Should an Oral Dialogue System be Modular?

Laurent Romary, Jean-Marie Pierrel

▶ **To cite this version:**

Laurent Romary, Jean-Marie Pierrel. Should an Oral Dialogue System be Modular?. European Conference on Speech Communication and Technology (EUROSPEECH 89), Sep 1989, Paris, France. pp.2569-2572, 10.21437/Eurospeech.1989-300 . hal-00521601

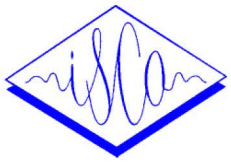
HAL Id: hal-00521601
https://hal.science/hal-00521601

Submitted on 17 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ISCA Archive
http://www.isca-speech.org/archive



First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

SHOULD AN ORAL DIALOGUE SYSTEM BE MODULAR ?

Laurent ROMARY*, Jean-Marie PIERREL**

CRIN INRIA-Lorraine, B.P. 239, 54506 Vandœuvre Lès Nancy Cedex, France
*Supelec Metz - **Université de Nancy I

Abstract : We discuss in this paper the problems bound to modular architectures, which lie at the root of most current speech understanding systems. We see why the use of different linguistic models brings about communication problems inside such systems and therefore, we detail the possible steps that could lead to an integration of this different types of knowledge. Finally, we propose a structure for a new architecture, illustrating its feasability through the study of temporal information in man-machine dialogues.

1 INTRODUCTION.

Recent developments in speech understanding systems and more precisely in man-machine dialogue systems have shown that there exists a great gap between task restricted applications using an artificial language and real natural dialogues that would involve a human and a computer. On the one hand, specific systems are concerned with situations where the potential users are supposed to be aware of possible instructions and it is thus easy to reduce the complexity of the task to be achieved. This explains that such systems have now reached a pre-industrial development in such applications as sonar console controling (1) or ground control simulation (2).

On the other hand, designing an interface using natural language and commonsense reasonning doesn't seem to be possible in a close perspective by merely enlarging the size of the language and the database to be used.

This first originates in the obvious fact that we can't lay confidence in present acoustic-phonetic decoder yet, but moreover, because the system architectures developed so far doesn't suit the purpose defined. We will try in this paper to analyse the situation thus created as follows :

- we first present the constraints for a real natural dialogue and we see why modular architectures can't meet these issues.
- we then analyse two steps towards the integration of the different levels of information in a dialogue system, as we consider it the basis for a future architecture evolution.
- we propose a frame for the analysis of man-machine dialogue, focusing our attention upon the problem of time representation and we expose the foundations of a new architecture.
- we finally see how the problem of modularity can be re-examinated in the scope of the above developments.

2 AN OVERVIEW OF DIALOGUE SYSTEMS AND MODULARITY.

2.1 MODULARITY IN CURRENT UNDERSTANDING SYSTEMS.

Understanding sentences and, more significantly, dialogues, implies that various kinds of knowledge should be taken into account : acoustic, prosodic, phonetic, phonological, syntactic, semantic, pragmatic... It is crucial for good performance of speech understanding systems to choose adequate models to represent and implement this different knowledge sources. We can find two opposite ways of defining such a system (8) :

- a single structure is defined into which every knowledge sources are integrated.
- each knowledge source works in full independance, in which case we observe a great modularity.

The first solution has been chosen for the Harpy system (12). In this system, every pieces of available knowledge from acoustic one to semantic are integrated in a single precompiled network which models the whole possible sentences, together with their phonological variants.

This technique that integrates several kinds of information proved to be particularly efficient and shows, if necessary, the advantage of precompiled knowledge in relation to interpreted one. A similar solution was chosen by Groc and Tuffeli (7) and we can finally establish a parallel with such approaches as stochastic models (10) or hidden Markov models (17).

However, this solution remains very specific and presents some major drawbacks :

- the least change in the language induces a long learning phase and a new compilation of the network,
- there is no possibility for the system to be parameterized by any of the information sources,
- finally, we can't contemplate treating that way subparts of natural language : the size of the network soon becomes unacceptable and moreover, the fact that, a priori, all variants of possible sentences are memorized implies a rigidity that is really not compatible with the freedom of natural communication.

On the other hand, a possible solution consists in defining a structure into which each knowledge source acts as an independent module that can communicate with all the others. But designing a pure heterarchical model with full interconnection of all modules is nearly impossible because of its complexity. Furthermore, although the best way for implementing those asynchronously working modules would be to make use of plain parallelism, it seems that we are not yet able to address this last issue on a computational point of view.

Several compromises have thus been proposed to allow both a modular structure and an interaction between different kinds of knowledge. Among the most important ones, we can mention :

(i) the blackboard model, in which the different knowledge sources communicate with each other through a complex database : the *blackboard*. The Hearsay II system (11) was designed that way.

(ii) the hierarchical model, that works under the control of a supervisor. The different modules, each corresponding to a knowledge source, are triggered by the supervisor and no more through a data-driven mechanism as in the preceding model. HWIM (18) makes a good exemple of such systems with explicit hierarchy.

(iii) the multi-agents model, where the modules communicate two by two through exchange processes like the producer-consumer protocol, as it is the case in the DIAL system of the CRIN.

2.2 ADVANTAGES AND DRAWBACKS OF MODULAR SYSTEMS.

The modular approach, that involves either a heterarchical, hierarchical or multi-agents model shows many advantages that have to be preserved. Among them, we can distinguish :

- the possibility of parameterizing the system at the level of each knowledge source.
- the gradual definition of those systems : each module can be defined independently of the others, as soon as its communications (data and results) are specified.

Still, a maximal cooperation between the various knowledge sources and modules lies at the root of a full success for those understanding systems. Unfortunately, this is not always possible, far from it, as the results of the

different modules appear in representations that are rarely compatible.

Whatever model is chosen, including blackboard (as shown in figure 1), each module can be finally seen as a translator from one representation (maybe incomplete) of the sensed reality into another.

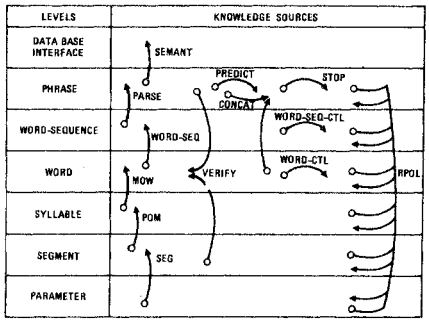


Figure 1 : The levels and knowledge sources [of Hearsay II] are indicated by vertical arcs with the circled ends indicating the input level and the pointed ends indicating output level (5).

In this way, the acoustic-phonetic module transforms the speech signal into phonemes ; the lexical module, the phonemes into words or conversely ; the syntactic module, words into structures or conversely etc... At the end of the processing, only the most abstract representation - the semantic one, for exemple - is given as a result of the recognition stage.

As a matter of fact, in order to understand and manage a dialogue, the sole vision of the most abstract level (semantic or pragmatic) is not sufficient. It would be desirable to harmonize the results representation of the different modules so as to obtain a close and efficient collaboration between each knowledge source for the making of the final analysis.

As a conclusion, even if it seems preferable to preserve a modular definition of each knowledge source, we think that it is necessary to work for a more integrated approach that leads to a cognitive representation of the results given by different kinds of linguistic knowledge. This representation shall at the same time :

- maintain a single structured memory of the results inferred by each knowledge source.
- trigger inductive resonning (bottom-up) as well as deductive ones (top down).
- obtain finally a variable depth understanding.

3 DIFFERENT STEPS FOR INTEGRATING KNOWLEDGE REPRESENTATIONS.

The general objective that we have exposed in the preceding section seems too far to be reach in a close future. However, we now show why unifying the different knowledge sources in a dialogue understanding system can already be understood and justified, considering two steps that can lead us to our purpose.

3.1 FROM WORD TO DIALOGUE.

At the base of any speech understanding system are perceptual elements given by the first stage of signal treatment and decoding, which are usually obtained as phonemes or occasionally as syllables. Those first treatments are generally mathematical and it is easy to consider them as a whole on which the other levels will construct their analyses. If we fix the phonemes as the basic elements to be manipulated, we can see that the different linguistic analyses that we may introduce in a system rest on the same kinds of mechanisms. We can put it into details for the different levels that are commonly observed.

Firstly, if we consider the recognition of words along a phonetic lattice, we see that this operation consists in retrieving specific schemes that those words are made of. Results given by psychologists show that this activity is much influenced by word frequency (6) and that it is, before all, made in a sequential way, as Marslen-Wilson (13,16) points it with his cohort model .

So, if we have to use a specific word, it can either be considered as a single conceptual entity, or, if we *decompose* it as a *sequence* of phonetic patterns, or as a sequence of phonemes at a lower level. At that point, the choice of one representation or another depends on the operation that has to be done : decomposition if a recognition is to follow or preservation as a single entity when the word is to be combined with others at a higher level.

As a matter of fact, the same reflexion can be done at the syntagmatic level. There too, psychology shows the importance of word patterns for the recognition of syntagmatic groups as well as phonemes do for words, by means of syntactic or semantic priming mechanisms (9). Once again, there exists an alternative between using a syntagm as a whole or as a sequence of following words for recognition purposes.

At a upper level, syntagmatic groups are conceptual chunks of which each sentence is made. The constraints within a sentence, either called syntactic or semantic, are definitely structural relations, resulting from usage of specific patterns. In fact those constraints can be easily compared to the ones that relate phonemes to words since, if we don't precise any pragmatic context, only usage can say wether a sequence is acceptable or not (linguistic tests are effectively based on this fact).

Finally, let us analyse the final decomposition step in a man-machine dialogue interpretation, the dialogue itself. Figure 2 shows a possible dialogue in the context of an administrative information questionning, which is the task domain of the DIAL system. This dialogue shows a general theme that caracterizes it, but it is also made of a sub-dialogue (E2) where the system is asking some precise information to the user. At each level of this dialogue, we see that there exists either structural constraints that act horizontally - for example, a question/answer scheme - or a possible vertical decomposition of a dialogue into sub-dialogues or simply into utterances.

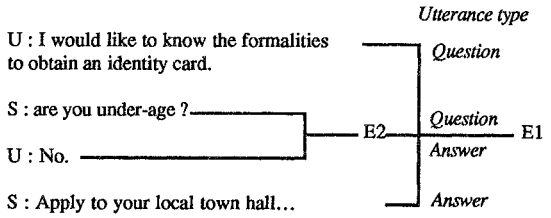


Figure 2 : An short dialogue.

Whereas those different linguistic levels are commonly treated in dialogue understanding systems as different types of information, that are represented independently, we have been able to present the important similarities that could urge us to consider them in a unified approach. Treating them that way would enormously reduce the different drawbacks of modular architectures that we have already brought to notice. However, a problem could remain with the transformation of linguistic information into conceptual one for reasonning objectives. Let us see if there exists some arguments that could favor a more advanced unification of knowledge representation that could take into account this information.

3.2 LANGUAGE AND WORLD REPRESENTATION.

Analysing utterances in a dialogue can be separated into two principal steps : first recognizing the linguistic parts of this utterance, and then, effectively understanding this utterance to situate it within the local context of a dialogue. We have already seen how the information resulting from the first stage is based on regular mechanisms, independently of the level from which it originates. Understanding a sentence is a complementary activity needed by the fact that the system has to reason on the different events that are successively exposed by the utterances of the user. As a matter of fact, it is obvious that a system cannot reason directly on linguistic information which only represents a reduced vision of situations that are always complex. Moreover, it would surely be impossible to systematically express in natural language any knowledge of the universe that would be needed by the system, especially in the case of iconic knowledge for instance.

On a pure computational point of view, this implies that an understanding system has to possess a translator of linguistic information into conceptual one and conversely, in order to present a proper functionning. Consequently, as we put it forward in our discussion on modular architecture, this can be seen a new barrier for a natural behavior, since linguistic information is no more available for a reasonning component that only manipulates conceptual data. Getting rid of this last barrier implies finding a common representation for the two types of information. From the discussion above, we can see some arguments for this solution. We can sum up the main points as follows :

- a pure psychological or physiological argument would be to admit that no difference is made in human brain between linguistic information and general conceptual information. However the discussion about human modularity is still too contrversial (3) for being addressed here. Moreover, as our aim is to design a dialogue system, it is preferable to limit ourselves on practical arguments,
- an important point in man-machine dialogue, which can be met in human dialogues too, is that preceding utterances are regularly referred to by following ones, either by means of anaphorical expressions or through meta-linguistic sentences aimed at managing the communication channel. Thus, if a system is to understand such expressions as "What is the cost *this paper*" or "Sorry, what did you say", it has to keep in memory information about the structure of the preceding utterances and not only their meaning,
- another argument concerns the construction of the semantical representation of an utterance that is not simply done after a complete linguistic analysis, but can be initiated at the very beginning of this analysis, as soon as some elementary units such as phonological marks or words are recognized. For example, let us consider the sentence : "Yesterday, he answered quickly". As a matter of fact, most parts of this sentence can lead to a semantical construction even if some other elements are still unknown. "Yesterday" can thus situate the temporal frame of the situation described or "quickly" can be linked for instance with a preceding discussion about some specific action that the speaker is supposed to refer to again in the present intervention. Those different constructions must be maintained together by referring to the corresponding part of the utterance that created them. So, linguistic knowledge must be kept in close relation with the conceptual information they bear.

As a partial conclusion, we see that there are several reasons for unifying not only the different kinds of linguistic information that appear in a speech understanding system, but more generally, any information needed by a system to reason on the discourse universe. However, this seems a very hard challenge to face and besides presenting it, it is necessary to propose some ways to reach it, at least partially. That is why we have chosen to treat a particular aspect that concerns the different levels discussed so far : time.

4 TOWARDS A COGNITIVE ANALYSIS OF MAN-MACHINE DIALOGUE.

4.1 GENERAL ASPECTS.

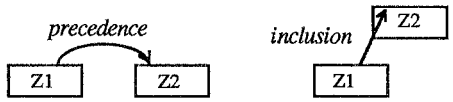
Integrating into a single representation the different information sources that appear in a man-machine dialogue means that you can actually find a common formalism that is general enough to express all of them and efficient however, so that every constraints at all levels can be expressed. There is several ways to contemplate such a formalism. We can divide them in two major approaches currently met in cognitive sciences :

- you may choose to decompose each concept into smaller entities whose various combinations will reflect the differences existing from on level to the other. This approach is the one currently taken by connexionnist models which don't make any difference between different concepts as they are all blurred along the network thus constructed.
- on the other hand we can find the symbolic approach, which tend to assign a particular entity to each concept existing in a system. The main advantages of this approach comparetively to the connexionnist one is that it is both easier to initialize a system with a priori informations (concepts and relations) and by the same way to follow the system behavior through the evolution of those concepts.

We base our research on this second approach as we state that each element in a dialogue system, either phoneme, words, dialogue or even high level concepts must be represented by a single entity : a symbol. Between those symbols, it is necessary to introduce some basic relation that will allow the system to structure its representationnal universe and, as a result, to introduce a differentiation within its concepts. Still, it seems difficult to present right now a complete and coherent list of those relations. That is why we present the results we have obtained for temporal information. Thanks to the model thus proposed, we will be able to present the first sketch of a new type of dialogue system architecture.

4.2 APPLICATION : TEMPORAL STUDY OF MAN-MACHINE DIALOGUE.

Considering temporal information means for us representing any chunk of knowledge as its projection on a temporal scale. As a matter of fact, we introdiuce a single type of objects : temporal zones. Between two zones can only exist one relation among the two we have defined : the inclusion and the precedence relation. Further on, those zones together with the relations will be schematized as follows :



The full detail of this model has been presented on other occasions (14,15) and we will essentially focus our attention on the important points for our discussion.

First of all, temporal zones can represent the different levels of linguistic information in a man-machine dialogue, since this information (phonetical, lexical etc...) can be seen as many effective linguistic acts that have a temporal aspect. Moreover each level can be decomposed, through the inclusion relation into elements of a lower level that can be occasionnaly related to each other thanks to the precedence relation as shown in figure 3.

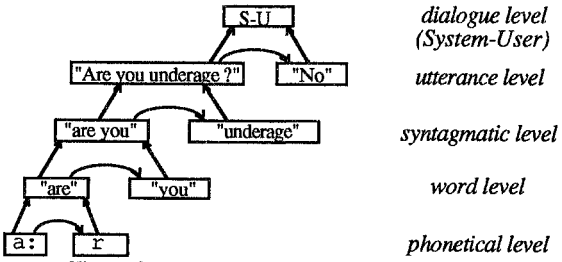
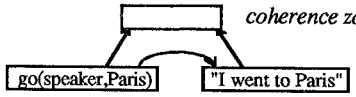


Figure 3 : illustration of a full temporal integration.

It is thus possible for each (temporal) knowledge source to bring in its contribution for the construction of a multi-level representation. Moreover, as any event expressed by the semantic of an utterance has a temporal aspect too, we can easily relate a discourse part with its semantical representation through a particular zone that we have called a *coherence zone*. For example, to the sentence "I went to Paris", can be associated a temporal zone representing the action of going to Paris, together with a temporal constraint between the time of utterance and the action time :



This association can be done at all levels of an utterance, such as phonological marks (-ed), words ("yesterday") adverbials ("in the night") as we have detailed it in (15).

We finally see that it seems possible, under certain restrictions, to exhibit a model that takes into account the different steps toward the integration of the representations usually met in a dialogue system. Obviously, the same analysis must be done on a larger scale so as to allow a real unification of this knowledge. There is still a great research field to be explored.

4.3 TOWARDS THE DEFINITION OF NEW ARCHITECTURES.

We can now wonder how such an approach can be integrated into a system architecture that puts into hand this unique knowledge representation. Such an architecture is straightforward to ellaborate and we propose a possible structure in figure 4.

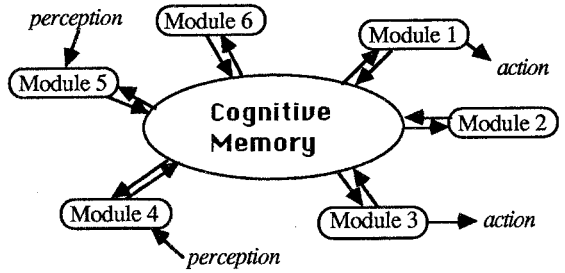


Figure 4 : a cognitive architecture.

In this figure, we can see six different modules that work on a same dynamic working memory (*cognitive memory*), into which each of them can either pick up data for an analysis or bring new knowledge to be combined with existing one so as to construct integrated representations like the one we have presented in figure 3 for temporal knowledge. In figure 4 appear several kinds of modules :

- input modules (4,5), that can integrate knowledge coming from perception levels,
- output modules (1,3), that can either operate on the universe or simply respond to the user,,
- basic modules (2,6), similar to the first two categories except that they can only base their analysis on the cognitive memory.

This architecture presents several advantages. First it avoids most of the drawbacks that we have put forward for classical modular approaches concerning the impossibility to keep in a single recognition result the different aspects that had generated it. At each level, it is not necessary that the recognition should be complete as soon as the combination of the different aspects renders it coherent enough to generate a response from an action-module. As a matter of fact, this is a very interesting aspect of this architecture to allow a close relation between the recognition and the generation phase, as those aspects are always separated in dialogue systems.

5 CONCLUSION.

In this paper we have tried to situate the present system architectures in relation to natural dialogues constraints in order to evaluate wether the modularity they present must be preserved for future systems. Several arguments proved that this problem is not straightforward. The main advantages of modular architectures are computational :

- it is easy to implement and to maintain.
- such systems can be developped without completely redesigning the whole structure.

On the other hand those architectures present some major drawbacks among which we can recall :

- the communication between the different modules is difficult because of the existence of different linguistic models and this implies that translators be put between them.
- it is difficult to combine the results given at different levels and this reduces the possibility of a global intelligent recognition of an utterance.

As a conclusion we think that there are two aspect for this problem that we can distinguish. First, modularity must be preserved on the basis of different recognition units that operate at different levels. However, it is necessary to reach an homogeneous form for knowledge representation to facilitate the construction of a single result for the recognition process that take into account the contribution of the different modules. This is why we proposed a cognitive architecture that can apparently take into account all these problems.

6 REFERENCES.

(1) P. Alinat, E. Gallais, J.P. Haton, J.M. Pierrel et P. Richard, 1987, "A continuous speech dialog system for oral control of sonar console", Proc. *IEEE ICASSP-87*, 1987.

(2) S. Bornerand, F. Néel, G. Sabah, 1988, "Un modèle de langage unifié dans un système de dialogue oral pilote/avion", *Actes des XVIIèmes J.E.P.*, Nancy, 20-22 Sept., 1988, pp.61-64.

(3) A.Caramazza, 1988, "Cognitive architecture and modularity : the view from neuropsychology", summarized in *Actes du 3ème colloque international de l'ARC*, Toulouse, 9-11 mars 1988.

(4) N. Carbonell, J.M.Pierrel, 1988, "Architecture and knowledge sources of a human computer oral dialogue system", in *Structure of multimodal dialogues*, M.M.Taylor, F.Neel and D.G.Bouwhuis Eds, North Holland, 1988.

(5) L.D.Erman and al, "The Hearsay II Speech understanding system : integrating knowledge sources to resolve uncertainty", *Computing survey*, Vol.12 n°2, pp.213-253.

(6) B. Gordon, 1985, "Subjective Frequency and the lexical decision latency function : implications for mechanisms of lexical access", *J. of memory and language*, 24, pp.631-645.

(7) B.Groc and D.Tuffelli, 1980, "A continuous speech recognition system for database consultation", *ICASSP 80*,Denver USA, PP.896-899.

(8) J.P.Haton, 1985, "Intelligence artificielle en compréhension automatique de la parole : état des recherches et comparaison avec la vision par ordinateur", *T.S.I.*, Vol.4 n°3, pp.265-287.

(9) K. den Heyer and A.Goring, 1985, "Semantic priming and word repetition : the two effects are additive", *J. of memory and language*, 24, pp.699-716.

(10) F. Jelinek, 1982, "Self-organized continuous speech recognition", in *Automatic speech analysis and recognition*, J.P.Haton Editor, D.Reidel.

(11) V.R.Lesser and al, 1975, "Organization of the Hearsay II speech understanding system", *IEEE Trans. ASSP*, 23(1),pp.11-23.

(12) B.T.Lowerre, 1976, "The Harpy speech recognition system", Tech.Report, Carnegie-Mellon University, Dept. of Computer Science.

(13) W.D.Marslen-Wilson, A.Welsh, 1978, "Processing interactions and lexical access during word recognition in continuous speech ", *Cognitive Psychology*, 10, pp.29-63.

(14) L. Romary, 1989a, "Vers la définition d'un modèle cognitif pour la représentation du temps dans un système de dialogue homme-machine", Thèse de l'université de Nancy I.

(15) L.Romary and J.M.Pierrel, 1989b, "Analyse cognitive d'expressions temporelles dans un dialogue homme-machine en langage naturel", submitted to AFCET-RFIA, Paris, 29 nov.-1 dec. 1989.

(16) J.Segui, 1988, "L'accès au lexique, données expérimentales et modèles", in *Calliope, la parole et son traitement automatique*, Masson (CNET-ENST).

(17) M.Weintraub and al, 1989, "Linguistic constraints in hidden markov model based speech recognition", *ICASSP-89*, Glasgow, pp.699-702.

(18) W.A.Woods and al, 1976, "Speech understanding system", Final Technical progress report, n°3438, I.V., BBN, 1976.