Interaction Models, Reference, and Interactivity in Speech Interfaces to Virtual Environments

Jussi Karlgren, Ivan Bretan, Niklas Frost, Lars Jonsson
Swedish Institute of Computer Science, KISTA, Stockholm, Sweden
diverse@sics.se

Abstract. The enhancement of a virtual reality environment with a speech interface is described. Some areas where the virtual reality environment benefits from the spoken modality are identified as well as some where the interpretation of natural language utterances benefits from being situated in a highly structured environment. The issue of interaction metaphors for this configuration of interface modalities is investigated.

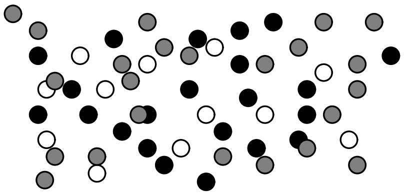
1 Introduction

Virtual reality interfaces sometimes seem to be thought of as embodying a return to a natural way of interaction – the way we interact with the real world[1]. The interaction metaphors already introduced for VR (with some trimming and tuning and the addition of proper tactile feedback...), would then be sufficient for interaction. No learning would be required, as opposed to traditional interfaces – the natural interaction mechanisms are all there. This is a familiar mistake: it has been made repeatedly in the natural language-processing community. Not until recent years has it been widely acknowledged that conventions from other human activities do not always carry over directly to interactions with computer systems. We will give some examples to show similar oversimplifications regarding virtual reality technology.

1.1 The Naming Of Things Is A Serious Matter

“This” and “that” used deictically are physical world concepts easily defined and formalized for virtual reality interfaces in the form of direct manipulation mechanisms. However, they constrain their users to the here and now, even if “here” and “now” may be defined differently than in the physical reality. Human languages are by design a step beyond “this”, “that”, “here”, and “now”. They allow the user to refer to entities other than concrete objects, using set conventions: abstract concepts (“reality”), actions (“eating”), objects that are not here (“the dog Pim”), objects that are not present now

[1] “[We are] on our own again, after the long mediation of top-down authored experience (...)”: Brenda Laurel, WIRED 1.6



"Select the grey marbles."

Figure 1: Just point and click.

("last month's salary"), objects that cannot exist ("perpetuum mobile"), and objects selected for a property ("slow things"). In general, rendering the domain of interaction in terms of physical objects is not always appropriate – many things are difficult to portray[2].

"Where is the paper about virtual reality I sent to CHI last fall?"

Figure 2: Try this with gestures.

1.2 Virtual metaphors are conventions

The virtual world does not need to obey the laws of the physical: in the real world, language is a means to change the world, and in a virtual world the world will be easier to change. Take something as simple as a virtual table. Unlike its physical relative, it can change to accommodate the preferences of the user. Similarly, the virtual world can be instructed to transport us to somewhere in the virtual space. Naturally, metaphors – a virtual saw, a virtual pot of paint, a flying carpet, superpowers – to do this with could be introduced, but they will not be more natural or less conventionally bound than use of language would be, on the contrary.

"Paint the table red and make it round."
"Take me to the moon."

Figure 3: Manipulating the world with language.

2 System sketch

Our system – DIVERSE (DIVE Real time Speech Enhancement) – is a speech interface to a generic virtual environment based on DIVE (Distributed Interactive Virtual Environment) that can be used with complex worlds modelled in a variety of formats [8]. DIVERSE allows a user to select and manipulate objects in the world and move about

[2] This is the point of playing charades.

in it. DIVERSE is implemented as a cascaded sequence of components. Speech recognition is done by means of a Hidden Markov Model system – HTK – which has been trained for the domain [21]. Text processing is performed by a general-purpose surface syntactic processor – ENGCG – which identifies syntactic roles and dependencies in the text [16, 17]. A resulting dependency graph is translated to a logical representation, which in turn is inspected for references to entities and objects and matched to the set of conceivable and possible actions. The resulting queries or commands are then sent to DIVE which manipulates or queries the world accordingly.

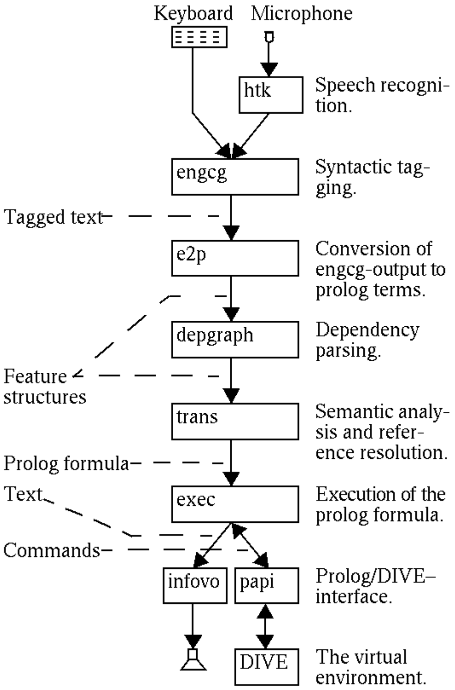


Figure 4: System architecture.

3 Interaction Metaphor

There is no obvious counterpart to the user for dialog with a system in a speech controlled virtual environment. There are several conceivable interaction models:

The basic metaphor of virtual environments is that of **Personal Presence**: the user is embodied in the real world through an actor or entity in it. This model poses problems for speech interaction – who will the user address? (“I now want to paint the house red...”) This metaphor can be extended to that of **Proxy**, where users in effect ride on the back of a virtual entity. Users share the perspective, and can address and control their proxies at will “Sindbad: paint the house red!”. An alternative similar to that of the proxy are the closely related metaphors of **Divinity**, where users give commands *as* a god to no obviously present counterpart but instead to the world itself: "Paint the house red!" or even "Let the house be red!"; or that of **Prayer** where users address commands in a similar fashion *to* a god.

Another extension of the basic metaphor of personal presence is that of **Telekinesis** where the objects and entities of the world themselves can be counterparts and interlocutors to users: "House, open your door!". Drawbacks include (1) the ability of an object or set of objects to participate in a dialog is far from obvious; (2) talking to objects not yet in the world will not be natural: "Three small red cubes, create yourselves!"; and (3) the need for object independent communication "Take me home". Of course, the last types of message could be addressed to some type of meta-object: a creation object or transportation object – in any case, the counterpart would be highly convention-bound.

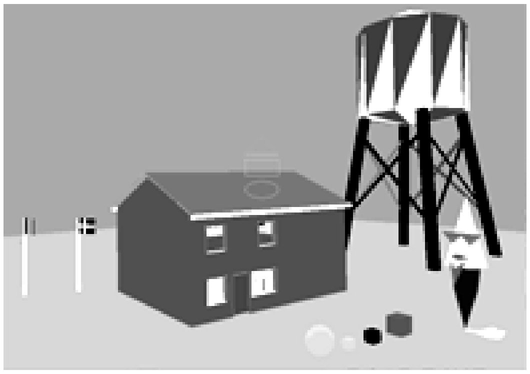


Figure 5: Interface snapshot with agent to the right.

A different type of interaction metaphor is that of an **Agent**. The agent model is different from other models in that it requires a separately rendered autonomous entity with communicative capabilities. The users will find a virtual, visually present, assistant or agent to interact with. This is necessary to be able to integrate visual and spoken feedback naturally; with no feedback or interlocutor, the interaction situation would most likely be very unfamiliar and difficult to make use of. This is the interaction model we have chosen for our implementation of DIVERSE. A consequence of machine use of a single interlocutor is that the system's linguistic competence can be modelled in this agent through its visual characteristics, its gestures, its language, and so on – this will encourage convergence in one direction. Accordingly, the DIVERSE agent has been provided with a simple vocabulary and a small set of gestures.

4 Reference resolution – pragmatics

One of the most challenging problems of language understanding is that of reference resolution: of tracking what referents referential expressions refer to.

We are not even sure of what the characteristics of referents are: we have reasonable evidence from text studies that referring expressions in the text do not refer directly to other expressions in the text itself, but to referents outside it (see e.g. Brown & Yule, [7]); similarly we have reasonable evidence that referring expressions do not refer directly to the "world", "knowledge base" or whatever we posit be the "reality" that the discourse is "about", but to some intermediate level, usually referred to as *discourse referents* [18]. We will make no claims about the characteristics of such referents: in our implementation, with the exceedingly simple task and object structure, we have yet had no need to implement an intermediate level. Our operations apply directly to the world. We may well have to add to the discourse representation in this respect if we try to add competence to the system beyond what we have now: the problems we are addressing at present will remain the same.

Resolving which discourse referent a speaker or writer refers to is non-trivial: usually there are several possible candidates. In the general case, knowledge of the domain in addition to syntactic information and access to the discourse and other aspects of the situation that the language use occurrs in are usually necessary. Brown and Yule, e.g., mention several approaches involving multiple knowledge sources [7]; an implementation by LuperFoy lists nine different sources her algorithms utilize, including Recency, Global Focus, various grammatical and lexical features, and some knowledge oriented features [20].

The knowledge sources used in the various approaches can roughly be categorized into two types: 1) situation specific features: recency, focus, and formal features of the referring expression; and 2) encyclopædic features, involving different kinds of world knowledge.

In DIVERSE we only have partial encyclopædic information. We have full knowledge of what objects exist in the world, and we have a certain hierarchical organization of objects with subparts, but there is no representation of object relations, roles, and world characteristics. We put most of our work into discourse tracking, to analyze multimodal focus.

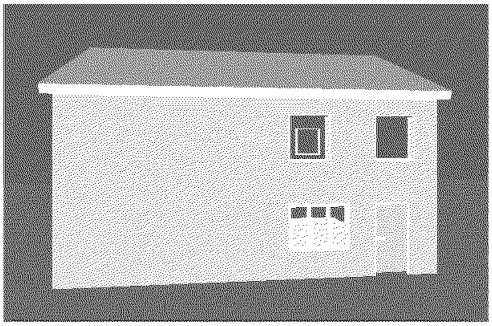


Figure 6: "Paint the house black." – What does "the house" refer to?

To concretize, the problem we need to solve is that of resolving what the referring expression “the house” in the user utterance “Paint the house black.” refers to as in figure 6, and what the referring expressions “a cube” and “it” refer to in figure 7. This is not simple in a purely text based system. Imagining that the picture were not available in figure 7: this would leave the discourse state much less explicit, and assuming a referent for “a cube” and “it” would be a risky prospect. In a visually oriented situation such as with DIVE, the attentional state of the system can be modeled by using the visual focus and highlighting mechanism of DIVE; this means that where a pure text based system might have to deliberate about different candidate cubes a multimodal system may have a less vague situation using the mutually salient information in the pictorial accompaniment.

In DIVERSE we give each object in the world a focus grade, based on recent mention, highlightedness, gestural manipulation by the user, and above all, visual awareness. So, primarily, if an object is in the perceptual focus in the virtual environment, i.e. the agent has a high degree of *awareness* of it [1, 2, 3, 12], it is a prime candidate for reference while it is visible. This effect declines rapidly when the object is not visible any more.

One of the actions available to users is to *manipulate* or *point at* an object. An object which the user points at gets a high focus grade, with a rapid rate of focus decline after the pointing gesture has been completed. Similarly, the command “Select *object*!” or even just “*Object*!” highlights the object. This is intended to be a method for users to pick out referents before issuing commands that process them.

Thirdly, we keep track of which objects have been referred to recently. If an object is in the textual discourse focus, i. e. in the recent *dialog history* it is a strong candidate for reference. An important design issue is how the dialog history is represented. To encourage users to refer to previously mentioned or manipulated objects, the discourse history can be made explicit: presumably the representations of likely candidates for reference will influence the actual references made. This will be studied empirically, with various varieties of DIVERSE implementations being compared to one another. The current version of the implementation shows a list of references above the agent’s head, as can be seen in figure 8.

The evidence from gestures, awareness status, previous commands, and discourse history is weighted together to determine which object is the one most likely to have been referred to. We expect that it will be near impossible to find a weighting of these different factors that will satisfy all users performing all kinds of tasks: instead of aiming at an “optimal” weighting we will work to find a way communicating the system evaluation to the user. We expect this to be much more efficient than trying to tune the system to accommodate users with potentially very disparate preferences and needs.

Typical problems for text based reference studies are that the prototypical case, where a definite noun phrase refers to previously introduced referents and indefinites introduce new referents, is not that frequent [13]. Thus, any algorithm for finding a referent for a definite noun phrase will need a fair amount of world knowledge to pick a contextual sponsor or anchor for the referential expression. We have found that the visual awareness factor overrides the importance of most other channels, so that in an interaction, objects can be introduced as salient just by looking at them. If the user moves to look at a tree, and then says “Move the tree to the left.” it is clear which tree is meant. And, if the visual awareness is given priority over other sources, the feedback

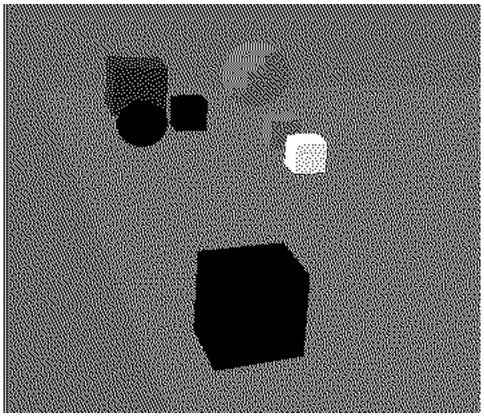


Figure 7: “Move me to a cube. Paint it black.”– Now, what does “a cube” refer to?

given the users will always give users information of what is going on.

A typical view of the drawbacks of natural language as an interface tool, be it keyboard entered or spoken, compared with direct manipulation is given by Cohen: “... another disadvantage [of natural language input] is that reference resolution algorithms do not always supply the correct answer in part because systems have underdeveloped knowledge bases, and in part because the system has little access to the discourse situation the user finds himself in, *even if the system’s prior utterances and graphical presentations have created that discourse situation*. ... These ... world knowledge limitations undermine the search for referents of anaphoric expressions and provide another reason that natural language systems are usually designed to confirm their interpretations.” [10].

Bos *et al* have implemented EDWARD, a text and direct manipulation operating system for workstations [4]. They note that users sometimes lose track of selected objects: “we found ... users not always being aware of the state of the model world: the markedness of objects selected a while ago was sometimes forgotten or overlooked.” In DIVERSE we may be able to expect slightly better user attention – visual awareness is much better determined; the view is fixed in EDWARD, whereas the user can change the view in DIVERSE, and as the visual focus overrides selection and highlighting of objects, a DIVERSE user can be expected to be more aware of the state of the model world and markedness of objects. Whatever the case may be on that count, Bos *et al* note that the mistakes the system makes do not seem to faze users; the errors are interactive enough for the user to accept them. Thus they partly answer Cohen’s objections: in a highly interactive environment, errors do not matter; at least if the interface is honest about its abilities and cooperative as to displaying them. In our design, feedback is not a matter of asking the user for confirmation, but a view of system actions.

5 Errors do not matter

The interactive design of the DIVERSE interface is related to recent trends in natural language interface research, where the underlying problem of interactive interfaces, especially natural language interfaces, today is identified as that of a low degree of interactivity or "one-shot"-interaction, where users believe – regardless of system competence – that systems expect them to pose queries in one go [5].

The conversational competence users expect from computers is extremely simple, which has been shown in a number of studies of natural language interfaces. This is specifically true for discourse structure, which has been shown to be modellable by an exceedingly simple dialog grammar, by examining the discourse structure of material obtained in Wizard of Oz simulation studies [11]. This can be explained by a fundamental *asymmetry of beliefs* between user and system [14]. Users do not expect computer systems to take responsibility for the coherence of a discourse, but expect to take full responsibility for the discourse management themselves. This is in contrast with naturally occurring dialog which is not only interactive but also *incremental*, i.e. in a form where both parties cooperatively build up referents and references during the course of a discourse.

To change this, the system must somehow display and make explicit what information it has for the user to refer to, and what assumptions about user intentions it makes; at the current point of sophistication, a high degree of interactivity and added communication channels to the system is arguably a better tool for raising system usefulness than adding functionality or intelligence to the existing channel, be it text, speech, or a rule based system [9, 15, 19].

As indicated in the previous section, in DIVERSE we make use of the errors-do-not-matter principle to the extent that we will not worry about the system misinterpreting the occasional user utterance: as long as the interface is interactive we do not expect misinterpretations to be too crucial a problem. More important than error handling is a broad acceptance of user utterances: every utterance should produce some effect.

The representation of the utterance is matched to representations of possible actions in the domain. If no good match is found, any referents that have been identified in the utterance are highlighted anyway, to facilitate users to continue the discourse, rather than starting from square one again. This is similar to recent ideas about how to generally design a natural language interface, using "non-threatening error messages that reiterate vocabulary and phrases the processor understands." as formulated by Zoltan-Ford [22].

6 Conclusions

Language is not only about conveying information[3]: it is a tool for acting in the world. Without immediacy with respect to the world it is used in, it is not natural language. Conversely, VR interaction without language does not take place in a natural or intuitive world. We are working on overcoming some of the most fundamental weaknesses of these two areas of interactive system design – through merging them.

[3] In fact, as an experiment, the reader is invited to approximate how large a percentage of language use the reader personally uses for conveying information.

7 References

1. Benford, Steve, John Bowers, Lennart Fahlén, and Chris Greenhalgh. 1994. “Managing Mutual Awareness in Collaborative Virtual Environments” *Proceedings of VRST’94,* Singapore. New York: ACM.

2. Benford, Steve, John Bowers, Lennart Fahlén, and Chris Greenhalgh. 1995. “User Embodiment in Collaborative Virtual Environments” *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI’95),* Boston. New York:ACM.

3. Benford, Steve and Lennart Fahlén. 1993. “A Spatial Model of Interaction in Large Virtual Environments” *Proceedings of 3d ECSCW* Milan: Kluwer.

4. Bos, Edwin, Carla Huls, and Wim Claassen. 1994. “EDWARD: full integration of language and action in a multimodal user interface” *International Journal of Human-Computer Studies,* **40**:473-495.

5. Bretan, Ivan and Jussi Karlgren. 1993. “Synergy Effects In Natural Language-Based Multimodal Interfaces” *Proceedings of 1993 ERCIM Workshop on Multimodal Human-Computer Interaction*, Nancy:INRIA. (also available as *SICS Research Report R94:04*.

6. Bretan, Ivan and Jussi Karlgren. 1994. “Worlds without Words”, *Proceedings of ERCIM Workshop on VR*, Stockholm:SICS.

7. Brown, Gillian and George Yule. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

8. Carlsson, Christer and Olof Hagsand. 1993. “DIVE, a Platform for Multi-User Virtual Environments”, *Computers & Graphics,* **17**:6.

9. Chandrasekar, R. and S. Ramani. 1989. “Interactive communication of sentential structure and content: an alternative approach to man-machine communication”, *International Journal of Man-Machine Studies* **30**:121-148.

10. Cohen, Philip. 1992. “The Role of Natural Language in a Multimodal Interface”, *In Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, Monterey: ACM.

11. Dahlbäck, Nils, Arne Jönsson, and Lars Ahrenberg. 1993. “Wizard-of-Oz Studies — Why and How”, *Proceedings of the 1993 International Workshop on Intelligent User Interfaces,* Orlando:ACM.

12. Fahlén, Lennart, Charles G. Brown, Olov Ståhl, Christer Carlsson. 1993. “A Space Based Model for User Interaction in Shared Synthetic Environments” *Proceedings of the ACM Conference on Human Factors in Computing Systems (InterCHI’93)* Amsterdam:ACM.

13. Fraurud, Kari. 1990. “Definiteness and the Processing of NP’s in Natural Discourse.” *Journal of Semantics* **7**:395-433.

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

15. Karlgren, Jussi, Kristina Höök, Ann Lantz, Jacob Palme, and Daniel Pargman. 1994. "The Glass Box User Model for Information Filtering", *Proceedings of the 4th International Conference on User Modeling* Cape Cod:ACM. (A longer version available as SICS Technical Report T94:09).

16. Karlsson, Fred. 1990. "Constraint Grammar for Parsing Running Text". *Papers presented to the Thirteenth International Conference On Computational Linguistics (COLING -90)*, H. Karlgren (ed.), Helsinki:University of Helsinki.

17. Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (eds.) 1995. *Constraint Grammar* Berlin: Mouton de Gruyter.

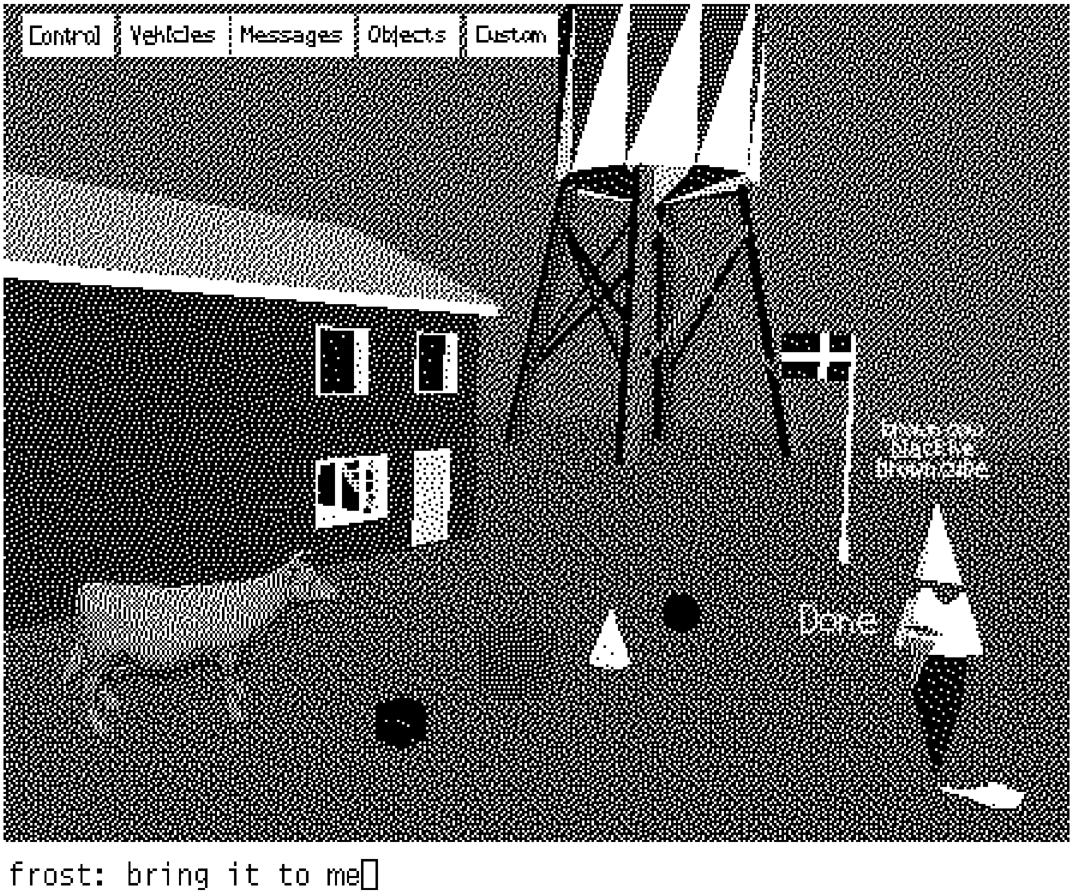
18. Lauri Karttunen. 1969 (1976). "Discourse Referents". Paper presented to the International Conference On Computational Linguistics (COLING -69), Sånga-Säby. Stockholm:KVAL. Also in James D. McCawley (ed.) Notes from the Linguistic Underground. Syntax and Semantics, Vol 7. Pp. 363-385. New York:Academic Press.

19. Lemaire, Benoît and Johanna Moore. 1994. "An Improved Interface for Tutorial Dialogues: Browsing a Visual Dialogue History". *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'94)*, Boston:ACM.

20. LuperFoy, Susann. 1991. *Discourse PEGS: A Computational Analysis of Context-Dependent Referring Expressions*. Ph D Dissertation. Austin:University of Texas at Austin.

21. Woodland, P.C., J.J. Odell, V. Valtchev and S.J. Young. 1994. "Large Vocabulary Continuous Speech Recognition Using HTK". *Proceedings of ICASSP'94*, Adelaide.

22. Zoltan-Ford, Elizabeth. 1991. "How to get people to say and type what computers can understand". *International Journal of Man-Machine Studies* **34**:527-547.



frost: bring it to me

Figure 8: Snapshot of a DIVERSE scenario.