Multimodal Unification-based Grammars

Michael Johnston

Center for Human-Computer Communication
Oregon Graduate Institute
P.O. Box 91000, Portland, OR 97291
johnston@cse.ogi.edu

Multimodal interfaces enable more natural and efficient interaction between humans and machines by providing multiple channels through which input or output may pass. Specifically, our research concerns interfaces which support simultaneous input from speech and pen. Such interfaces have clear task performance and user preference advantages over speech only interfaces, in particular for spatial tasks such as those involving maps (Oviatt 1996).

In order to realize their full potential, multimodal systems need to support not just input from multiple modes, but synchronized integration of modes. The design of representations and mechanisms to support multimodal integration is the central challenge in the development of next-generation multimodal systems.

Our position is that typed feature structures (Carpenter 1992) provide a desirable underlying common meaning representation for multiple modes. Multimodal integration can then be modelled by unification of typed feature structures (Johnston et al 1997).

This approach has a number of advantages. Many previous approaches (e.g. Neal and Shapiro 1991) treat gesture as a secondary dependent mode and integration of gesture is triggered by the appearance of expressions in the speech stream whose reference needs to be resolved (e.g. 'this one'). Unlike these speech-driven approaches, our approach is *fully multimodal* in that all elements of a command can in principle originate in either mode. Typed feature structures are formally well understood and unification provides a declarative well defined mechanism for multimodal integration. Another significant advantage of typed feature structures is that they allow for representation of partial meanings through underspecified feature structures and type constraints. Furthermore, this approach to multimodal language processing fits well with contemporary work in natural language processing, where unification-based formalisms are common.

Our approach to multimodal integration is implemented as part of QuickSet (Cohen et al 1997), a working system which supports dynamic interaction with maps and other complex visual displays. In the example in Figure 1, the user interacts with a map in order to coordinate disaster relief. In this case, the user has said '`flood zone`' and has specified its extent by drawing an area. We exemplify below how typed features structures are used in this system to represent the contributions of individual modes and to provide a declarative statement of the grammar of multimodal utterances.

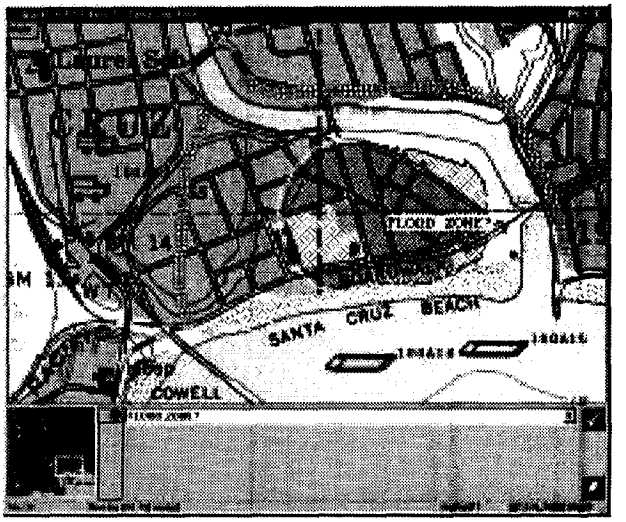

Figure 1: User Interface

Spoken and gestural input are assigned typed feature structures which specify their category and their semantic contribution, along with their probability, temporal extent and so on. The range of potential multimodal expressions is described using a unification-based multimodal grammar augmented with functional constraints. This grammar consists of a series of rule schemata, themselves encoded as typed feature structures, which express potential multimodal integration strategies. In the example above, '`flood zone`' is assigned the representation in Figure 2. The area gesture has the representation in Figure 3.

The rule responsible for basic two element multimodal integrations such as this is given in Figure 4. It states that a *located_command*, of which *area_type*

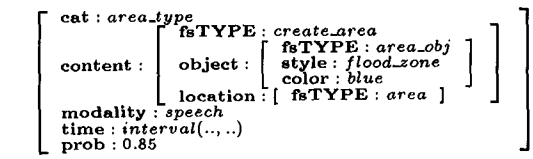

Figure 2: Spoken Input: '**flood zone**'

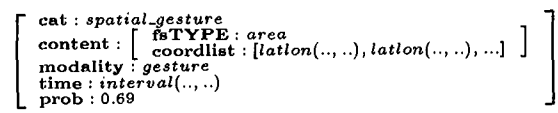

Figure 3: Gesture Input: area

is a subtype, can combine with a spatial gesture so long as the location feature of the *located_command* unifies with the content of the gesture. The resulting multimodal constituent inherits its content from the *located_command.*

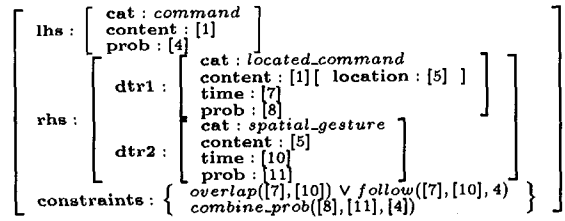

Figure 4: Basic Integration Rule Schema

The **constraints** feature specifies a number of functional constraints which must be met in order for the rule to apply. In this case, the first of these specifies that speech must either overlap with or start within four seconds of gesture. The second calculates the joint probability to be assigned to the result.

Multimodal integration is achieved through a multidimensional chart parsing process (Johnston 1998) which combines the inputs in accordance with rule schemata like that in Figure 4. Along with multimodal combinations, this approach to parsing supports unimodal gestures, unimodal speech, and visual parsing of multiple gestures. Furthermore, it is not limited to speech and gesture input and extents readily to other combinations of modes.

Multimodal Subcategorization

In order to account for multimodal utterances in which more than two elements are combined, such as '**sandbag wall from here to here**' with two point gestures, a form of multimodal subcategorization is employed. This draws on the lexicalist treatment of verbal subcategorization in unification-based approaches to grammar such as HPSG (Pollard and Sag 1994). Just as a verb subcategorizes for its complements, we can think of a terminal in the multimodal grammar as subcategorizing for the edges with which it needs to combine. For example, '**sandbag wall from here to here**' (Figure 5) subcategorizes for two gestures. This multimodal subcategorization is specified in a list valued **subcat** feature, implemented using a recursive **first/rest** feature structure (Shieber 1986).

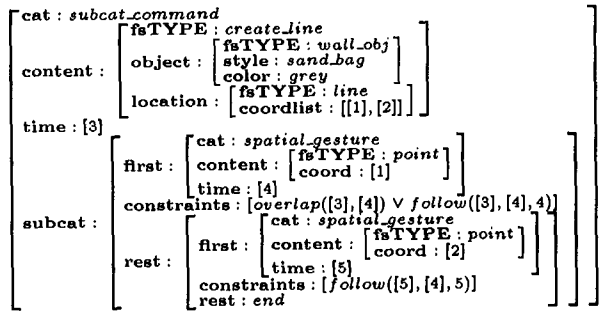

Figure 5: '**sandbag wall from here to here**'

Subcategorizing expressions are parsed using a pair of general combinatory schemata which incrementally remove elements from the subcat list and combine them with appropriate expressions from other modes.

The Evolution of Multimodal Systems

Current development of multimodal architectures is following a trajectory with parallels in the history of syntactic parsing. Initial approaches to multimodal integration were largely algorithmic in nature. The next stage is the formulation of declarative integration rules (phrase structure rules), then comes a shift from rules to representations (lexicalism, categorial and unification-based grammars). The approach outlined here is at the representational stage, although given the frequency of constructional meaning in multimodal utterances it is desirable to allow for specification of integration strategies by rule as well as in the 'lexicon' of multimodal utterances.

The next phase, which syntax is undergoing, is the compilations of rules and representations back into faster, low-powered finite state devices. At this early stage, we believe a high degree of flexibility is needed. In the future, once it is clearer what needs to be accounted for, the next step will be to explore compilation of multimodal grammars into lower power devices.

References

Carpenter, R. 1992. The logic of typed feature structures. Cambridge University Press, Cambridge, England.

Cohen, P. R., M. Johnston, D. McGee, S. L. Oviatt, J. A. Pittman, I. Smith, L. Chen, and J. Clow. 1997. QuickSet: Multimodal interaction for distributed applications. In Proceedings of the Fifth ACM International Multimedia Conference. ACM Press, New York, 31–40.

Johnston, M. 1998. Unification-based multimodal parsing. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 98).

Johnston, M., P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith. 1997. Unification-based multimodal integration. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, 281-288.

Neal, J. G., and S. C. Shapiro. 1991. Intelligent multi-media interface technology. In J. W. Sullivan and S. W. Tyler (eds.) Intelligent User Interfaces, ACM Press, Addison Wesley, New York, 45–68.

Oviatt, S. L. 1996. Multimodal interfaces for dynamic interactive maps. In Proceedings of Conference on Human Factors in Computing Systems: CHI '96, Vancouver, Canada. ACM Press, New York, 95–102.

Pollard, Carl and Ivan Sag. 1994. Head-driven phrase structure grammar. University of Chicago Press. Chicago.

Shieber, S. M. 1986. An Introduction to unification-based approaches to grammar. CSLI Lecture Notes Volume 4. Center for the Study of Language and Information, Stanford.