35. NATURAL LANGUAGE IN MULTIMEDIA/MULTIMODAL SYSTEMS

Elisabeth André

Abstract

Recent years have witnessed a rapid growth in the development of multimedia applications. Improving technology and tools enable the creation of large multimedia archives and the development of completely new styles of interaction. This chapter provides a survey of multimedia applications in which natural language plays a significant role. It addresses the following three issues: (1) How to integrate multimedia input including spoken or typed language in a synergistic manner? (2) How to combine natural language with other media in order to generate more effective output? And (3), how to make use of natural language technology in order to enable better access to multimedia archives? The chapter pleads for a generalization of techniques developed for natural language processing to help overcome some of the deficiencies of current multimedia technology.

35.1 Introduction

Multimedia applications are finding their way into nearly every area of our daily life, such as education, entertainment, business and transport. A walk through any computer fair shows that many manufacturers have already enriched their product lines with multimedia technology. The place of natural language as one of the most important means of communication makes natural language technologies integral parts of multimedia interfaces. An apparent advantage of natural language is its great expressive power. Imagine, for instance, the difficulties I would encounter if I had to provide this survey relying entirely on non-verbal media. Being one of the most familiar means of human interaction, natural language can significantly reduce the training effort required to enable communication with machines. On the other hand, the coverage of current natural language dialogue systems is still strongly limited. This fact is aggravated by the lack of robust speech recognizers. The integration of non-verbal media often improves the usability

and acceptability of natural-language components as they can help compensate for the deficiencies of current natural language technology. In fact, recent empirical studies by Oviatt (1999) show that properly designed multimedia systems have a higher degree of stability and robustness than those that are based on speech input only. From a linguistic point of view, multimedia systems are interesting because communication by language is a specialized form of communication in general. Theories of natural language processing have reached sufficiently high levels of maturity so that it is now time to investigate how they can be applied to other media, such as graphics, or pointing gestures.

The objective of this chapter is to investigate the use of natural language in the context of a multimedia environment. To start with, we shall clarify the basic terminology. The terms **medium** and **modality** especially, have been a constant cause of confusion due to the fact that they are used differently in various disciplines. In this paper, we adopt Maybury's distinction between **medium**, **mode** and **code** (see Maybury 1999). The term **mode** or **modality** is used to refer to different kinds of perceptible entities (e.g., visual, auditory, haptic, and olfactory) while the term **medium** relates to the carrier of information (e.g. paper or CD-ROM), different kinds of physical devices (e.g., screens, loudspeakers, microphones and printers) and information types (e.g., graphics, text and video). Finally, the term **code** refers to the particular means of encoding information (e.g., sign languages and pictorial languages).

Multimedia/multimodal systems are then systems that are able to *analyze* and/or *generate* multimedia/multimodal information or provide support in *accessing* digital libraries of multiple media.

Multimodal input analysis starts from low-level sensing of the single modes relying on interaction devices, such as speech and gesture recognizers and eye trackers. The next step is the transformation of sensory data into representation formats of a higher level of abstraction. In order to exploit the full potential of multiple input modes, input analysis should not handle the single modes independent of each other, but fuse them into a common representation format that supports the resolution of ambiguities and accounts for the compensation of errors. This process is called **modality integration**

or **input fusion**. Conventional multimodal systems usually do not maintain explicit representations of the user's input and handle mode integration only in a rudimentary manner. In Section 35.2, we will show how the generalization of techniques and representation formalisms developed for the analysis of natural language can help to overcome some of these problems.

Multimedia generation refers to the activity of producing output in different media. It can be decomposed into the following sub tasks: the **selection** and **organization of information**, the **allocation of media** and **content-specific media encoding**. As we do not obtain coherent presentations by simply merging verbalization and visualization results into multimedia output, the generated media objects have to be tailored to each other in such a way that they complement each other in a synergistic manner. This process is called **media coordination**. While the automatic production of material is rarely addressed in the multimedia community, a considerable amount of research effort has been directed towards the automatic generation of natural language. Section 35.3 surveys techniques for building automated multimedia presentation systems drawing upon lessons learned during the development of natural language generators.

Multimedia access to digital data is facilitated by methods for document classification and analysis, techniques to condense and aggregate the retrieved information as well as appropriate user interfaces to support search tasks. Most contemporary multimedia retrieval systems do not aim at a deeper analysis of the underlying information, but restrict themselves to classifying and segmenting static images and videos. In Section 35.4, we argue that the integration of natural language technology can lead to a qualitative improvement of existing methods for document classification and analysis.

35.2 Analysis of multimodal/multimedia input including language

Based on the observation that human-human communication is multimodal, a number of researchers have investigated the usage of multiple input devices for man-machine interaction. The first systems in this area accept written or spoken natural language input in combination with pointing gestures. Examples include: "Put-That-There" (Bolt

1980) and CUBRICON (Neal and Shapiro 1991) that operate on maps, and XTRA (Allgayer et al. 1989), an expert system for tax forms. One of the limitations of these early systems is the fact that they only allow for a limited set of deictic gestures in combination with language. In contrast, Koons and colleagues (1993) developed two prototype systems which are able to analyze simultaneous input from hand gestures, gaze and speech. The first system allows for interaction with a two-dimensional map. The second system enables users to manipulate objects in a 3-dimensional blocks world and includes not only deictic, but also iconic and pantomimic gestures. In QuickSet (Cohen et al. 1997), the user interacts with a map by drawing directly on the map displayed on a wireless hand-held and simultaneously uttering commands via speech. Drawing gestures in QuickSet include: map symbols, editing gestures and spatial features (see Fig. 35.1). Similar interaction styles are supported by the Map-based Tourist Information System described in (Cheyer and Julia 1995).

Figure 35.1: QuickSet User Interface (Figure Used with Permission of OGI)

35.2.1 Natural language technology as a basis for multimodal analysis

Most systems rely on different components for the low-level analysis of the single modes, such as eye trackers, speech and gesture recognizers, and make use of one or several mode integrators to come up with a comprehensive interpretation of the multimodal input. This approach raises two questions: How should the results of low-level analysis be represented in order to support the integration of the single modalities? How far should we process one input stream before integrating the results of other modality

analysis processes? On the one hand, it does not make sense to merge gesture data and voice data on the level of pixels and phonemes. On the other hand, ambiguities of one modality should be resolved as early as possible by considering accompanying modalities.

CUBRICON and XTRA rely on parsers that have originally been developed for the analysis of natural language: CUBRICON uses an Augmented Transition Network (ATN) parser, XTRA a unification-based chart parser. To employ such parsers for the analysis of pointing gestures accompanied by natural language, additional grammatical categories, such as "deictic", have been introduced. However, both parsers just map deictic gestures onto the corresponding categories, and additional gesture analyzers are necessary in order to get to a more fine-grained representation of the non-verbal input. Since no uniform grammar formalism is used for the fusion of modalities, the integration of additional modes is relatively difficult.

In contrast to this, Koons and colleagues propose frames as a uniform representation format for speech, gestures and gaze. For natural language input, a parse tree is created first and then transformed into a system of connected frames that represent the categories and properties of the single tokens as well as timing information. Frames with timing information are also created for eye and hand motion. Extracted features for motions of the eye include: fixations, saccades and blinks while gestures are characterized by: posture, orientation and motion. Even though Koons and colleagues present a uniform representation formalism, their approach is still lacking of a declarative method for modality integration.

Johnston and colleagues (1997) propose an approach to modality integration for the QuickSet system that is based on unification over typed feature structures. The basic idea is to build up a common semantic representation of the multimodal input by unifying feature structures which represent the semantic contributions of the single modalities. For instance, the system might derive a partial interpretation for a spoken natural language reference which indicates that the location of the referent is of type "point". In this case, only unification with gestures of type "point" will succeed. Their approach also allows for the compensation of errors in speech recognition by gesture

and vice versa (see Oviatt 1999). For instance, QuickSet may select speech recognition alternatives on the basis of gesture recognition and discard wrong hypotheses.

One limitation of the original unification-based approach as described in (Johnston et al. 1997) lies in the fact that it can only handle combinations of single spoken phrases with single gestures. In order to account for a broader range of multimodal combinations, Johnston (1998) later combines the unification-based approach with a generalized chart parser that enables the integration of multiple elements which are distributed across two or three spatial dimensions, a temporal and an acoustic dimension.

35.2.2 Integration of modalities

In the ideal case, multimodal systems should not just accept input in multiple modalities, but also support a large variety of mode combinations. This requires sophisticated methods for modality integration.

An important prerequisite for modality integration is the explicit representation of the multimodal context. For instance, the interpretation of a pointing gesture often depends on the syntax and semantics of the accompanying natural language utterance. In "Is this number <pointing gesture> correct?", only referents of type "number" can be considered as candidates for the pointing gesture. The (semantic) case frame indicated by the main verb of a sentence is another source of information that can be used to disambiguate a referent since it usually provides constraints on the fillers of the frame slots. For instance, in "Can I add my travel expenses here <pointing gesture>?", the semantics of *add* requires a field in a form where the user can input information.

A fundamental problem of most systems is that there is no declarative formalism for the formulation of integration constraints. A noteworthy exception is the approach used in QuickSet which clearly separates the statements of the multimedia grammar from the mechanisms of parsing (cf. Johnston 1998). This approach enables not only the declarative formulation of type constraints, such as "the location of a flood zone should be an area", but also the specification of spatial and temporal constraints, such as "two regions should be a limited distance apart" and "the time of speech must either overlap with or start within four seconds of the time of the gesture". The basis for the temporal

constraints are empirical studies by Oviatt and colleagues (1997).

35.3 Generation of multimedia output including language

In many situations, information is only presented efficiently through a particular media combination. Multimedia presentation systems take advantage of both the individual strength of each media and the fact that several media can be employed in parallel. Most systems combine spoken or written language with static or dynamic graphics, including bar charts and tables, such as SAGE (Roth et al. 1991) and MAGIC (Dalal et al. 1996), maps, such as AIMI (Maybury 1991) and CUBRICON (Neal and Shapiro 1991) and depictions of three-dimensional objects, such as COMET (Feiner and McKeown 1991), WIP (Wahlster et al. 1993) and PPP (André et al. 1999). There are also systems which integrate natural language with hypertext. ALFRESCO (Stock et al. 1997) generates text with entry points to an underlying preexisting hypermedia network while PEA (Moore and Swartout 1990), ILEX (Knott et al. 1996) and PEBA-II (Dale and Milosavljevic 1996) make use of an hypertext-style interface to present the generated text.

35.3.1 Natural language technology as a basis for multimedia generation

Encouraged by progress achieved in Natural Language Generation (see Chapter 15 for an introductory overview), several researchers had tried to generalize the underlying concepts and methods in such a way that they can be used in the broader context of multimedia generation.

A number of multimedia presentation systems make use of a notion of schemata (see Chapter 15) based on the original proposal by McKeown (1985) for text generation. Schemata describe standard patterns of discourse by means of rhetorical predicates which reflect the relationships between the parts of a multimedia presentation. Examples of systems using a schema-based approach are COMET (Feiner and McKeown 1991) and an earlier prototype of SAGE (Roth et al. 1991). SAGE only relies on schemata to organize the textual parts of a multimedia presentation which makes the handling of

structural dependencies between different media more difficult. Unlike SAGE, COMET employs schemata to determine the contents and the structure of the overall presentation. The result of this process is forwarded to a media coordinator which determines which generator should encode the selected information.

In the last few years, operator-based approaches similar to those used for text generation have become increasingly popular for multimedia presentation. Examples include AIMI (Maybury 1991), MAGIC (Dalal et al. 1996), WIP (Wahlster et al. 1993), PPP (André and Rist 1996), and a recent extension of SAGE (Kerpedjiev et al. 1997). The main idea behind these systems is to generalize communicative acts to multimedia acts and to formalize them as operators of a planning system. Starting from a presentation goal, the planner looks for operators whose effect subsumes the goal. If such an operator is found, all expressions in the body of the operator will be set up as new subgoals. The planning process terminates if all subgoals have been expanded to elementary generation tasks which are forwarded to the medium-specific generators. The result of the planning process is hierarchically organized graph that reflects the discourse structure of the multimedia material (see Fig. 35.2 for a presentation plan generated by the WIP system).

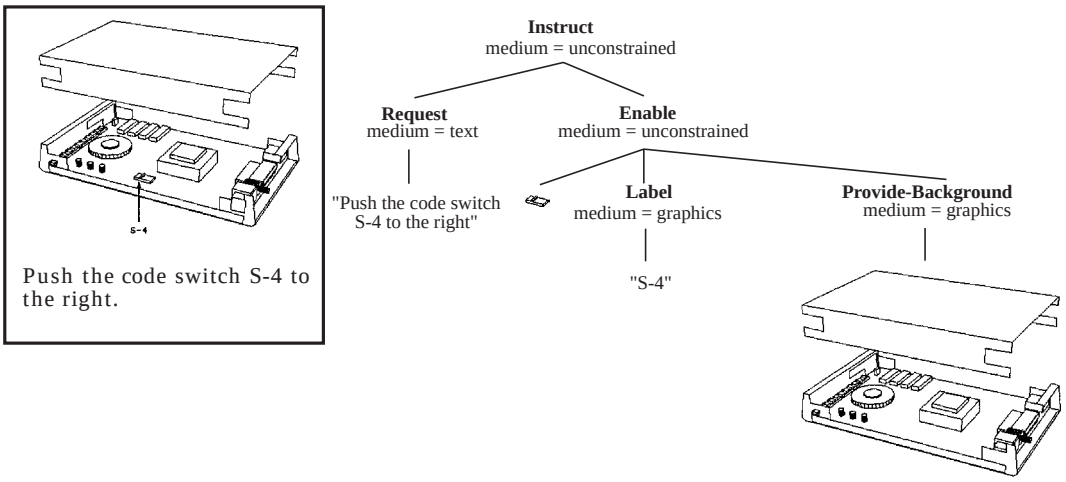

Figure 35.2: Sample Document Generated by WIP and Corresponding Discourse Structure

An advantage of an operator-based approach is that additional information concerning media selection or the scheduling of a presentation can be easily incorporated and

propagated during the content selection process. This method facilitates the handling of dependencies, as medium selection can take place during content selection and not only afterwards, as is the case in COMET (cf. André and Rist 1995).

Operator-based approaches have also proven useful for the automated generation of hypertext. The basic idea is to refine certain parts of the discourse structure only on demand, namely if the user clicks on the corresponding hyperlink. Such an approach was used, e.g., in PEA (Moore and Swartout 1990).

35.3.2 Media coordination

Multimedia presentation design involves more than just merging output in different media; it also requires a fine-grained coordination of different media. This includes distributing information onto different generators, tailoring the generation results to each other, and integrating them into a multimedia output.

35.3.2.1 Media allocation

The media allocation problem can be characterized as follows: Given a set of data and a set of media, find a media combination which conveys all data effectively in a given situation. Earlier approaches rely on a classification of the input data, and map information types and communicative functions onto media classes by applying media allocation rules. Examples of media allocation rules are as follow (see André and Rist 1993):

1. Prefer graphics over text for spatial information (e.g., location, orientation, composition) unless accuracy is preferred over speed, in which case text is preferred.

2. Use text for quantitative information (such as most, some, any, exactly, and so on)

3. Present objects that are contrasted with each other in the same medium.

This approach can be generalized by mapping features of input data to features of media (e.g. static–dynamic, arbitrary–non-arbitrary). An example of such a mapping rule is:

Data tuples, such as locations, are presented on planar media, such as graphs, tables, and maps (cf. Arens et al. 1993).

However, since media allocation depends not only on data and media features, media allocation rules have to incorporate context information as well. Arens and his colleagues (Arens et al. 1993) proposed representing all knowledge relevant to the media allocation process in And-Or- Networks like those used by Systemic Functional linguists (see also Chapter 15) to represent grammars of various language in a uniform formalism. Presentations are designed by traversing the networks and collecting at each node features which instruct the generation modules how to build sentences, construct diagrams, and so on.

35.3.2.2 Tailoring output in different media to each other

To ensure the consistency and coherency of a multimedia document, the media-specific generators have to tailor their results to each other. An effective means of establishing coferential links between different media is the generation of **cross-media referring expressions** that refer to document parts in other presentation media (cf. André and Rist 1994). Examples of cross-media referring expressions are "the upper left corner of the picture" or "Fig. x". Another important task is the coordination of picture and sentence breaks (see Feiner and McKeown 1991). Constraints from graphics have to be considered when determining sentence size and the other way round. If we include too many objects in one graphics, the individual objects may be rendered too small to be recognizable. On the other hand, shortening a sentence may lead to ungrammatical text if obligatory case roles are left out. To support media coordination, the systems WIP and COMET rely on a common data structure which explicitly represents the design decisions of the single generators and allows for communication between them.

35.3.2.3 Spatial and temporal coordination of the output

Media coordination is also needed when integrating the individual generator results into a multimedia output. This includes the spatial arrangement of text blocks and graphics by means of a layout component. A purely geometrical treatment of the layout task

would, however, lead to unsatisfactory results. Rather, layout has to be considered as an important carrier of meaning. To account for this, the WIP system maps coherence relations between presentation parts (such as sequence or contrast) onto geometrical and topological constraints (e.g., horizontal and vertical layout, alignment, and symmetry) and uses a finite domain constraint solver to determine an arrangement that is consistent with the structure of the underlying information (cf. Graf 1992).

If information is presented over time, layout design also includes the temporal coordination of output units. The PPP system generates a temporal schedule from a complex presentation goal to synchronize speech with the display of graphical elements and annotation labels. Basically, PPP relies on the WIP approach for presentation planning. However, in order to enable both the creation of multimedia objects and the generation of scripts for presenting the material to the user, the following extensions have become necessary (cf. André and Rist 1996): (1) the specification of qualitative and quantitative temporal constraints in the operators and (2) the development of a mechanism for building up presentation schedules. A similar mechanism is used in MAGIC that synchronizes spoken references to visual material with graphical highlighting.

35.4 Language processing for accessing multimedia data

Rapid progress in technology for the creation, processing and storage of multimedia documents has opened up completely new possibilities for building up large multimedia archives. While the necessary infrastructure is already in place, we still need tools for making information accessible to users in a beneficial way. Methods for natural processing facilitate the access to multimedia information in at least three ways: (1) Information can often be retrieved more easily from the audio or closed caption streams (2) natural language access to visual data is often much more convenient since it allows for a more efficient formulation of queries; and (3) natural language provides a good means of condensing and summarizing visual information.

35.4.1 NL-based video/image retrieval

Whereas it is still not feasible to analyze arbitrary visual data, a great deal of progress has been made in the analysis of spoken and written language. Based on the observation that a lot of information is encoded redundantly, a number of research projects rely on the linguistic sources (e.g., transcribed speech or closed captions) when analyzing image/video material. Unlike the approaches discussed in Section 35.2, systems for NL-based video/image retrieval do not aim at a complete syntactic and semantic analysis of the underlying information. Instead, they usually restrict themselves to tasks, such as image/video classification and video segmentation, employing standard techniques for shallow natural language processing, such as text-based information retrieval (see Chapter 29 for an introduction) and extraction (for an introduction, we refer to (Appelt 1999) and Chapter 30).

Sable and Hatzivassiloglou (1999) show how information found in associated text sources can be used for effectively classifying photographs as indoor or outdoor. Their classifier is based on information retrieval measures of text similarity which they adapted to their particular task by evaluating several variants of the standard approach, such as limiting their analysis to targeted parts of the surrounding text or certain word classes. They could show that their text-based classification methods clearly outperform competing image-based approaches and nearly approach human accuracy.

Jones et al. (1997) combine speech recognition with information retrieval in order to analyze video mail. Since most images in their application just consist of "talking heads", they exclusively focus on the linguistic channel. Even though the authors had to cope with a number of problems that do not exist in text-based retrieval systems, such as the unreliability of the available speech recognition technology, they could achieve a retrieval performance between 75% and 95% of the performance that can be achieved with transcribed text depending on the generality of the underlying language model.

Whereas the approaches mentioned above focus on the task of image/video classification and do not aim at a deeper analysis of the visual material, the Broadcast News Navigator (BNN) developed by MITRE performs a segmentation of videos into

different topics (see Merlino et al. 1997). Besides video and audio cues, such as scene and speaker changes, the system looks for discourse cues embedded in a broadcast's transcribed speech or closed caption streams. By making use of information extraction methods, such as token detection and named entity tagging, the system can recognize typical language patterns that indicate topic shifts, such as explicitly stated reporter welcomes. For instance, it tries to detect names of persons, organizations and locations which frequently occur in speaker introductions, such as "This is Britt Hume, CNN, Washington".

35.4.2 NL access to image and video databases

Direct manipulation interfaces often require the user to access objects by a series of mouse operations. Even if the user knows the location of the objects he or she is looking for, this process may still cost a lot of time and effort. Natural language supports direct access to information and enables the efficient formulation of queries by using simple keywords or free form text.

The vast majority of information systems allows the user to input some natural language keywords that refer to the contents of an image or a video. Such keywords may specify a subject matter, such as "sports" (cf. Aho et al. 1997), but also subjective impressions, such as "buzzing sound" (cf. Blum et al. 1997) or "romantic image" (cf. Kato 1992). The Informedia system (Hauptmann and Witbrock 1997) also accepts free form typed or spoken natural language, but relies on similar methods for query processing as the other systems. It simply eliminates all stop-words from the user's query and starts an index-based search.

The ALFRESCO system (Stock et al. 1997) combines the benefits of hypermedia-style interaction and natural language queries to provide the user access to a database of frescoes. An interesting feature of the system is that it enables the user to continuously shift between browsing and querying. For instance, if the user asks: "Tell me something about Amborgio Lorenzetti", the system comes up with a generated text that contains links to a hypertext. The user is then free to browse through the hypertext or to ask natural language follow-up questions.

Even though natural language offers a number of advantages for query formulation, it might be hard in some cases to specify heterogeneous multimedia information exclusively via verbal means. Therefore, some systems are investigating additional means of accessing information, such as visual or example-based querying.

35.4.3 NL summaries of multimedia information

One major problem associated with visual data is information overload. Natural language has the advantage that it permits the condensation of visual data at various levels of detail according to the application-specific demands. Indeed, a number of experiments performed by Merlino and Maybury showed that reducing the amount of information (e.g. presenting users just with an one-line summary of a video) significantly reduces performance time in information seeking tasks, but leads to nearly the same accuracy (cf. Merlino and Maybury 1999).

The Columbia Digital News System (CDNS, Aho et al. 1997) provides summaries over multiple news articles by combining methods for text-based information extraction and text generation (see McKeown et al. 1998). The basic idea is to use an information extraction system to deliver template representations of the events mentioned in the articles. These templates are then transformed into natural language using various summarization techniques, such as the merge of templates or the aggregation of linguistic phrases (see Chapter 31 for a general overview of summarization techniques). The system does not summarize images, but makes use of image classification tools to select a representative sample of retrieved images that are relevant to the generated summary.

While BNN and CDNS only partially analyze image or video material and assume the existence of linguistic channels, systems, such as VITRA-SOCCER, start from visual information and transform it into natural language (cf. Herzog and Wazinski 1994). Here, the basic idea is to construct a symbolic scene representation from a sequence of video images which is then transformed into conceptual units of a higher level of abstraction. This process includes the explicit representation of spatial configurations by means of spatial relations, the interpretation of objects movements, and even the automatic recognition of presumed goals and plans of the observed agents. A similar

approach has been taken to generate natural language descriptions for games of the RoboCup simulator league (cf. André et al. 2000).

35.5 Language in inhabited multimedia environments

Recent years have seen a trend towards more intuitive interfaces which go far beyond the traditional user interfaces in which people interact via keyboard or computer mouse. Animated characters play an important role in such interfaces since they allow for the emulation of communication styles common in human-human communication (see Elliott and Brzezinski 1998 for an overview). Examples include lifelike characters that guide users through media spaces, tutor agents in pedagogical environments, and embodied personalized information agents.

Fig. 35.3 shows a conversational agent currently under development at DFKI GmbH in the role of a receptionist. Cyberella^{TM} runs on an info-terminal within the DFKI entrance. Her task is to welcome visitors, business partners, and students to DFKI and to answer questions on a wide range of dialogue topics covering news, research projects, and people within DFKI.

Figure 35.3: Conversational Agent in the Role of a Receptionist

Bringing a character like Cyberella to life, is not just a challenge from the point of view of computer graphics and animation. To come across as socially believable, char-

acters need to be realized as unique individuals with their own personality. According to their functional role in a dialogue, they must be able to exhibit a variety of conversational behaviors. In particular, they have to execute gestures that express emotions (e.g., happiness or anger), convey the communicative function of an utterance (e.g. warning the user by lifting its index finger), support referential acts (e.g., look at object and point at it), regulate the interaction between the character and the user (e.g., establishing eye contact with the user during communication), and articulate what is being said.

For research on natural language processing, this has the following consequences. First of all, the generation of language should not only be driven by the goal of information delivery, but also be influenced by social and psychological factors. In order to support the full bandwidth of human-human communication, conversational models have to encompass not only speech, but also intonation, facial expressions, gaze and body gestures.

From a technical point of view, it makes no difference whether we plan presentation scripts for the display of static and dynamic media, or communicative acts to be executed by animated characters. Basically, we can rely on one of the temporal planners presented in Section 35.3.2.3 if we extend the repertoire of plan operators by including operators which control a characters' conversational behavior. Such an approach has been used in the PPP system to determine high-level presentation acts for an animated presenter, the so-called PPP Persona. The planning approach also allows us to incorporate models of character's personality and emotions by treating them as an additional filter during the selection and instantiation of plan operators. For instance, we may define specific plan operators for characters of a certain personality and formulate constraints which restrict their applicability.

While planning techniques have proven to be useful for the specification of high-level conversational acts, the generation of immediate reactions and smooth animation sequences requires a method which is computationally less expensive. One solution is to precompile declarative behavior specifications into finite-state machines (cf. Ball et al. 1997), which are also a suitable mechanism for synchronizing character behaviors. For

instance, Cassell and colleagues (Cassell et al. 1994) use the so-called parallel transition networks (PaT-Nets) to encode facial and gestural coordination rules as simultaneously executing finite-state automata.

One of the most important communication channels is a character's face. To describe possible facial motions performable on a face, most systems rely on the facial action coding system (FACS, Ekman and Friesen 1978) or MPEG-4 facial animation parameters (FAPs). For instance, both Nagao and Takeuchi (Nagao and Takeuchi 1994) and Cassell and colleagues (1994) map FACS actions, such as Inner Brow Raiser, onto communicative functions, such as Punctuation, Question, Thinking or Agreement.

The believability of a lifelike character depends on the quality of the output speech. Unfortunately, most lifelike characters today simply use the default intonation of a speech synthesizer (see Chapter 17 for an introduction to speech synthesis). The integration of natural language generation and speech synthesis technology offers a great potential for improvement, since a natural language generator may provide the knowledge of an utterance's form and structure that a speech synthesizer needs to in order to produce good output. Prevost (1996) provides an example of such an approach: this work uses a Categorial Grammar to translate lexicalized logical forms into strings of words with intonational markings. Another noteworthy approach to speech synthesis is that adopted by Cahn (1990), which also addresses the affective impact of an utterance.

35.6 Conclusion

Multimedia systems pose significant challenges for natural language processing, which focuses on the analysis or generation of one input or output medium only. A key observation of this chapter is that methods for natural language processing may be extended in such a way that they become useful for the broader context of multimedia as well. While unification-based grammars have proven useful for media orchestration and analysis, text planning methods have successfully been applied to multimedia content selection and structuring. Work done in the area of multimedia information retrieval demonstrates that the integration of natural language methods, such as named entity recognition, en-

ables a deeper analysis of the underlying multimedia information and thus leads to better search results.

The evolution of multimedia systems is evidence of the trend away from procedural approaches towards more declarative approaches, which maintain explicit representations of the syntax and semantics of multimedia input and output. While earlier systems make use of separate components for processing multiple media and are only able to integrate and coordinate media to a limited extent, more recent approaches are based on a unified view of language and rely on a common representation formalism for the single media. Unfortunately, a generic framework providing a bi-directional view of multimedia communication is currently not yet available. There has been an international initiative towards the development of a standard reference architecture for multimedia systems, however, it has been focusing on generation aspects only (cf. Bordegoni et al. 1997).

Further reading and relevant resources

The original versions of many papers that have been discussed here can be found in (Maybury 1993), (Maybury 1997) and (Maybury and Wahlster 1998). A more detailed overview on multimedia generation systems with a special focus on natural language is provided by (André 2000). (Cassell et al. 2000) contains a comprehensive collection of papers on embodied conversational agents. I also recommend the following Special Issues: (Mc Kevitt 1994-1996), (Oviatt and Wahlster 1997), (Rist 1998) and (André 1999a). A useful web site with conference announcements and downloadable information sources is the electronic colloquium on Intelligent User Interfaces of the Electronic Transactions of Articificial Intelligence (ETAI) (http://www.dfki.de/etai/colloqb.html).

Glossary

- **Medium**

  carrier of information (e.g. paper or CD-ROM), physical device (e.g., screens, loudspeakers, microphones and printers) or information type (e.g., graphics, text and video).

- **Mode** or **Modality**

  perceptible entity (e.g., visual, auditory, haptic, and olfactory).

- **Multimedia/Multimodal Systems**

  systems that are able to *analyze* and/or *generate* multimedia/multimodal information or provide support in *accessing* digital libraries of multiple media.

- **Modality Integration** or **Input Fusion**

  process of transforming input in different modalities into a common representation format

- **Media Coordination**

  process of tailoring different media to each other during the generation process, includes subtasks such as media allocation, the generation of cross-media references, and the determination of the spatial and temporal layout

- **Cross-media References**

  references from one medium to document parts in other presentation media, such as "the upper left corner of the picture" or "Fig. x".

Biography

Elisabeth André is a principal researcher at the department of Intelligent User Interfaces at DFKI GmbH in Saarbrücken, Germany where she has been leading various academic and industrial projects on multimedia communication and lifelike characters. She chairs the ACL Special Interest Group on Multimedia Language Processing (SIGMEDIA). She is on the Editorial Board of Artificial Intelligence Communications (AICOM), Universal Access in the Information Society (UAIS), and Cognitive Processing (International Quarterly of Cognitive Science), and she is the Area Editor for Intelligent User Interfaces of the Electronic Transactions of Artificial Intelligence (ETAI).

35.7 Acknowledgments

This work has been supported by the BMBF under the contract 9701 0. I would like to thank Hamish Cunningham, Thierry Declerck, Michael Johnston, Peter Makumbi, Ruslan Mitkov and Thomas Rist for their valuable comments and corrections of an earlier draft of this chapter.

References

Aho, A.V., S. Chang, K. McKeown, D.R. Radev, J.R. Smith, and K.A. Zaman. 1997. "Columbia Digital News System: An Environment for Briefing and Search over Multimedia Information". *International Journal of Digital Libraries*, 1(4), 377–385.

Allgayer, J., K. Harbusch, A. Kobsa, C. Reddig, N. Reithinger, and D. Schmauks. 1989. "XTRA: A Natural-Language Access System to Expert Systems". *International Journal of Man-Machine Studies*, 31, 161–195.

André, E., W. Finkler, W. Graf, T. Rist, A. Schauder, and W. Wahlster. 1993. "WIP: The Automatic Synthesis of Multimodal Presentations". *Intelligent Multimedia Interfaces* ed. by M. T. Maybury, 75–93. Menlo Park: AAAI Press.

André, E. and T. Rist. 1993. "The Design of Illustrated Documents as a Planning Task". *Intelligent Multimedia Interfaces* ed. by M. T. Maybury, 94–116. Menlo Park: AAAI Press.

André, E. and T. Rist. 1994. "Referring to World Objects with Text and Pictures". *Proceedings of the 15th International Conference on Computational Linguistics (Coling '94)*, 530–534. Kyoto, Japan.

André, E. and T. Rist. 1995. "Generating Coherent Presentations Employing Textual and Visual Material". *Artificial Intelligence Review, Special Volume on the Integration of Natural Language and Vision Processing*, 9(2-3), 147–165.

André, E. and T. Rist. 1996. "Coping with temporal constraints in multimedia presentation planning". *Proceedings of the 13th National Conference of the American Association on Artificial Intelligence (AAAI-96)*, Volume 1, 142–147, Portland, Ore-

gon.

André, E. (Editor). 1999a. *Applied Artificial Intelligence Journal, Special Double Issue on Animated Interface Agents*, Vol. 13, No. 4-5.

André, E., T. Rist, and J. Müller. 1999. "Employing AI Methods to Control the Behavior of Animated Interface Agents". *Applied Artificial Intelligence Journal*, 13(4-5), 415–448.

André, E. 2000. "The Generation of Multimedia Presentations". *A Handbook of Natural Language Processing: techniques and applications for the processing of language as text* ed. by R. Dale, H. Moisl, and H. Somers, 305–327, New York: Marcel Dekker Inc.

André, E., K. Binsted, K. Tanaka-Ishii, S. Luke, G. Herzog, and T. Rist. 2000. "Three RoboCup Simulation League Commentator Systems". *AI Magazine*. 21(1):57-66.

Appelt, D. 1999. 1999. "An Introduction to Information Extraction". *AI Communications*. 12(3):161-172.

Arens, Y., E. Hovy, and M. Vossers. 1993. "Describing the Presentational Knowledge Underlying Multimedia Instruction Manuals". *Intelligent Multimedia Interfaces* ed. by M. T. Maybury, 280–306. Menlo Park: AAAI Press.

Ball, G., D. Ling, D. Kurlander, J. Miller, D. Pugh, T. Skelly, A. Stankosky, D. Thiel, M. van Dantzich, and T. Wax. 1997. "Lifelike Computer Characters: The Persona Project at Microsoft". *Software Agents* ed. by J. M. Bradshaw, 191–222. Menlo Park: AAAI/MIT Press.

Blum, T., D. Keislaer, J. Wheaton, and E. Wold. 1997. "Audio Databases with Content-Based Retrieval". *Intelligent Multimedia Information Retrieval*, ed. by M. T. Maybury, 113–135. Menlo Park: AAAI Press.

Bolt, R.A. 1980. "Put-That-There: Voice and Gesture at the Graphics Interface". *Computer Graphics*, 14, 262–270.

Bordegoni, M., G. Faconti, S. Feiner, M. T. Maybury, T. Rist, S. Ruggieri, P. Trahanias, and M. Wilson. 1997. "A Standard Reference Model for Intelligent Multimedia Pre-

sentation Systems". *Computer Standards and Interfaces: The International Journal on the Development and Application of Standards for Computers, Data Communications and Interfaces*, 18(6-7), 477–496.

Cahn, J. 1990. "The Generation of Affect in Synthesized Speech". *Journal of the American Voice I/O Society*, 8, 1–19.

Cassell, J., C. Pelachaud, N.I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. 1994. Animated conversation: Rule-based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. In *Proceedings of Siggraph'94*, Orlando, 1994.

Cassell, J., J. Sullivan, S. Prevost and E. Churchill. 2000. *Embodied Conversational Agents*. Cambridge, MA: The MIT Press.

Cheyer, A. and L. Julia. 1995. "Multimodal maps: An agent based approach". *Proceedings of the International Conference on Cooperative Multimodal Communication*, 103–114, Eindhoven, The Netherlands.

Cohen, P.R., M. Johnston, D. McGee, and S. Oviatt. 1997. "Quickset: Multimodal Interaction for Distributed Applications". *Proceedings of the 5th International Multimedia Conference (Multimedia 97)*, 31–40. ACM Press.

Dalal, M., S. Feiner, K. McKeown, S. Pan, M. Zhou, T. Höllerer, J. Shaw, Y. Feng, and J. Fromer. 1996. "Negotiation for Automated Generation of Temporal Multimedia Presentations". *Proceedings of the 4th International Multimedia Conference (Multimedia 96)*, 55–64. ACM Press.

Dale, R. and M. Milosavljevic. 1996. "Authoring on Demand: Natural Language Generation in Hypermedia Documents". *Proceedings of the First Australian Document Computing Symposium (ADCS'96)*, 20–21, Melbourne, Australia.

Ekman, P. and W.V. Friesen. 1978. *Facial Action Coding*. Consulting Psychologists Press Inc.

Elliott, C. and J. Brzezinski. 1998. "Autonomous Agents as Synthetic Characters". *AI Magazine*, 19(2), 13–30.

Feiner, S.K. and K.R. McKeown. 1991. "Automating the Generation of Coordinated Multimedia Explanations". *IEEE Computer*, 24(10), 33–41.

Graf, W. 1992. "Constraint-Based Graphical Layout of Multimodal Presentations". *Advanced Visual Interfaces (Proceedings of AVI '92, Rome, Italy)* ed. by M. F. Costabile, T. Catarci, and S. Levialdi, 365–385. Singapore: World Scientific Press.

Hauptmann, A.G. and M.J. Witbrock. 1997. "Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval". *Intelligent Multimedia Information Retrieval* ed. by M. T. Maybury, 214–239. Menlo Park: AAAI Press.

Herzog, G. and P. Wazinski. 1994. "Visual translator: Linking perceptions and natural language descriptions". *Artificial Intelligence Review, Special Volume on the Integration of Natural Language and Vision Processing*, 8(2-3), 175–187.

Johnston, M. 1998. "Unification-Based Multimodal Parsing". *Proceedings of the International Conference on Computational Linguistics and the 36 th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-98)*, 624–630, Montreal, Canada.

Johnston, M., P.R. Cohen, D. McGee, S.L. Oviatt, J.A. Pittman, and I. Smith. 1997. "Unification-Based Multimodal Integration". *Proceedings of the International Conference on Computational Linguistics and the 35 th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-97)*, 281–288, Madrid, Spain.

Jones, G., J. Foote, K. Spärck Jones, and S.J. Young. 1997. "The Video Mail Retrieval Project: Experiences in Retrieving Spoken Documents". *Intelligent Multimedia Information Retrieval* ed. by M. T. Maybury, 191–214. Menlo Park: AAAI Press.

Kato, T. 1992. "Cognitive view mechanism for content-based multimedia information retrieval". *Proceedings of the First International Workshop on Interfaces to Database Systems (IDS92)* ed. by R. Cooper. New York: Springer.

Kerpedjiev, S., G. Carenini, S.F. Roth, and J.D. Moore. 1997. "Integrating Planning and Task-Based Design for Multimedia Presentation". *Proceedings of the 1997 International Conference on Intelligent User Interfaces*, 145–152, Orlando, Florida.

Knott, A., C. Mellish, J. Oberlander, and M. O'Donnell. 1996. "Sources of Flexibility in Dynamic Hypertext Generation". *Proceedings of the 8th International Workshop on Natural Language Generation*, Sussex, England.

Koons, D.B., C.J. Sparrell, and K.R. Thorisson. 1993. "Integrating Simultaneous Input form Speech, Gaze, and Hand Gestures". *Intelligent Multimedia Interfaces* ed. by M. T. Maybury, 257–276. Menlo Park: AAAI Press.

Maybury, M. T. 1999. "Multimedia Interaction for the New Millenium". *Proceedings of Eurospeech 99*, 357-364, Budapest, Hungary.

M. Maybury (Editor). 1993. *Intelligent Multimedia Interfaces.* Menlo Park: AAAI Press: The MIT Press.

Maybury, M. T. (Editor). 1997. *Intelligent Multimedia Information Retrieval.* Menlo Park: AAAI Press/The MIT Press.

Maybury, M. T. 1991. "Planning Multimedia Explanations Using Communicative Acts". *Proceedings of the 8th National Conference of the American Association on Artificial Intelligence (AAAI-91)*, 61–66, Menlo Park: AAAI Press.

Maybury, M. T. and W. Wahlster (Editors). 1998. *Readings in Intelligent User Interfaces.* San Mateo, CA: Morgan Kaufmann Publishers.

McKeown, K., S. Feiner, M. Dalal, and S. Chang. 1998. "Generating Multimedia Briefings: Coordinating Language and Illustration". *AI Journal*, 103(1-2), 95–116.

McKeown, K. R. 1985. *Text Generation.* Cambridge, MA: Cambridge University Press.

P. Mc Kevitt (Editor). 1994-1996. *Artificial Intelligence Review, Special Issues of on the Integration of Natural Language and Vision Processing*, Volume 8, No: 2-3, No: 4-5, Vol. 9, No: 2-3, No: 5-6, Vol. 10, No: 1-2, No: 3-4. Boston: Kluwer, also available as Volume I (The BLUE book), Volume II (The BLACK book), Volume III (The GREEN book) and Volume IV (The GREEN book).

Merlino, A., and M. T. Maybury. 1999. "An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News". *Automated Text Summarization* ed. by I. Mani and M. T. Maybury. MIT Press.

Merlino, A., D. Morey, and M. T. Maybury. 1997. "Broadcast News Navigation Using Story Segmentation". *ACM Multimedia 97*, 381–391. ACM Press.

Moore, J. D. and W. R. Swartout. 1990. "Pointing: A Way Toward Explanation Dialogue". *Proceedings of 7th National Conference of the American Association on Artificial Intelligence (AAAI-90)*, 457–464, Menlo Park: AAAI Press.

Nagao K., and A. Takeuchi. 1994. "Social Interaction: Multimodal Conversation with Social Agents". *Proceedings of the 32 nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, volume 1, 22–28.

Neal, J. G. and S. C. Shapiro. 1991. "Intelligent Multi-Media Interface Technology". *Intelligent User Interfaces* ed. by J. W. Sullivan and S. W. Tyler, 11–43. New York: ACM Press.

Oviatt, S. L. 1999. "Mutual Disambiguation of Recognition Errors in a Multimodal Architecture". *Proceedings of Conference on Human Factors in Computer Systems (CHI 99)*, 576–583.

Oviatt, S. L., A. DeAngelie, and K. Kuhn. 1997. "Integration and Synchronization of Input Modes During Multimodal Human-Computer Interaction". *Proceedings of Conference on Human Factors in Computer Systems (CHI 97)*, 415–422.

Oviatt, S. and W. Wahlster (Editors). 1997. *Human-Computer Interaction, Special Issue on Multimodal Interfaces*, Vol. 12, No. 1-2.

Prevost, S. 1996. "An Information Structural Approach to Spoken Language Generation". *Proceedings of the 34 th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, 294–301, Santa Cruz, CA.

Rist, T. (Editor). 1998. *Computer Standards & Interfaces, Special Double Issue on Intelligent Multimedia Presentation Systems.* North-Holland, Vol. 18, No. 5-7.

Roth, S. F., J. Mattis, and X. Mesnard. 1991. "Graphics and Natural Language as Components of Automatic Explanation". *Intelligent User Interfaces* ed. by J. W. Sullivan and S. W. Tyler, 207–239. New York: ACM Press.

Sable, C. and V. Hatzivassiloglou. 1999. "Text-Based Approaches for the Categorization

of Images". *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, Paris, France.

Searle, J. R. 1980. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, England: Cambridge University Press.

Stock, O., C. Strapparava and M. Zancanaro 1997. "Explorations in an Environment for Natural Language Multimodal Information Access". *Intelligent Multimedia Information Retrieval* ed. by M. T. Maybury, 381-398. Menlo Park: AAAI Press.

Wahlster, W., E. André, W. Finkler, H.-J. Profitlich, and T. Rist. 1993. "Plan-Based Integration of Natural Language and Graphics Generation". *AI Journal*, 63, 387–427.

Index