Investigating Quality Raters' Performance Using Interface Evaluation Methods

Michael Röder[1,3], Maximilian Speicher[1,2], and Ricardo Usbeck[1,3]
[1]R&D, Unister GmbH, Leipzig, Germany
[2]Chemnitz University of Technology, Germany
[3]Leipzig University, Germany
{michael.roeder, maximilian.speicher, ricardo.usbeck}@unister.de

1 Introduction

In today's IT industry, it is common to be faced with quality evaluation tasks such as rating automatically extracted word sets regarding their coherence, which are difficult to solve for computers. Tasks like these can be solved by human quality raters (QRs), who are different from average crowd workers (cf. [How06]) in the sense that they have more expertise. In the context of this paper, we assume QRs to be in-house workers that do not require additional quality assurance contrary to, e.g., the Find-Fix-Verify crowd programming pattern [BLM+10].

There are two ways of posing a total of n items for evaluation to a QR, i.e., either rating one item at a time or pairwise comparisons of all items. The first approach requires n ratings/steps for processing all items, but the QR lacks a broader overview of the presented data. The second approach is supposed to generate higher-quality results [Saa08], but requires $\frac{n(n-1)}{2} > n$ pairwise ratings to process all items.

To investigate a trade-off between the above options, we created an extension to the first approach showing three items per step (Fig. 1, top) and compared it against a single-item interface (Fig. 1, bottom) using state-of-the-art interaction tracking methods. Results show significant differences between ratings of the same items in the different interfaces.

2 Task and Design of Tools

The investigated use case is the evaluation of word sets comprising five words regarding their coherence[1]. The task of the QRs was to rate the understandability of a word set as either *good*, *neutral* or *bad*. The QR tool created for this task featured two different interfaces showing one or three word sets per step respectively (cf. Fig. 1).

[1]The complete word sets can be obtained from `http://topics.labs.bluekiwi.de/data/gi2013`.

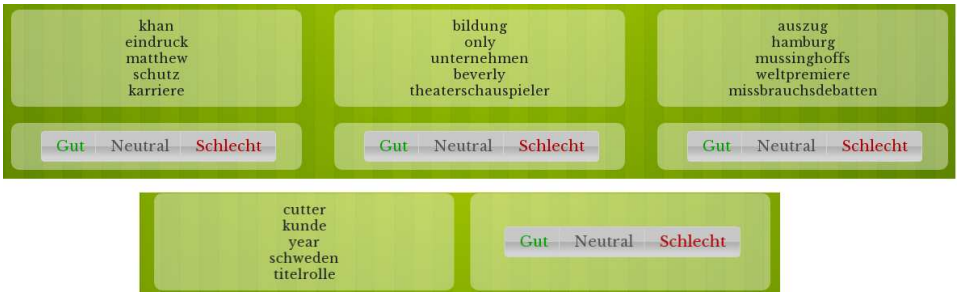

Figure 1: Interface 1 features three items per step (top) while Interface 2 features only one item per step (bottom).

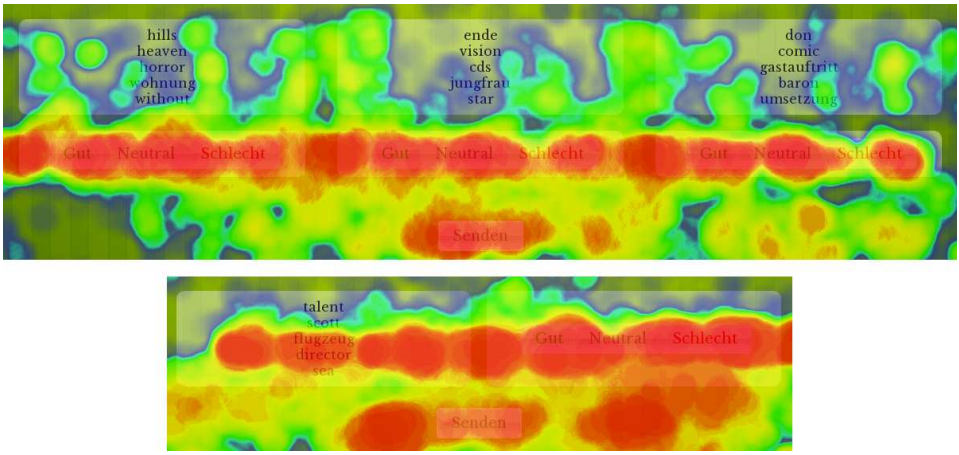

Figure 2: Cursor interaction heat maps for Interface 1 (top) and Interface 2 (bottom).

3 Method & Evaluation

For evaluation, we performed a split test by showing each interface to the same amount of QRs. We engaged state-of-the-art interaction tracking based on existing research [AWS06, NC12, SBG13] for quantitative and qualitative analysis and comparing QR behavior.

The average working time per word set (≈12 s) did not show significant differences between the two interfaces. Also, we tracked nearly the same average numbers of rating button clicks per word set (≈2.4–2.5). More than one click per word set implies that QRs changed their decision. This indicates that QR efficiency is not significantly influenced by a small difference in the number of displayed word sets. A heat map analysis underpinned this finding by showing two homogeneous usage patterns with no larger deviations between interfaces (cf. Fig 2). That is, most interactions happen on the rating buttons while interactions with the word sets are slightly more dense in Interface 2. This is most proba-

bly due the facts that a) users can more easily focus on a single word set and b) the single word set is vertically aligned with the rating buttons. In contrast to efficiency, comparing ratings of the same word sets (N=578) showed significant differences (Wilcoxon signed rank test: V=3381, $p<0.001$, α=0.05), with more *good* ratings when using Interface 1 and more *neutral/bad* ratings when using Interface 2.

4 Conclusion

In a real-world setting we showed that different interfaces w.r.t. the number of displayed word sets affect the results produced by QRs while maintaining their efficiency. However, we can not yet decide how much this difference is based on human factors rather than the interface alone. We intend to investigate this in the future by, e.g., changing the set-up of the split testing environment. The presented methods will support the adaptation of QR tools to human needs and showcase their potential for research as well as industrial use.

Acknowledgments This work has been supported by the ESF and the Free State of Saxony.

References

[AWS06] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. Knowing the User's Every Move – User Activity Tracking for Website Usability Evaluation and Implicit Interaction. In *Proc. WWW*, 2006.

[BLM+10] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A Word Processor with a Crowd Inside. In *Proc. UIST*, 2010.

[How06] Jeff Howe. The Rise of Crowdsourcing. *Wired*, 14(6), 2006.

[NC12] Vidhya Navalpakkam and Elizabeth F. Churchill. Mouse Tracking: Measuring and Predicting Users' Experience of Web-based Content. In *Proc. CHI*, 2012.

[Saa08] Thomas L. Saaty. Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1):83–98, 2008.

[SBG13] Maximilian Speicher, Andreas Both, and Martin Gaedke. TellMyRelevance! Predicting the Relevance of Web Search Results from Cursor Interactions, 2013. Submitted for review.