Cohesion in multi-modal documents: Effects of cross-referencing

Cengiz Acartürk

In this paper, we concern ourselves with the type of signaling or reference used in the text to introduce the graphic material. This form of (deictic) cross-referencing can be understood as a way to establish coherence, through the use of cohesive links (Halliday & Hasan, 1976). Extensive research in the Hallidayan tradition has shown that cohesive links in the text contribute to the perceived coherence of a document. We would like to pursue this view of cohesive linking further, and extend it to the reference created between text and other pictures, Mayer and colleagues have shown that placing the picture and the text that refers to the picture close together leads to a better learning outcome (Mayer, 2009).

Taking the perspective of information design, explicit reference from the text to the figure (i.e., signaling) can be seen as a specific subtype of referring, taking different forms in communication settings (Clark, 2003; Brennan, 2005; Heer & Agrawala, 2008). In a spoken communication setting, for instance, this form of deictic reference aims to attract visual attention of the communication partner to a referred entity in the environment. In a multimodal document, the referred entity is depictive material in the multimedia document. Accordingly, the

exophoric one, where the cohesive item in the text points to something outside the text proper (to the depictive element). However

sentence *Kepler found the answer to this by the method shown in Figure 1*. In the *elliptic* condition, the corresponding sentence was *Kepler found the answer to this*, without any reference to the figure.[6]

The experiment was conducted in single sessions. Each participant was randomly assigned to one of the three experimental conditions and the presentation of the stimuli was randomized. The participants were asked to read and understand the stimuli by imagining themselves reading the lecture summaries of a set of missed classes in a summer school. Participants' eye movements were recorded by the eye tracker during their reading and inspection of the stimuli. The experiment took approximately 30 minutes.

4.2. Results

As the first step of the analysis, the prior knowledge tests were analyzed to reveal if there was a difference between the participant groups in the three experimental conditions. The participants assessed their knowledge about astronomy, physics, economics, biology, thermodynamics and linguistics by giving a score between *1* and *5*. The results of an analysis of variance test revealed no significant difference between the participants in the three experimental conditions in terms of their reported prior knowledge.[7]

4.2.1. Eye movement parameters

Data from one participant were not included into the analysis due to a total calibration problem in the eye tracker. A total of 1,638 eye movement protocols were recorded with the remaining 91 participants (18 multimodal stimuli x 91 participants). Out of the 1,638 eye movement protocols, 66 of them were not included into the analysis due to partial calibration problems.

The mean number of participants' gaze

Gaze Time (s)	*Directive*	*Descriptive*	*Elliptic*
Total	55.2* (6.68)	49.4* (4.94)	52.5* (6.78)
Figure	10.5 (3.59)	9.5* (3.09)	11.0 (3.53)
Text	44.7* (4.70)	39.9* (3.36)	41.4* (4.50)

*p < .05

The results revealed significant differences between the three experimental conditions, for the total gaze time on

The posttest included 18 multiple-choice questions, one per multimodal stimulus screen. Almost all posttest questions asked about the information content presented both in the text and in the figure part of the multimodal stimuli. Participants either selected from one of the two answers, or they selected the "*I do not remember*" answer. The percentages of the "*I do not remember*" answer in the directive, the descriptive and the elliptic condition are 23.3%, 27.8% and 23.0% respectively. For the analysis of the results, each correct answer was given a score of 1, and each wrong answer and each "*I do not remember*" answer were given a score of 0. The results revealed that the participants in the descriptive condition received lower posttest scores (*M* = .50, *SD* = .29) compared to the participants both in the directive condition (*M* = .59, *SD* = .27) and the participants in the elliptic condition (*M* = .59, *SD* = .27). A further analysis of correlation coefficients revealed a correlation of posttest scores with gaze times (i.e., study times) on figure (*r* = .21, *p* = .05), showing that longer inspections of the figure part of the material led to higher posttest scores.

The participants reported their own judgments on the effort needed to understand the experimental stimuli by using a nine-point scale ranging from 1 (*very low effort*) to 9 (*very high effort*) indicating subjective judgments of mental effort invested in learning (Paas, 1992). For each of the 18 multimodal stimuli, they gave a score by answering the question "*How much mental effort did you spend while you read the presented material?*" The results revealed a significant difference in the reported mental effort scores between the experimental conditions, showing that the participants in the directive condition reported lower scores (*M* = 4.17, *SD* = 1.49) compared to both the participants in the descriptive condition (*M* = 4.50, *SD* = 1.58) and the participants in the elliptic condition (*M* = 4.46, *SD* = 1.44). No correlation of the mental effort scores was obtained with either posttest scores or eye movement parameters.

5. Discussion

The aim of the present study was to investigate how different types of signaling (i.e., reference) versus the lack inspecting the material are measures of the difficulty in the integration of the information contributed by the text and the figure. Accordingly, the findings indicate that humans spend less effort to integrate the information contributed by the two modalities (i.e., the text and the figure) when a descriptive reference, rather than a directive reference is used in the text. A low number of gaze shifts is observed when there is no reference in the text to introduce the figure, as well. However, this finding alone does not tell much about the integration of the material in different modalities. A more salient finding is the longer fixation duration on the figure when there is no explicit reference compared to the use of an explicit reference (either descriptive or directive reference) in the text. This finding suggests that the lack of explicit reference in the text results in high effort for the integration of the information in different modalities.

Moreover, the analysis of answers to posttest questions reveals a low retention of the material when a descriptive reference is used in the text, compared to the use of a directive reference or the lack of reference link in the text. In other words, both the directive reference and the lack of reference are correlated with better retention compared to the descriptive reference. Since posttest scores correlate with the time spent on the material, the lower retention score with a descriptive reference may be an outcome of spending less time to study the material, thus leading to shallow processing of the material, when a descriptive reference is used in the text and vice versa for the directive reference and for the lack of a reference.

In summary, we have shown that eye movement measures point to descriptive reference as the type that results in the least integration effort, although there seems to be a reduced retention of the document content with a descriptive reference. The trade-off between time and effort spent on task versus retention of the content is important in many areas of document design, including the creation of educational materials. It seems that descriptive reference is ideal from the point of view of immediate integration, but, if retention is desired, maybe other methods for achieving that goal need to be used.

The participants reported low judgment scores for difficulty in understanding the material (i.e., mental effort scores) when a directive reference is uw; γ spent the longest gaze time on the document when the document consisted of a line graph rather than a bar chart or a pictorial illustration. We note that in line with the focus of the present study on signaling, we performed comparisons among the conditions by keeping the multimodal documents the same across the conditions, except for the signaling. However, the observed differences between the types of the depictive material is an outcome of the intricate relationship between the text and the depictive material, and they are bound by the practical limitations of the study, such as the use of a limited set of stimuli as representations of different types of depictive material. Future research could investigate these further, also in relation to the signaling conditions.

As we mentioned in the text, captions undoubtedly play an important role in the integration of main text and depiction, and follow-up work should concentrate on how the wording and placement of captions influences integration and recall. Finally, we are also interested in the placement of the depictive material with respect to the text. In our experiments, the depictions were all located in the same place, to the rigc4nq(ang (en a(n)6(tD)-3(atio)-ce17(f)8( )hgctex)5(t,

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

Carroll, P. J., Young, R. J., & Guertin, M. S. (1992). Visual analysis of cartoons: A view from the far side. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 444-461). New York: Springer.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: where language, culture, and cognition meet* (pp. 243-268). London: Erlbaum.

Delin, J., Bateman, J., & Allen, P. (2002). A model of genre in document layout. *Information Design Journal*, 11(1), 54-66.

Dolk, S., Lentz, L., Knapp, P., Maat, H. P., & Raynor, T. (2011). Headline section in patient information leaflets: Does it improve reading performance and perception? *Information Design Journal*, 19(1), 46-57.

Elzer, S., Carberry, S., Chester, D., Demir, S., Green, N., Zukerman, I., et al. (2005). Exploring and exploiting the limited utility of captions in recognizing intention in information graphics. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 223-230). Ann Arbor, MI.

Fan, X., Aker, A., Tomko, M., Smart, P., Sanderson, M., & Gaizauskas, R. (2010). Automatic image captioning from the web for GPS photographs. In *Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 445-448). Philadelphia, PA.

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1239-1249). Uppsala, Sweden.

Garcia, M. R., & Stark, P. A. (1991). *Eyes on the news*. St. Petersburg, FL: The Poynter Institute.

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language, 31*, 129-151.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge: Cambridge University Press.

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction, 20*, 167-171.

Mittal, V., Moore, J. D., Carenini, G., & Roth, S. (1998). Describing complex charts in natural language: A caption generation system. *Computational Linguistics, 24*, 431-468.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive load approach. *Journal of Educational Psychology*, 84, 429-434.

Paraboni, I., & van Deemter, K. (2002). Towards the generation of document-deictic references. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in language generation and interpretation* (pp. 329-354). Stanford, CA: CSLI.

Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Peebles, D. J., & Cheng, P. C.-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, *45*(1), 28-35.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422.

Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied, 7*, 219-226.