REPRESENTING THE CONTENTS OF PRESENTATIONS IN MULTIMODAL GENERATION SYSTEMS *

Harold J. Paredes-Frigolett  
IIG  
Institute of Computer Science and Social Research  
Albert-Ludwig University of Freiburg  
Friedrichstraße 50  
D-79098 Freiburg  
Germany  

e-mail: paredes@cognition.iig.uni-freiburg.de

Abstract

Recently developed multimodal generation systems use terminological and feature structure logics ([3], [18]) for both, knowledge and semantic representation. In this article, we argue that terminological and feature structure systems are inadequate for knowledge and semantic representation in multimodal presentation systems. The main reason for this is the fact that they are not expressive enough to represent the semantics of multimodal presentations involving natural language text. We show this for a multimodal discourse fragment taken from a task-oriented domain of information presentation. We present our proposal for representing multimodal presentations combining natural language text and illustrated graphics. In our account, we integrate main features of situation semantics and feature logics without restraining the expressiveness of the framework to a subset of first order logic. The representational framework proposed is supported by the denotational and computational semantics put forth with Episodic Logic.

*The work presented here was supported by IIG, the Institute of Computer Science and Social Research in Freiburg, and the German National Science Foundation in Bonn, Germany. Currently, the author has an appointment as visiting researcher in the AI Lab of the University of Rochester, where he is involved in knowledge engineering projects that have been developed for the Boeing Co. in USA and Canada.
1 Introduction

Feature structure systems and terminological systems in the KL-ONE tradition ([18], [3]) for knowledge representation are formally supported by a well-defined denotational semantics which corresponds to a subset of first order logic ([14]). What is even more important from an engineering point of view, current implementations of term subsumption and feature structure logics for knowledge representation are available to researchers in various areas of artificial intelligence ([7], [18]). This has led researchers and developers to make use of them not only in knowledge-based language processing systems, but also in multimodal presentation systems ([12], [19]). Unfortunately, the huge gap between the expressiveness of these frameworks and the expressiveness required to represent the contents of natural language text and multimodal presentations ultimately leads to separating the knowledge and semantic representation tasks and formalisms, providing a well-defined semantics for the former and an ill-defined operational semantics for the latter. It has long been argued within the knowledge representation community in favor of restraining the expressiveness of these knowledge representation systems in order to maintain soundness and completeness of the representational frameworks. Also, some extensions to the expressiveness of terminological systems have been proposed so as to represent natural language semantics. But even in those systems ([18], [7]), little natural language and multimodal discourse fragments are actually expressible due to the fact that the expressiveness of the employed logics not even extends to full first order logic ([16], [11]). We give arguments in favor of using more expressive logics for representing the semantic contents of both, natural language and multimodal presentations, and present a representational framework for multimodal generation.

2 A sample multimodal discourse

Consider the following multimodal discourse fragment, MMDF, taken from the application domain of the WIP system ([19]):

**MMDF**: Lift the lid and pour in three cups of water in the water container. Turn on the blue on/off switch in the picture in order to make espresso.

For our discourse we require a logical framework capable of representing:

1. a complex plan consisting of states of affairs, events and actions,
2. a speech act involving the system requesting the user to accomplish a plan,
3. the temporal relation between the situations described by the sentences to lift the lid, to pour in three cups of water in the water container, and to turn on the blue on/off switch in the picture,
4. the causal relation between the situations described by the sentences to turn on the blue on/off switch and to make espresso,
5. the expression of the mass term espresso,
6. the generalized quantifier *three cups of water*,

7. the cross-modal referent *blue on/off switch in the picture*,

8. the reified property expressed by the infinitive *to make*.

Terminological logics and feature logics do not provide the expressiveness required to represent the semantic phenomena 1, 2, 4 and 8 listed above. Some extensions to terminological logics have been proposed for both, representing simple plans using role value maps, temporal projection and plan subsumption, and for representing generalized quantifiers and mass terms ([7]). But since the expressiveness of terminological and feature logics is *per definitionem* a subset of first order logic ([14]), these extensions just contribute to reduce the reliability of these representational frameworks by postulating an operational semantics for these extensions. A much richer ontology of individuals is needed to cope with the representational problems posed by natural language and multimodal discourse. For the representation of our multimodal discourse fragment in particular, we need explicit situational variables for representing events and actions, sentence and predicate modifiers, sentence and predicate reification as well as kinds of individuals. The expressiveness required for the representation of our fragment goes far beyond the one of first order logic.

3 The representational framework

New grammar formalisms in computational linguistics make use of feature logics to describe the syntax of natural language ([6], [3]). This approach to describe syntactic information has also been proposed to describe semantic information in both, natural language understanding and generation systems ([14], [15]), as well as in recently developed multimedia generation systems ([12], [19]). The decision to make use of such an attribute-value representation is mainly based on the fact that although not all of the text generation process is monotonic and therefore well-suited to be modeled and implemented using unification, main processes of the so-called how-to-say component of a text generator ([10]) are in fact monotonic and can advantageously be implemented using a unification-based approach. A central claim in this article is that this also applies for the process of generating multimodal documents.

We strive for a semantic representation in which we first distinguish a core multimodal logical form conveying the semantic content of the utterance and the graphics to be generated. This representation conveys just semantic information. Then we postulate an augmented multimodal logical form consisting of the core multimodal logical form to which we add systemic and pragmatic features. To provide a formal and computational semantics for our representational framework, we propose EPISODIC LOGIC as a semantic and knowledge representation for multimodal discourse and endow the representational framework with further features necessary for supporting the multimodal generation tasks.

3.1 Episodic Logic

As a semantic and knowledge representation system, we propose EPISODIC LOGIC (EL) and its computational implementation EPILOG. EL is a recently developed, highly expressive
second order computational logic for natural language understanding. For a detailed description of EL and EPILOG the reader is referred to [8] and [17]. EL has a rich ontology of individuals including the following:

1. ordinary individuals,
2. possible situations (involving exhaustive situations, possible times, possible worlds, and moments of time),
3. propositions and facts (consistent propositions),
4. kinds of ordinary individuals, kinds of actions and kinds of situations,
5. n-dimensional regions and space-time trajectories,
6. collections and n-vectors.

This rich ontology of individuals allows for straightforward interpretation and simplifies the translation from natural language into logical form. EL has been successfully used for story understanding ([8]), as a natural language interface for the TRAINS planning system at the University of Rochester ([11]), and for message processing applications at the Boeing Co. ([13]). As presented in [16], EL enables us the representation of:

1. events, actions, kinds, attitudes, modification and donkey sentences,
2. generalized quantifiers,
3. lambda abstraction,
4. sentences and predicate modifiers, sentence and predicate reification,
5. intensional predicates,
6. unreliable generalizations and explicit situational variables.

### 3.2 A representational framework for multimodal generation

Table 3.1 shows the general characteristics of the multimodal representation.

<table>
<thead>
<tr>
<th>Language</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>general format</td>
<td>feature structures</td>
</tr>
<tr>
<td>representation form</td>
<td>attribute-value graphs</td>
</tr>
<tr>
<td>augmented multimodal logical form</td>
<td>semantic, pragmatic, and systemic information</td>
</tr>
<tr>
<td>core multimodal logical form</td>
<td>semantic information</td>
</tr>
<tr>
<td>syntax</td>
<td>context-free grammar</td>
</tr>
<tr>
<td>semantics</td>
<td>model-theoretic, denotational and computational</td>
</tr>
</tbody>
</table>
The representation consists of an attribute-value graph where semantic, pragmatic and systemic information is conveyed. The so-called core multimodal logical form, \( \text{CMMLF} \), conveys semantic information about the multimodal document to generate and corresponds to the filler of the attribute \text{content} in an augmented multimodal logical form, \( \text{AMMLF} \). This in turn contains additional pragmatic and systemic features. The pragmatic features describing the situation of presentation is \text{presentation act} which has as possible values either a speech act or a picture act. The pragmatic features are summarized in Table 3.2. Table 3.3 describes the speech acts involved with an informal description of their meaning. Table 3.4 describes the picture acts involved with an informal description of their meaning.

### Table 3.2: Features conveying pragmatic information

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{presentation act}</td>
<td>\text{speech act, picture act}</td>
</tr>
<tr>
<td>\text{speech act}</td>
<td>\text{assertion, description, order, hypothesis, introduction, annotation, request}</td>
</tr>
<tr>
<td>\text{picture act}</td>
<td>\text{background, identification, introduction, attention}</td>
</tr>
<tr>
<td>\text{referents}</td>
<td>\text{the discourse referents in a representation}</td>
</tr>
<tr>
<td>\text{configuration}</td>
<td>\text{the roles occurring in the representation}</td>
</tr>
<tr>
<td>\text{focus}</td>
<td>\text{the role under focus in the representation}</td>
</tr>
<tr>
<td>\text{distance}</td>
<td>\text{to be introduced, just introduced, actual, near, far}</td>
</tr>
</tbody>
</table>

### Table 3.3: Speech acts involved in the representations

<table>
<thead>
<tr>
<th>Speech-act</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{assertion}</td>
<td>\text{a general assertion}</td>
</tr>
<tr>
<td>\text{description}</td>
<td>\text{a description of a process or an object in the domain}</td>
</tr>
<tr>
<td>\text{recommendation}</td>
<td>\text{a recommendation from the system to the user}</td>
</tr>
<tr>
<td>\text{order}</td>
<td>\text{an order from the system to the user}</td>
</tr>
<tr>
<td>\text{request}</td>
<td>\text{a request from the system to the user}</td>
</tr>
<tr>
<td>\text{introduction}</td>
<td>\text{a concept definition from the system to the user}</td>
</tr>
<tr>
<td>\text{hypothesis}</td>
<td>\text{a hypothetical event in the domain}</td>
</tr>
<tr>
<td>\text{annotation}</td>
<td>\text{a nominal phrase serving as annotation for an object introduced using graphics}</td>
</tr>
</tbody>
</table>

### Table 3.4: Picture acts involved in the representations

<table>
<thead>
<tr>
<th>Picture act</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{background}</td>
<td>\text{a depiction providing the background for describing an object of the domain}</td>
</tr>
<tr>
<td>\text{introduction}</td>
<td>\text{a depiction introducing an object of the domain}</td>
</tr>
<tr>
<td>\text{attention}</td>
<td>\text{a depiction used to attract the attention of the user}</td>
</tr>
<tr>
<td>\text{identification}</td>
<td>\text{a depiction used to help to identify an object in a picture}</td>
</tr>
</tbody>
</table>
The meaning of the speech and picture acts as well as the world knowledge of a particular application domain are formally described using meaning postulates and world knowledge axioms in EL. Table 3.5 describes the features used to annotate the contents in the semantic representation that is generated using text, graphics, animation, or gestures. The attribute media is filled with a mark representing either a particular modality or a combination of modalities. Table 3.6 describes the speech acts and picture acts.

Table 3.5: Features conveying information for media coordination

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>media</td>
<td>text, graphics, gestures, animation, text and graphics, ...</td>
</tr>
<tr>
<td>cross-modal referents</td>
<td>the referents introduced in other modalities</td>
</tr>
</tbody>
</table>

Table 3.6: Types of speech and picture acts involved in the representations

<table>
<thead>
<tr>
<th>Values</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>picture</td>
<td>an object or group of objects in the knowledge base</td>
</tr>
<tr>
<td>attribution</td>
<td>the value of a role of an object in the knowledge base</td>
</tr>
<tr>
<td>equation</td>
<td>identity between the role values of two knowledge base entities</td>
</tr>
<tr>
<td>action, event, plan</td>
<td>a situation/situations related by causal and temporal relation</td>
</tr>
</tbody>
</table>

4 A sample representation

In this section, we show what the representation of our little multimodal discourse looks like. In doing so, we introduce the so-called augmented multimodal logical form, \( \text{AMMLF} \), and the core multimodal logical form, \( \text{CMMMLF} \). The former corresponds to a logical form containing the systemic, pragmatic and semantic features necessary for supporting the generation process, describing the situation of presentation, and specifying the semantic content concerning the events and actions involved. The latter corresponds to a semantic representation of single sentences or objects in the multimodal discourse.

4.1 Augmented multimodal logical form

The augmented multimodal representation shown below can be decomposed in two presentation acts, the speech act request and the picture act background. The speech act request involves both a precondition and a postcondition. The precondition and the postcondition consist of composite actions, \( e_1 \) and \( e_2 \), related by the temporal relation meets. The picture act background involves a depiction\(_1\) and a depiction\(_2\). The value of the attribute refers ents corresponds to a set of variables denoting situation individuals, ordinary individuals, location individuals and picture individuals. For the speech act request, we introduce the referents \( e_1 \) and \( e_2 \). The two episodic variables \( e_1 \) and \( e_2 \) refer to composite actions and are further decomposed in two subepisodes, respectively. For the picture act background, we introduce the referents \( p_1 \) and \( p_2 \).
$\textbf{AMMLF} \ 3.1_{e_0} :$

\begin{align*}
\text{illlocutionary force} & : \text{request} \\
\text{type} & : \text{composite action} \\
\text{contents} & : [\text{rel} \text{ meets } \text{precondition}] \\
\text{configuration} & : \{\text{precondition, postcondition}\} \\
\text{focus} & : \text{precondition} \\
\text{media} & : \text{text} \\
\text{referents} & : \{e_1, e_2\} \\
\text{cross-modal referents} & : \{p_1, p_2\} \\
\text{distance} & : \text{to be introduced}
\end{align*}

$\textbf{AMMLF} \ 3.1_{p_0} :$

\begin{align*}
\text{pictorial force} & : \text{background} \\
\text{type} & : \text{composite picture} \\
\text{contents} & : [\text{rel} \text{ left-of} \text{depiction}_1 \text{ depiction}_2] \\
\text{configuration} & : \{\text{depiction}_1, \text{depiction}_2\} \\
\text{focus} & : \text{depiction}_1 \\
\text{media} & : \text{graphics} \\
\text{referents} & : \{p_1, p_2\} \\
\text{cross-modal referents} & : \{} \\
\text{distance} & : \text{actual}
\end{align*}

$\textbf{AMMLF} \ 3.1_{e_1} :$

\begin{align*}
\text{type} & : \text{composite action} \\
\text{contents} & : [\text{rel} \text{ meets } \text{postcondition}] \\
\text{configuration} & : \{\text{action}_1, \text{action}_2\} \\
\text{focus} & : \text{action}_1 \\
\text{media} & : \text{text} \\
\text{referents} & : \{e_3, e_4\} \\
\text{cross-modal referents} & : \{} \\
\text{distance} & : \text{to be introduced}
\end{align*}
In representation $\text{AMMLF } 3.1_{e_0}$ we introduce the situation individuals $e_1$ and $e_2$ of which the temporal relation \textit{meets} holds. That is, episode $e_1$ is immediately followed by episode $e_2$ in the discourse. They are both subepisodes of a coextensive episode $e_0$. Similarly, in representation $\text{AMMLF } 3.1_{p_0}$ we introduce the picture individuals $p_1$ and $p_2$ of which the spatial relation \textit{left-of} holds. That is, picture $p_1$ is located to the left of picture $p_2$. They are both subpictures of a coextensive composite picture $p_0$. In our representations, the feature \textit{ind} stands for individual and it is used to introduce multimodal discourse referents.

The information episodes $e_1$ and $e_2$ describe is given by introducing a set of restraining conditions under the feature \textit{cond}. We describe these conditions in $\text{AMMLF } 3.1_{e_1}$ and $\text{AMMLF } 3.1_{e_2}$, respectively.

\textbf{AMMLF } 3.1_{e_2}:

\begin{align*}
&\text{type} \quad \text{composite action} \\
&\quad \text{rel} \quad \text{cause-of} \\
&\quad \arg_{cause} \quad \text{ind} \quad e_5 \\
&\quad \arg_{effect} \quad \text{cond} \quad \text{AMMLF } 3.1_{e_5} \\
&\text{configuration} \quad \{\text{cause, effect}\} \\
&\text{focus} \quad \text{cause} \\
&\text{media} \quad \text{text} \\
&\text{referents} \quad \{e_5, e_6\} \\
&\text{cross-modal referents} \quad \{\} \\
&\text{distance} \quad \text{to be introduced}
\end{align*}

\textbf{AMMLF } 3.1_{e_3}:

\begin{align*}
&\text{type} \quad \text{action} \\
&\text{content} \quad \text{CMMLF } 3.1_{e_3} \\
&\text{configuration} \quad \{\text{theme, loc-episode}\} \\
&\text{focus} \quad \text{theme} \\
&\text{media} \quad \text{text} \\
&\text{referents} \quad \{\text{ind}_1, \text{ind}_{loc_3}\} \\
&\text{cross-modal referents} \quad \{\text{ind}_1\} \\
&\text{distance} \quad \text{to be introduced}
\end{align*}

\textbf{AMMLF } 3.1_{e_4}:

\begin{align*}
&\text{type} \quad \text{action} \\
&\text{content} \quad \text{CMMLF } 3.1_{e_4} \\
&\text{configuration} \quad \{\text{theme, loc-episode, loc}\} \\
&\text{focus} \quad \text{theme} \\
&\text{media} \quad \text{text} \\
&\text{referents} \quad \{\text{ind}_2, \text{ind}_3, \text{ind}_{loc_4}, \text{ind}_{loc_7}\} \\
&\text{cross-modal referents} \quad \{\text{ind}_2, \text{ind}_3, \text{ind}_{loc_7}\} \\
&\text{distance} \quad \text{to be introduced}
\end{align*}
Note that in representations $\text{AMMLF} \ 3.1_{e_1}$ and $\text{AMMLF} \ 3.1_{e_2}$ new discourse referents have been introduced which correspond to episodes $e_3$ and $e_4$, and $e_5$ and $e_6$. They are subepisodes of episodes $e_1$ and $e_2$, respectively. Representations $\text{AMMLF} \ 3.1_{e_3}$ and $\text{AMMLF} \ 3.1_{e_4}$ describe the simple actions involved in the composite action of representation $\text{AMMLF} \ 3.1_{e_1}$. Similarly, representations $\text{AMMLF} \ 3.1_{e_5}$ and $\text{AMMLF} \ 3.1_{e_6}$ describe the cause-effect event involved in the composite action of representation $\text{AMMLF} \ 3.1_{e_2}$. For the sake of space, we omit representations $\text{AMMLF} \ 3.1_{e_3}$ and $\text{AMMLF} \ 3.1_{e_4}$.

In representations $\text{AMMLF} \ 3.1_{e_3}$ and $\text{AMMLF} \ 3.1_{e_4}$, we introduce the ordinary individual referents $ind_1, ind_2$ and $ind_3$ that refer to the lid, three cups of water and the water container of the espresso machine, respectively. We also introduce the location individuals $ind_{loc_3}$ and $ind_{loc_4}$ that correspond to the spatiotemporal locations of the episodes $e_3$ and $e_4$, respectively, and the location individual $ind_{loc_7}$ denoting a localization in the picture which is in the water container. Similarly, in the augmented multimodal logical forms $\text{AMMLF} \ 3.1_{e_5}$ and $\text{AMMLF} \ 3.1_{e_6}$ we introduce the ordinary individual referents $ind_4$ and $ind_5$ to refer to the blue on/off switch in the picture and espresso, respectively. The spatiotemporal locations of the episodes $e_5$ and $e_6$ are $ind_{loc_8}$ and $ind_{loc_9}$, respectively. Finally, we introduce the location individual $ind_{loc_8}$ that refers to a salient localization in the picture.

\begin{align*}
\text{AMMLF} \ 3.1_{e_3}:
\begin{array}{|c|c|}
\hline
\text{type} & \text{content} \\
\text{configuration} & \text{action} \\
\text{focus} & \text{CMMLF} \ 3.1_{e_3} \\
\text{media} & \{\text{theme, loc-episode}\} \\
\text{referents} & \text{theme} \\
\text{cross-modal referents} & \{\text{text}\} \\
\text{distance} & \{\text{ind_4, ind_{loc_3}, ind_{loc_4}}\} \\
\hline
\end{array}
\end{align*}

\begin{align*}
\text{AMMLF} \ 3.1_{e_4}:
\begin{array}{|c|c|}
\hline
\text{type} & \text{event} \\
\text{content} & \text{CMMLF} \ 3.1_{e_4} \\
\text{configuration} & \{\text{theme, loc-episode}\} \\
\text{focus} & \text{theme} \\
\text{media} & \text{text} \\
\text{referents} & \{\text{ind_4}\} \\
\text{cross-modal referents} & \{\text{ind_4}\} \\
\text{distance} & \text{just introduced} \\
\hline
\end{array}
\end{align*}

4.2 The core multimodal logical form

The value of the attribute content in our last representations corresponds to the core multimodal logical form containing the semantic information generated by the multimodal generators for single sentences or objects. These representations are not augmented since they do not convey pragmatic and systemic information. The format of the episodic multimodal logical forms given below resembles situation schemata ([5]). Here, the feature rel stands for a relation corresponding to an n-ary predicate taking a certain number of arguments.
and having the polarity \( \text{pol} \). This polarity can assume the value either 1 or 0, according to which the predicate holds of its arguments or not. We omit positives polarities in the representations. For the sake of space we present here just the core multimodal logical form needed for the representation of the sentence press the blue on/off switch in the picture.

\[ CMMLF \ 3.1_{e_5} : \]

\[
\begin{align*}
\text{rel} & \quad \text{press} & \quad \text{ind} & 4 \\
\text{spec} & \quad \text{the} & \quad \text{has-localization} & \quad \text{cond} \\
\text{argtheme} & \quad \text{ind} & \quad \text{indloc} & \quad \text{in} \\
\text{argloc-episode} & \quad \text{indloc} & & \\
\end{align*}
\]

Here, we assume that the cross-modal referent \( \text{ind}_4 \) has been generated using graphics according to representation \( AMMLF \ 3.1_{p_2} \). We just present the part of \( AMMLF \ 3.1_{p_2} \) displaying the semantic and sistemic information available for \( \text{ind}_4 \):

\[ AMMLF \ 3.1_{p_2}(\text{ind}_4) : \]

\[
\begin{align*}
\text{type} & \quad \text{picture} & \quad \text{ind} & \quad \text{ind}_4 \\
\text{content} & \quad \text{cond} & \quad \{ \text{theme} \} \\
\text{configuration} & \quad \text{theme} \\
\text{focus} & \quad \text{graphics} \\
\text{media} & \quad \{ \text{ind}_4 \} \\
\text{referents} & \quad \{ \text{ind}_4 \} \\
\text{cross-modal referents} & \quad \text{actual} \\
\text{distance} & \quad \{ \text{theme} \} \\
\end{align*}
\]

The attribute value \( \text{actual} \) of the sistemic feature \( \text{distance} \) tells us that the cross-modal referent \( \text{ind}_4 \) has been recently introduced in the multimodal discourse and is being displayed in the picture. A presentation planning module is able to use this information in order to annotate representation \( CMMLF \ 3.1_{e_5} \) with the attribute value pair \( \text{spec} \ \text{the} \). This will allow the text generator to choose a definite description for the noun phrase involved. Similarly, the attribute value \( \text{graphics} \) of the sistemic feature \( \text{media} \) tells us that \( \text{ind}_4 \) has been generated using graphics. The planner uses this additional information in order to expand the semantic representation with additional information regarding the localization of the cross-modal
referent $i_{d_4} \text{ (in the picture)}$ in representation $C_{MMLF} \ 3.1_{c_5}$. The text generator will use this information as a modifier of the noun phrase blue off/on switch. The information that tells the text generator that the individual $i_{d_4}$ corresponds to an instantiation of a blue on/off switch is already available in representation $A_{MMLF} \ 3.1_{p_2(i_{d_4})}$ and can be accessed using the co-reference $[i_4]$ in representation $C_{MMLF} \ 3.1_{e_5}$.

5 Conclusions

Since the limited expressiveness of current mainstream knowledge representation systems based on terminological logics ([14]) is shown to be inadequate for the representational phenomena encountered in natural language and multimodal discourse, we have proposed **EPISODIC LOGIC** ([16]) and its computational implementation **EPILOG** ([17]) for knowledge and semantic representation in natural language and multimedia generation systems. There are two main reasons for this decision. First, **EPISODIC LOGIC** is the most expressive logic for knowledge representation available to artificial intelligence researchers interested in knowledge-based natural language processing and multimodal generation systems. Second, it has been already implemented and applied for story and text understanding ([8]), as a natural language interface for the TRAINS planning system at the University of Rochester ([1]), and for message processing applications at the Boeing Co. ([13]).

We have presented a formal framework to successfully deal with the representational problems posed by multimodal discourse. The main claim of the proposal is that multimodal presentations are construed as situations or episodes. Unlike other approaches based on situation semantics ([2]), our approach proposes the use of explicit situational or episodic variables that can be related by temporal and causal relations. The use of **EPISODIC LOGIC** allows us to endow the representational framework with a formal denotational and computational semantics. Also, the semantic and knowledge representation borrows many of the characteristics of current precomputational semantic theories ([2], [9]).

We have not described **EPISODIC LOGIC** here, the reader has been referred to [8], [17], [16] and [13] for a detailed description of the logic and its implementation. The distinction between augmented multimodal logical forms and multimodal logical forms has been made for expository reasons and should not be understood as we were trying to define two different kinds of representations. Also, the use of feature structures has been motivated by the possibility of using the representational framework proposed here in concert with large grammars often described using unification-based formalisms ([6], [4], [15]). The important issue here is that the semantics that formally supports the language is not based on typed feature logics, but on **EPISODIC LOGIC**.

References


