's Alpha, varying from 0.69 to 0.86. Cronbach's Alpha provides an estimate of the internal consistency among test scores and a consistency of 0.7 $\leq \alpha$ is commonly acknowledged as acceptable. Therefore, we conclude that the results of Schrepp et al. are sufficient enough for us to use the UEQ in our user test.

1.2 Procedure for UEQ

Test sessions were conducted in person or via a video connection that showed the face of the participant as well as the screen. In all cases, participants were in (semi)-private spaces, such as at home or a quiet workspace.

Quality	Scale	Description
Attractiveness	Attractiveness	General impression of the product. Do users appreciate the product?
Pragmatic	Perspicuity	Is the product easy to use? Is it easy to learn to use the product?
Pragmatic	Efficiency	Can users perform tasks efficiently?
Pragmatic	Dependability	Do users have the feeling to have control over the product?
Hedonic	Stimulation	Is the product appealing and motivating to use?
Hedonic	Novelty	Is the product innovative? Do users have an interest in the product?

Table 1. UEQ measurement scales

UEQ Scale	Mean	Standard deviation	Confidence (p=0.05)	Comparison to UEQ benchmark	Interpretation
Attractiveness	1.44	0.57	0.29	Above average	25% of results better, 50% of results worse
Perspicuity	0.62	0.79	0.40	Bad	In the selection of the 25% worst results
Efficiency	1.58	0.65	0.33	Good	10% of results better, 75% of results worse
Dependability	0.78	0.80	0.41	Below average	50% of results better, 25% of results worse
Stimulation	1.32	0.69	0.35	Good	10% of results better, 75% of results worse
Novelty	1.28	0.78	0.39	Good	10% of results better, 75% of results worse

Table 2. UEQ Results for N=15 compared to UEQ benchmark

Participants first were given a general introduction into the task of relation extraction from text, without them seeing the system (no tutorial how the system work). After they obtained a basic level of knowledge of relation extraction, participants were given two tasks. The first task was to verify whether an already verified triple was correct for a given sentence. After this task, participants immediately went to the second task, relation extraction. The task of relation extraction was to extract as many relations they could find. If they either ran out of possible relations or got bored with the sentence, they had permission to go to the next sentence. All participants were shown the same sentences to minimize a possible difference in the difficulty of a sentence.

After playing three sentences, the observer asked the participants to fill in the UEQ. The UEQ was conducted immediately after the performed tasks and participants were asked to answer the questions conforming their experience and perception. After the UEQ, the semi-structured interview was conducted.

1.3 Data Analysis

The UEQ is analyzed with an Excel-based analysis tool, developed by the UEQ community (www.ueq-online.org). To detect potential random answers given to the UEQ, we checked how much the best and worst response contributing to a UEQ scale differed. When there was a big discrepancy in answers (¿3), we examined this as an indicator for ambiguous data. Based on this heuristic, we removed one response from a total of 16 responses, leaving 15 responses for further analysis. To determine the precision of the UEQ answers relative to our sample size, we examined the confidence of the answers, as presented in table 2. We report that the 95% confidence interval for the mean (μ) of each scale ranges from about $\mu \pm 0.29$ to about $\mu \pm 0.41$. Therefore, we conclude that our current sample size mirrors an acceptable indication of the UEQ scale scores.

Furthermore, table 2 shows the mean and the standard deviation (SD) for each scale. According to the UEQ, a mean between -0.8 and 0.8 represents a neutral evaluation. Scores ¿ 0.8 describe a positive evaluation and scores ¡ -0.8 represent a negative evaluation. The range of the different scales varies from -3 (extremely bad) to +3 (extremely good). Because the mean is calculated from the results of all participants with divergent opinions, it is unlikely to get a score above +2 or below -2.

To interpret these results, we use a UEQ benchmark, as developed by Schrepp et al. We consider this benchmark useful since this is our first product evaluation and therefore we do not have comparison material. The UEQ benchmark exists of data from 246 product evaluations that used the UEQ in a wide range of applications, such as, but not limited to, business applications (100) and web services or shops (64). In total, the UEQ benchmark consists of 9905 responses. The sample size differs per evaluation from 3 to 1.390 participants. The average amount of participants is 40.26. The feedback of the UEQ benchmark limits to five different categories: excellent, good, above average, below average and bad. Table 2 shows the exact numbers.