Text-Picture Relations in Multimodal Instructions

Ielka van der Sluis Anne Nienke Eppinga
Gisela Redeker
Center for Language and Cognition Groningen
University of Groningen
{i.f.van.der.sluis}{g.redeker}@rug.nl

Abstract

This paper presents a multi-method approach to the description and evaluation of multimodal content in situated contexts. First, a corpus of 15 simple first-aid instructions that include texts and pictures is shown to exhibit variation in content and presentation. We then report a user study in which four versions of a tick-removal instruction were used to test the effects of the relative placement of text and pictures in a particular instruction. The participants' processing of the instruction and their task performance were video-recorded and registered with an eye tracker. Questionnaires and interviews were used to measure comprehension, recall and the instruction's attractiveness. Results show that users first read at least some of the text before looking at the pictures, and prefer to have the pictures placed to the right or below the text.

1 Introduction and Background

In document design research the combination of text and pictures has been noted but not so much investigated in terms of function and content (cf. Schriver, 1997). Although useful starting points have been provided (Bateman, 2014; Aouladomar, 2005), we are unaware of a standard methodology to describe and evaluate text-picture relations in situated use. This leaves document designers without specific guidelines, while users may experience difficulties in effectively processing multimodal content due to mismatches with their expectations and cognitive capacities. Various models (Schnotz et al., 2017; Sweller, 2016; Mayer, 2005; Schnotz, 2005) and empirical studies (van Hooijdonk and Krahmer, 2008; Houts et al., 2006; Katz et al., 2006; Maes et al., 2004) suggest that functional pictorial content benefits processing content in general. In this paper we explore a method in which we first describe and subsequently evaluate multimodal content to specify the exact text-picture relations that serve this enhanced processing. The research question that we aim to answer is: How can we identify the relations between text and pictures that influence the quality of multimodal content presentations?

In this paper we concentrate on multimodal first-aid instructions. Multimodal instructions (MIs) include pictures and text and instruct users to perform procedural tasks. Consequently, MIs allow evaluation of comprehension and recall not only through reading and judging, but also via assessment of situated use (cf. the discussion in Van der Sluis, Leito & Redeker, 2016). We present a corpus analysis (Section 2) to explore the variation between existing MI designs and to determine which text-picture relations may affect the MI quality. We present a user study (Section 3) in which multiple methods are employed to evaluate particular document designs and to inform annotation models for multimodal content. We conclude and discuss the outcomes and future directions (Section 4).

2 Corpus Study

A small corpus of 15 Dutch tick-removal instructions was collected from public online sources. The instructions were selected to contain both text and pictures, i.e. they were multimodal instructions (MIs). For comparability, the selected MIs described and depicted the procedure with tweezers, not

with dedicated tick-removal tools. Figures 1 and 2 present two examples to illustrate the variation in the corpus (e.g., position of text and pictures, number of steps, amount of text per step, inclusion of additional information in either text or pictures). The corpus annotation was developed and refined in close discussion between the authors of this paper, with the second author taking the lead in the majority of the analyses in the context of her MA thesis (Eppinga, 2017).

Figure 1: MI14 - www.ggddrenthe.nl

Figure 2: MI15 - www.serviceapotheek.nl

2.1 Description of Text

The MI texts were analysed in terms of composition (e.g., textual elements, layout), general characteristics (e.g., title, number of steps, number of sentences), actions (status: preparatory, core or closing; aspect: process or result) and control information (e.g., warning, conditional, motivation, explanation, extra information).

Composition: Three genre-specific textual elements were distinguished: preamble, actual instruction, and closing. Figure 2, for instance, includes a preamble (ie., the first two sentences), the actual instruction (here, five steps) and a closing (ie., two sentences that advise the user to visit a docter if the victim feels ill in the three months after the tick was removed). Only three of the 15 MIs in our corpus did not start with a preamble; in all others, all three textual elements were identified. Text-structuring layout elements such as white space and horizontal lines, and paragraph numbering were noted.

General Text Characteristics: All fifteen MIs in our corpus include a title. These titles usually concern the removal of a tick (e.g., MI1: 'Wat te doen bij een tekenbeet', what to do in the case of a tick; MI11 'Hoe een teek verwijderen', How to remove a tick). Sometimes however the title addresses a particular audience, for instance, in the case of MI13 'Instructie over risico's bij werken in het groen' (Instruction about risks for gardeners). MI13 also includes a subtitle that is more specific ('Ziekte van Lyme: een teek verwijderen', Lyme disease: removing a tick). No other instances of subtitles were discovered. Sometimes the title is quite generic (e.g., M17: Wat kunt u zelf doen om overlast te voorkomen?', What can you do to avoid any inconvenience?). The number of steps in the MIs varies from two to ten ($M = 5.5$, $std = 1.68$). The number of sentences varies from seven to ten ($M = 11.4$, $std = 3.22$). The number of words varies from 110 to 275 ($M = 163.01$, $std = 43.29$).

Actions: The actions included in the actual instruction were classified as preparatory actions (e.g., pick up tweezers), core actions (e.g., grab the tick, pull the tick out) and closing actions (e.g, disinfect the wound, write down the date). Table 1 presents the nine actions and their frequencies in the corpus. The only action that occurs in all MIs is 'pull the tick out'. All actions in the corpus are expressed as processes with an imperative verb. Most MIs ($N = 7$) contain five actions which the user should carry out to perform the task 'neem' (take), 'pak vast' (grab), 'trek uit' (pull out), 'ontsmet' (disinfect) and 'noteer' (write down). Some MIs omit the instruction to take the tweezers (MI8 and MI10), some MIs include other actions, for instance to wash hands after contact with the tick, to take a photograph of the wound, to monitor the victim's health or to visit a GP. In four MIs alternative actions are offered for cases in which a particular action cannot be performed. These cases usually include a conditional. For instance MI18 suggests to grab the head of the tick as near to the skin as possible, if the user has no tweezers at hand, while MI13 instructs the victim to visit a GP if the tick was not removed successfully.

Control information: Control information is included about seven times per MI, usually in the form of warnings and explanations, incidentally also conditionals or motivations occur. Warnings include negated actions like, not to damage the tick (M18), not to drip any liquids on the tick (MI15), or not to crush the tick (M13). Explanations discuss the dangers of a Lyme disease infection or inform the user that it is not harmful if parts of the mouth of the tick are left behind in the skin as a result of pulling the tick out.

2.2 Description of Pictures

The MI pictures were analysed in terms of general characteristics (e.g., type, visualised objects), actions (aspect: process or result) and control information (e.g., warning, explanation, extra information). All MIs in the corpus use five drawn pictures with visualisations similar to the ones presented

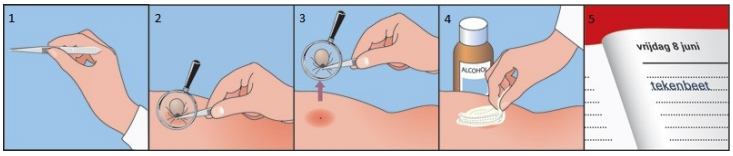

Figure 3: Visualised actions (3,4), and results of actions (1,2,5) in https://thci.nl/teek-pak-beet/.

in Figure 3. These pictures are usually positioned horizontally above or below the instruction. Only in MI13 the pictures are presented vertically to the right of the text. The pictures visualise a number of objects (i.e., tweezers, magnifying glass, tick, wound, flask with alcohol, wipe, agenda, human hands, skin). The magnifying glass is used to show the tick, how to grab it and how to pull it out (see Figure 3). In one MI, MI15 (see Figure 2), an inset in the second picture is used to enlarge the tweezers and the tick as well as to add an exclamation mark to indicate that care should be taken while performing this action. In the picture that visualises the action 'pull the tick out', all MIs use an arrow to indicate the movement as well as the direction in which to pull. The action 'disinfect the wound' is also visualised as a process in which a hand wipes the wound with cotton. However, no arrows appear in any of the pictures to indicate this movement. Not all actions are visualised in the MI pictures as processes. Instead, the result of the action is visualised for the actions: 'pick up tweezers', 'grab the tick' and 'write down the date', where respectively a hand already holds the tweezers, the tick is already grabbed and the date is already noted. Control information in the pictures usually occurs once per MI, in the sense that a magnifying glass is used to provide more detail. The exclamation mark in Figure 2 is an exceptional occurence of a visualised warning.

	Action	**Aspect of Action in Text**	**Number of Verbalised Actions**	**Aspect of Action in Pictures**	**Number of Visualised Actions**
Preparatory Actions	Pick up the tweezers	Process	9	Result	15
Core Actions	Grab the tick	Process	14	Result	15
	Pull the tick out	Process	15	Process	15
Closing Actions	Desinfect the wound	Process	14	Process	15
	Wash hands	Process	1	-	-
	Write down the date	Process	13	Result	15
	Take a photograph of the wound	Process	1	-	-
	Monitor health	Process	1	-	-
	Visit GP	Process	2	-	-

Table 1: Steps in removing a tick as presented in 15 MIs (control information excluded).

2.3 Text-Picture Relations

Table 1 presents an overview of the actions in the corpus to show how text and pictures are related. For each MI all actions are only counted once; repetitions like the alternative way to pull the tick out without tweezers are not counted. The pictures present the same information in all MIs. In ten cases,

the text does not mention an action that is depicted in one of the pictures. Seven of these omissions concern details of the actual pulling-out action: ‘pick up the tweezers’ (6 cases), ‘grab the tick’ (1 case). The remaining three omissions concern closing actions: ‘disinfect the wound’ (1 case), and ‘write down the date’ (2 cases). Conversely, there are five cases where a verbalised action is not shown in picture. They all concern closing actions, most of which would be hard to visualise in the instruction (e.g. ’monitor health’, ’visit GP’).

3 User Study

We illustrate the use of text-picture analysis for user-based evaluation of document design with a user study, conducted by the second author as part of her MA thesis research (Eppinga, 2017).

3.1 Setup

The study investigates how horizontal (H) or vertical (V) positioning of the pictures (P) and the text (T) affects the comprehension, recall, performance and attractiveness of a tick-removal instruction. The setup four conditions: HPT ($N = 5$) and HTP ($N = 6$), VPT ($N = 6$), VTP ($N = 6$). The participants in the user study were 22 mothers with at least one child less than 16 years old and with a mean age of 41.2 years. The participants did not hold a first-aid certificate and had no experience with removing ticks. Twelve of the participants worked at the university, including eight in the linguistics department. The education levels of the participants varied, but were balanced across the conditions. Figure 4 displays the four MI versions that were presented to the participants. The MI, with five written instruction steps accompanied by five pictures, presents the basic procedure derived from the results of our corpus analysis. Figure 5 shows the setting in which the participants were recorded on video while reading one of the four MI versions on the screen of an eye tracker and carrying out the instruction using the materials on the tray in front of them: tweezers, cotton, alcohol, pen, agenda, and a puppet with a tick (Figure 6). To avoid ethical issues, the tick was represented by a headed pin located in the armpit of the puppet.

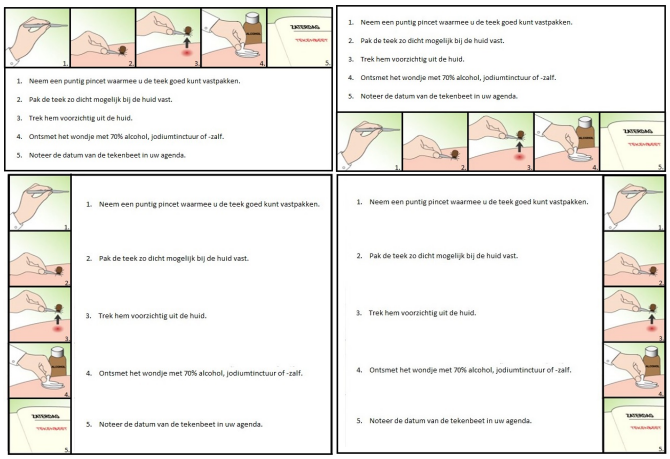

Figure 4: Four versions of the MI to remove a tick f.l.t.r. HPT, HTP, VPT and VTP.

Figure 5: Experiment setting.

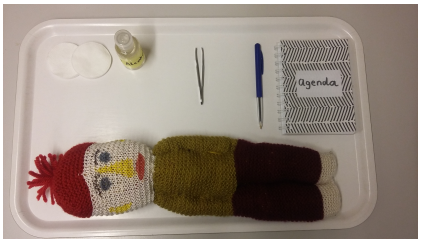

Figure 6: Tick removal materials.

3.2 Procedure

Participants in the study were received in our eye tracking lab and invited to read the introduction to the study. All participants signed a consent form to indicate their voluntary participation and to allow the use of the data they provided for research purposes. Participants first filled out a demographic questionnaire and then sat down in front of the eye tracker and camera. Participants were free to choose their own strategy in carrying out the instruction; in particular, they could switch freely between the screen and the tray. As the researcher was seated out of sight, participants were asked to indicate when they were finished executing the instruction. Once finished, the participants filled out a questionnaire for measuring comprehension and recall. Subsequently, participants were shown all four variations of the instruction and were asked to indicate which one they would prefer to use in real life. Finally, a short semi-structured interview was conducted to gain further insights in the participants experience. The sessions were closed with a debriefing.

3.3 Results

Video analysis and the questionnaires: Table 2 shows that it took participants approximately 70 seconds to process the instruction and carry out the task. No errors were made in the sequence of the instructed actions, however some actions were omitted (1 point) and sometimes the performance quality was suboptimal (0.5 points), or the materials were not properly used (1 point). In total, 2 errors were made in the action 'disinfect the wound' (once omitted and twice suboptimally performed) and 3.5 errors were made in 'write down the date' (twice omitted and three times suboptimally performed), 1 error was due to improper use of attributes. The mean number of switches participants made in looking from the instruction to the task and vice versa, were highest in the conditions HPT and VPT, where the pictures were positioned on the left or above the text. When asked to describe the instruction after finishing it, participants usually recalled four of the five actions. Participants were also able to answer most of the seven cued recall questions correctly and indicated that the MIs were comprehensible. No significant differences were found in these measurements between the conditions. When comparing the four versions of the MI, participants found the vertically oriented designs (VPT and VTP) more attractive than the horizontally oriented ones (HPT and HTP). In the interview participants commented that they considered 'grab the tick' and 'pull it out' as one action instead of two. Moreover, they found the instructions comprehensible and easy to perform, but difficult to recall.

		HPT	**HTP**	**VPT**	**VTP**
Performance	*Duration in seconds*	68.9	71.8	69	71.1
	Total number of errors	0.5	3	2	1
	Number of switches	7.8	5.2	7.8	4.8
Recall	*Free (5 actions)*	4.4	4.6	3.7	4.5
	Cued (7 questions)	4.4	5.2	5.2	5.3
Comprehension	*4 questions*	6.5	6.5	6.7	6.4
Attractiveness (N of choices)	*Most attractive*	4 (18%)	0 (0%)	8 (36%)	10 (45%)
	Least attractive	4 (18%)	16 (72%)	1 (4.5%)	1 (4.5%)

Table 2: User study results (means unless otherwise indicated; comprehension was measured on a scale from $1 = incomprehensible$ to $7 = comprehensible$).

Eye movements: Due to technical issues, the eye tracker data of one participant in condition VTP was discarded. In general participants spent 11.42 seconds reading the text and 2.09 seconds viewing the pictures. Fixation on areas of interest was measured for each verbal step and each picture. In general the text was better studied than the pictures (98.25% versus 80.75%). Interestingly, the picture fixations show that only the pictures in the VTP condition were fully studied (100%), while the pictures in the other conditions drew considerably less attention (HPT= 64%, HTP= 72% and VPT= 87%). All participants started the task execution only after they had fully studied the instruction. The eye tracker data shows that participants always started with reading the first step of the instruction, even when there were pictures placed to the left of or above the text. In the conditions HPT and HTP the pictures were only studied after participants had read the verbal instruction at least as far as step 3. In general, the pictures were studied more thoroughly by the participants in the conditions in which the pictures were placed to the right of or below the text.

4 Conclusion and Discussion

The studies described in this paper present a worked example of annotating and evaluating multimodal content. As the possibilities to describe multimodal content are infinite, we advocate corpus studies and user studies to determine the relevance of annotation categories in situated contexts. Although small scale, results of studies like these allow fine-tuning of annotation models for multimodal content in terms of text, pictures and their relations as implemented in the PAT Workbench (Van der Sluis, Kloppenburg & Redeker, 2016). We are currently developing more fine-grained models that allow specification of functional and content relations between e.g., words, clauses and sentences, and between text and pictures. The application of those models to a wide variety of first-aid instructions (several hundred in our PAT corpus) and their tests in controlled user studies, will inform the development of an evaluation system and design guidelines for MIs.

References

Aouladomar, F. (2005). A semantic analysis of instructional texts. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS-6)*.

Bateman, J. (2014). Using multimodal corpora for empirical research. In *The Routledge handbook of multimodal analysis*, pp. 238–252. Routledge, London.

Eppinga, A. N. (2017). Van voor naar achter of van links naar rechts? Een onderzoek naar de positionering van tekst en beeld in tekenbeetinstructies. Master's thesis, Communication and Information Sciences, University of Groningen.

Houts, P., C. Doak, L. Doak, and M. Loscalzo (2006). The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling 61*(2), 173–190.

Katz, M., S. Kripalani, and B. Weiss (2006). Use of pictorial aids in medication instructions: a review of the literature. *American Journal of Health-System Pharmacy 63*(23), 2391–7.

Maes, A., A. Arts, and L. Noordman (2004). Reference management in instructive discourse. *Discourse Processes 37*(2), 117–144.

Mayer, R. (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In *The Cambridge handbook of multimedia learning*, pp. 49–70. New York.

Schnotz, W., I. Wagner, F. Zhao, M. Ullrich, H. Horz, N. McElvany, A. Ohle, and J. Baumert (2017). Development of dynamic usage of strategies for integrating text and picture information in secondary schools. In D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Eds.), *Competence Assessment in Education*, pp. 303–313. Springer International Publishing.

Schriver, K. A. (1997). *Dynamics in document design: Creating text for readers*. Wiley, New York.

Sweller, J. (2016). *Cognitive Load Theory: What We Learn and How We Learn*. Springer International Publishing.

van der Sluis, I., L. Kloppenburg, and G. Redeker (2016). PAT Workbench: Annotation and evaluation of text and pictures in multimodal instructions. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*, pp. 131–139.

van der Sluis, I., S. Leito, and G. Redeker (2016). Text-picture relations in cooking instructions. In H. Bunt (Ed.), *Proceedings of the Twelfth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-12) at LREC 2016*, Volume 16, pp. 22–27.

van Hooijdonk, C. and E. Krahmer (2008). Information modalities for procedural instructions: The influence of text, pictures, and film clips on learning and executing RSI exercises. *IEEE Transactions on Professional Communication 51*(1), 50–62.