Transparent Natural Language Interaction through Multimodality

Ivan Bretan and Jussi Karlgren
ivan@sics.se, jussi@sics.se
Swedish Institute of Computer Science
Box 1263, 164 28 Kista, Sweden

August 31, 1993

Abstract

A scheme for integration of a natural language interface into a multimodal environment is presented with emphasis on the synergetic results that can be achieved, which are argued to be:

1. Complementary expressiveness.
2. Mutual illumination of language syntax and semantics.
3. Robust pragmatics and graceful recovery from failed natural language analyses through the reification of discourse objects to enable user control of discourse management.

Introduction

The main point with providing interaction through multiple modalities in the first place is that the total usability thus obtained is higher than the usability of each individual modality. In general, this is true since certain modalities support the realization of different types of communicative intentions with varying degrees of "appropriateness", which will be illustrated. In addition, complementary modalities can serve to mutually illuminate each other's characteristics and limitations if cross-modal translations of expressions are possible.

Pragmatics and robustness are difficult notions to tackle in natural language processing (NLP), where robustness generally follows from the application of sound pragmatics. It seems almost impossible to formalize pragmatics successfully within a natural language interface (NLI) although this does not seem to be an insurmountable issue in systems based on "simpler" modalities, such as graphically based direct manipulation interfaces. We will argue that through making pragmatics more explicit and user-controlled we can overcome some of the difficulties with handling natural language discourse, both ill- and well-formed.

The Role of Natural Language in Multimodal Interaction

Ever since the earliest days of computational linguistics, researchers have devoted substantial efforts to the development of interactive natural language interfaces. Only in the last few years, the value and naturalness of the *teletype* interaction normally taken as the goal of these projects have been questioned by NLP researchers. This approach puts strong emphasis on the self-sufficiency of (typed) natural language as an interaction mode for all tasks. For example, in a natural language interface (NLI) for database querying, not only would all queries be written in natural language,[1] but also meta-tasks such as inquiring about the competence of the system, graphical formatting of the data or perhaps even file management would be specified in natural language (*"...and save the answer in a file"*). Most existing systems do draw the line between the intended use of natural language and coexisting modalities — but normally on a low level, i.e., the level where the general (graphical) environment of which the interface is part can be used. The idea of devoting costly linguistic development to providing an NLI with the capability of understanding *"Make this window a little bigger"* or *"Move the mouse pointer two centimeters to the right"* when there is already a superior way of specifying this in the graphical environment is an idea which seems somewhat bizarre.

Instead, we now see the emergence of a line of research which aims to use the modality of written or spoken language only as a component in a larger, multimodal context in order to achieve an "artificial naturalness" in interactive systems — what Oliviero Stock refers to as the "third modality of natural language" (1992).

Natural language interaction, even in the stifled form that current NLP technology can support, has several obvious advantages. Any interface language has to be learned, and will be easier to learn if it resembles a language you are already familiar with. *Naturalness* is the principal feature of natural language which distinguishes it from other modes of interaction, and enables *learnability* of NL-like languages. Interaction which is natural in this sense can free the user from having to ponder the lower-level organization and processing of data. This naturalness extends to the provision of expressive constructs for e.g., quantification, negation and contextual references, which are not easy to find natural realizations of in other, more artificial languages. Ambiguity, vagueness, metonymy, and metaphor are other features of natural languages which enable us to communicate efficiently, which are not easily incorporated into non-natural languages.

Successful exploitation of these linguistic phenomena require "natural natural language", i.e., interaction where communicative intentions are not only conveyed through literal meaning of words, but through references to the surrounding context, relations to previous discourse, gestures, prosody, relying on mutual knowledge and so on. Is the teletype approach realistic in (implicitly) assuming that we can do without these dimensions of communication without sacrificing efficient interaction?

Interestingly enough, Wizard-of-Oz-style user studies of mock-up NLP systems (Dahlbäck and Jönsson, 1986, 1991) have shown that the type of language typically used when communicating with an alleged computer is impoverished with respect to grammar, dialogue, use of contextual references and relying on mutual knowledge. The conclusion of these studies is that the limited functionality of NLP systems currently existing and to be developed in the foreseeable future is still useful. So why indeed incorporate natural language into a multimodal interface in the first place? Because less natural interfaces are more difficult to learn. Users participating in Wizard-of-Oz evaluations may stick to a simplistic register[2] which fulfills their needs, but it will always happen to be the *right* register, because of the adaptability of the wizard, which a real NLI will lack. We believe that integration of natural language interaction into a multimodal framework can compensate for this lack of flexibility, and also stimulate the user to a more discourse-oriented, incremental way of interacting — in contrast to the observed retrograde behaviour in the Wizard-of-Oz studies. More specifically, an NLI integrated with alternate modalities can:

[1] Although certain types of information retrieval tasks could certainly be performed more efficiently by e.g., browsing or keyword searches.

[2] A variety of language according to use.

Figure 1: Visual language query

We shall not delve into the details of the syntax and semantics of visual languages of this sort, what the expressive power is and how easy they are to learn. Suffice to say, this type of visual query would typically be constructed by choosing the three entity icons corresponding to noun phrases above and then linking them together with the appropriate relations, corresponding to verbs. Operators such as the universal quantifier can then be applied to the expression, with well-defined "edit semantics". Graphical representations of quantifiers are not trivial to design, but the above mentioned visual languages all have innovative mechanisms to deal with quantification. Their ease of use remain to be evaluated empirically.

These visual languages are certainly significantly more formal than natural language. In fact, they can be seen as visual notations for a logical language. However, they have some other interesting properties as well. Since actions and objects are explicit, it is almost impossible to generate an expression with illegal syntax or semantics. Choices and actions are easily reversible, so that the user can incrementally work until the desired result has been obtained. In general, the interaction in a visual language interface is much more guided than in a conversational interface. For instance, in the above example, one could imagine that the relationships "sell" and "supply" with corresponding entities were selected by means of the user navigating through a concept graph representing the universe of discourse, marking the objects of interest. This is a highly interactive mode of communication, which may not be suitable for the experienced user who wants to input as much information as possible in one "chunk", who knows exactly what to say and how to say it. Also, as noted above, complex (quantificational) information may have very concise natural language formulations which are hard to match in a visual language, since, as Cohen *et al.* (1989, 1992) point out, visual languages are best suited for selection and manipulation of expressions which directly lends themselves to visualizations, e.g., in the form of icons, whereas natural language excels when it comes to indirect, abstract descriptions of information. Perceptually grounded characteristics of visual languages, such as size, colour, relative location etc. are of course interesting to exploit whenever possible (as in ACORD, Bes and Guillotin, 1991), but we are here concerned with visual languages that can express more abstract information.

Most existing visual languages are much simpler than the ones discussed above. Normally, there is no way of reference through specification of properties, only through explicit selection of a graphical object. In a GUI provided with a typical operating system, one file icon means one file, and two hundred file icons could perhaps happen to cover all the files on the disk, but there is no expression equivalent to "all the files on the disk" in this type of simplistic language. Of course, in these cases, the virtues of natural language become even more important. However, cross-modal translation of arbitrary expressions becomes impossible in all but the most trivial cases, and we will assume the existence of a visual language with relatively high expressive power to achieve the synergy effects of multimodal integration that we discuss in this paper. Less expressive visual languages may still provide for some (lesser) degree of synergy.

Mutual Illumination of Language Characteristics

Natural language systems are opaque to the user: it is not obvious what language they actually handle. The main problem with natural language interaction is how to teach the user the language the system uses. Natural language systems generally are uncooperative in this respect.

Humans have a natural tendency to mimic their counterparts' language (Ray and Webb, 1966; Levelt and Kelter 1982; Isaacs and Clark, 1987; Fussell and Krauss, 1990, 1992). This can be exploited in the design of interactive systems (Brennan, 1991; Zoltan-Ford, 1991; Karlgren 1992) and has been used in the implementation of IBM SAA LanguageAccess (Sanamrad and Bretan, 1992), for its Model Help Component: when users posed queries about a term in the lexicon, the Model Help Component output sentences in which the term was used, as a way of familiarizing the user with the content of the conceptual model and of the coverage of the grammar (Petrović, 1992).

In short, the solution to the problem of the non-transparency of natural language systems' linguistic competence is to use natural mechanisms of the user to have the user learn the system competence. If the system produces the kind of language it understands, the users will pick it up and recycle it. This kind of solution may prove unwieldy in a pure teletype environment, however. Generating paraphrases and spurious natural language feedback may produce too much feedback text for it to have any effect.

Similar learning problems can be suspected to arise in a visual language environment, since there are no naturally dedicated mechanisms of learning the manipulation of visual symbols and graphs to rely on, as in the case of natural language. An indication of what problems may show up is an empirical result which shows that users have more trouble remembering icons than command names (Black and Moran, 1982).

In a multimodal environment we can make use of both mechanisms, and use one to teach the workings of the other. Assume that the type of visual expression exemplified in figure 1 can be translated into a logical form of the same type as the ones a natural language interface uses as internal representations.[3]. Provided that the natural and visual language processors include generation as well as analysis components this makes it possible for visual expressions to be paraphrased in natural language and vice versa, shedding light on the way one would use the other modality to express the same message. For instance, the visual query in figure 1 could be paraphrased by the natural language sentence *"Show me the departments..."* which would indicate to the user a way of formulating queries in the NLI. Preferably this paraphrasing is done by means of a bidirectional analysis/generation grammar framework (such as in the Core Language Engine, Alshawi 1992) to guarantee that the language that is generated is actually accepted by the analysis phase.

This translation establishes links between natural language words and phrases on the one hand and visual language objects on the other. Since the visual language is closely modelled around the internal representation language that underlies all user- and system-generated expressions (regardless of which modality they originally were formulated in), some of the opacity obscuring how the natural language processor actually "understands" the user input is dispersed.

Interactive natural language systems often echo back to the user a natural language paraphrase of the interpretation of the user's utterance for confirmation, and possibly also disambiguation (which in the simplest case is done through the selection of one out of several different paraphrases). This gives the user some possibilities for control of the interpretation process, but no possibilities for *modification*. Say, for instance, that an almost correct interpretation of a sentence was produced. The only way to obtain the correct one is to reformulate the entire question (which the user may or may not manage to do so that the system can process it correctly).

A similar situation is where a follow-up question or assertion needs to be made, which involves the same or almost the same concepts and relationships as in the previous utterance. In a visual language which reveals more of the internal representations than natural language does, such modifications are likely to be much smaller (since syntax is simplified, while semantics remains equally powerful), and what is more important, they can be performed incrementally through direct manipulation of the visual expression. In such a framework we envisage translating a natural language expression (or parts of it) into a directly manipulable form thus enabling incremental communication. Of course, at any point it should be possible to translate the modified visual expression back into natural language, so that its meaning can be expressed in more accessible terms.

[3] This is a reasonable assumption as long as the visual interface and NLP system have a way of modelling the domain in common, such as a *conceptual schema*.

Discourse Management and Cross-Modal Natural Language Analysis Recovery

Due to the restricted linguistic competence of current NLP systems, analysis failures are common. This should not be regarded as an anomaly. Natural discourse between humans is also rife with analysis failures and misunderstandings on different levels of linguistic processing. These failures seldom lead to failures in discourse, but are starting points for further dialogue. The crucial point is that natural dialogue is not only interactive but also *incremental*[4] — a structure which human-computer dialogue in today's systems does not support. In human-human conversation, both parties take the responsibility to maintain this incrementality.

A problem with human-machine dialogue can be framed as the "one-shot"-problem: systems expect users to pose queries in one go, rather than discuss a topic until consensus is reached. This is not a natural way of using natural language. The whole premise that a user would be able to frame a query with a well-defined content in terms of the system knowledge base without a discourse context is alien to the nature of natural language — one-shot dialogues occur rarely in natural discourse.

As a special case of the fact that users expect little linguistic competence from computer systems, (Malhotra, 1975; Thompson, 1980; Wachtel, 1986; Guindon 1987; Kennedy *et al* 1988) user expectations on the discourse competence of computer systems is low. The dialogue between user and system can be modelled by an exceedingly simple dialogue grammar, by examining the discourse structure in the material obtained by carefully designed Wizard-of-Oz studies. This can be explained by a fundamental asymmetry of beliefs between user and system as posed by Aravind Joshi (1982). Users do not expect computer systems to carry on a coherent discourse, but, on the contrary, expect to take full responsibility for the discourse management themselves.

This expectation can be utilized to aid user interaction. The system will have to produce such information to the user that will support user decision making by displaying as much as possible of the system's knowledge structures. This is still just one-shot from the system's point of view: what it does in this scenario is simply to leave the responsibility to maintain dialogue structure to its users, and as an aid to the users, give them a good basis for making decisions about where to go. One approach to how this is done is *mumbling* (Karlgren, in preparation) or *commenting* (Claassen *et al* 1993), both involving output of natural language for the user to inspect. There is considerable risk that such approaches produce too much information, as already was noted in the previous section. In a multi-modal framework, there is a better chance of keeping the output within reasonable limits.

Failures to analyse an utterance in human-machine discourse should analogously be taken as situations where the user can learn something about which constructions are correctly processed and which are not. This learning process is automatic in humans, but needs to be supported by the system providing as much information as possible about the system's linguistic competence. This is what humans do in normal discourse, and the recovery of failed analyses can be regarded as a special case of normal discourse management and feedback.

For instance, if a sentence cannot be recognized as such, normally there are several parts of it that are recognized as noun phrases, verb phrases etc. How can this information be presented to the user in a comprehensive manner? In the CLARE system (Alshawi *et al* 1992) a "segmented" version of the sentence is presented to the user indicating what partial constituents were recognized. For example, the sentence

[4] As pointed out by Victor Zue.

Figure 2: Visual language query fragments

ported a “close-to” relation that could connect the two entities, creating the intended interpretation, the final visual query could subsequently be paraphrased in natural language as:

```
What colleges are close to
the centre of the city?
```

And the user would have learned about a limitation in the linguistic coverage of the system and a way to get around it.

One central feature of the multimodal discourse model is that it supports the incremental nature of natural discourse through the *persistence* of discourse objects. The objects can be made visible and accessible for subsequent discourse turns, keeping them in attentional focus available as potential *discourse sponsors* of referring expressions. In “real” dialogue we of course have other mechanisms of keeping discourse objects on top of the focus stack — the point is that there must be ways to achieve changes in attentional focus in dialogue. Linguistic objects which do not give rise to discourse objects that are kept in focus will decay and disappear as possible antecedents to referring expressions relatively quickly, so multimodality offers an extra dimension of managing the focus stack. Susann LuperFoy (1992) describes how discourse objects and linguistic objects differ. One consequence of this difference is that they support different kinds of referring expressions. For instance, definite NPs are naturally discourse sponsored, whereas anaphoric pronouns tend to be linguistically sponsored. So, an object introduced into the discourse model through the visual modality normally does not spon-

Figure 3: Visual answer

A follow-up question could then refer to these colleges, or a subset constrained as specified by the user, for instance through some type of pointing gesture (Kobsa *et al*, 1986, describe how to interpret different types of deictic pointing gestures). This gesture could be temporally synchronized with the follow-up query (complementing or substituting a referring expression), which would correspond to how pointing normally works in real-life discourse[5] But this type of gesture could have a more general function if we allow it to constitute a marking of the entities of interest to the user — the current attentional focus. This action could then precede a segment of natural language discourse, allowing for subsequent non-gestural deictic expressions:

`What are the addresses of` *these*?

The demonstrative would then be sponsored by antecedents on the focus stack introduced in the graphical modality. In fact, definite NP references, which are directly or indirectly discourse sponsored, would be possible to interpret the same way, as in:

`What are the addresses?`

Haddock (1992) also suggests enabling user control of attentional focus, through mouse clicks and enlarging and reducing graphical objects. Cohen *et al* (1989) achieve a similar effect through presenting a menu with entities that can be put in focus together with the answer to user queries.

We have now approached the notion of *foregrounding* or *thematization* in theoretical pragmatics. Some human languages have syntactic constructions that follow the topic-comment or given-new distinction reasonably closely, whereas in other languages the distinction is made by different means. In multimodal discourse this distinction will seem very natural: the user "sets the stage", or puts entities on top of the focus stack by calling them forth — either by direct manipulation or by requesting them through verbal commands — and then produces a comment about them, again by either modality.

Conclusions

We have discussed a framework for multimodal intergration of a visual language modality with natural language analysis and generation, with emphasis on the synergetic results that can be achieved, rather than on implementation details. The framework includes a visual language with a high expressive power (close to first order logic) and a cross-modal translation mechanism. Special attention has been payed to how to address problems with robustness and pragmatics through unconventional methods which aim to enable user control of the discourse management process. The ideas described here will be put to use in a number of different multimodal projects at SICS starting this fall, where the NLP component employed probably will be the Core Language Engine.

[5] Nigay and Coutaz (1993) refer to this type of concurrent, multimodal interaction as *synergistic*, and also point out that such systems, while powerful, require a sophisticated technical architecture.

References

Hiyan Alshawi, (ed.), *The Core Language Engine*, MIT Press, 1992.

Hiyan Alshawi, David Carter, Richard Crouch, Steve Pulman, Manny Rayner, and Arnold Smith, *CLARE: A Contextual Reasoning and Cooperative Response Framework for the Core Language Engine*, SRI International Report, 1992.

G.G. Bes and T. Guillotin (eds.), *A Natural Language and Graphics Interface.* ESPRIT Research Report, Project 393, ACORD, 1991.

John B. Black and Thomas P. Moran, "Learning and Remembering Command Names", *Proceedings of Human Factors in Computer Systems* Gaithersburg, 1982

Susan E. Brennan, "Conversation with and through Computers", *User Modeling and User-Adapted Interaction* 1, 1991, 67-86.

Ivan Bretan, "Enhancing the Usability of Natural Language Interfaces with Direct Manipulation", *Master's Thesis at the department of Computational Linguistics*, University of Gothenburg, Gothenburg, 1990.

Ivan Bretan, "The Role of Natural Language in Multimodal Interaction", *Proceedings of the first NORFA Doctoral Symposium on Computational Linguistics*, University of Copenhagen, Copenhagen, 1992.

Wim Claassen, Edvin Bos, Carla Huls, and Koenraad De Smedt, "Commenting on Action: Continuous Linguistic Feedback Generation", *Proceedings of the 1993 International Workshop on Intelligent User Interfaces*, ACM, 1993.

Philip R. Cohen, "The Role of Natural Language in a Multimodal Interface", *Proceedings of the ACM Symposium on User Interface Software and Technology*, Monterey, 1992, 143-150.

Philip R. Cohen, Mary Dalrymple, Douglas B. Moran, Fernando C. N. Pereira, Joseph W. Sullivan, Robert A. Gargan Jr, Jon L. Schlossberg, and Sherman W. Tyler, "Synergistic Use of Direct Manipulation and Natural Language", *Proceedings of CHI '89*, Austin, 1989, 227-233.

Nils Dahlbäck, *Representations of Discourse — Cognitive and Computational Aspects*, Linköping Studies in Arts and Science 71, Linköping Studies in Science and Technology 264, Doctoral dissertation at Linköping University, 1991.

Nils Dahlbäck and Arne Jönsson, "A System for Studying Human Computer Dialogues in Natural Language", *Research Report*, Department of Computer and Information Science, Linköping University, LiTH-IDA-R-86-42, 1986.

Susan R. Fussell and Robert M. Krauss, *Coordination of Knowledge in Communication: Effects of Speakers' Assumptions about Others' Knowledge*, manuscript from Columbia University, Department of Psychology, New York, 1990.

Raymonde Guindon, Kelly Shuldberg, and Joyce Conner, "Grammatical and Ungrammatical Structures in User-Adviser Dialogues: Evidence for the Sufficiency of Restricted Languages in Natural Language Interfaces to Advisory Systems", *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, 1987.

Nicholas J. Haddock, "Multimodal Database Query", *Proceedings of COLING '92*, Nantes, 1992, 1274-1278.

Ellen A. Isaacs and Herbert H. Clark, "References in Conversation Between Experts and Novices", *Journal of Experimental Psychology: General* 116:1, 1987, 26-37.

Aravind Joshi, "Mutual Beliefs in Question-Answering Systems", in N. V. Smith (ed), *Mutual Knowledge*, Academic Press, London, 1982.

Jussi Karlgren, *The Interaction of Discourse Modality and User Expectations in Human-Computer Dialog*, Licentiate Thesis at the

Department of Computer and Systems Sciences, University of Stockholm, 1992.

Jussi Karlgren, "Mumbling — User-driven Cooperative Interaction", *Manuscript in preparation*, SICS, Stockholm.

Alan Kennedy, Alan Wilkes, Leona Elder, and Wayne S. Murray, "Dialogue with Machines", *Cognition* 30:1, 1988, 37-72.

Alfred Kobsa, Jürgen Allgayer, Carola Reddig, Norbert Reithinger, Dagmar Schmauks, Karin Harbusch, and Wolfgang Wahlster, "Combining Deictic Gestures and Natural Language for Referent Identification", *Proceedings of COLING '86*, Bonn, 1986, 356-361.

Robert M. Krauss and Susan R. Fussell, "Perspective-taking in Communication: Representation of Others' Knowledge in Reference", in press, *Social Cognition*, 1992.

Willem J. Levelt and Stephanie Kelter, "Surface Form and Memory in Question Answering", *Cognitive Psychology* 14:1, 1982, 78-106.

Susann LuperFoy, "The Representation of Multimodal User Interface Dialogues Using Discourse Pegs", *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, 1992.

Ashok Malhotra, "Knowledge-based English Language Systems for Management Support: An Analysis of Requirements", *Proceedings of International Joint Conference on Artificial Intelligence*, 1975, 842-847.

Laurence Nigay and Joëlle Coutaz, "A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion", *Proceedings of InterCHI '93*, 1993, Vol. I, 172-178.

Sanja Petrović, "Providing Help in a Natural Language Query Interface to Relational Databases", *Yugoslav Journal of Operations Research*, 2:2, 1992, 207-218.

Michael L. Ray and Eugene J. Webb, "Speech Duration Effects in the Kennedy News Conferences", *Science* 153, 1966, 899-901.

Mohammad A. Sanamrad and Ivan Bretan, "IBM SAA LanguageAccess: A Large-Scale Commercial Product Implemented in Prolog", *Proceedings of the 1st International Conference on the Practical Application of Prolog*, 1992.

Kent Saxin Hammarström and Robert Nilsson, "A Query Interface for Visual and Natural Language", *Master's Thesis at the department of Computer Science*, University of Uppsala, Uppsala, forthcoming.

Oliviero Stock, "A Third Modality of Natural Language", *Proceedings of the 10th European Conference on Artificial Intelligence*, 1992.

Bożena Thompson, "Linguistic Analysis of Natural Language Communication with Computers", *Proceedings of COLING '80*, Tokyo, 1980, 190-201.

Tom Wachtel, "Pragmatic Sensitivity in Natural Language Interfaces and The Structure of Conversation", *Proceedings of COLING '86*,Bonn, 1986, 35-41.

Kyu-Young Whang, Ashok Malhotra, Gary H. Sockut, Luanne Burns, and Key-Sun Choi, "Two-Dimensional Specification of Universal Quantification in a Graphical Database Query Language", *IEEE Transactions on Software Engineering*, 1992, 18:3, 216-224.

Elizabeth Zoltan-Ford, "How to Get People to Say and Type What Computers Can Understand", *International Journal of Man-Machine Studies* 34, 1991, 527-547.