UNDERSTANDING OF SPONTANEOUS UTTERANCES IN HUMAN–MACHINE–DIALOG

Wieland Eckert

Lehrstuhl für Mustererkennung (Informatik 5),
Universität Erlangen–Nürnberg,
Martensstraße 3, D-91058 Erlangen, Germany
email: wieland.eckert@informatik.uni-erlangen.de

Abstract

In this paper discuss three basic problems of current spoken dialog systems: (1) the problem of *understanding* speech, (2) the additional problems imposed by *spontaneous* speech, and (3) the problem of *dialog* processing. We describe different definitions of the term *understanding* and propose some theses for an interpretation system. We show two principal methods for enhancing the robustness against phenomena of spontaneous speech. Then we discuss several definitions of the term *dialog* used within the speech community. Interpretation of utterances within a dialog context is necessary to resolve ambiguities. After that we discuss some factors of dialog control that have great influence on the next user utterance. Since evaluation of dialog systems is not yet standardized, we show the definition of a measure for the systems capabilities to understand utterances. This measure can be calculated automatically. Finally, we report some figures for our own spoken dialog system. The references given in this paper are expected to be a quite comprehensive starting point for further readings.

1 Introduction

Currently, we see a fairly large number of automated systems upcoming which pretend to guide a natural spoken dialog with a human [28, 32, 5, 30, 2, 3, 7, 9, 14, 17, 19, 24]. Unfortunately, everyone has got his/her own meaning for important things like *interpretation*, *understanding*, *dialog* or *quality measures*. These different views of the world are mainly caused by the fact that there is no commonly agreed standard for applications nor a common benchmark for dialog systems.

In this paper we want to shed some light on the terms *understanding* and *dialog*. While some of the theses given in the paper might be provocative, we aim to start discussions in the community about proper (and commonly agreed) definitions. In the following section we discuss possible definitions and properties of the term *understanding* in the context of spoken utterances. Aspects of understanding *spontaneous* speech are treated in section 3. In section 4 we discuss the term *dialog* in the context of spoken human–machine dialogs. After that we summarize in section 5 the current approaches to rate various aspects of dialog systems. In section 6 we present a short description of our own demonstration system and present some evaluation results.

2 Understanding Speech

A typical spoken dialog system consists of a word recognizer, a parser, and a dialog manager. While there exist different approaches (like an integrated knowledge base coupled with a search mechanism, compiled network, blackboard architecture), most of the systems mentioned above utilize this kind of modular approach. Considering these modules we can ask the questions: *what is understanding* and *in which module is it performed*? Experts in different fields have different views of the understanding process. Some of them are:

Understanding = recognizing the word sequence. An utterance might be called understood when the correct (i.e. actually spoken) sequence of words was recognized. From this point of view the words are to be identified with their meaning.

Understanding = building internal structure. Here a sequence of words is seen as just

a carrier to transport intentions from the speaker to the hearer. This means that words (or acoustic waveforms) have no meaning *per se* but could be replaced by other words (or word sequences) which carry the same intention.

Understanding = deducing all consequences. While an internal structure is just a collection of data, the effect of this data has to be considered. A suggestive example is the usage of performative verbs like in I judge you guilty!. Consequences are known and must be deduced in order to obtain all implications of an utterance.

Understanding = appropriate and intelligent reaction. In a theory of black boxes where we are just tired to explain the invisible understanding process, we can say a person or system has understood if it reacts in an appropriate and intelligent way. This might include performing actions (e.g. Stop!) or replying to a statement (e.g. Answer this question!).

This set of possibilities to define the term *understanding speech* is not intended to be complete. It should show that a variety of plausible descriptions is available and reside in different scientific fields, from philosophy to engineering. But this variety does not help any further since they *do* provide a verbal description, but they *don't* specify any methods for implementations. Since we are interested in having an operational system, we need more concrete guidelines for specifying the understanding process.

From the system engineers point of view we came across a set of heuristics which are shown below. They constitute the foundation of our demonstration dialog system described later. While these theses are to some extent *ad hoc*, unscientific, and unproven, they still seem to work quite well.

These 1 *Understanding speech is based on the interpretation of semantic information. Speech is just a carrier for pieces of semantic information.*

Everything below semantics is not considered. While the syntax is relevant to analyze an utterance, for understanding the intentional content we can safely ignore all morphological and syntactic information like words or NPs, VPs and so on. By definition the semantic content abstracts from the actual wording. We are interested in some kind of meaning, not in the surface form.

These 2 *We need a formalism to* represent *semantic information. We need a mechanism to* interpret *semantic information. We need to separate data and algorithms.*

We assume that a parser returns (mainly/only) semantic information about the utterance. Thus the information exchanged between the parser and the dialog manager has to be represented in some semantic language. On the other hand the interpretation process that "makes sense" of the users utterance has to consider dialog context and world knowledge. Therefore an interpretation process utilizing different knowledge sources is started on the semantic representation — resulting in the understanding of the utterance. Coupling the representation formalism and the interpretation mechanism in a compound knowledge base might cause difficulties regarding the maintenance of the system.

These 3 *For semantic description of utterances we need an adequate level of representation — not too simple and not too complicated.*

We hope that nearly every knowledge engineer would agree that finding a proper representation formalism is not a science but an art. In principle all representation formalisms are supposed to be of equal power. But there is never the *right* one.

These 4 *It is not useful to represent or interpret* all *possible relations of objects. Only a small part of them is meaningful and relevant in the dialog context.*

While there are approaches to make up the most general knowledge base of the world, we think that spoken dialog system does not really need to deduce everything. Applying large amounts of world knowledge would lead to increasing sets of ambiguities which are meaningless within the current domain of the system. Restriction of the system capabilities to a certain (small) world increases the effectiveness for "proper" dialogs that do not leave the application domain.

These 5 *Idioms and phrases need to be described as a whole. Ambiguities that are generated by taking the verbal interpretation are (usually) unintended.*

Idiomatic and phrasal expressions are used that often in (spontaneous) speech that they deserve simplified processing and could be easily excluded from ordinary linguistic analysis. A simple pattern matcher can assign a semantic interpretation to phrases like May I ask you a question? without performing expensive analysis steps.

These 6 *Utterances containing the same meaning should have the same representation, utter-*

ances containing a similar meaning should have a similar representation.

Apart from the idiomatic and phrasal expressions, the composition principle is a basic property of the language: further descriptions of objects are simply performed by attaching PPs or by relative clauses. Therefore the resulting semantic representation should reflect the minimal change imposed by this description by having only small parts modified. Thus, the composition principle should be employed into the semantic representation formalism.

These 7 *Understanding utterances can be performed by simple deduction rules with local scope, together with the generation of references into some environment.*

An environment is a suitable place to store initial world knowledge as well as dynamic referents. Given a structured semantic representation induced by the composition principle, we claim the existence of *simple* deduction rules. The interpretation process is applying these rules and results in chains of deduction steps.

These 8 *For the representation of elementary actions a small number of primary types is sufficient. Further distinction is performed by additional attributes.*

In a particular application domain we just need to represent a few relevant types of actions. According to [27] it is appropriate to have only 12 of them. While this might be a quite domain dependent design decision, we still believe that a small number of primary actions is sufficient.

These 9 *Not every ambiguity has to be resolved. Ambiguity in utterances might be present but irrelevant. Ambiguity might be used intentionally or systematically and must be preserved in that case.*

A well known example for a structural ambiguity is the sentence I saw the man with the telescope. While it is hard to tell the owner of the telescope, this sentence can be translated easily into, for instance, German — retaining this ambiguity. However it is a quite hard problem to decide automatically, whether some ambiguity has to be resolved or might/must remain present in the resulting semantic description.

These 10 *An understanding system is nonmonotonic. There is no "proven" knowledge; "facts" make only sense with respect to their context.*

Utterances or even parts of a single utterance might be contradictory. Since a speaker could not be forced to talk in first order logic, we have to expect contradictions. Self repairs within spontaneous speech (cf. next section) are a special case of contradiction. In the interpretation mechanism there must be provisions to revise or even "forget" objects or attributes.

These 11 *An interpretation system is incomplete. There are always propositions which could not be interpreted.*

Since individuals have different models of the world, there is currently no chance to find *the* most general model[1]. Considering the current state of the art it is useful to limit the systems capabilities to a certain small domain and a simple task. We need to accept that a spoken dialog system is allowed to fail.

These theses have quite some impact on the resulting system. By considering them we get a clearer idea of the capabilities and limits of the overall system as well as its components. Obviously, some of these theses could be discussed. This is what they are made for!

3 Spontaneous Speech

Spontaneous speech differs from *clean* language and in the theses shown above there was no provision to deal with specific phenomena observed in spontaneous speech. Common effects are (cf. [25, 33]): elliptic utterances, irregular word order, self corrections, restarts, or utterances containing multiple sentences. These effects are more often observed than regular, grammatical utterances. Everyday speech does not follow the hard rules of grammar. A more detailed analysis shows that:

- prosody and speaking speed differ from read speech,
- utterances follow a quite simple pattern with low linguistic complexity,
- users' creativity in building new utterances is very limited, they use the same words as the system (parrot syndrome) or they complete system utterances using ellipses, and
- effects of false starts, hesitations and self corrections are not systematic — they can happen at every word position within an utterance.

[1] This model must include the idea of self reference — another difficult problem.

Considering these findings, we need special provisions to automatically understand spontaneous utterances. First of all, the word recognizer has to be trained with real data, i.e. data containing an appropriate amount of these irregular utterances. Both steps of training the word models as well as the language models benefit from a sample of spontaneous data. Variations in prosody and speed are mainly incorporated into the word models. The other effects mainly influence the resulting language models. Recent advances in the field of speech recognition show that the language models have a large impact on the recognizers accuracy.

A major problem is the linguistic analysis of the effects of spontaneous speech described above. We assume that the linguistic analysis is performed by a parser, which utilizes a lexicon and a grammar. For the analysis of ungrammatical input there are two different directions:

- All variations of expected ungrammaticality are analyzed and a grammar of spontaneous speech is build by merging these additional rules with a grammar of written language.
- The grammar only contains proper rules and all ungrammaticality has to be handled by the parsers ability to deal with partial parses.

The first case seems computational expensive since it allows nearly arbitrary combinations. Using a search mechanism we have to find the best parse out of many different "ungrammatical" (but modeled!) continuations. Considering the possibility of misrecognition within the acoustic recognizer, this approach would always find an interpretation — even when processing garbage input. Actually, this approach is counterproductive when we consider the word recognizer to deliver not only the best word string but a word lattice or a word graph. We can easily image that due to a few rules of spontaneous phenomena the search space explodes.

A robust linguistic processor needs to be able to analyze partial utterances and to represent partial parses. In this case the grammar contains the clean theory not extended with rules to cover spontaneous phenomena. It is up to the parser to find maximal consistent subsequences in the recognized word string or word graph. For that purpose the parser utilizes a lexicon and a set of grammar rules that define possible combinations of words as well as the semantic representation of the phrases resulting from these combinations. Traditionally, a parser can either analyze a given input with respect to the underlying grammar or it fails. Thus, all spontaneous speech phenomena that are to be understood by the system have to be modeled in the grammar. Apart from the fact that it is quite unrealistic to foresee all types of errors, corrections etc. this approach becomes prohibitive when the word recognizer delivers a word graph instead of the best word string. Using such an interface it is the task of the parser to find the best scoring grammatical(!) path through the graph. But if the grammar models ungrammatical strings that may occur in spontaneous speech, the grammar becomes worthless for separating grammatical from ungrammatical paths through the graph. Furthermore, this approach is computationally too expensive since it allows nearly arbitrary combinations.

Therefore a less rigid parser must be used which allows partial parsing if the grammar does not permit a complete analysis of the input. Such a robust parser does not fail if no result spanning the whole input can be generated but delivers one or more partial results instead. These partial results represent grammatically well-formed utterance fields. A sequence of such utterance field objects (UFO) can then be handed over to the dialogue manager which tries to combine the parts using contextual knowledge.

4 Human–Machine–Dialog

In this section we want to clarify the third of the keywords given in the title of this paper. Please keep in mind that we restrict ourselves to task oriented spoken dialogs between a human and a machine. For the moment human–human dialogs or multimodal dialogs are not considered.

4.1 What is a Dialog?

In a dictionary [1] you get two[2] basic definitions of the word dialog: *a conversation between two or more people* and *an exchange of ideas or opinions*. Since the term *dialog* is explained by other terms (*conversation* or *exchange*) we can not make use of this definition for our problem.

On the other hand, several so called dialog systems have already been built. They are not comparable since every system developer has a different view of a dialog regarding quality of a dialog or power of their systems. In the following we distinguish three classes of dialog systems. Their

[2] The other definitions are not relevant in our context.

Type	Initiative	Response
Menu	system	direct answer, no history
Q&A	user	direct answer, limited history
Conv.	mixed	answers, questions of both partners, large history

Table 1: Different types of dialog systems.

main difference lies in the role of the dialog initiative and the sort of expected response. An short overview is given in Table 1.

Menu Systems are controlled by the system. All system utterances are of the kind Do you want choice a, b, or c? and the user is just allowed to answer directly. A typical example for this kind of systems is the automated hotline telephone service of a large company, which is typically based on touch tone recognition in spite of the presence of speech recognition technology.

Question & Answer Systems are designed in a way that the user takes the initiative and formulates a complete request. The response is a set of data that fulfills the request. In rare cases it is possible to refer to the result of the previous request. Typically, a dialog consist of one or two user turns. The original ATIS systems meet this definition.

Conversational Systems have the ability to move the initiative from one dialog partner to the other one (and back again, of course). Dialogs contain several turns and they contain requests, answers, clarifications, confirmations, and so on. In conversational systems the reaching of a complex goal is split into several steps which are performed in sequence, while in Q&A systems these steps are combined into a single exchange. The later conversational ATIS and the family of the SUNDIAL systems are typical examples for this class.

The given order shows increasing dialog complexity. The systems tasks are extended, too: the interpretation process needs to apply more knowledge for understanding an utterance, and the conversational capabilities of the dialog manager require more elaborated dialog models and access to the dialog history. Obviously, the class of conversational systems is the desired solution for speech communication systems. For other modalities we already have some examples of successful systems (e.g. automatic cash machines, flight

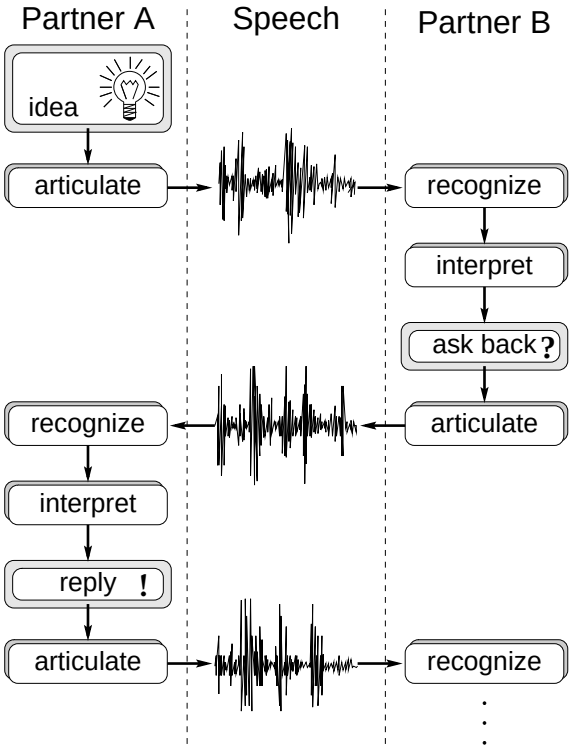

Figure 1: Understanding demonstrated by dialog behavior.

schedules information) in our everyday life. In the following we concentrate on conversational dialog systems.

With this view of a dialog system we get back to the four different definitions of the term "understanding" shown in section 2. The only plausible definition remaining is that of a proper system reaction. Figure 1 illustrates this point of view. In the following two subsections we discuss the remaining problems of interpreting an utterance and finding a proper system reaction.

4.2 INTERPRETING UTTERANCES

Interpretation of an utterance can be separated into several steps:

- A sequence of semantic descriptions is constructed by a parser which applies linguistic knowledge to the recognized word string or word graph. As shown in sections 2 and 3, in general there is no chance to find a parse covering the whole utterance in every case. Information about the order of the semantic units is needed for processing of some of the spontaneous effects, like self repairs.

- These semantic descriptions have to be embedded into an environment containing the dialog context (*anchoring*). The dialog context is used for the disambiguation of ellipses. The environment has to be dynamic (cf. These 10), it defines the focus of the current interpretation [20].
- The final step in the interpretation process is to apply the deduction rules (cf. These 7) to the anchored semantic objects. These rules represent the world knowledge and the domain knowledge of the system, and their task is to extract the meaning[3] of the utterance.

As a result we get the pragmatically relevant information conveyed in the utterance. Based on that and the current dialog state an appropriate system reaction has to be planned.

4.3 System Reaction

It remains to find a proper system reaction. When we assume that the interpretation process found the correct meaning of an utterance, we need to specify the mechanism to generate a system utterance given a dialog state and an user utterance. We just mention, that there have been proposals for rule based systems and finite automata to accomplish this task. For a user it is irrelevant by which means the system utterance was decided on. However, there are some factors to be considered which influence the users opinion about the system:

Confirmation strategies specify if the system has to ask the user for confirmation of parameters. With the current state of the art we must consider misrecognition and misunderstanding. A way to limit the bad effects of misunderstanding is to show the user parts of the internal state, i.e. to present the pieces of information found in the user utterance: Look, this is what I understood. Concerning the eloquence of the system, there are different strategies of confirmation possible: no confirmation at all, isolated confirmation of single parameters, confirmation of several parameters, and confirmation of parameters together with a new system initiative. Depending on the overall system performance one of these *static* strategies might be selected manually. Moreover, a "smart" dialog system could try to figure out the *current* understanding performance based on the number of rejections or corrections within the user utterances. This kind of *dynamic* adaptation of the confirmation strategy shows the users that the system has problems or recovers from trouble in understanding the user.

Initiative strategies have already been discussed in section 4.1: a sophisticated system is supposed to perform a mixed initiative dialog. Nevertheless, the initiative strategies might be changed dynamically according to the current understanding performance, too. A conversational system is expected to guide the dialog when the user is not doing so. An active user leading the conversation should not be restricted. Thus, the initiative strategy has to be adapted to the user according to his abilities.

Formulation of system utterance is well known to affect the users behavior. Given the same informational content, different wordings can make the system look smart or dumb. Even the quality of the synthesized speech affects the users utterances, both in content and in appearance: some users tend to mimic the systems utterances using the same words and the same prosody. A lot of these effects are already reported from WOZ experiments [11, 23].

Currently, there are no sufficient examinations of the effect of the different strategies. However, with the number of demonstration systems the corpora of spoken human machine dialogs is growing rapidly. A systematic variation of the strategies outlined above is worth to be performed. This will lead to much better models of real users and their behavior.

In order to model the system reaction, there seems to be agreement in the community to use *dialog acts* [6] to describe the users and systems intentions. In a very simple interpretation system it is sufficient to extract the parameters of each utterance that are relevant for the task. When a system evolves towards conversational capabilities, the representation of conversational intentions benefits from the usage of dialog acts. However, there is still no commonly agreed definition of the term dialog act [8, 26, 31, 22]. While we would appreciate a proper definition we still doubt that a comprehensive and complete list of dialog acts is possible and would be accepted by everyone. There are always excuses to use a different ontology or methodology.

On the other hand there seems to be agreement that the dialog planning process is determined by the most recent user utterance, the dialog state (i.e. all user and system utterances of the cur-

[3] According to [1] the *meaning* is *something that one wishes to convey, esp. by language.*

rent dialog), and the static strategy parameters. Therefore we can see the dialog state as a discrete point in the space of possible dialogs, and the generation of an system utterance is a transition in this dialog space. It is the goal of a dialog step function DSTEP to specify the subsequent dialog state for each particular point in the dialog space. The DSTEP function represents the systems dialog model, and it contains the effect of the dynamic strategies. Again, there is no common agreement on how to implement this transition function.

5 Evaluation Methodology

After building a spoken dialog system, we want to find a rating whether it is a good or a bad system. There are two principal approaches to system evaluation [29]: the *black box* evaluation methodology only considers input output behavior of the whole system, whereas in the *glass box* evaluation the intermediate results of modules are analyzed.

When analyzing the systems behavior, the crucial problem is that there is no single "reference" dialog. Judging the appropriateness of a system utterance has to be performed manually by a referee. This independent expert has to provide an annotation of each system utterance in the context of the current dialog. There could not be a reference answer since several different system reactions might "make sense" and are permitted as proper system utterances. When the annotation of each dialog is performed, the corresponding rating of appropriate system reactions can be calculated automatically. The next and most important measure for system evaluation is the resulting dialog success rate, i.e. finding out whether the users general request was satisfied. Finally, a dialog is supposed to be better when it was performed faster, both in the number of turns and the time elapsed to get the information. Unfortunately, these measures differ largely for different domains and tasks. Thus, a comparison of different systems is not easy to accomplish. First approaches to standardized system evaluation are reported in [29, 12, 16].

Evaluation of single components requires access to internal protocols of the dialog system, e.g. at module interfaces. As described above, the typical result of a word recognizer is a word string or word graph, the result of a parser is a semantic description. Since we want to deal with larger corpora of data, we prefer to have an automatic method to calculate the performance of a module. Apart from the dialog manager the other modules can be evaluated by comparing their actual result against a reference result, i.e. for the word recognizer we need the transliteration and for the parser we need a semantic annotation of the users speech. As argued above, the dialog manager could not be evaluated by comparing the system utterance with some reference utterance. For the recognizer and the parser this approach is feasible and leads to ratings that could be compared with other systems.

Word Accuracy (WA) is a widely accepted evaluation measure for word recognizers. The automatic calculation of WA for a given set of recognition results requires the existence of reference transliterations for all spoken utterances. The reference answers consist of a transcription of what was actually spoken. WA is calculated as a percentage using the formula

$$WA = 100\left(1 - \frac{W_S + W_I + W_D}{W}\right)\% \qquad (1)$$

where W is the total number of words in the transliteration, and W_S, W_I, W_D are the number of reference words which were substituted, inserted, and deleted in the recognized string, respectively. This measure is easily extended to rate the accuracy of word graphs considering their density.

Accordingly, we define the quality of a parser by calculating the *semantic concept accuracy* (CA) which considers only the information content represented by semantic units (SU):

$$CA = 100\left(1 - \frac{SU_S + SU_I + SU_D}{SU}\right)\%, \qquad (2)$$

where the semantic units are attribute-value pairs that are present in the semantic annotation. The substitutions, insertions, and deletions are counted in analogy to (1). The definition of the attributes relevant for understanding is determined by domain dependent *task parameters* which reflect the functionality of the system, and by dialog control markers for words and phrases like yes, no, good morning, could you repeat etc.

As an intermediate result we can calculate the *coverage* of the parser by measuring the semantic concept accuracy obtained on the transliteration, i.e. assuming to have a perfect word recognizer. This gives an indicator of the parsers ability to deal with phenomena of spontaneous speech[4].

[4] Interestingly, the parser does *not* need to find correct parses for all utterances. Actually there might be parts which could not be parsed. If these parts do not contain

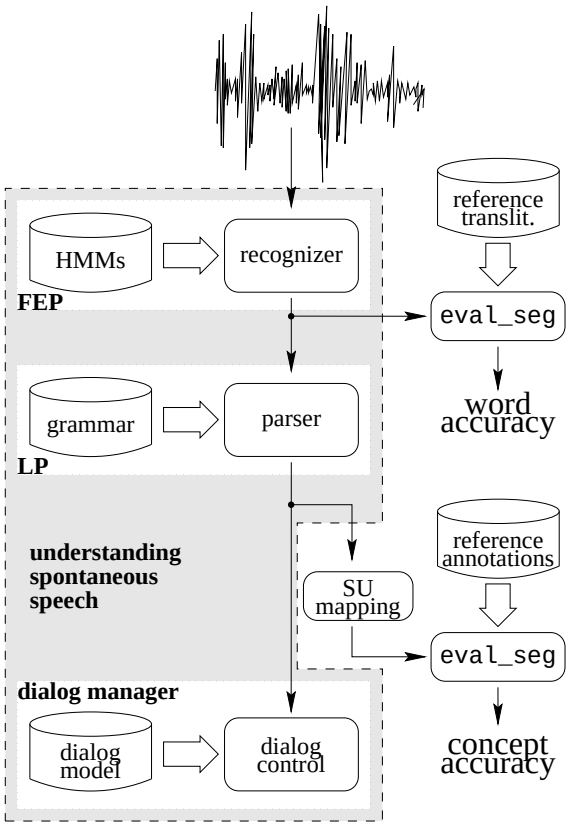

Figure 2: Architecture of the automatic evaluation system.

In order to obtain a measure for the *understanding accuracy*, the CA is calculated when feeding actual recognizer output into the parser. Thus, the chain of word recognizer and parser are evaluated as a whole which is a step further towards black box evaluation of a system. As a side effect the comparison between WA, coverage and the obtained understanding accuracy shows the robustness of the parser against recognition errors.

Utilizing the reference transliterations and the semantic annotations all these measures can be calculated in an automatic way. Figure 2 shows this procedure, where the program `eval_seg` is used to calculate the Levenshtein distance. A mapping procedure might be necessary to perform format conversion between the parser output and the semantic annotation format.

6 A Spoken Dialog System

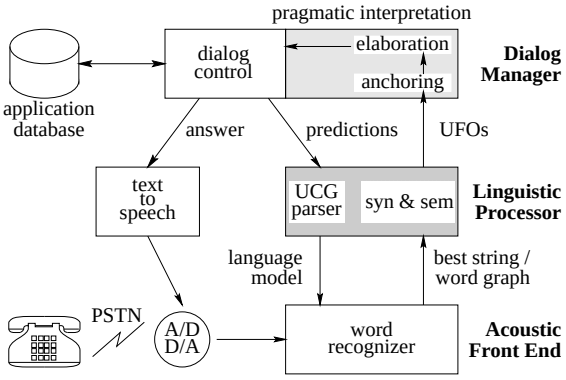

Figure 3: Architecture of our spoken dialog system.

At Erlangen University we have installed a train time table information system (T^3I) which is able to answer inquires about German intercity train connections. This system is shown in Figure 3 and contains the three major modules *word recognizer*, *parser* and *dialog manager*. This system is connected to the public telephone network and was used to collect dialogs with naive users. A first collection of about 1000 dialogs was used as a test corpus to evaluate the system and the modules (cf. [14, 4]). Table 2 gives an overview of the test corpus.

A semantic annotation scheme was developed containing the task parameter names and the dialog control marker as attributes. Together with their corresponding values they make up the semantic units for equation (2). Additional attributes had to be introduced for underspecified (elliptic) utterances, e.g. when replying a sole city name to a question on the location of departure. Disambiguation is performed by the dialog manager using the dialog context, so the semantic annotation needs to represent this ambiguity. Thus we obtained a total of 38 different classes of semantic units.

Total number of dialogues	1092
Total number of utterances	10114
Total number of words	33477
Average of words per utterance	3.31
Total number of semantic units	14584
Different classes of semantic units	38

Table 2: Figures of the test corpus.

Discussions arose when calculating the semantic concept accuracy since the semantic units consist of the *pair* of a parameter and its value. Lets

semantic information, the retrieved semantic units are not distorted and the resulting CA is not affected.

say, for example, the user had uttered

Ich möchte morgen nach Bonn
(I want to go to Bonn tomorrow) (3)

and the correct semantic annotation consists of the two SUs

[goalcity : Bonn, date : tomorrow] . (4)

A substitution of morgen (tomorrow) with the word morgens (in the morning) results in the semantic units

[goalcity : Bonn, partofday : morning] (5)

leading to a misunderstanding of both the semantic concept and its value. On the other hand the substitution of Bonn with Berlin would result in

[goalcity : Berlin, date : tomorrow] (6)

with only the value of the parameter goalcity being misunderstood. One could argue that the latter case is more severe than the previous, but the definition (2) judges both as equal[5].

Since we consider both the parameter name and the value as properties of the semantic unit, the whole unit must be recognized correctly. A quick comparison with possible word recognizer errors shows, that (2) is an appropriate measure. When counting the word errors we do *not* consider homophones to be less severe (e.g. I look in your [eyes | ice]), and we do *not* consider a mismatch in tense or gender as a less severe error. All of them are just wrong. The calculation of the semantic concept accuracy is performed in analogy resulting in an error no matter how close the result is.

First results using this dialog corpus have already been reported in [14]. In [13] we found that, while the system evolved, 53.1% of all dialogs were finished successfully. An average dialog took 154 seconds of connection time and contained 9.2 user utterances. Using this corpus with our current word recognizer, we obtained a WA of 79.1% [18]. The parser has a coverage of 92.8% on the transliterations of spontaneous speech. The sequence of word recognizer and parser results in a CA of 79.8% [4]. We found that in our case the relation between WA and CA is nearly linear, which means that the word recognizer and the parser are well matched.

7 Summary

In this paper we tackled the difficulties of defining the term *understanding*. We presented a set of theses which we think are worth to be considered when building a speech understanding system. While robustness against phenomena of spontaneous speech might be modeled explicitly, we favor the approach of generating partial descriptions. Three different definitions of the term *dialog* were presented and only the *conversational system* was found challenging for further research. Final steps of understanding an utterance are the anchoring within a dialog context and the contextual interpretation utilizing world knowledge. A dialog control mechanism which specifies the system reaction is seen as the application of a dialog step function. In this model, every dialog state is represented as a discrete point within a space of possible dialogs. Dialog strategies, e.g. regarding the confirmation or dialog initiative, are represented as parameters of the dialog step function. Experiments have shown that the actual wording of the system utterance is an important strategy parameter, too. For evaluation of understanding systems, we presented the measure *semantic concept accuracy* which is calculated in analogy to word accuracy. A short description of our own spoken dialog system, some effects observed, and the resulting figures complete this paper.

[5] Actually, it does not matter whether only the parameter name, only the value, or both are misunderstood.

8 Acknowledgements

We are very grateful to our colleagues at the University and at FORWISS, Erlangen. Parts of the research presented in this paper are funded by the Daimler-Benz Research Institute, Ulm, in the project SYSLID. Parts of this work were supported by the German Research Foundation (DFG) under contract number 810 830-0.

References

[1] *American Heritage dictionary.* Houghton Mifflin Company, second college edition, 1985.

[2] H. Aust and M. Oerder. Dialogue Control in Automatic Inquiry Systems. In Dalsgaard et al. [10], pages 125–128.

[3] A. Baekgaard et al. The Danish Spoken Language Dialogue Project — A General Overview. In Dalsgaard et al. [10], pages 89–92.

[4] M. Boros et al. Towards understanding spontaneous speech: Word accuracy vs. Concept accuracy. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, Sept. 1996. (submitted).

[5] A. Brietzmann et al. Integration of Acoustics-linguistics for a Robust Speech Dialogue System.

In *Proc. Int. Conf. on Spoken Language Processing*, Yokohama, Sept. 1994.

[6] H. Bunt. Context and Dialogue Control. *THINK*, pages 19–31, May 1994.

[7] H. Bunt et al. Cooperative Multimodal Communication in the DenK Project. In H. Bunt et al., editors, *Proc. Int. Conf. on Cooperative Multimodal Communication CMC/95*, pages 79–102, Eindhoven, May 1995.

[8] H. C. Bunt. Rules for the interpretation, evaluation and generation of dialogue acts. In *IPO annual progress report 16*, pages 99–107. Technische Universiteit, Eindhoven, 1981.

[9] R. Carlson et al. Dialog management in the Waxholm system. In Dalsgaard et al. [10], pages 137–140.

[10] P. Dalsgaard et al., editors. *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, Vigsø, Denmark, June 1995.

[11] L. Dybkjær and H. Dybkjær. Wizard of Oz Experiments in the Development of the Dialogue Model for P1. Technical Report 3, Center for Cognitive Informatics, Roskilde University, 1993.

[12] EAGLES — Spoken Language Systems. DRAFT – work in progress, Oct. 1994. (in preparation).

[13] W. Eckert. *Gesprochener Mensch–Maschine–Dialog.* PhD thesis, Universität Erlangen–Nürnberg, 1995.

[14] W. Eckert et al. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In Dalsgaard et al. [10], pages 193–196.

[15] *Proc. European Conf. on Speech Communication and Technology*, Berlin, Germany, Sept. 1993.

[16] N. Fraser. Quality Standards for Spoken Dialogue Systems: a report on progress in EAGLES. In Dalsgaard et al. [10], pages 157–160.

[17] N. M. Fraser and J. H. S. Thornton. VOCALIST: A Robust, Portable Spoken Language Dialogue System for Telephone Applications. In *Proc. European Conf. on Speech Communication and Technology*, pages 1947–1950, Madrid, Spain, Sept. 1995.

[18] F. Gallwitz et al. Integrating Large Context Language Models into a Real Time Word Recognizer. In N. Pavesic and H. Niemann, editors, *3rd Slovenian-German and 2nd SDRV Workshop.* Faculty of Electrical and Computer Engineering, University of Ljubljana, Ljubljana, Apr. 1996. (submitted).

[19] E. Gerbino et al. Analysis and Evaluation of Spontaneous Speech Utterances in Focused Dialogue Contexts. In Dalsgaard et al. [10], pages 185–188.

[20] B. J. Grosz. The Representation and Use of Focus in Understanding Dialogs. In Grosz et al., editors, *Readings in Natural Language Processing.* Morgan Kaufman Publishers, 1986.

[21] B. Hildebrand. *Struktur und Bedeutung temporaler Konstituenten in einem sprachverstehenden Dialogsystem.* PhD thesis, Universität Bielefeld, 1995.

[22] S. Jekat et al. Dialogue Acts in VERBMOBIL. VM–Report 65, Universität Hamburg, Apr. 1995.

[23] C. MacDermid. Features of Naive Callers' Dialogues with a Simulated Speech Understanding and Dialogue System. In EUROSPEECH 93 [15], pages 955–958.

[24] M. Naito et al. A real-time speech dialogue system for a voice activated telephone extension service. In Dalsgaard et al. [10], pages 129–132.

[25] D. O'Shaughnessy. Analysis and Automatic Recognition of False Starts in Spontaneous Speech. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 724–727, Minneapolis, 1993.

[26] M. D. Sadek. Dialogue Acts are Rational Plans. In M. M. Taylor, editor, *Proc. of the Venaco II workshop on "The structure of multimodal dialogue"*, La Maratea, Italy, 1991.

[27] R. C. Schank. Computers, Primitive Actions, and Linguistic Theories. In P. Eisenberg, editor, *Semantik und künstliche Intelligenz.* de Gruyter, 1977.

[28] S. Seneff et al. Interactive Problem Solving and Dialogue in the ATIS Domain. In *Proc. Speech and Natural Language Workshop*, pages 354–359, San Mateo, California, Feb. 1991. Morgan Kaufman.

[29] A. Simpson and N. Fraser. Black Box and Glass Box Evaluation of the SUNDIAL System. In EUROSPEECH 93 [15], pages 1423–1426.

[30] R. W. Smith and D. R. Hipp. *Spoken Natural Language Dialog Systems.* Oxford University Press, New York, 1994.

[31] D. R. Traum and E. A. Hinkelman. Conversation Acts in Task–Oriented Spoken Dialogue. Technical Report RR-93-32, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Dec. 1993.

[32] W. Wahlster. Verbmobil — Translation of Face-to-Face Dialogs. Technical Report RR-93-34, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, June 1993.

[33] W. Ward. Understanding Spontaneous Speech. In *Speech and Natural Language Workshop*, pages 137–141. Morgan Kaufmann, Philadelphia, 1989.