A CONVERSATIONAL AGENT TO NAVIGATE IN VIRTUAL WORLDS

Pierre Nugues, Christophe Godéreaux, Pierre-Olivier El Guedj, & Frédéric Revolta
Institut des Sciences de la Matière et du Rayonnement
GREYC
6, boulevard du Maréchal Juin
F-14050 Caen, France
eMail: pnugues@greyc.ismra.fr

ABSTRACT

This paper describes the prototype of a spoken conversational agent embedded within a virtual reality environment. This prototype – *Ulysse* – accepts utterances from a user enabling him or her to navigate into relatively complex virtual worlds. The paper first describes what we can expect from such an interface in the communication quality between a user and virtual worlds. Then it describes Ulysse’s architecture which includes a speech recognition device together with a speech synthesizer. Ulysse consists in a chart parser for spoken words; a semantic analyzer; a reference resolution system; a geometric reasoner, and a dialogue manager. Ulysse has been integrated in a virtual reality environment and demonstrated.

INTRODUCTION

User interaction in virtual environments has almost always been undertaken with more or less sophisticated pointing devices. These devices enable to move in horizontal and vertical planes and to rotate. They also enable to point at a specific object and to “teleport” the user to it. Finally, they enable to interact with objects of the virtual world: to move them, trigger them, etc.

Navigating in virtual worlds – virtual reality – with devices such as mice, space balls, is one of the trickiest issues for new users. Certain motions are difficult and a novice user can easily get seasick with her/his “body” upside-down within a two minutes session. However, in many situations, pointing devices enable a fast and accurate interaction.

Speech interfaces are beginning to appear in virtual or simulation environments to ease interaction (Karlgren 1995; Bolt 1980; Ball 1995; Everett 1995). Spoken interaction in a virtual environment requires to complement conventional pointing devices, to coordinate both means of interaction and to leave the user the choice – the initiative – of interacting means she/he wants to use. While it does not seem desirable to try to substitute completely these devices – it is sometimes much easier to point at an object than to describe it in a verbal way – there are some situations where we prefer “to say it” rather than to “do it”.

Human-machine dialogue requires several relatively generic linguistic modules or devices such as speech recognition systems and speech synthesizers, syntactic parsers, semantic analyzers, and dialogue managers (Allen 1994a). In a virtual environment, speech is only one mode of interaction – possibly a minor one – and some adaptations must be made to classical dialogue architectures. Pointing devices must be smoothly integrated with speech. This notably implies means to resolve deictic references that is coordinated with pointing devices and hence to reason about the geometry of the scene. Beside, the architecture must be complemented by an action manager that will make the user feel comfortable with her/his “body” motion in the virtual world.

COOPERATIVE WORK, TELECONFERENCING AND, VIRTUAL REALITY

Computer Supported Cooperative Work (CSCW) gave the framework of the Ulysse project and was part of the European commission COST-14 project

(CoTech 1995). CSCW research attempts to determine how a computer can help people better work together on a project, to design a product, to take a decision, etc. across a network. CSCW tools enable notably to share documents with multi-user editing tools, to discuss design strategies using shared white boards, to communicate using real time teleconferencing: text, audio, or video.

Text teleconferencing is now widespread on the Internet and notably consists of forums you can connect to and participate in to work, discuss ideas, make friends, etc. Video teleconferencing that broadcasts participants' image is certainly an improvement over text and audio provided there are only two parties. When the number of users increase, the screen gets cluttered with the faces of the different parties and the communication tends to be difficult*: who is talking to whom?*

Spatial metaphors have been identified as a mean to improve the comprehension of teleconferencing and Internet forums. It resulted into the re-creation of meeting rooms or more complex scenes using virtual environments. Such environments enable users to meet in a virtual room, to move about or get from one room to another. There, participants are embodied within these virtual worlds using more or less realistic 3-D icons called embodiment, representation, or avatars (Benford 1995). As a result, users can immediately realize the complexity of a situation. The counterpart is that it is much more difficult to interact with the interface.

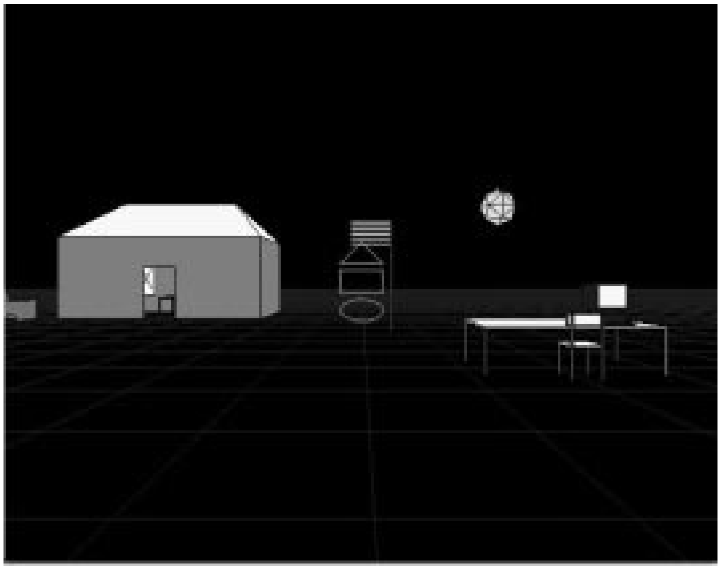


Figure 1 A Snapshot of the Ithaque World

THE TASK

We investigated spoken interaction in a virtual environment using the Distributed Interactive Virtual Environment (DIVE) (Andersson 1994) from the Swedish Institute of Computer Science (SICS). DIVE enables to build virtual worlds where users can connect from a remote location, move into, and meet other participants. Participants share the same geometric model of the world with a different point of view. Modifications of the world from user interactions are replicated to the other participant sites to keep the world consistent.

We collected a corpus of dialogues involving two experienced and two novice users. We recorded these dialogues from four interaction sessions (Godéreaux 1994) in a world – Ithaque – similar to that on Fig. 1. Each dialogue involved two participants: the interacting user and another who played the role of the agent by acting on the virtual world. We plotted two scenarios. In the first one, novice users had to move and discover the world and in the second one, more experimented users had to discover a treasure hidden in the world. Users were supposed to be alone in the world – with no other connected participants.

In comparing mouse and speech interactions, we found that mouse navigation was a major difficulty. In subsequent sessions, we even realized that most novice users were unable to go around the house. More precisely, if navigation can be relatively easy in a given plane, it is much more difficult to align or to carry out a circular motion. In addition, it is impossible to look at a specific location while moving using a single mouse. This makes some motions clumsy, for instance when the user is going round an object.

In contrast, many motions are easy to formulate verbally (Table 1; Table 2) and coordination of voice and mouse input enables a user to roam all the recesses of the virtual world. In summary, we found that dialogue interfaces can improve the usability of virtual environments. They ease navigation and bring a new channel of interaction.

A	*nous sommes connectés au monde robot.*
U8	*tourne sur toi-même.*
A	*vers la droite ou vers la gauche?*
U9	*vers la droite.*
A	*voilà.*
U10	*prends de la hauteur.*
U11	*arrête de monter.*
U12	*monte.*
A	*oui.*
U13	*stop.*

Table 1 Dialogue Excerpt

U45	*va jusque là.*
A	*je me dirige vers la montagne.*
U46	*fait le tour de la montagne.*
A	*oui.*
U47	*retourne sur l'île précédente*
A	*je ne connais pas l'île précédente.*
U48	*regarde à droite.*
A	*voilà.*
U49	*encore.*
U50	*c'est ici.*
A	*je me dirige vers la montagne.*

Table 2 Dialogue Excerpt

ULYSSE'S ARCHITECTURE

Ulysse takes the form of a conversational agent that is incorporated within the user's embodiment. Ulysse's overall structure is similar to that of many other interactive dialogue systems (Allen 1994b). It is inspired by a prototype we implemented before (Nugues 1993; Nugues 1994) and features speech recognition and speech synthesis devices, a syntactic parser, semantic and dialogue modules. Ulysse's architecture is also determined by the domain reasoner and the action manager. At the difference of TRAINS (Allen 1994b, p. 18), deindexing is closely tied to dialogue and to reasoning capabilities.

Ulysse's capabilities are relatively specific and concern only navigation. Ulysse assists the user within the world by responding positively to motion commands. Ulysse acts consequently and transports the user within the virtual environment on her/his behalf. In other projects such as (Karlgren 1995) and (Everett 1995), more general capabilities are implemented that allow the user to talk to the "world". The corresponding agents act upon the context and usually navigate (move the user embodiment), manipulate virtual objects, or answer to queries.

Understanding navigation commands requires to resolve the many deictic references that occur in the conversation and to reason about the geometry of the world. Ulysse's architecture is complemented by a reference resolver that works in coordination with the user's gestures enabling her/him to name and point at objects and a geometric reasoner to understand the world. The navigation is completed by an action manager that brings the user in a relatively continuous motion where she/he wants to go.

SYNTACTIC PARSING

Speech is recognized using the IBM's VoiceType commercial device. VoiceType is operating on isolated words – the speaker must pause between words – and is primarily intended for report dictation. We have chosen this device because it can process French and can recognize several other European languages with a vocabulary of up to 30,000 words. A chart parser is connected to the recognition device output and takes up the words. This chart (El Guedj 1994) adopts a classical bottom-up algorithm with a dual syntactic formalism: It can operate using phrase-structure rules and a dependency formalism (Tesnière 1957).

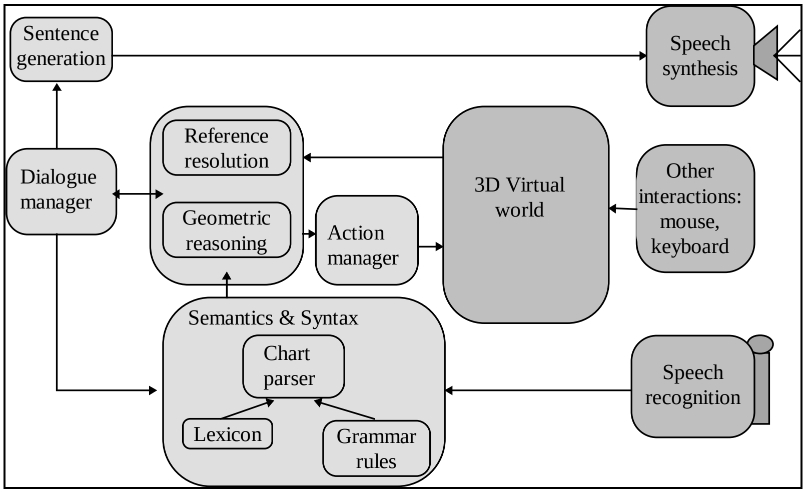


Figure 2 Ulysse System Architecture

A constituent grammar was used to encode the lexicon – 350 distinct words – and phrase-structure rules accepting all the 400 utterances of the corpus (Godéreaux 1994; Godéreaux 1996). The lexicon is using parts-of-speech that are a variation of Multext categories (Véronis 1995). We retained as features only those that were relevant for French.

Phrase-structure rules are rewriting the utterance structure using unification constraints and non terminal categories such as noun groups, verb groups, prepositional groups, determiner groups, adverb groups, adjective groups, etc. Rules were adapted to accept missing and unknown words. They include a large number of prepositional, adverbial, and demonstrative locutions that are ubiquitous in spoken language.

Utterances correspond to four main clause types: orders, questions, statements, and subordinate clauses, and also to phrases without a verb. The user segments her/his utterances using a “push-to-talk” scheme and signals the end of them by pressing a button. The analysis results in parse trees – up to eight in our corpus –. They reflect the syntactic or semantic ambiguity of the utterance.

SPEECH ACTS AND SEMANTICS PROCESSING

Utterance parse trees are first mapped to speech acts representing mainly navigation commands, such as *va dans la maison*. Other conversation acts that are identified by Ulysse are:

- deictic clarifications, such as *celle-ci*
- motion modifications, such as *plus vite*
- motion repetitions, such as *encore*

Semantic interpretation considers only navigation commands. It splits the utterance into clauses, and tags constituents from the chart parse tree with syntactic functions. Functions correspond to classical subject, object, or adjunct that are sub-classified using ontological categories. This stage also attaches modifying adverbs to their head words: verbs or other adverbs. Semantic annotation of verbs is related to the motion – the navigation – that is desired by the user and to space description. Considering our corpus and lexical sources (Bescherelle 1980), we divided them into five main navigation categories:

1. go (*aller, avancer, entrer, monter, sortir*, etc.) corresponds to a change of location with a possible rotation of the embodiment;
2. return (*revenir, retourner*, etc.) in this category, the object visibility does not matter;
3. rotate (*se tourner, regarder, pivoter*, etc.) corresponds to the rotation of the user’s embodiment head;
4. stop (*arrêter, stopper*, etc.)
5. continue (*continuer*)

Assigning these semantic tags is sometimes ambiguous. Compare: *retourne-toi* that belongs to the 3rd category and *retourne dans la maison* that belongs to the 2nd. We carried out disambiguation using verb syntactic sub-categories: i.e. transitive, intransitive, or pronominal, that we encoded as unification constraints.

As a result of this stage, each sentence is transformed in a list with as many items as there are clauses. Each clause is mapped to a structure whose members are the subject, verb group, and a list of complements. Each complement being annotated with a semantic tag: time, manner, location, etc. Verbs groups are also annotated with a motion tag and packed with possible adverbs and clitic pronouns. Ulysse maps to the same command wording differences such as:

Avance (go on)

Je veux avancer (I want to go on)

Peux-tu avancer? (Can you go on?)

Je veux que tu avances (I want you to go on)

When utterances consist into several clauses, they are concatenated and possibly rearranged according to "connectors". These connectors are associated with list operators such as append, delete, replace, or insert. Connectors can be adverbs, conjunctions, or syntactic forms. For instance negation adverb *not* in the sentence *Monte* ***non*** *descends* results into the replacement of first verb. Adverb *puis* in the sentence: *Monte sur la maison* ***puis*** *va devant l'ordinateur* results into the appending of the second action. Gerund *en passant* in the sentence: *Va vers la maison* ***en passant*** *devant le drapeau* results into the insertion of the last motion before the first one.

The logical form list is post-processed to relate it to a sequence of basic actions. According to the verb type, a clause can be expanded in one or several basic actions (up to three). For example:

monte sur le drapeau, corresponds to

1. go onto OBJECT (flag)

retourne dans la maison, corresponds to

1. turn back
2. go into OBJECT (house)

GEOMETRIC REASONING AND REFERENCE RESOLUTION

The reference resolution module de-indexes the sequence of action predicates resulting from the semantic interpretation. Object references are ubiquitous in the corpus and in all the subsequent experiments we conducted. It includes specific parts, such as: *va devant l'ordinateur, rentre dans la voiture à gauche de la maison*, plurals: *dirige-toi vers les cubes*, multiple choices: *va dans la maison* – with several houses –, and deictic sentences such as: *va ici*. References must take into consideration the state of the world database, the user's position in the world, together with the interaction history.

Associating a name to an object is sometimes tricky. Users can have a different wording to designate the same thing. Geometric databases may also consider certain objects as compounds or hierarchy although they form unique entities in the user's mind. For instance a house can be represented as a single entity, as windows, doors, walls, etc. or as a set of polygonal lines. In addition, it is important to differentiate objects that have a front and a back from other non oriented objects.

In the present prototype, we addressed the naming problem by carefully associating a name with the entities of the world database. We structured the database to keep the most consistent relations between names and world entities according to our corpus. We also gave a main orientation to each object if it could have one and references axes originating at its gravity center. The overall shape and gravity center of objects enable to compute a kind of acceptable distance to position a user relatively to an object when she/he wants to move to it: close for a small object, farther for a bigger one. Gravity centers enable also to approximate a group of objects to a unique entity.

Ulysse references objects by constructing a list of compatible entities from the geometric database when an object name occurs in an utterance. When several objects are candidate, it resolves the ambiguity using a salience algorithm similar to that of Huls (1995). Two criteria are taken into account according to the verb type:

- Visibility of objects from the user point of view;
- Focus coefficients that reflect object interaction histories.

Visibility of an object results from the intersection of the user's visibility cone with the world database. If there are still several objects that remain candidate, Ulysse considers the Karlgren's focus (1995) that is attached to each object and retain the greatest. The focus of an object is incremented each time the user interacts with it – mentions it or points at it – to become the greatest of all the foci. Although apparently simple, this resolution scheme yields accurate results.

DIALOGUE PROCESSING

The dialogue module monitors the turn taking and the sequencing of Ulysse modules. It corresponds to getting the utterances, processing them, and executing them. The dialogue module manages the syntactic ambiguities by sequentially providing the semantic interpreter with the parse trees until it finds a correct one. It then passes the clause list to the reference resolution manager.

If the references can be resolved, and there is only one solution, the list of actions is passed to the action manager. If there are several possibilities, the situation is clarified by the dialogue manager and notified to the user using a spoken message. We implemented a simple scheme to handle multi-modal clarifications using the focus coefficient. The dialogue manager asks for a pointing designation and the referencing process is repeated with the last semantic interpretation. The pointed object is designated without ambiguity since its focus is the highest of the list. If there is no referencing solution, and if there are other syntactic parse trees, the dialogue manager gets another semantic interpretation and passes it to the referencing module until the parse tree list is empty. The user is then signaled of the failure.

If the action can be completed (corresponds to implemented navigation commands), the action manager will go on and move the user's embodiment. Otherwise the dialogue manager restarts the process with the next parse tree. An example of it is the sentence *Prends de la hauteur* (Gain height) that can be a locution and also be interpreted as a transitive verb and a direct object (Take height) that can not be executed.

If an utterance corresponds to executable commands, the system will acknowledge them using a random positive message while carrying them out. Otherwise, once all the possibilities are exhausted, the dialogue manager rejects the utterance, indicating the cause. The natural language generator uses template messages and possibly selects a random one. The parser is also used by the generator to check the correctness of the agent answer.

THE ACTION MANAGER

The Action manager queries the Geometric reasoner to convert the referenced list of actions into a sequence of position coordinates. The reasoning is based on the verb category of each item of the action list. These verbs corresponds roughly to the categories Change of location and Change of posture described by (Sablayrolles 1995). According to the category, different kinds of actions are undertaken:

- go corresponds to a change of location and to a sequence of space positions.
- turn corresponds to a rotation of the whole body or of the head.
- return corresponds to a turn and to a change of location.
- stop will stop the action
- continue will resume the action

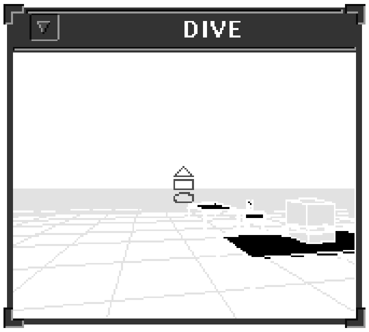
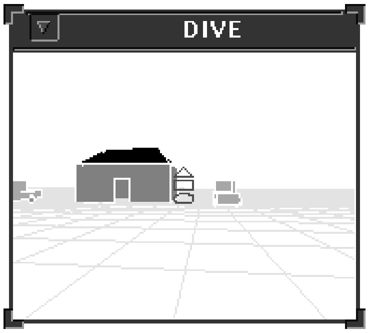
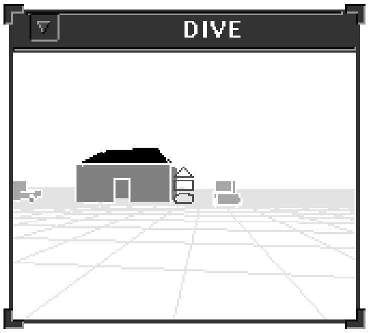
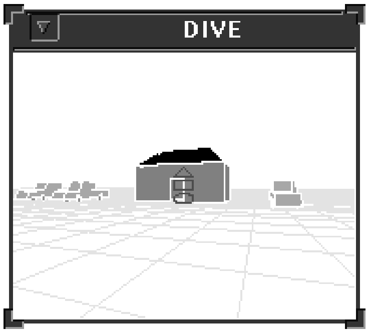
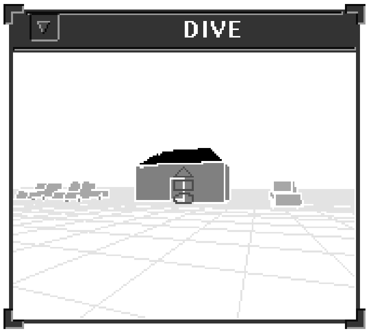
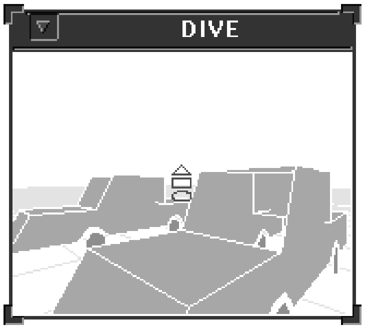
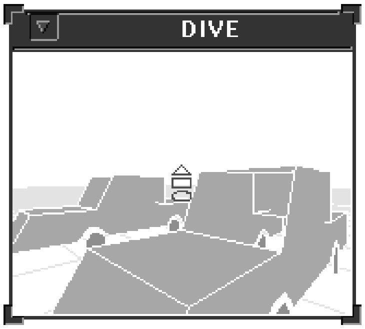
Computation of space positions takes into account the triplets verb, preposition, and object shape. It enables a user to go to a house and to stop while this house is still largely visible and go to a chair in a similar way but to get in fact closer.

In addition, head and body are articulated and rotated separately. This enables the user to move in a direction while looking in a different one. When going around a house, a simple motion could implement a four corner motion with a reasonable distance, but the user would loose the eye from the object he/she probably wants to consider. In our prototype, the sight is directed on the main object of the utterance, for example when going around a house, the user will keep an eye it. This makes the user feel more comfortable with her/his embodiment.

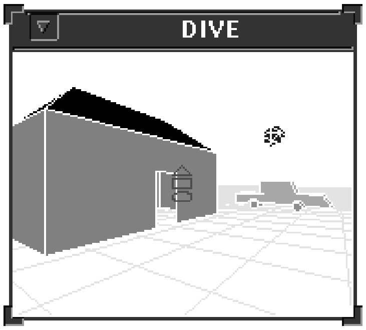
The action are implemented by dividing the action into small sub-motions using a callback function. The callback adjusts the length of the sub-motion and enables to vary the speed.

A DIALOGUE EXAMPLE

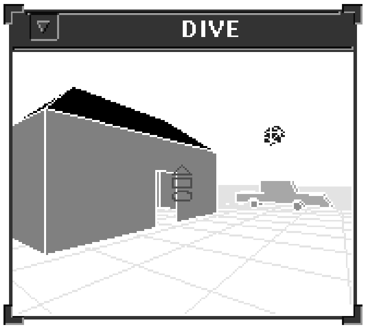
The action manager enables exchanges such as the sequence on Fig. 3.

User and agent utterances	**Snapshots**
Bonjour Fred, bienvenue dans le monde “*Ithaque*”	
Retourne toi	
Voilà	
Regarde la maison	
D’accord	
Va vers les deux voitures à gauche de la maison	
Voilà	
Tourne à droite	

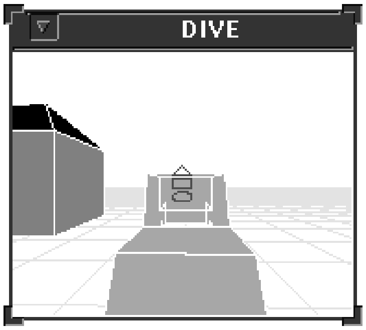
Oui



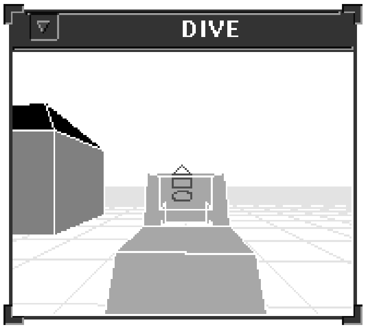
Va devant cette voiture



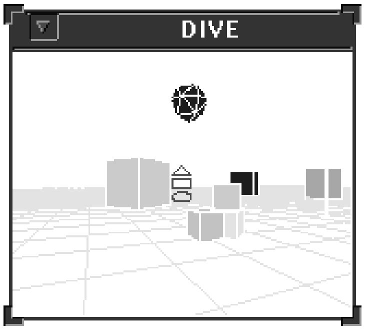
Voilà



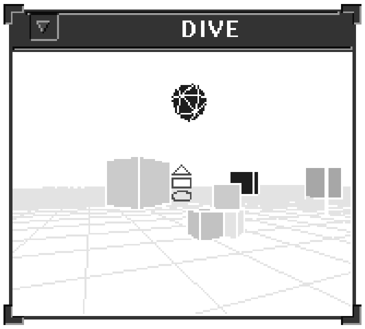
Tourne à droite



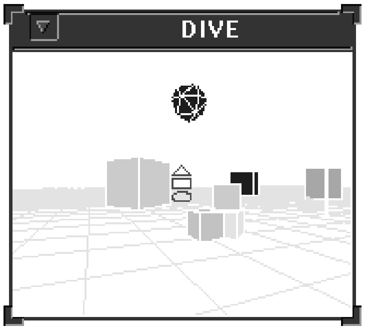
Voilà



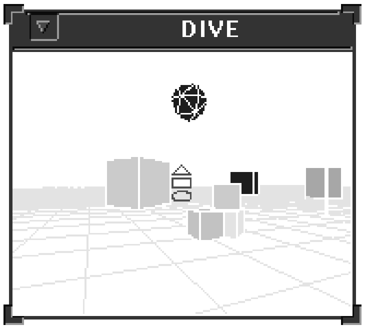
Va vers le cube



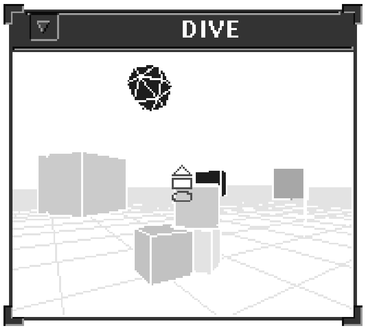
Il y en a plusieurs



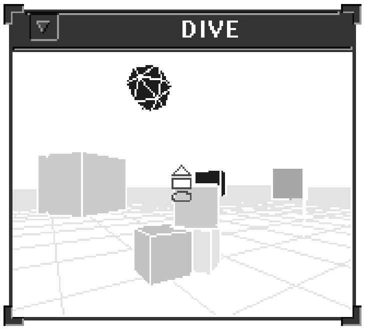
Va vers les petits cubes



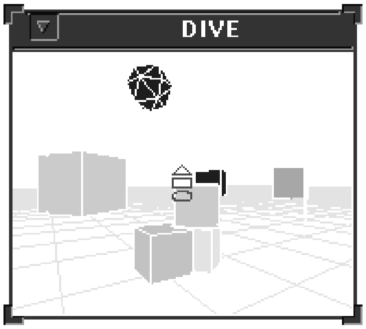
Voilà



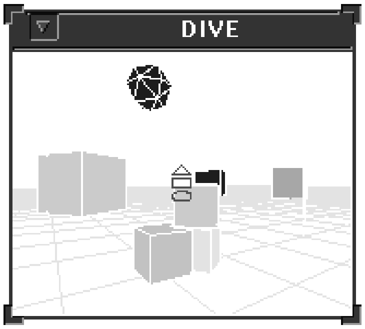
Voilà



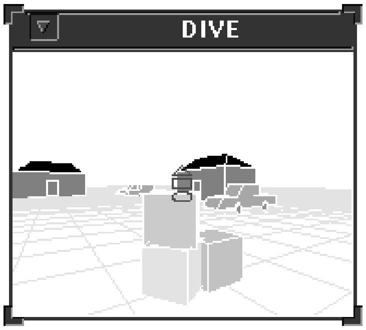
Va derrière



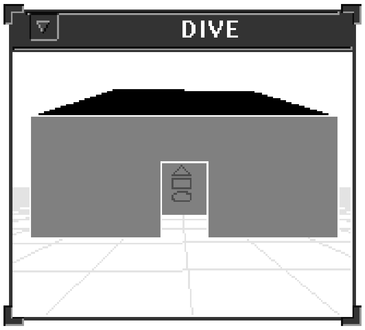
Retourne devant la maison



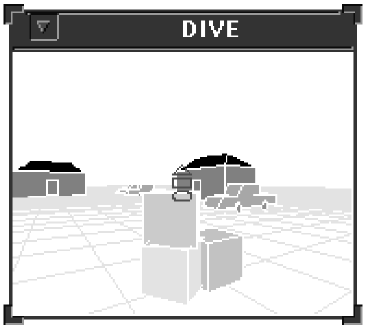
Voilà



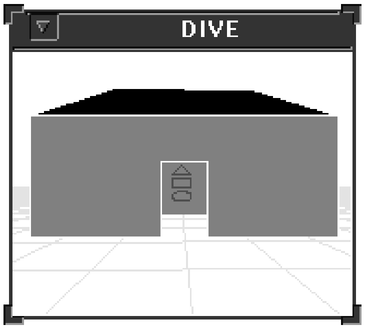
Voilà

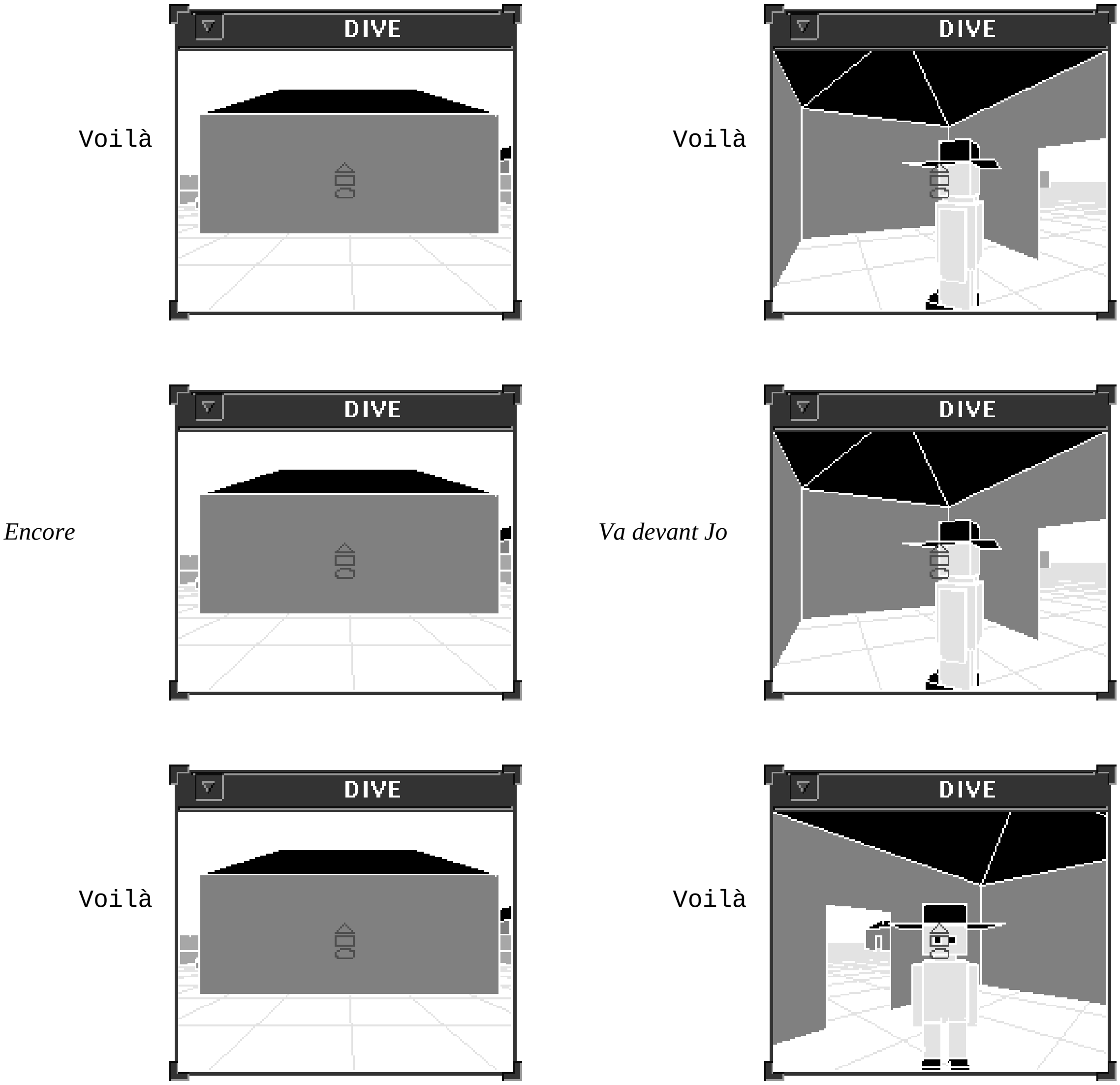


Encore



Va derrière





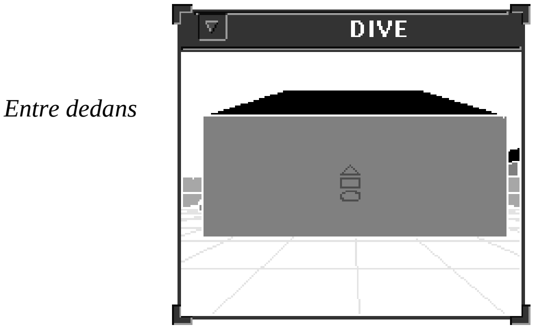


Figure 3 A Dialogue Example

CONCLUSION AND PERSPECTIVES

We have presented a conversational agent – *Ulysse* – that enables a user to carry out relatively complex motions within a virtual environment using voice. Our prototype consists in a commercial speech recognition device, together with a speech synthesis circuit. It relies on a modular architecture embedded within a virtual environment. The entities of the prototype are to process syntax and semantics, together with dialogue and actions that are resulting from them. We demonstrated the prototype to various people, notably at *La Science en Fête*, the national science open day in France where it gathered the enthusiasm of a young attendance.

This project has been developed within the framework of the COST-14 program on CSCW tools from the European commission. We think this type of conversational agent has other application perspectives. Virtual reality environments are blooming – they are now included in many Web browsers – and spoken interfaces could complement and sometimes substitute conventional pointing devices.

At Caen, we plan to adapt our agent to the virtual reconstruction of the Ancient Rome that is being undertaken from a plaster model. We are also adapting the agent to the spoken manipulation of brains reconstructed from MRI images. In conclusion, we think that this kind of agent offers perspectives to experiment tools and theories in dialogue and space linguistics.

REFERENCES

ALLEN, J.F.(a): *Natural Language Understanding*, Second edition, Benjamin/Cummings, 1994.

ALLEN, J.F., SCHUBERT, L.K., FERGUSON, G., HEEMAN, P., HEE HWANG, C., KATO, T., LIGHT, M., MARTIN, N.G., MILLER, B.W., POESIO, M. AND TRAUM, D.R. (b): The TRAINS Project: A case study in building a conversational planning agent, TRAINS Technical Note 94-3, University of Rochester, New York, September 1994.

ANDERSSON, M., CARLSSON, C., HAGSAND, O. AND STÅHL, O.: DIVE, The Distributed Interactive Virtual Environment, Technical Reference, Swedish Institute of Computer Science, Kista, Suède, March 1994.

BALL, G. ET AL: Likelike Computer Characters: The Persona Project at Microsoft Research, in *Software Agents*, J. Bradshaw ed., MIT Press, To appear.

BENFORD S., BOWERS J., FAHLÉN L, GREENHALGH C., AND SNOWDON D.: User Embodiment in Collaborative Virtual Environments, CHI'95, May 1995, Denver, Colorado, 1995.

BESCHERELLE, *L'art de conjuguer*, Hatier, 1980.

BOLT, R.A.: Put That There: Voice and Gesture at the Graphic Interface, Computer Graphics, vol. 14, n° 3, pp. 262-270, 1980.

COTECH: Minutes of the COTECH Workgroup: Virtual and Augmented Environments for CSCW, Department of Computer Science, University of Nottingham, Nottingham, England, 1995.

EL GUEDJ, P.O. ET NUGUES, P.: A chart parser to analyze large medical corpora, Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Baltimore, pp. 1404-1405, November 1994.

EVERETT S., WAUCHOPE K., PEREZ M.A.: A Natural Language Interface for Virtual Reality Systems, *Technical Report of the Navy Center for Artificial Intelligence*, US Naval Research Laboratory, Washington DC, 1995.

GODÉREAUX, C., DIEBEL, K., EL-GUEDJ, P.O., REVOLTA, F. ET NUGUES, P.: Interactive Spoken Dialogue Interface in Virtual Worlds, One-Day Conference on Linguistic Concepts and Methods in Computer-Supported Cooperative Work, London, November 1994, Proc. to appear Springer Verlag.

GODÉREAUX C., EL GUEDJ P.-O., REVOLTA F., & NUGUES P.: Un agent conversationnel pour naviguer dans les mondes virtuels, *Humankybernetik*, Band 37, Heft 1, pp. 39-51, 1996.

HULS, C., BOS, E. AND CLAASSEN, W.: Automatic Referent Resolution of Deictic and Anaphoric Expressions, Computational Linguistic, vol. 21, n° 1, pp. 59-79, 1995.

KARLGREN, J., BRETAN, I., FROST, N. AND JONSSON, L.: Interaction Models, Reference, and Interactivity in Speech Interfaces to Virtual Environments, 2nd Eurographics Workshop, Monte Carlo, Darmstadt, Fraunhofer IGD, 1995.

NUGUES, P., GODÉREAUX, C., EL GUEDJ, P.O. AND CAZENAVE, F.: Question answering in an Oral Dialogue System, In: Proceedings of the 15th Annual International Conference IEEE/Engineering in Medicine and Biology Society, Paris, vol. 2, pp. 590-591, 1993.

NUGUES, P., CAZENAVE, F., EL GUEDJ, P.O. AND GODÉREAUX, C.: Un système de dialogue oral guidé pour la génération de comptes rendus médicaux, In: Actes du 9e congrès de l'AFCET-INRIA Reconnaissance de Formes et Intelligence artificielle, Paris, vol. 2, pp. 79-88, janvier 1994.

SABLAYROLLES, P.: The Semantics of Motion, Proceedings of the 7th conference of the EACL, Dublin, 1995.

TESNIÈRE, L.: *Éléments de syntaxe structurale*, Klincksieck, 1957.

VÉRONIS J., KHOURI L.: Étiquetage grammatical multilingue : le projet MULTEXT, *Traitement Automatique des Langues*, vol. 36 n°1-2, pp. 233-248, 1995.