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

Cengiz Acartürk
Informatics Institute
Middle East Technical University
Ankara, Turkey
acarturk@metu.edu.tr

Maite Taboada
Department of Linguistics
Simon Fraser University
Burnaby, Canada
mtaboada@sfu.ca

Christopher Habel
Department of Informatics
University of Hamburg
Hamburg, Germany
habel@informatik.uni-hamburg.de

Abstract. In multimodal documents, different types of cohesive or cross-reference link (i.e., signaling) are used in the text to link verbally coded content with the graphical material. In this study, we identify three types of reference, within the framework of previous work on cohesion (Halliday & Hasan, 1976): directive signaling, descriptive signaling, and no signaling in the text to introduce the figure. In an experimental study, we use eye tracking to investigate how those three reference types influence the processing of the material by humans. The results reveal differences between the reference types both in terms of eye movement parameters and retention of the material.

Keywords: Multimodal documents, cross-reference, text-figure relations, cohesion

1. Introduction

Much of recent research in learning, educational materials and multimedia concentrates on the relationship between pure text and depictive material, that is, material presented through other modalities. Research on document design and learning has been trying to elucidate what kind of impact multimodal material has on the reader. Mayer (2009), for instance, reports on years of studies showing that students learn better when they learn from words and pictures than when they learn from words alone. The learning, however, is improved only when certain principles in the presentation of the multimedia material are followed. The principles refer to the order of presentation, the coherence of the text and pictures, and the type of cross-reference used (i.e., the reference in the text to depictive material).

Studies have considered whether to use multimodal material or not, where to place it, and what effect captions or other verbal information surrounding such material have on the reader (Acartürk et. al., 2008; Acartürk, 2010, and references therein). Less frequently discussed is the nature of the reference that the text establishes with respect to the depictive material (but see Slough et al., 2010, for a description of cross-referencing in science textbooks). References in the text such as *see Figure 1* or descriptions such as *In Figure 3 we show…* help guide the reader in performing the integration of text and depictive material. Beyond exhortations in style manuals to always reference figures and tables (particularly in scientific writing) as a matter of good style, the effect that different types of reference have on the reader has not been widely studied.

* Acartürk, C., M. Taboada and C. Habel (2013) Cohesion in multimodal documents: Effects of cross-referencing. *Information Design*. 20 (2): 98-110. (Pre-publication version).

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 modalities. Our hypothesis is that certain forms of reference will lead to better integration (expressed in different measures) of text and other modalities.

We are concerned mostly with figures (flow diagrams, bar or other charts, system overviews, maps, etc.) and tables.[1] We will often use the generic word "figure" or "graph" to refer to the types of illustrations discussed, as a whole, or the more abstract term "depiction" for modalities other than text (Kosslyn et al., 2006), and will use it as a cover term for pictures, graphs, figures and tables. In this paper, we will tend to ignore the large differences across all these different types of depictive material, in how they present information, how they make use of both symbolic and visual codes, and how they are deployed by authors.

In order to quantify the effect of reference between text and depictive material, we first classified different types of reference, following work on cohesion by Halliday & Hasan (1976). We present a summary of cohesion in text, and our classification of reference types, in Section 3. Before that, in Section 2, we provide a very brief overview of previous work on multimodality and the types of relations established across different modes. The main body of the paper is in Section 4, where we present the experimental investigation that we carried out, testing the integration of text and depictive material under different conditions, using eye-tracking methodology to reveal the difference in reading behavior between the conditions.

2. Integration of Text and Depictive Material in Multimodal Documents

The design of multimodal documents has been investigated by many scholars in different disciplines, ranging from the analysis of design elements in scientific text (e.g., Slough et al., 2010), newspapers and leaflets (Holsanova et al., 2006; Bouayad-Agha et al., 2000; Dolk et al., 2011) and genre-based approaches (Delin et al., 2002; Bateman, 2008) to investigations of the integration of information contributed by text and figures from the perspective of human comprehension (e.g., Glenberg & Langston, 1992; Hegarty & Just, 1993) and learning (e.g., Ayres & Sweller, 2005; Mayer, 2009, among many others). From the latter perspective, a common view is that readers construct referential links—thus inter- and intra-representational coherence—among entities that are contributed by text and depictive material at a conceptual level (Seufert, 2003; Bodemer & Faust, 2006). Among many factors that specify how humans construct referential links—such as the prior knowledge and the verbal and/or spatial abilities of the readers—the design of multimodal documents has received most of the attention in many of the relevant disciplines. For instance, in their investigation of spatial contiguity between text and

[1] In particular, we do not consider the role of photographic and other 'realistic images', as their comprehension involves an additional semiotic layer.

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 reference in the text aims to attract visual attention of the reader to the depictive material. A methodologically sound means to measure the shift of visual attention, which is an inner cognitive process, is to record and analyze the gaze shifts of humans during reading the text and inspecting the depictive material. Based on a set of assumptions, (e.g., the mandatory shift hypothesis, Pashler, 1998; the eye-mind hypothesis, Just & Carpenter, 1980), it is usually assumed that a gaze shift from an entity to another entity indicates a corresponding shift of attention from the former entity to the latter. Therefore, consecutive gaze shifts between text and depictive material in a multimodal document can be taken as a path of loci of attention.[2] A methodologically appropriate way to analyze how different types of signaling in text influence the readers' processing of text and depictive material is then to record and analyze eye movements of the readers during comprehension of documents that involve text and depictive material.

Eye tracking has been used widely as a research methodology in reading research and scene viewing research in the last two decades. Eye movements are analyzed as an indication of moment-to-moment cognitive processes such as the cognitive processes that underlie anaphora resolution and lexical and syntactic ambiguity in reading or the interaction between data-driven visual properties of a scene of top-down knowledge during scene viewing (see Richardson, Dale & Spivey, 2007 for a review). The most widely used measures of eye movement for those purposes (i.e., eye movement parameters) have been fixation duration (the mean duration of gaze fixation), gaze time (the sum of the fixation durations on a specific region), and fixation count (the number of fixations on a specific region). From the perspective of the information processing approach, longer gaze times, higher fixation counts and longer fixation durations on certain regions (e.g., a part of text) over the others have been used as a measure of processing difficulty in humans.

The eye movement methodology has been frequently used in research where readers/learners have to integrate the information contributed by text and figures, in domains such as advertising (e.g., Rayner et al., 2001), newspaper reading (e.g., Garcia & Stark, 1991; Holsanova et al., 2006), cartoon reading (e.g., Carroll et al., 1992), in learning from multimedia instructions (e.g., Hegarty & Just, 1993; see Mayer, 2010, for a review), and in comprehension of statistical graphs (e.g., Peebles & Cheng, 2003). A common finding in these studies is that the inspection of figures is largely directed by text. A set of different signaling techniques have been employed, such as the use of spoken language to accompany pictures and the use of color coding (see Tabbers et al., 2004, for a review on cueing in multimedia instructions). However, as far as we know, no study has investigated how

[2] This path is *partial* because readers can shift their attention without shifting their gaze. On the other hand, in natural and unconstrained settings, the position of the gaze and the visual attention are usually coupled (Richardson, Dale & Spivey, 2007).

different types of signaling used in the main text (i.e., paragraphs) to introduce the figure, as described below, help the integration of the text and depictive material by humans.

3. Cohesion Between Text and Depictive Material

To understand the references in the text that help integrate depictive material, we make use of the classical Halliday & Hasan framework (Halliday & Hasan, 1976), with some changes. According to Halliday & Hasan, the property of being a text is called *texture*. Texture is what distinguishes a text from a non-text, and it is derived from the fact that the text functions as a unity with respect to its environment. Texture is realized in relations existing between parts of a text. For instance, in Example (1), there is a relation between *six cooking apples* and *them*. It is the relation between those two phrases that makes the two sentences become a text, because they hang together as one unit.

(1) Wash and core six cooking apples. Put them into a fireproof dish.

This relation is a *cohesive relation*, and the pair of related items is a *cohesive tie*. Obviously, no single item is cohesive in itself. Although we will be referring to particular linguistic expressions (or the lack thereof) in the text, and categorizing them as a particular type of cohesion, we should always bear in mind that cohesiveness is established through the relation between the two items, i.e., the text and the depictive material, not by one item in isolation.

The meaning of a cohesive relation is that the two items refer to the same thing; they share corefentiality. Identity of reference is not the only type of cohesive relation; there also exist relations of similarity. But in a more general sense, cohesion occurs when the interpretation of some element in the discourse depends on the interpretation of another one, whether preceding or following. In our example above, *them* presupposes *six cooking apples*. We need to refer to the latter in order to resolve the presupposition. The fact that it is resolved successfully establishes the cohesive relation between the two sentences.

Another important aspect of Halliday & Hasan's model is that cohesive ties are not reduced to two elements. Reference is obviously established throughout the text among multiple entities with the same reference, or among entities that are related through some lexical relation, such as synonymy or hyponymy. Elements that are so related form a *cohesive chain*. In multimodal documents, it is often the case that a figure is referred to multiple times throughout the text, sometimes as a whole, or through reference to particular aspects of the figure, table or graphical material. In our experiments we included a single reference (or no reference) to the depictive material, but a worthwhile extension of this research would consider multiple references and cohesive chains—or even cohesive graphs—that consist of both text and depictive material.

Cohesive ties are established between elements in the text (endophoric reference), not with elements that have their referent outside the text (exophoric reference). In adapting this view of discourse to the relationship between text and depictive material, we are faced with the possibility that the relation that we are analyzing is actually an

exophoric one, where the cohesive item in the text points to something outside the text proper (to the depictive element). However, we prefer to consider this as an endophoric relation, because text and depictive material form part of the same document, and the document can be understood as a text composed of different modalities.

Cohesion can be used as a measure of readability in text (Graesser et al., 2007). Previous studies of multimodal cohesion have studied a wide array of cohesive links. André & Rist (1994) discuss "cross-media referring expressions", which include many different types of links, with relations between parts of the text and parts of the figure. Similarly, Paraboni & Van Deemter (2002) discuss "document deictic references", which includes reference to depictive material, but also the references typically found in a document to other sections, examples, etc.

Liu & O'Halloran (2009) extend Halliday & Hasan's notion of cohesion to create what they term intersemiotic cohesion, the links established between the representations that both text and depictive material evoke relations of parallelism, polysemy or hyponymy.

In our proposal we consider exclusively linguistic realizations of cohesion, where the main text contains a cross-reference to the depictive material, or some part of it. These are almost deictic or pointing links, where the text invites the reader to direct their attention to the depiction.

Upon initial study of the types of cohesive or cross-reference links in text that refer to the depictive material, we discovered a paucity of types. We thus propose three different types of links:

- No link. The figure is present on the page, but no explicit mention of it is made in the text. This is, in fact, ellipsis in the Halliday & Hasan classification. Therefore, we also use the term 'elliptic reference' throughout the paper.
- Personal reference with a directive signaling. The depictive material is referred to with a form of personal reference, proper noun most often which is an explicit mention of the figure, with a directive to examine it (henceforth, directive signaling). For example:

  See Figure x.

  For an example of y, see Figure x.

  In this category, we also include a form of implicit directive, in the form of a parenthetical:

  To isolate the impact of the NaCl binary constraints (Table 1), we built the SPEC2000…
- Personal reference with a descriptive signaling. The depictive material is referred to with a personal reference, without a directive but rather a description of what the figure represents, and perhaps how it relates to the text (henceforth, descriptive signaling).

  Figure 2 shows the architecture of a real-time business intelligence system.

Our labels are self-explanatory, and the classification was straightforward in all cases. We argue that *'(Figure 1)'* is directive, rather than descriptive, because it provides no information on how the relation is to be interpreted, merely directing the reader to examine the figure instead. This distinction also builds on research on cross-referencing in text in general, such as Swales' classification of text references into integral and non-integral

(Swales, 1986, 1990)[3]. Integral (or extensive) citations are those that are included as part of the grammatical structure of a sentence, whereas non-integral (or short) citations tend to be placed in brackets. Swales and others have analyzed the distribution of both types of citations, and their overall function, with integral citations typically being central to the argument, either as support or as part of an attack on previous work. Non-integral citations, on the other hand, tend to be an acknowledgement of previous work, and an expression of command of the field in which the article is written. The sociological and field-specific implications of citations are not our concern here, but we do see a parallelism in the way the references are integrated (or not) in the text. The directive references have little or no integration with the rest of the text, whereas the descriptive references participate more fully in the syntactic and discourse structure of the text.

Our initial hypothesis was that the more information about the depictive material in the reference, the easier the integration between text and depictive material. That is, we postulated that descriptive signaling would lead to better integration than directive, and that directive signaling would be, in turn, better than no signaling. (In the next section, we describe in more detail what we mean by better integration, and the measures that we used to assess integration.) Our hypothesis was based on the principle that telling the reader *how* to integrate is more informative than just telling them *to* integrate, and will thus lead to better comprehension. We will see that the experimental results reveal a more complicated picture than we anticipated.

4. Experiment

In an experimental investigation, we analyzed how lack of signaling versus the two types of signaling, namely descriptive and directive signaling, influenced the reading characteristics of humans in multimodal documents and their retention after reading all the documents.

4.1. Participants, Materials and Design

Ninety-two participants (mean age = 21.5, SD = 2.66) from the Middle East Technical University, Turkey participated in the experiment. The participants were presented 18 single-screen, multimodal documents that involved a combination of text and figure, a prior knowledge test, a posttest to measure retention, and a mental effort judgment test to elicit a subjective report of mental effort. The multimodal stimuli were presented on a Tobii T120, a non-intrusive, 120 Hz eye tracker, integrated into a 17" TFT monitor with 1024x768 pixels; the tests were presented on paper. Spatial resolution and accuracy of the eye tracker was about 0.30° and 0.50° degrees respectively. The stimuli were prepared based on a set of original sources, including published popular science books, course books and institutional reports. The original sources covered a wide range of domains involving astronomy, physics, economics, biology, thermodynamics and linguistics. The experiment was conducted with participants who had Turkish as their native language. Bilingual original sources, published both in English and in Turkish were selected for designing the experimental stimuli.[4] Accordingly, the multimodal

[3] The classification of citations in terms of the rhetorical relations they hold with the works cited has been undertaken by Teufel and colleagues, in part with the goal of extracting information from scientific articles (Teufel & Moens, 2002; Teufel et al., 2009).
[4] Follow-up work will be carried out with native English speakers; that is why bilingual sources were used.

stimuli that were used in the experiment were published Turkish translations of the original English sources, modified according to the conditions described below. The text part of the stimuli included an average of 105.9 (*SD* = 2.17) words (based on the word count in the elliptic condition). Three sub-types of figures (six pictorial illustrations, six bar graphs, and six line graphs) were used in the stimuli set.

The experiment was a between-subject design with three experimental conditions. The conditions were specified by the type of the *reference from the text to the figure*, as described below.

i. Condition 1. *Directive signaling* within the text, specified by phrases in brackets, such as '*(Figure 1)'*.
ii. Condition 2. *Descriptive signaling* within the text, specified by phrases without brackets, such as *Figure 1 shows that …*
iii. Condition 3. Lack of signaling, specified by no link to the figure (henceforth, the *elliptic condition*).

There were 31 participants in condition 1 and in condition 2, and 30 participants in condition 3. In all stimuli, the text was presented to the left, with the figure to the right, in a split landscape page. [5] Figure 1 exemplifies the multimodal stimuli that were used in the experiment.

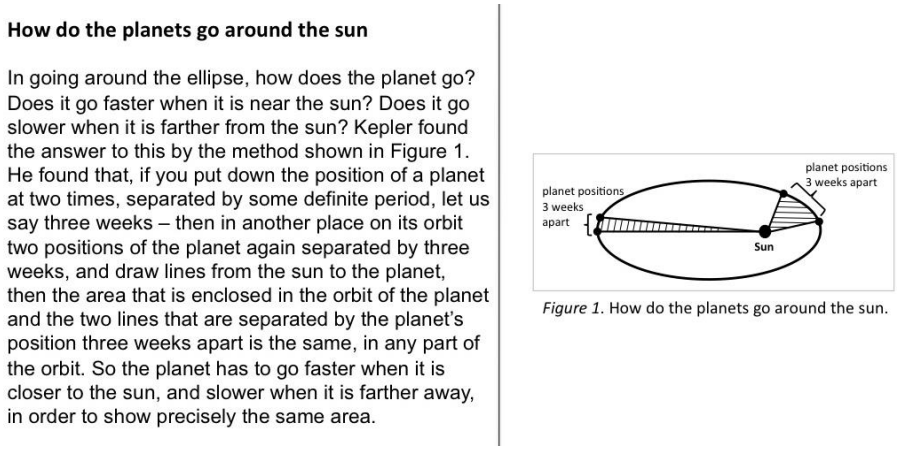

Figure 1. Sample multimodal stimuli from the experiment. The original text in English is used in the text part of the stimuli and the figure part is redrawn for clarity. Original source: *The Character of Physical Law by Richard Feynman* published by the MIT Press, Twelfth Printing, 1985 (First MIT Press paperback edition, March 1967).

The stimulus in Figure 1 exemplifies the experimental stimuli in the *descriptive* condition. The corresponding stimulus in the *directive* condition involved the sentence *Kepler found the answer to this (Figure 1)* instead of the

[5] Captions for figures are undoubtedly an element to bear in mind. Extensive research has shown that they are used to bridge main text and images, and to process the individual parts of images (Elzer et al., 2005). There is also an active area of research in automatic generation of captions (Fan et al., 2010; Feng and Lapata, 2010; Mittal et al., 1998), or annotation and extraction of images based on their captions (Barnard et al., 2003). Because, in our experiments, captions remained constant, and the only variable we observed was reference to the depictive material in the text, captions will not be considered in this paper.

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 shifts was calculated for each protocol. The term 'gaze shift' is used for participants' shift of gaze between the text part of the stimuli and the figure part of the stimuli (the figure caption was also taken as part of the figure part of the stimuli). The results revealed a significant difference between the three conditions: The participants in the directive condition performed more gaze shifts (M = 5.60, SD = 1.54) compared to both the participants in the descriptive condition (M = 4.71, SD = 1.34) and the participants in the elliptic condition (M = 4.80, SD = 1.63).

Participants' mean total gaze time on (i.e., the study time of) the stimuli and the mean gaze time on the figure- and text-part of the stimuli were calculated for each eye movement protocol. The results for the three experimental conditions are shown in Table 1.

Table 1: Mean of the gaze times of the participants on the figure, on the text, and the total gaze time of the participants on the multimodal document (in seconds).

[6] We thank our reviewers, who noted that a more informative variation of the descriptive cross-reference is also possible, as in "Kepler found the answer to this by comparing the distance traversed by a planet along two segments of the orbit within the same duration, namely 3 weeks (Figure 1)." The scope of this study was limited to the investigation of the aforementioned three types of reference, leaving other types for further research.

[7] All the presented results in the present study are statistically significant ($p < .05$), except the ones explicitly mentioned.

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

*p < .05

The results revealed significant differences between the three experimental conditions, for the total gaze time on the stimuli, for the gaze time on the figure, and for the gaze time on the text. Further analyses revealed the following conclusions: The total gaze time on the stimuli was highest in the directive condition, then the elliptic condition and then the descriptive condition. Moreover, the participants in the descriptive condition spent less time both on the figure and on the text compared to the participants in the other two conditions. In addition, the analyses of fixation counts revealed exactly the same pattern of results as gaze times.

The last eye movement parameter in the analysis was fixation duration. The term 'fixation duration' is used for the duration of each single fixation on the experimental stimuli. Participants' mean fixation durations on the multimodal stimuli (overall), the mean fixation durations on the figure and the mean fixation durations on the text were calculated for each eye movement protocol. The results are shown in Table 2.

Table 2: Mean fixation duration of the participants on the figure, on the text, and the overall mean fixation duration on the multimodal document (in seconds).

Fixation Duration (s)	*Directive*	*Descriptive*	*Elliptic*
Figure	0.338	0.337	0.352*
	(0.041)	(0.041)	(0.036)
Text	0.276	0.273*	0.277
	(0.006)	(0.007)	(0.005)
Overall	0.286*	0.283*	0.291*
	(0.010)	(0.010)	(0.008)

*p < .05

The results revealed significant differences for the overall mean fixation duration on the stimuli, for the mean fixation duration on the figure and for the mean fixation duration on the text: The participants in the elliptic condition exhibited the longest mean fixation duration on the figure. Concerning the mean fixation durations on the text part of the stimuli, the mean fixation duration in the descriptive condition was shorter than the other two conditions. Moreover, the overall mean fixation duration on the stimuli in the elliptic condition was longer than the overall mean fixation duration in the directive condition, which was longer then the overall mean fixation duration in the descriptive condition.

4.2.2. Answers to Posttest Questions and Subjective Reports of Mental Effort

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 thereof influence humans' reading of multimodal documents that involve text and figures. Following the studies on cohesion by Halliday & Hasan (1976), we identified three reference types: *directive reference* (e.g., *see Figure 1* or a parenthetical reference such as '*(Figure 1)*', *descriptive reference* (e.g., *Figure 1 shows the architecture of a real-time…*) in which the citation is included as part of the grammatical structure of the sentence, and *elliptic reference* in which there is no explicit mention of the figure in the text (i.e., no link). Therefore, it is on the one hand a type of reference, but on the other hand, it needs more effort for the integration of text and depictive material, similar to processes of ellipsis. In an experimental investigation, three groups of participants were presented documents that included a combination of text and one figure, one type of reference per group. During their reading of the material, eye movements were recorded and then analyzed to reveal differences in reading behavior under the three experimental conditions. We also measured retention of the participants as well as their subjective reports of mental effort.

We take the set of eye movement parameters as a measure of processing characteristics: We assume that the higher number of gaze shifts between the main text and the figure and the higher time spent on reading and

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 used. We note that the mental effort scores reflect participants' subjective report of the difficulty of the material, and it is not as reliable a measure of comprehension difficulty as eye movement measures and retention scores. Therefore, we will leave the interpretation of this finding to further research, which may employ objective measurements of cognitive resources, such as dual-task approaches that involve secondary tasks, interference tasks, and preload tasks for measurement of the working memory components (Schüler et al., 2011, for a review). Future research should also address the investigation of how the location of the reference in the text and how the way the figure was described as a part of the grammatical structure of the sentence influence participants' own judgment of the difficult of understanding the text.

We observed, in the current experiments, some differences across the types of depictive material, whether pictorial illustrations, bar charts or line graphs: The participants exhibited higher retention of the content when the figure part of the multimodal document was a pictorial illustration rather than a bar chart or a line graph. They also found the documents with line graphs more difficult to understand than the documents with bar graphs, and the documents with bar graphs more difficult than the documents with pictorial illustrations. Finally, they

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 right of the text and centered with respect to it. It would be interesting to study how placement affects integration and long-term recall, especially when longer texts are involved.

Acknowledgements. We thank the Middle East Technical University HCI Research and Application Laboratory for their technical support in the experiments. This research was carried out in part while Maite Taboada was a visiting researcher at the University of Hamburg, thanks to a Fellowship from the Alexander von Humboldt Foundation.

References

Acartürk, C. (2010). Multimodal comprehension of graph-text constellations: An information processing perspective. Unpublished Ph.D. dissertation. University of Hamburg, Hamburg. http://ediss.sub.uni-hamburg.de/volltexte/2010/4501/pdf/dissertation.pdf

Acartürk, C., Habel, C., Cagiltay, K. & Alacam, O. (2008). Multimodal comprehension of language and graphics: Graphs with and without annotations. *Journal of Eye Movement Research*, 1(3):2, 1-15.

André, E., & Rist, T. (1994). Referring to world objects with text and pictures. In *Proceedings of COLING 1994: The 15th International Conference on Computational Linguistics* (pp. 530-534). Kyoto, Japan.

Ayres, P., & Sweller, J. (2005). The Split-Attention Principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135-146). Cambridge, MA: Cambridge University Press.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. d., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research, 2*, 1107-1135.

Bateman, J. (2008). *Multimodality and genre: a foundation for the systematic analysis of multimodal documents*. London: Palgrave Macmillan.

Bodemer, D., & Faust, U. (2006). External and mental referencing of multiple representations. *Computers in Human Behavior, 22*, 27-42.

Bouayad-Agha, N., Scott, D., & Power, R. (2000). Integrating content and style in documents: A case study of patient information leaflets. *Information Design Journal*, 9(2), 161-176.

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.

Graesser, A.C. ,Jeon, M., Yan, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, 15, 199-213.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Heer, J., & Agrawala, M. (2008). Design considerations for collaborative visual analytics. *Information Visualization, 7*, 49-62.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Holsanova, J., Rahm, H., & Holmqvist, K. (2006). Entry points and reading paths on newspaper spreads: Comparing a semiotic analysis with eye-tracking measurements. *Visual Communication, 5*(1), 65-93.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford: Oxford University Press.

Liu, Y., & O'Halloran, K. (2009). Intersemiotic texture: Analyzing cohesive devices between language and images. *Social Semiotics, 19*(4), 367-388.

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.

Richardson , D. C., Dale, R., & Spivey, M. J. (2007). Eye movements in language and cognition: A brief introduction. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson & M. J. Spivey (Eds.), *Methods in Cognitive Linguistics* (pp. 325-346).

Schüler, A., Scheiter, K., & van Genuchten, E. (2011). The role of working memory in multimedia instruction: is working memory working during learning from text and pictures? *Educational Psychology Review*, DOI 10.1007/s10648-011-9168-5.

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction, 13*, 227-237.

Slough, S.W., McTigue, E.M., Kim, S., & Jennings, S.K. (2010). Science textbooks' use of graphical representation: A descriptive analysis of four sixth grade science texts. *Reading Psychology, 31*(3), 301-325.

Swales, J. M. (1986). Citation analysis and discourse analysis. *Applied Linguistics, 7*(1), 39-56.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tabbers, H. K., Martens, R. L., & Van Merrienboer, J. J. G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology, 74*, 71-81.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical structure. *Computational Linguistics, 28*(4), 409-445.

Teufel, S., Siddharthan, A., & Tidhar, D. (2009). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (pp. 80-87). Sydney, Australia.