MOG 2007

Workshop on Multimodal Output Generation

CTIT Proceedings of the Workshop on
Multimodal Output Generation

Ielka van der Sluis, Mariët Theune, Ehud Reiter and Emiel Krahmer (eds.)

1495 University of Aberdeen

Tilburg University

University of Twente
Enschede - The Netherlands

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Sluis, van der. I., Theune, M., Reiter, E., Krahmer, E.

MOG 2007
Proceedings of the Workshop on Multimodal Output Generation
I. van der Sluis, M. Theune, E. Reiter, E. Krahmer (Eds.)
University of Aberdeen, United Kingdom
25-26 January 2007

ISSN 1574–0846
CTIT Workshop Proceedings Series WP07-01

trefwoorden: multimodality, natural language generation, multimodal generation,
modality choice, human modalities.

Book orders:
Ms. C. Bijron
University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
Human Media Interaction
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893680
fax: +31 53 4893503
Email: c.g.bijron@ewi.utwente.nl

Druk- en bindwerk: Reprografie U.T. Service Centrum, Enschede

On Reciprocal Improvement in Multimodal Generation: Co-reference by Text and Information Graphics[*]

Christopher Habel and Cengiz Acartürk
Department of Informatics
University of Hamburg
D-22527 Hamburg, Germany
{habel / acarturk }@informatik.uni-hamburg.de

Abstract

In this paper, we argue that in the production of complex documents combining text and information graphics, improvement—seen as a specific revision process—has to be considered as an independent module in a multi-pass architecture. Inside improvement, graph comprehension and text comprehension modules interact in building up a common content representation based on a conceptual inventory specified via topological and geometrical concepts. During concurrent comprehension the reciprocal improvement module inspects possible gaps in co-reference and coherence and decides which gaps should be filled. We exemplify these tasks and processes with an analysis of an excerpt from a business-news article in the New York Times.

Keywords: Multimodal generation, graph comprehension, conceptual representations, spatial concepts, information graphics, business news

1 COMBINING TEXT AND INFORMATION GRAPHICS

Documents containing modalities such as figures (graphs, drawings, photographs), tables, equations etc. together with text are wide-spread in print media as well as in electronic media. The most frequently cited argument for combining modalities—we call it the *argument of division of labor between representational modalities*—can be exemplified by a characterization given in the *Publication Manual* of the American Psychological Association (APA):

> Tables are often preferred for the presentation of quantitative data in archival journals because they provide exact information; figures typically require the reader to estimate values. On the other hand, figures convey at quick glance an overall pattern of results. They are especially useful in describing an interaction—or lack thereof—and nonlinear relations. A well-prepared figure can also convey structural or pictorial concepts more efficiently than can text. (4th ed., 1994, p. 141)

Although scientific texts—in particular those in a specific field—are the main focus of interest in this characterization, it is also applicable to multimodal documents in non-scientific journals or newspapers etc., as well as to corresponding documents accessible via the Internet. Interpretation of graphics is crucial in many areas such as trends in economy, diagnosis and medical treatment, time-series analysis of experimental data, computer-based instructional material in early school education, among others. Graph

[*] 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).

comprehension research literature emphasizes a broad range of factors playing important role in interpretation of graphical data. For example, Peebles and Cheng (2001) specifies information retrieval from graphics as the interaction between visual properties of graph, cognitive abilities of graph user and requirements of task. There are also studies suggesting methods and design guidelines for improvements to facilitate reader's interpretation of graphics and identification of trends (e.g. Kosslyn, 1994, 2006). Nevertheless, there are inconsistencies in interpretation of the same graphical data (e.g. trends in time-series graphics) even among expert scientists (DeProspero and Cohen, 1979). Such limitations, caused by the implicit nature of diagrammatical representations to present certain aspects in interpretation of quantitative data as well as perceptual difficulties during early stages of information extraction bring the need for multimodal documents including both graphics and text.

While diagrammatical representations are accepted to be computationally more efficient than sentential representations in specific tasks and domains (Larkin and Simon, 1987), in addition to the proposals that use of multiple modalities facilitates learning (Ainley et al. 2000, Winn 1991), the principal pros of combining information in different modalities are opposed to—possibly—additional cognitive efforts of producers and recipients in processing cross-modal relations (e.g. co-reference, coherence etc.) and in considering cross-modal dependencies. Thus, systematic investigations of such relations and dependencies are fundamental for any approach to comprehension or production of multimodal documents combining text and graphics. In the present paper, we exemplify the additional cognitive tasks with the case of *co-reference*, which was also in the focus of some NLG approaches to generation of multimodal output, such as WIP or AutoBrief (for an overview see André, 2000).

2 REFERENCE AND CO-REFERENCE IN MULTIMODAL DOCUMENTS

From a linguistic point of view, the basic type of *referring* is constituted by a *referential expression* that *refers* to an *entity* of the domain of discourse. *Co-reference*, the backbone of text coherence has to be established by speaker and hearer employing internal—conceptual—representations, which mediate between the language and the domain of discourse. In processing multimodal documents, additional types of reference and co-reference relations have to be distinguished. Foremost, there exist corresponding referential relations (reference links) between graphical entities and entities in the domain of discourse.[1] Figure 1 shows a hand-drawn sketch map: some of the *lines* refer to rivers, to roads or to parts of the costal line, i.e. they refer to entities in the geographical world, whereas other lines constitute a *rectangle*, i.e. member of another class of graphical entities, which refers to a region, namely Aberdeen University's 'Old Aberdeen Campus', or other regions, such as part of a harbour or the North Sea (see Figure 1). In other words, the systems of regularities for combining atomic graphic entities to complex, meaningful configurations behave similar to grammatical systems.

When the producer of this sketch map explains the environment using the map, for example by saying "The rail station is between the red lines left of the harbour", the recipient has to integrate referential links of different types. Firstly, there are links between linguistic expressions and geographical entities, e.g. a reference relation between "*rail station"* and a *building in the town of Aberdeen*. Secondly, there are referential links between graphical entities, e.g. *lines*, and geographical entities, e.g. *streets*, and thirdly there are links between linguistic expressions and graphical entities, e.g. between "*red lines*" and some *red lines on a sheet of paper.* The composition—in the mathematical sense of composition of maps or relations—of the language-to-graphics reference and the graphics-to-domain reference leads to a language-to-domain reference, which we call in the following *implicit reference.* In particular, the implicit reference link mentioned above connects the phrase "*between the red lines*" to a region in the town of Aberdeen, which was not explicitly mentioned in the text.

[1] In the present paper we focus exclusively on co-reference in text–graphics multimodality; other types of multimodal communication, e.g. the combination of language and gesture, show similar phenomena and problems (see Beun and Cremers, 2001). Graphical entities that hold meaning with respect to the domain, and thus can be seen as corresponding to words, phrases or sentences in language, is one the central research topics in the psychology of graph comprehension (cf. Kosslyn, 1989; Shah and Hoeffner, 2002).

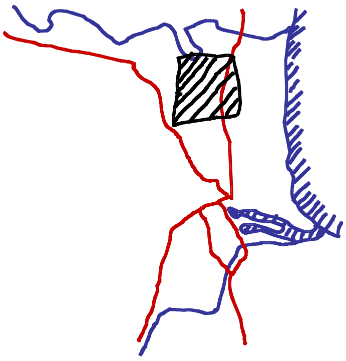


Figure 1: Sketch map of Aberdeen (pure depiction without textual elements).

The types of reference and co-reference relations we exemplified above with the class of (sketch) maps are central for analyzing the reference relations that are involved in the processing of text-depiction documents in general. (See Figure 2, which depicts the structure of different types of referential links and their composition.) Tappe and Habel (1998) describe that people producing verbal descriptions of the drawing of sketch maps employ two conceptual layers of representation: a layer corresponding to graphical entities and a layer corresponding to domain entities (entities of the real world referred to by text and sketch map). The empirical data of their experimental study supports the assumption that both layers are simultaneously accessible during speech production. Furthermore, different experimental settings of the verbalization task leads to different use of referential types, which results in different usage of words and phrases, correspondence to graphic entities vs. correspondence to real-world entities. In contrast to Tappe and Habel (1998), which focus on the relations between the external representations (language and graphics) and the layers of conceptual representations active in language processing, Tabachneck-Schijf, Leonardo and Simon's (1997) CaMeRa model of *Computation with multiple representations*, which proposes 'referent ties' as a major means to link different layers of internal representations, namely between pictorial and verbal short-term memory, focuses on the connections between the layers of short-term memory and of long-term memory.

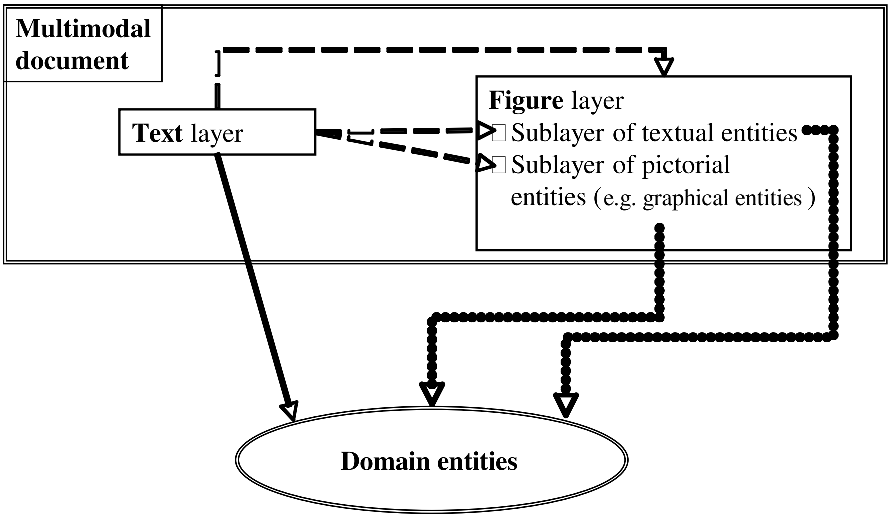

Multimodal document
Text layer
Figure layer
Sublayer of textual entities
Sublayer of pictorial entities (e.g. graphical entities)
Domain entities

Figure 2: Explicit and implicit reference links in multimodal documents (figures with enclosed textual elements).

We will now exemplify the general character of the *reference type* and *representation layer* presented above with text–graphics combinations from the domain of meteorology:[2] referential and co-referential links going out from linguistic entities can consider, firstly, figures as a representational whole, as *'Figure 3'* in (1), secondly, graphical entities, which are constitutional parts of a figure, as*'peak'* in (2), and thirdly, entities, which are represented by graphical entities—as 'warmer period' in (2) or 'norm temperature in (1).

(1) *Figure 3* shows the deviation of the average temperature in August from the *norm temperature* w.r.t. the period 1961–1990.

(2) The warmer period starting in the eighties culminates in the peak at 2003.

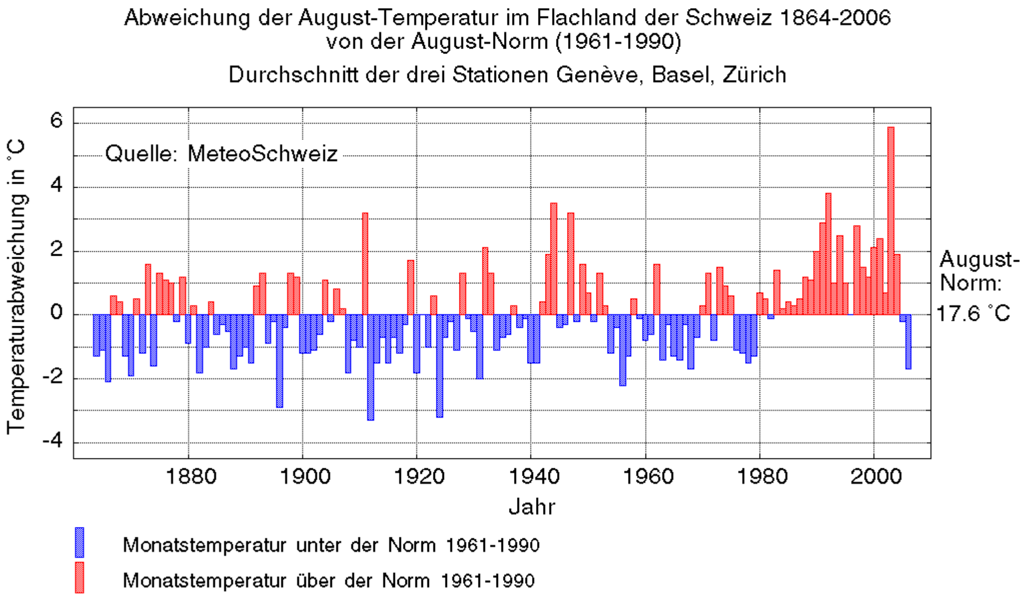


Figure 3: August-temperature chart of Switzerland (©MeteoSwiss / Federal Office of Meteorology and Climatology).

A major difference between maps and information graphics considers the ontological inventory of the domain 'accessible by pictorial elements': whereas maps can be seen as—more or less veridical—pictures of geographic space, information graphics are especially used as visualizations of abstract entities, i.e. as externalization of our mental conceptualization of the external world. The domain entities referred to—from sentences (1) and (2) as well as from the information graphics (Figure 3)—have a more or less abstract nature: beyond the geographical frame of Switzerland with some weather station, the temporal dimension of the domain considers 'calendar entities' as years and months, periods of temporal entities. Furthermore, the temperature dimension is in the focus of the document, both on the textual and on the graphical layer: the document refers to temperatures, average temperature, deviation of average temperature, etc. (The abstract character of the entities referred to by information graphics will be discussed in more detail in Section 3.)

In comprehending multimodal documents recipients have a range of freedom when to turn to text and when to turn to figures. For example, a figure can get their attention immediately after turning over the

[2] The graphics presented here was published by MeteoSwiss, the Swiss Federal Office of Meteorology and Climatology, on 01.09.2006 as part of a multimodal document with the title "Wie kalt war der August 2006 wirklich?" (How cold was it really during August 2006?). The textual examples used in the present paper are either translations—by the first author—from the original or slight modification to exemplify a phenomenon with respect to the graphics.

page, but the recipients can also neglect focusing on figures until these become relevant to construction of meaning from the text or explicit reference to figures or graphical entities is given verbally. With respect to a particular word or phrase in the text it is difficult to decide which type of reference—to the figure or to the domain—is at issue only by observing the behavior of people comprehending text-figure configurations. For example, 'the peak' in (2) can be interpreted as a *peak in the graph* or as a *peak temperature*, thus resulting in an explicit language-to-graphics reference (to a part of Figure 3) or in an implicit language-to-domain reference via a mediating language-to graphics link. Since the use of definite article in 'the peak' should—according to many approaches to resolving co-reference links—cause the readers to search a peak representation preexisting or easy to find in the discourse model, the two above mentioned alternatives of reference relations can be characterized as 'finding a peak in the accompanying graph' vs. 'constructing (via a bridging inference) a peak in the course of temperature'.[3]

Conversely, producers of multimodal documents have the task to give explicit or implicit hints in the text to lead the recipient's attention as early as necessary to the figure—or even to that parts of the figure—relevant for understanding a sentence or paragraph. Furthermore, producers have to decide in which cases the implicit, mediated language-to-domain reference would not be sufficient and explicit language-to-domain reference has to be realized.

3. IMPROVEMENT: AN ADDITIONAL STEP IN PRODUCING TEXT–GRAPHICS DOCUMENTS

When humans produce multimodal discourse containing language and figures, there exists a spectrum of different grades of coupling the processes of language production, on the one hand, and of the design and realization of figures, on the other hand. Whereas lecturing using chalk and blackboard, e.g. in mathematics or economics—the latter task is used as domain by Tabachneck-Schijf, Leonardo and Simon (1997)—mostly is a one-pass multimodal production process, in which the producer generates speech, writes text and sketches diagrams or graphs in an integrated manner, the production of a newspaper article containing graphics—such as the example discussed in Section 3.1—often is a multi-pass production process, in which the tasks of text production, of graphics production and of text-graphics integration can even be distributed to different people or institutions, in particular, if they possess specific expertise; this case of distributed production corresponds to *type 3* in André's classification (André, 2000, p.309).

Since the dichotomy 'one-pass production process' – 'multi-pass production process' is basic, we will shortly discuss those aspects that are essential for the following. Human production of speech is a prototypical case of a one-pass production process: even repair processes can be seen as part of the only pass (cf. the status of self-perception and self-repair in Levelt's (1999) 'blueprint of the speaker'). When we see co-production of speech and writing or drawing on a blackboard as one task, which can be performed by concurrent processes, then the one-pass perspective is taken with respect to the temporal granularity of speech perception processes. In contrast to this, for humans the production of a route-instruction combining written text and graphics is—on the level of the motor actions of writing and drawing—not realized via concurrent processes. Nevertheless, on the level of designing and realizing a multimodal document, which fulfills specific communicational goals, it is appropriate to see this generation process also as a one-pass process. In particular, since the realizing constraints of humans—on the motoric level—are not relevant for machines, it is appropriate to use one-pass architectures for corresponding computer systems, e.g., the direction services of Mapquest.com. Integrated multimodal generation has been successfully realized using architectures of the 'pipeline' style (Reiter, 1994), or kindred architectures as in WIP (Wahlster et al. 1993) or AutoBrief (Green et al. 2004), in which monitoring and repairing are generation-internal subtasks.

[3] Although cognitive psychology researchers have obtained many fundamental insights in multimodal construction of internal models from text and diagrams (see Glenberg and Langston, 1992; Hegarty and Just, 1993), their models and experiments do seldom focus on the analytical level of the semantics and formal pragmatics of referring expressions, which is in the focus of our research. Thus, we plan empirical investigations—e.g. in the eye-movement paradigm—to support our theoretical models described here.

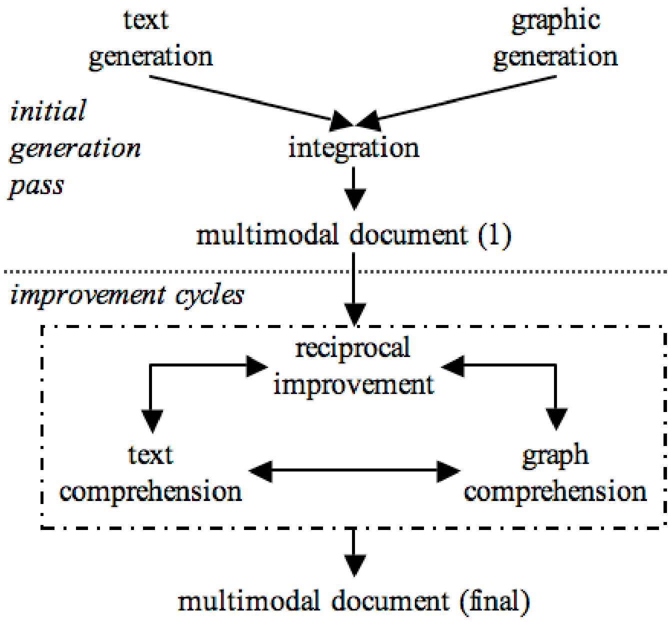


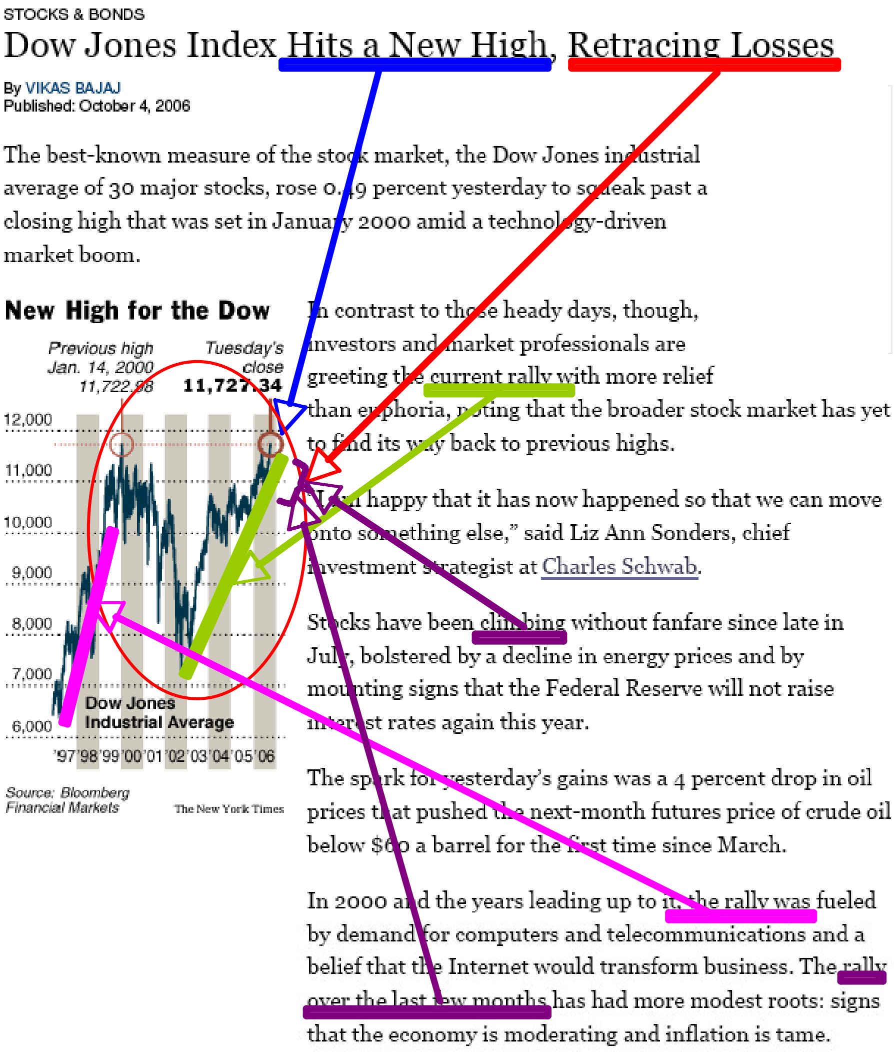
Figure 4: Multi-pass production architecture with improvement module

In contrast, multi-pass production that contains explicit, independent phases of improvement (some authors prefer *revision*) is successful in human text production (cf. Butterfield et al, 1996). Furthermore, Shah, Mayer and Hegarty (1999) show that redesign of graphics increases the quality of viewer's interpretations. Based on these insights from humans' generation of monomodal documents, we propose a multi-pass approach containing an independent 'improvement'—module (see Figure 4) for complex multimodal production tasks. In particular we focus in the present paper on generation, in which modality specific expertise is contributed by loosely coupled, distributed modules or agents.

Improvement of a text-graphic constellation requires comprehension (often called 'interpretation') of both ingredients; in other words, text comprehension and graph comprehension are two basic processes of revising the document (cf. Leinhardt, Zaslavsky and Stein (1990) on the role of interpretation in the construction of graphs). Graph comprehension includes processing of spatial as well as propositional information in different levels by different processes such as preattentive visual processes (Cleveland and Mcgill 1984, 1985, 1987) and interpretive processes (Carpenter and Shah 1998). In multimodal documents, since text comprehension and graph comprehension should lead to an integrated understanding, these processes have to be based on a common system of semantic/conceptual representations, which we will not discuss in detail in the present paper. (Although Green et al. 1998 focus exclusively on the production perspective, their 'content language' is kindred to our representations mentioned above.)

3.1 Improving co-reference by text and information graphics: A case study

In the current section, we discuss multimodal co-reference constellations in an article published in the New York Times on October 4, 2006, to exemplify the requirements on and the tasks of reciprocal improvement in multi-source production of multimodal documents. An excerpt of the text, which was produced by a New York Times author, and the chart provided by Bloomberg Financial Markets, augmented by depictions of referential and co-referential links is depicted in Figure 5. In the following subsection 3.2 we will complement the description of the phenomena with a detailed discussion of the role of conceptual representations in multimodal integration.



STOCKS & BONDS

Dow Jones Index Hits a New High, Retracing Losses

By VIKAS BAJAJ
Published: October 4, 2006

The best-known measure of the stock market, the Dow Jones industrial average of 30 major stocks, rose 0.19 percent yesterday to squeak past a closing high that was set in January 2000 amid a technology-driven market boom.

In contrast to those heady days, though, investors and market professionals are greeting the current rally with more relief than euphoria, noting that the broader stock market has yet to find its way back to previous highs.

"I am happy that it has now happened so that we can move onto something else," said Liz Ann Sonders, chief investment strategist at Charles Schwab.

Stocks have been climbing without fanfare since late in July, bolstered by a decline in energy prices and by mounting signs that the Federal Reserve will not raise interest rates again this year.

The spark for yesterday's gains was a 4 percent drop in oil prices that pushed the next-month futures price of crude oil below $60 a barrel for the first time since March.

In 2000 and the years leading up to it, the rally was fueled by demand for computers and telecommunications and a belief that the Internet would transform business. The rally over the last few months has had more modest roots: signs that the economy is moderating and inflation is tame.

Figure 5: Dow Jones Index Hits a New High, Retracing Losses, by Vikas Bajaj, published on October 4, 2006 (©The New York Times).

The first part of the article title, i.e. "Dow Jones Index Hits a New High", refers with the object-NP 'New High' either—in the language-to-domain way—explicitly to a domain-entity of the conceptual type '*VALUE OF AN INDEX*'[4] or —in the language-to-graphics way—to a maximum point (graphical entity) of the graph. Nevertheless, the verbal attribute 'new' induces that a 'former high' exists, which has—compared with the 'new high'—only minor salience at this stage of comprehension. On the other hand, in the graph exist two small circles that mark maximum points. To sum up, whereas the former high is only implicitly mentioned in the text—presupposed via the phrase 'New High'—it is explicitly presented in the graphics.

The remaining part of the title, "Retracing Losses" co-refers with a salient V-shaped structure of the graph to a complex two-phase event of '*LOSING*' and subsequent '*RECOVERING FROM A LOSS*'. A detailed interpretation of the graph provides readers with information given neither in the title nor in the further article, namely, about the '*AMOUNT OF THE LOSS*'. In interpreting the Dow Jones Chart, the graph-

[4] We use small italic capitals to denote entities of the conceptual representation layers. According to our prior remarks on the abstractness of some domain entities referred to by text and or information graphics, we characterize in this description abstract domain entities referred to by using the same typographical style as to denote conceptual representations. The role of the conceptual layers in multimodal comprehension and improvement will be discussed in more details in Section 3.2.

comprehension sub-module should detect the V-shaped depiction of the *LOSING & RECOVERING* event in question, which contains the '*DEEP OF 2002*' as a salient protagonist not referred to verbally in the text. The improvement process then could react to this analysis by making changes in one or both modalities, for example, by mentioning the topic '*LOSING & RECOVERING*' verbally either in the title or in a subsequent paragraph, or by editing the V-shape structure in the graph to make the topic more explicit there; furthermore, the improvement process can reject to make any changes.

The second paragraph of the excerpt, (3), includes the phrase 'current rally', which refers to the increase resulted in the second high. The lexeme 'rally' corresponds—in its general meaning—to a process of increasing; in the specialized terminology of stock markets the term refers to a rise or recovery in stock prices. In both cases, 'rally' corresponds to a—co-referring—graph-structure of a specific shape, i.e. the right hand site of the V-shape discussed above. Note that in these comprehension steps different types of knowledge interact: on the one hand, general knowledge about graphs, which is used to detect *gestalts* in graphs, and which is kindred to Pinker's (1990) conception of *graph schemata*, and, on the other hand, knowledge of a sublanguage, namely that of stock markets and charts.

(3) In contrast to those heady days, though, investors and market professionals are greeting the current rally with more relief than euphoria, noting that the broader stock market has yet to find its way back to previous highs.

The third paragraph contains information that cannot be deduced from the graph, i.e. the terms included in this paragraph do not have co-reference relations with the components of the graph, i.e. gestalts in the graph. However, at the beginning of the next paragraph, the focus of the reader is shifted back into the graph: "Stocks have been climbing without fanfare since late in July, ...". Although the term 'climbing' corresponds a specific shape (of the type *INCREASE*), which is similar to the right hand side of a V-shape, the granularity of the figure is poor. As a result, there is no distinguished entity to be identified as a relevant object in the chart. This could prompt the improvement module to modify the chart, either by replacing it with another graph of higher resolution, or by graphical highlighting.

The last paragraph of the excerpt, the first two sentences of which are given in (4), mentions two rallies. The 'rally' in the first sentence refers to sequential increases in year 2000 and previous years. The 'rally' in the following sentence refers to the increase in the rightmost part of the graph, not distinguishable in the graph due to low resolution. Remember that the same '*INCREASE*' was previously referred to in the sentence "Stocks have been climbing without fanfare since late in July, ...". Since the '*RALLY STARTING IN JULY*' becomes now a focused object, it would be appropriate for the improvement process to revise the graphical parts of the document, for example by inserting an additional 6-months-chart.

(4) In 2000 and the years leading up to it, the rally was fueled by demand for computers and telecommunications and a belief that the Internet would transform business. The rally over the last few months has had more modest roots: signs that the economy is moderating and inflation is tame.

3.2 THE CONCEPTUAL LEXICON AS BASIS OF MULTIMODAL INTEGRATION OF TEXT AND GRAPHICS

In particular from the perspective of communication, conceptual representations are the pivot of human cognition. The level of conceptual representations, which encodes meaning independent from any particular language, is the content-specifying level in language comprehension as well as in language production (in psycholinguistics this level in the production process often is called as 'preverbal messages'; cf. Levelt 1999). Furthermore, conceptual structures are an essential part of the interfaces between language and perception as well as action (cf. Jackendoff's interface architecture 1997, 2002), and additionally, conceptual representations are the material for many types of thinking and problem solving. Jackendoff (1997, 2002) uses in his framework the notion 'conceptual structure', which is—by the theoretical and empirical context of his approach—more constrained than the term 'conceptual representation', which has framework-specific variants in interpretation. In the present paper we use these terms as quasi-synonyms, since we cannot specify our framework of conceptual representations for natural language processing in detail here (see Tschander et al. (2002) on CRIL (Conceptual Route Instruction Language), an internal language connecting natural language and action plans, and Guhe et al. (2004) on the use of conceptual representations in language generation).

As we will argue for in this section, conceptual representations play a central role also in multimodal communication. In particular, the interaction between language and graphics is supported by shared conceptual representations. Since conceptual structures are independent from individual languages, the correspondence between lexemes and concepts is usually not of the one-to-one type. In other words, the conceptual counterpart to a lexeme is in most cases a complex structure of conceptual building blocks. The reciprocal assignment of lexemes to conceptual structures and vice versa is fundamental for language comprehension as well as for language production. The relevant knowledge source for these assignments is the lexicon (in Jackendoff's notion, 1997) or the lexical network (using Levelt's terminology, 1999).

In the following, we exemplify the nature of lexeme to conceptual structure relations on a coarse level with the phrase "retracing losses" discussed in Section 3.1. The noun 'loss' (and in a similar manner, the verb 'lose') 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 a value argument, which can be filled by an entity of an ordered structure, which functions as the domain of the value. By using such abstract representations, which generalize over different value domains, it is possible to catch the common properties 'loss of money', 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'. (Note: this does not mean that the producer or the recipient knows how to anchor this interval in physical time.) 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.[5] Such abstract topological and geometrical structures are relevant building blocks of conceptual representations in general, not only needed for communication about physical space, but also for types of using what often are called 'figurative language' (cf. Habel, 1990; Habel and Eschenbach, 1997; Eschenbach et al., 1998; Eschenbach et al., 2000).

Let us now look on the lexeme 'retrace', which in some dictionaries is paraphrased as "trace back or trace again". A corresponding concept representation is $\mathit{INVERSE}(\__{\mathit{PATH}})$. The syntactic and semantic analysis of 'retracing losses' specifies that *PATH* corresponding to $\mathit{DECREASE\_OF\_VALUE}(\__{\mathit{TEMP}},\__{\mathit{VALUE}},\ldots)$ is the conceptual argument for $\mathit{INVERSE}(\__{\mathit{PATH}})$. The next step in conceptualizing goes as follows. The inverse following of the value path leads to an $\mathit{INCREASE\_OF\_VALUE}(\__{\mathit{TEMP}},\__{\mathit{VALUE}},\ldots)$ structure. Since a retracing has to occur after the process that is the argument of, the temporal ordering between the two intervals in question, namely the time of the losses and the time of retracing, is inferable: *LOSS* A *RETRACE*. Except from the relation between initial value and final value during the loss-interval, namely *VALUE(BEGIN(LOSS))* > *VALUE(END(LOSS))*, the details of time–value correspondences are not mentioned explicitly in the text. But by a reasonable inference the recipient can infer the following:

$$\mathit{VALUE(BEGIN(LOSS))} \approx \mathit{VALUE(END(RETRACE))}\ \mathit{VALUE(END(LOSS))} = \mathit{VALUE(BEGIN(RETRACE))}^{6} \quad (5)$$

In Section 3.1, we proposed a multimodal co-reference constellation constituted by the phrase 'retracing losses' and a V-shaped structure in the chart. At this point of processing the conception of *graph schema* comes into the play (cf. Pinker, 1990; Lohse, 1993): graph schemata provide knowledge to locate and decode information presented in the graph. Whereas Pinker and Lohse focus on the procedural character of graph schemata—i.e. they take a perspective, kindred to Ullman's (1984) *visual routines* approach—we emphasize the abstract spatial properties of graph schemata. Two of the most relevant *gestalt atoms* for line graphs are *increasing* and *decreasing paths*, where 'path' is used as the technical term for (ordered) sequences of line segments. The corresponding entities on the level of conceptual representations are

[5] *Paths* are directed linear entities (cf. Habel, 1990; Eschenbach et al., 2000).

[6] Please remind that we do not specify the conceptual structures in detail in the present paper. Therefore, (5) should be read as 'pseudo-code' and not as formal specification of conceptual meaning.

denoted by $\mathit{INCREASE\_P}(\_{}_{\mathit{PATH}}, \_{}_{\mathit{SRS}})$ and $\mathit{DECREASE\_P}(\_{}_{\mathit{PATH}}, \_{}_{\mathit{SRS}})$ specifying the particular property for a path argument with respect to a 'spatial reference system' (*SRS*). Using an additional concatenation concept $\mathit{CONCAT}(\_{}_{\mathit{PATH}}, \_{}_{\mathit{PATH}}, \_{}_{\mathit{SRS}})$ a V-structure, built by two paths anchored in the same reference system, can be specified by

$$\text{V-STRUCTURE } (\mathit{CONCAT}(\mathit{PATH}_1, \mathit{PATH}_2, \mathit{SRS})) \Leftrightarrow_{\text{def}} \mathit{DECREASE\_P}(\mathit{PATH}_1, \mathit{SRS}) \wedge \mathit{INCREASE\_P}(\mathit{PATH}_2, \mathit{SRS}) \quad (6)$$

Furthermore, the prototypicality of V-structures is determined with respect to additional properties concerning for example the *amount* of increase and decrease, the *order of magnitudes* between these amounts, and their *steepness*. Beyond specific graph schemata, as that of *INCREASE*, *DECREASE* and V-STRUCTURE, there exist more general graph schemata, for example considering the standard *spatial reference system* of line graphs, namely co-ordinate systems with scaled axes, i.e. axes possessing an ordering structure. Exactly the ordering of the y-axes is considered by the graph schema routines, determining what an *INCREASE* and what a *DECREASE* is.

Now, we come back to the process of multimodal comprehension. The phrase 'retracing losses' introduces the internal proxies for two succeeding processes into the discourse model, whose conceptual representations contain *DECREASE_OF_VALUE* and *INCREASE_OF_VALUE* components. Having a complementary line graph in the multimodal document, the graph reader's comprehension component can start a query with respect to a V-structure, corresponding to the characterization (6), which also specifies the correspondence between the conceptual structure built up by language comprehension and the graph schema. The graph schema routines triggered by the abstract spatial description can process the actual graph in goal directed manner; in particular they can be based on using the most salient discourse entities first. In this case, the co-reference of 'new high' and of the induced 'former high' with two small circular entities in the graph should have been built up in the prior step. Thus the 'retracing losses' – graphical V-structure co-reference relation is easy to construct, and additional information can be attached to the discourse units underspecified by the textual information, for example concerning the temporal location of the 'depth', i.e. the point of maximal losses and start of retracing.

4. CONCLUSION AND FUTURE WORK

In the present paper we proposed 'reciprocal improvement' as a means to generate high-quality combinations of text and graphics. These types of revision in later stages of a multi-pass architecture are in particular relevant, if text and graphics are produced by different agents (people, institutions, systems), as in the example we discussed in Section 3.

Since both comprehension modules, which are the base of improvement, namely text comprehension and graph comprehension, employ the same type of conceptual (concept) representation, the lexical and the ontological analysis of terms which refer to 'graphical entities' (e.g., *peak, increase*, etc.) and their verbal encoding in specific domains (e.g., *rally*) are essential. We will perform this analysis via corpus studies as well as by language production experiments with human subjects.

REFERENCES

Ainley, J., Nardi, E. and Pratt, D. (2000). The construction of meanings for trend in active graphing, *International Journal of Computers for Mathematical Learning* 5(2), 85–114.

André, E. (2000). The generation of multimedia presentations. In R. Dale and H. Moisl and H. Somers (eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. (pp. 305–327). Marcel Dekker Inc.

Beun, R. J. and Cremers, A. H. M. (2001). Multimodal reference to objects: An empirical approach. In H. C. Bunt and R. J. Beun (eds.), *Cooperative Multimodal Communication*. (pp. 64–86). Berlin: Springer-Verlag.

Butterfield, E. C.; Hacker, D. J. and Albertson, L. R. (1996). Environmental, cognitive, and metacognitive influences on text revision: Assessing the evidence. *Educational Psychology Review, 8*. 239-297.

Carpenter, P. A., and Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4(2),* 75-100

Cleveland, W. S., and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 77,* 541–547.

Cleveland,W. S., and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science 229,* 828–833.

Cleveland, W. S. and McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society. Series A (General), 150 (3),* 192-229.

DeProspero, A. and Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*. 573–579.

Eschenbach, C.; Habel, Ch.; Kulik, L. and Leßmöllmann, A. (1998). Shape nouns and shape concepts: A geometry for ‚corner‘. In C. Freksa, Ch. Habel and K.F. Wender (eds.), *Spatial Cognition.* (pp. 177–201). Berlin: Springer.

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

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

Green, N.; Carenini, G.; Kerpedjiev, S.; Mattis, J.; Moore, J. and Roth, S. (2004). AutoBrief: an experimental system for the automatic generation of briefings in integrated text and information graphics. *International Journal of Human Computer Studies, 61*. 32–70.

Green, N.; Carenini, G.; Kerpedjiev, S.; Roth, S. and Moore, J.A. (1998). Media-independent content language for integrated text and graphics generation. *CVIR'98 Content Visualization and Intermedia Representations (*ACL-Coling98 workshop*).*

Guhe, M.; Habel, Ch. and Tschander, L. (2004). Incremental generation of interconnected preverbal messages. In T. Pechmann and C. Habel (eds.), *Multidisciplinary approaches to language production*. (pp. 7–52). Berlin: Mouton de Gruyter.

Habel, Ch. (1990). Propositional and depictorial representations of spatial knowledge: The case of *path* concepts. In R. Studer (ed.): *Natural language and logic*. (pp. 94–117). Lecture Notes in Artificial Intelligence. Berlin: Springer.

Habel, Ch. and Eschenbach, C. (1997). Abstract structures in spatial cognition. In Ch. Freksa, M. Jantzen and R. Valk (Eds.). Foundations of Computer Science – Potential – Theory – Cognition. (pp. 369-378). Springer: Berlin.

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

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT-Press.

Jackendoff, R. (2002). *Foundations of language. brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology, 3*. 185–226.

Kosslyn, S. (1994). Elements of graph design. New York: W.H. Freeman.

Kosslyn, S. (2006). Graph Design for the Eye and Mind. OUP.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11.* 65-99.

Leinhardt, G., Zaslavsky, O., and Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*, 1–64.

Levelt, Willem J.M. (1999). Producing spoken language: a blueprint of the speaker. In C.M. Brown and P. Hagoort (eds.), *The neurocognition of language.* (pp. 83–122). Oxford: Oxford University Press.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction, 8*, 353–388.

Peebles, D. J., and Cheng, P. C.-H. (2001). Graph-based reasoning: from task analysis to cognitive explanation. In J. D. Moore and K. Stenning (Eds.), *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 762-767). Mahwah, NJ: Lawrence Erbaum.

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

Publication Manual of the American Psychological Association (4th ed.). (1994). Washington, DC: American Psychological Association.

Reiter E. (1994) Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? *IWNLG-1994*, 163–170, Kennebunkport, ME.

Shah, P., Mayer, R. E., and Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, 91(4), 690-702.

Tabachneck-Schijf, H. J. M.; Leonardo, A. M. and Simon, H. A. (1997). CaMeRa: A Computational Model of Multiple Representations. *Cognitive Science, 21.* 305–350.

Tappe, H. and Habel, Ch. (1998). Verbalization of dynamic sketch maps: Layers of representation and their interaction. [Full version of one page abstract / poster at Cognitive Science Conference; Madison WI, August, 1.-4., 1998.] ftp://ftp.informatik.uni-hamburg.de/pub/unihh/informatik/WSV/TappeHabel_CogSci_1998.pdf

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

Ullman, S. (1984). Visual routines. *Cognition, 18.* 97–159.

Wahlster, W.; André, E.; Finkler, W.; Profitlich, H. -J. and Rist, Th. (1993). Plan-based integration of natural language and graphics generation. *Artificial Intelligence, 63.* 387–427.

Winn, W. (1991). Learning from maps and diagrams. *Educational Psychology Review*, 3, 211–247.