Eye-tracking evidence for multimodal language-graphics comprehension: The role of integrated conceptual representations

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

Cengiz Acarturk
University of Hamburg
Hamburg, Germany
acarturk@informatik.uni-hamburg.de

Abstract

In this paper, we propose a computational architecture for multimodal comprehension of text and graphics. A theoretical account of the integrated conceptual structures induced by linguistic and graphical entities is presented. We exemplify these structures with the analysis of an excerpt from a report published by Point Reyes Bird Observatory (PRBO). Experimental evidence, based on the analyses of subject's eye movement recordings was evaluated under the framework of the architecture.

1 Introduction

Multimodal communication combining language and graphics is a successful means to convey information: it includes *persistent* documents, such as newspaper articles, educational material and scientific papers in print media or in electronic media as well as *transient* oral presentations using power point or chalk-and-blackboard lectures.[1] Humans seem to integrate information provided by different modalities—as language and graphics—almost always based on unconscious cognitive processes. Whereas researchers from different disciplines investigated multimodal documents of different types in different domains, research on cognitive mechanisms underlying multimodal integration is currently in a less mature state and detailed computation models of language-graphic comprehension are rare.

The focus of the present study is multimodal comprehension of expository text accompanied by graphics of a specific type, namely line graphs of functions with *time*-arguments and numbers as values. Figure 1 shows an excerpt from a waterbird census report[2], which contained verbal information about the number of birds (1).

(1) Bolinas Lagoon Population Trends

> From a peak of about 60 wintering birds in 1976, numbers have declined to about 20 birds currently.

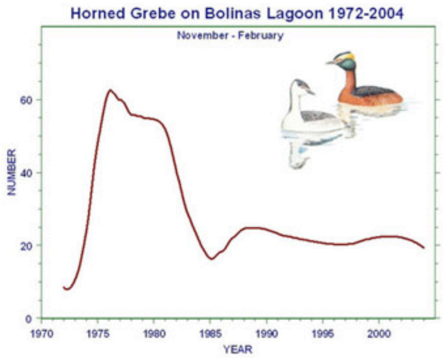


Figure 1. Trend graph depicting the number of wintering birds.

From a linguistic point of view, the process of *referring*, which is constituted by a *referential expression*, as 'peak of about 60', that refers to an *entity* of the domain of discourse, that can contain also abstract entities, as *numbers*, is the core of comprehension. Based on this, *co-reference*, the backbone of text coherence has to be established by speaker and hearer employing internal conceptual representations, which mediate between language and the domain of discourse. In processing text-graphics documents, in which both modalities contribute to a common conceptual representation, additional types of reference and co-reference relations have to be distinguished. Foremost, there exist correspond-

[1] In this paper, we use the term 'modality' as shorthand for 'representational modality'.

[2] "Waterbird Census at Bolinas Lagoon, Marin County, CA" by Wetlands Ecology Division, Point Reyes Bird Observatory (PRBO) Conservation Science: (http://www.prbo.org/cms/index.php, retrieved on 14 April 2009).

ing referential relations (reference links) between graphical entities and entities in the domain of discourse. Furthermore, there exist referential links between linguistic and graphical entities. To sum up, a layer of common conceptual representations is the place where *co-reference links* among conceptual entities introduced by various modalities are constructed where *inter-* and *intra-representational coherence* is established (Seufert, 2003).

A systematic investigation of multimodal comprehension of graph-text documents needs specification of referential link constructions between different representational formats, namely language and graphics. A graph-text document, either in printed or electronic media, is an external representation that includes graphical entities and textual entities.

The purpose of the present paper is to propose a computational model of integrated comprehension of language and graphics based on conceptual representations, which play the crucial role in interfacing between modalities (Jackendoff, 2007). The model is supported by experimental studies using eye-tracking methodology.

2 Integrated Comprehension of Language and Graphics

2.1 Comprehending Language and Comprehending Graphics

Language comprehension, in its most basic form, includes a set of processes that transforms *external* linguistic representations, such as words, phrases, sentences, into internal mental representations, in particular into conceptual structures and spatial representations (Jackendoff, 1996). Comprehension includes phonological, syntactic and semantic processes, which are governed by a set of rules and constraints, often called grammar, and processes of memory retrieval and reasoning to employ knowledge about the world. Furthermore, during the last two decades psycholinguistics has intensively investigated the interaction of—in particular, spoken—language comprehension and visual perception (Ferreira & Tanenhaus, 2007) giving clear evidence that concurrent perception can affect the interpretation of discourse. The ‘language module’ depicted in Figure 2 is based on Tschander et al. (2003); their approach focuses on ‘verbally instructed navigation’, i.e., on a language comprehension task, in which processing of spatial language and spatial knowledge is essential (details of the conceptual representation language presented in this paper is discussed in section 2.3). Therefore specific components to process spatial concepts and to match spatial representations with (idealized) visual percepts are foregrounded in their approach.

Comprehension of graphs, in a similar way to language comprehension, can be seen as a set of processes that transform *external* representations, namely graphics, consisting of axes, tick marks, graph lines, etc., into internal conceptual and spatial representations. Graphs, unlike pictorial representations and iconic diagrams, have grammatical structures. Thus graph comprehension involves—particularly in comprehension of statistical information graphics such as line graphs—perceptual, syntactic and semantic processes (Kosslyn, 1989).

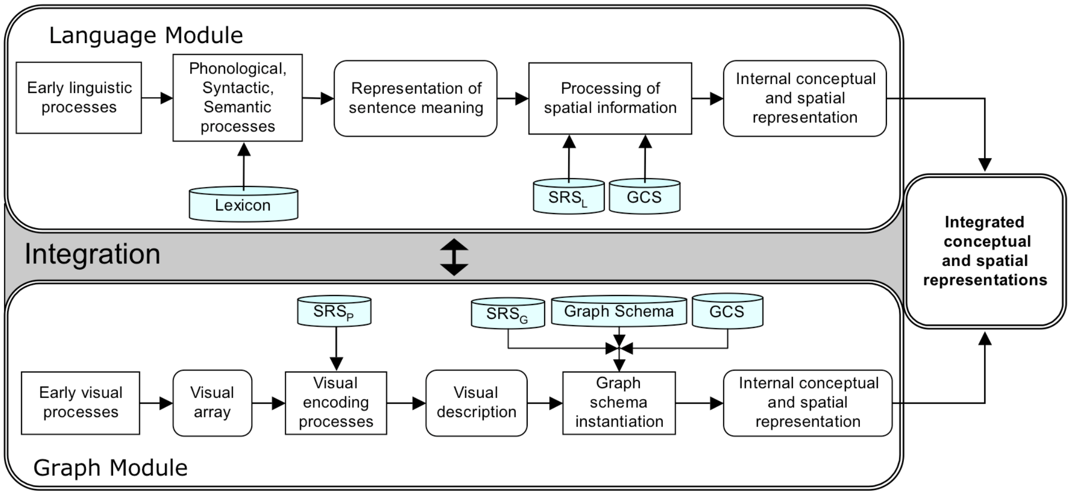

Language Module
Early linguistic processes
Phonological, Syntactic, Semantic processes
Representation of sentence meaning
Processing of spatial information
Internal conceptual and spatial representation
Lexicon
SRS$_L$
GCS
Integration
Integrated conceptual and spatial representations
SRS$_P$
SRS$_G$
Graph Schema
GCS
Early visual processes
Visual array
Visual encoding processes
Visual description
Graph schema instantiation
Internal conceptual and spatial representation
Graph Module

Figure 2. The three basic components of the information flow architecture.

The ‘graph module’ depicted in Figure 2 is an adaptation of Pinker’s (1990) graph comprehension architecture. It transforms the information induced by external graphical representations, such as shape and position of graph line segments, into *visual array* and then into *visual description* by employing visual encoding processes (c.f. *visual routines*, Ullman, 1984). Visual description represents information about relative spatial positions of graphical entities (e.g., horizontal and vertical lines as well as segmented graph lines) and textual entities (e.g., axis labels, value labels). Visual description is then transformed into internal conceptual and spatial representations via instantiation of *graph schemata*. The graph schema is a long-term memory structure that includes information for specifications of gestalt atoms in graphs. For example, for a line graph, these gestalt atoms are the diagonal lines ‘ / ’ and ‘ \ ’ leading to INCREASE and DECREASE concepts (see section 2.3). It is the graph schema that makes possible to process perceptual information provided by the lines on paper or on screen as entities belonging to a line graph. Whereas visual encoding corresponds to the phonological, morphological and syntactic stages of language comprehension, graph schema instantiation corresponds to the semantic and pragmatic stages.

2.2 Multimodal comprehension: Integration

Multimodal comprehension of a text-graphics document requires the integration of information contributed by both representational modalities, namely language and graphs, or in other words, the interaction between the language comprehension module(s) and the graph comprehension module(s) (cf. Schnotz, 2005, and Holsanova 2008, for kindred approaches). As discussed by Habel and Acarturk (2007), in processing text-graphics documents humans construct different types of reference and co-reference relations (cf. section 1). The underlying idea of the present study is that *integrated conceptual representations* mediate between language, graphics and domain entities in multimodal comprehension of language and graphics.

Figure 2 depicts the information flow between the modality specific modules and the integration processes as proposed in this paper. Since humans do language comprehension as well as graph comprehension incrementally—as empirical research in psychology and neuroscience convincingly argues for—the core research questions concerning the internal structure of the integration module are: (a) *which level of incremental entities are involved in integration?*, (b) *which types of representations are constructed by the modality specific modules to be transferred to integration?*, (c) *how are these representations constructed by modality specific modules be processed?*, and (d) *how do integrated representations influence modality specific comprehension.?*[3] In the present paper we focus on questions (b) – (d), in particular on the construction of referential and co-referential links.

2.3 The role of conceptual representations in integration

In a first step, we exemplify the construction of conceptual structures by the language module with example sentence (1).[4] The lexical information of ‘decline’ provides a conceptual representation containing a process concept

$$\mathit{DECREASE\_OF\_VALUE}(\_{\mathit{TEMP}}, \_{\mathit{VALUE}}, \ldots).$$

We focus here only on two arguments of this process, namely a temporal argument, which can be filled by an interval, and value argument, that can be filled by an entity of an ordered structure, which functions as the domain of the value, here the NUMBER-domain. By using such abstract representations, which generalize over different value domains, it is possible to catch the common properties ‘decline of number’, ‘loss of weight’, and others. The temporal argument, which is necessary for all process and event concepts, stands for the ‘temporal interval during which the whole process is occurring’; in sentence (1) the beginning of the interval is explicitly specified. Putting this together, the process concept *DECREASE_OF_VALUE* stands for a specification of a mapping from the temporal domain in the value domain, or—using the terminology of topology—for a ‘path’ in the value space. Moreover, the lexical information of ‘decline’ provides *SOURCE* and *GOAL* arguments to be filled optionally. Sentence (1) supplies ‘peak of about 60’ [via a *from-PP*] and ‘about 20’[via a *to-PP*].

The task of the second phase of line graph comprehension (as depicted in figure 2), the con-

[3] Figure 2 is undetermined with respect to the internal structure of the integration module as well as to the details of the interaction processes since these question are only partially answered up to now.

[4] The system of conceptual and spatial representations we use is a computation-oriented extension of Jackendoff’s conceptual semantics (see, Jackendoff 2007) described in Eschenbach et al. (2000) Tschander et al. (2003).

struction of structured visual descriptions, in particular contains this: descriptions of relevant parts of the graph line, their geometrical properties and spatial relations between these parts. In this step, the system of spatial representations plays the role of a descriptional inventory, which is accessed by visual routines. We exemplify this with salient parts of the trend graph for Horned Grebes (cf. figure 1). Visual segmentation of the line graph leads to—inter alia—a line, which overall direction is vertical and which possesses one local maximum of curvature. Figure 3.a depicts the correspondence between an idealized shape of this type and its structured description by a spatial representation. Figure 3.b depicts the correspondence between a complex part of the line graph, namely a sequence of line segments, which has an overall horizontal orientation, and a abbreviated spatial description of that graphical constellation.[5]

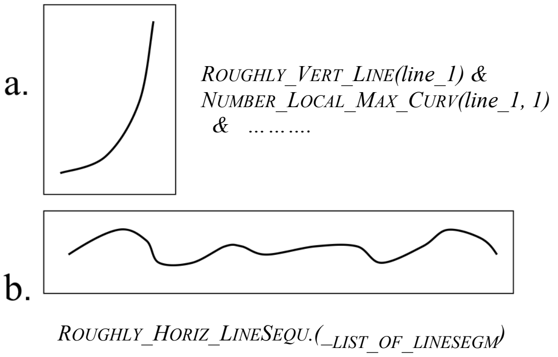

a.
ROUGHLY_VERT_LINE(line_1) &
NUMBER_LOCAL_MAX_CURV(line_1, 1)
& ……….
b.
ROUGHLY_HORIZ_LINESEQU.(_LIST_OF_LINESEGM)

Figure 3. A sample set of integral conceptual representation.

In contrast to the more general *visual encoding processes* the following sub-module, *graph schema instantiation* in Pinker's (1990) terminology, has the task to interpret elements as parts of graphs. In the Horned Grebe graph, for example, the vertically extreme *POINT_OF_MAXIMAL_CURVATURE*, which is characterized as connection point between two roughly vertically oriented lines, will be determined as a *PEAK* of graph line. Since the x-axis of the trend graph in question refers to the temporal domain, the 'natural order' of time, leads to an inherent orientation of the line segments: Thus the most left part of the trend graph has to be interpreted as an *INCREASE*, the following—after the *PEAK*—as a *DECREASE*.

As the Horned Grebe example shows, both modality specific comprehension via contribute via conceptual representations based on a common conceptual inventory and the referential links build up during comprehension to an integrated and—hopefully—coherent interpretation of the text graphics document: The verb 'decline' provides *DECREASE* conceptualizations, as well as the application of graph schemata. The linguistically mentioned 'peak' is source of two referential links, one the one hand, to a domain-entity, namely an approximate number of birds, on the other hand, to a graphical entity.

In the second part of this paper, we present empirical evidence for the process descriptions presented in this section.

3 Eye Movements in Multimodal Comprehension

The investigation of eye movement parameters has been a widely used research method for the investigation of online comprehension processes in psycholinguistics research (Staub & Rayner, 2007), graph comprehension research (Shah & Vekiri, 2005), as well as in multimodal discourse analysis (Holsanova, 2008). But, as of our knowledge, there is no systematic analysis of eye movement behavior on graph-text documents.

3.1 The Experiment

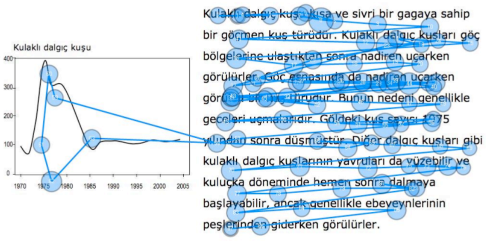
We conducted two experiments, both based on the material exemplified in Section 1. In Experiment 1, ninety-one graduate or undergraduate students were presented 42 graph lines in rectangular frame, without any labels or numbers. The graphs were redrawn based on the original source (see fn. 2). The subjects were informed that they would see a set of graphs on the screen, each for three seconds; and they were expected to inspect the graphs as they change automatically.

In Experiment 2, text-graph constellations were presented to 36 graduate or undergraduate students. Each subject was presented twelve text-graph documents, similar to the one in Figure 4. The figure also shows resulting eye movement patterns on the presented stimuli.

The stimuli were based on the texts and graphs in the original source, after redrawing of graphs and modifications of text for systematic investigation. There were two factors in Experiment 2: the shape of the graph line (with three condi-

[5] A detailed description of these steps is beyond the scope of this paper. De Winter & Wagemans (2006) give a thorough overview about segmentation processes in perceiving line drawings.

tions) and the number of graph-related sentences in the text (with four conditions). We call these *target sentences*. The text in each stimulus consisted of three parts in the mentioned order: (1) several sentences before the target sentences (namely, *pre-target sentences*). These were not related to the graph, presenting information such as breeding, migration etc. (2) The *target sentences*. (3) Several sentences after the target sentences (*post-target sentences*). These were not related to the graph. The subjects were informed that they would see inventory information about wintering birds; they were expected to investigate the presented information and then to answer some questions. A 50 Hz eye tracker recorded eye movements of the subjects in both experiments.



Kulaklı dalgıç kuşu kısa ve sivri bir gagaya sahip
bir göçmen kuş türüdür. Kulaklı dalgıç kuşları göç
bölgelerine ulaştıktan sonra nadiren uçarken
görülürler. Göç esnasında da nadiren uçarken
görülen bir kuş türüdür. Bunun nedeni genellikle
geceleri uçmalarıdır. Göldeki kuş sayısı 1975
yılından sonra düşmüştür. Diğer dalgıç kuşları gibi
kulaklı dalgıç kuşlarının yavruları da yüzebilir ve
kuluçka döneminde hemen sonra dalmaya
başlayabilir, ancak genellikle ebeveynlerinin
peşlerinden giderken görülürler.

Figure 4. Sample eye movement protocol.

3.2 Results

In this section, a partial summary of the results of the experiments is presented. First, we discuss the results concerning the characteristics of eye movement behavior in a qualitative manner. Based on the eye movement recordings, fixation maps can be drawn, as exemplified in Figure 5. This fixation map is based on the fixation counts of all the subjects on one of the graphs presented in Experiment 1. In this figure, the red, yellow and green regions show fixation distribution in decreasing order.

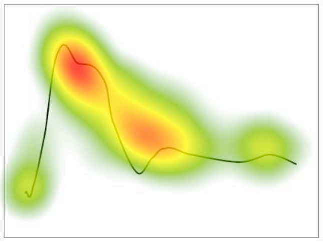


Figure 5. Sample fixation map.

Since the graphs were not accompanied by text in Experiment 1, the resulting fixation maps reflect the visually salient regions. In other words, these patterns were *not-linguistically-guided* fixation patterns. In Experiment 2, part of the stimuli of the first experiment was presented with accompanying text. We have divided the fixations on the graph region in Experiment 2 into three groups for analysis: (1) the fixations before the target sentences were read (namely, *pre-target phase* fixations). These occurred generally at the beginning of the reading of the text. (2) The fixations immediately after reading the target sentences (*target-phase* fixations). (3) The fixations after the target sentences were read. (*post-target phase* fixations). These occurred generally at the end of the reading phase. In this study, we focus on the first two types of fixations.

The fixations on the graph region were transcribed based on their location and total gaze time. A qualitative comparison of the fixation patterns, based on the exemplified stimuli in Figure 4 revealed that in Experiment 2, the *pre-target phase* fixation patterns were different than the *target-phase* fixation pattern*s*. Furthermore, the *target phase* fixation patterns of Experiment 2 were different than the *not-linguistically-guided* fixation patterns obtained in Experiment 1. In other words, different fixation maps were obtained in *linguistically-guided* and *not-linguistically-guided* inspection of the same graph.

A further analysis of the *target-phase fixations* in Experiment 2 was performed by quantitative comparisons of the fixation counts and gaze times on the graph proper (the fixations on the numbers and labels were excluded) that occurred after the two target sentences of the accompanying text: "*The number of birds declined after 1975*" and "*The number of birds remained stable around 100 after 1985*".[6] The results showed that after the 'decline' target sentence, the mean fixation count was higher on the decline-line than the mean fixation count on the remain-line of the graph, $t(16) = 4.76$, $p < .01$. On the other hand, after the 'remain' target sentence, the mean fixation count was higher on the remain-line than the mean fixation count on the decline-line of the graph, $t(11) = -5.70$, $p < .01$. On the other hand, in Experiment 1, there was no significant difference between the mean fixation count on the decline-line and the one on the remain-line, $t(90) = 1.86$, $p = .07$. Parallel results were found for gaze times. In summary, quantitative analyses revealed partial evidence that the linguistic repre-

[6] The number of target sentences had four conditions from one sentence to four sentences. For the purpose of this study, we compare the fixations only after the first and second target sentences.

sentations with different conceptual representations, in our case the '*decline*' and the '*remain*' sentences resulted in significant differences between mean fixation counts and gaze times.

4 Discussion

In the present paper we proposed a computational architecture for multimodal comprehension of text-graphics documents. We analyzed comprehension processes in terms of the interaction between the information induced by graphical and linguistic entities at conceptual level. We presented experimental support for the architecture by the analysis of eye movement patterns and parameters. First, we presented evidence for the difference between linguistically-guided and not-linguistically-guided inspection of graphs. Second, the findings of Experiment 2 revealed a difference between the fixations that followed the 'decline' sentence and the 'remain' sentence.

5 Conclusion

The interaction between language and graphs, as the two representational modalities, is not a well-investigated domain compared to research on multimodal comprehension of pictorial or diagrammatical illustrations. Methodologically, compared to research on eye movement control in reading, the studies that investigate eye movement behavior in multimodal documents have a relatively premature state due to abundant types of visual representations. This study contributes on both theoretical and experimental aspects of research on multimodal graph-text comprehension.

Acknowledgments
The research reported in this paper has been partially supported by the DFG (German Science Foundation) in ITRG 1247 'Cross-modal Interaction in Natural and Artificial Cognitive Systems' (CINACS). We thank the HCI Lab at the Middle East Technical University and the two anonymous reviewers for their helpful comments.

References

De Winter, J. & Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale integrative study. *Cognition, 99*. 275–325.

Eschenbach, C., Tschander, L., Habel, C., & Kulik, L. (2000). Lexical specifications of paths. In C. Freksa, W. Brauer, C. Habel & K. F. Wender (Eds.), *Spatial Cognition II* (pp. 127-144). Berlin: Springer.

Ferreira, F. & Tanenhaus, M.K. (2007). Introduction to special issue on language-vision interactions. *Journal of Memory and Language, 57*. 455-459.

Habel, C., & Acarturk, C. (2007). On reciprocal improvement in multimodal generation: Co-reference by text and information graphics. In I. van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds.), *Proceedings of the Workshop on Multimodal Output Generation: MOG 2007* (pp. 69-80). United Kingdom: University of Aberdeen.

Holsanova, J. (2008). *Discourse, vision, and cognition. Human Cognitive Processes 23*. John Benjamins Publishing Company: Amsterdam / Philadelphia.

Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In P. Bloom; M. A. Peterson; L. Nadel & M. F. Garrett (eds.), *Language and Space*. (pp. 1-30). Cambridge, MA: The MIT Press.

Jackendoff, R. (2007). Linguistics in cognitive science: The state of the art. *The Linguistic Review 24*. 347-401.

Kosslyn, S.M. (1989). Understanding Charts and Graphs. *Applied Cognitive Psychology, 3*. 185-226.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73-126). Hillsdale, NJ: Erlbaum.

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

Schnotz, W. (2005). An Integrated Model of Text and Picture Comprehension. In R.E. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning.* (pp. 49-69). Cambridge: Cambridge University Press.

Shah, P., Freedman, E., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 426-476). New York: Cambridge University Press.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In G. Gaskell (Ed.), *The Oxford* handbook *of psycholinguistics* (pp. 327–342). New York: Oxford University Press.

Tschander, L., Schmidtke, H., Habel, C., Eschenbach, C., & Kulik, L. (2003). A geometric agent following route instructions. In C. Freksa, W. Brauer, C. Habel & K. F. Wender (Eds.), *Spatial Cognition III*. (pp. 89–111). Berlin: Springer.

Ullman, S. (1984). Visual routines. *Cognition, 18*, 97-106.