DESIGNING THE DIALOGUE COMPONENT IN A SPEECH TRANSLATION SYSTEM

—

A CORPUS BASED APPROACH

Jan Alexandersson and Norbert Reithinger *
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
{janal,bert}@dfki.uni-sb.de

ABSTRACT

New and challenging requirements arise for the dialogue processing component in the speech-to-speech translation system VERBMOBIL. It has to cope with both unexpected and vague input as well as gaps in the input. The design is based on a large corpus of transliterated dialogues. A careful analysis of this corpus and of the requirements from other components of VERBMOBIL resulted in a hybrid approach consisting of both knowledge based as well as statistic based processing. In this paper, we present the design process and the resulting architecture. Using the corpus, we made various experiments to evaluate the first design of the component.

1 INTRODUCTION

The role of the dialogue processing component in a speech-to-speech translation system like VERBMOBIL [20, 7] differs in various respects from other natural language systems with typed or speech input. One important point is that the translation system is not an active dialogue participant, except in cases where clarification dialogues between the system and the user are necessary. The users of the system interact in English and activate VERBMOBIL only if the owner of VERBMOBIL lacks knowledge of English and demands the translation of utterances in her mother tongue.

In contrast, a system like SUNDIAL [1, 14] – where the user requests travel information – is a dialogue participant that has the ability to control the ongoing dialogue to fulfill its task. It does not only monitor the dialog, it also actively engages in the interaction. In such a system the dialogue component plays an important role in controlling the overall system and the dialogue.

The application context of VERBMOBIL sets up demanding requirements for a dialogue component. In this paper we present the design process and the resulting architecture for the component and show first results for the fully implemented system. We conclude the paper with a discussion and a suggestion for further topics.

2 DESIGNING THE SYSTEM

The first application scenario for VERBMOBIL is an appointment scheduling dialogue between two business persons, one of them

*This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01IV101K/1. The responsibility for the contents of this study lies with the authors.

German, where both are non-native speakers of English. If the German partner is not able to express himself adequately he can switch to his mother tongue indicating the need for translation by pressing a button. VERBMOBIL is then expected to translate the utterances into English.

The Corpus

The empirical basis for the development of the dialogue component was a corpus of speech data. For different purposes in the development of VERBMOBIL, e.g. training the speech recognizers, a large number of German-German scheduling dialogues has been collected and transliterated [8]. Like previous approaches for modeling task-oriented dialogues, we assume that a dialogue can be modeled by means of a limited but open set of dialogue acts (see e.g. [3], [10] for speech processing and [17] for the use for machine translation). We examined this corpus for the occurrence of dialogue acts as proposed by e.g. [2, 18] and for the necessity to introduce new, sometimes problem-oriented dialogue acts.

In a first step, we defined 17 dialogue acts together with semi-formal rules for their assignment to utterances [9]. Following the assignment rules, which also served as starting point for the automatic determination of dialogue acts within the semantic evaluation component, we hand-annotated over 200 dialogues with dialogue act information to make this information available for training and test purposes. After one year of experience with these acts, the users of dialogue acts in VERBMOBIL selected them as the domain independent "upper" concepts within a more elaborate hierarchy (see figure 1) that becomes more and more propositional and domain dependent towards its leaves [5]. Such a hierarchy is useful e.g. for translation purposes.

From the analysis of the annotated corpus we derived a standard model of admissible dialogue act sequences. Figure 2 shows our dialogue model which consists of a network representation of admissible sequences of speech acts. The model for the usual sequence of dialogue acts is described in the left network; digressions that can occur everywhere in the dialogue are displayed at the right side of the main net.

This dialogue model and the acts defined therein are the basic units for the processing in the dialogue component. Main input from the other modules is based on dialogue acts for an utterance, either determined during deep processing or while spotting the English parts of the dialogue. Also information for the other components is based on the dialogue acts.

Requirements from the other Components

In a system like VERBMOBIL that combines deep analysis for translation, shallow dialogue tracking by a keyword spotter, and speech as input, the tasks of a dialogue component are manifold. The three main requirements are

1. to provide and to store contextual information which is used by the linguistic modules of VERBMOBIL e.g. the transfer component
2. to provide top down expectations about what dialogue steps are most likely to follow. This information is used to support the analysis components for narrowing down the search space which is extremely important for speech processing systems (see e.g. [14]).
3. to integrate both modes of processing within a unified approach to get a hold on the overall flow of the dialogue

In contrast to the abovementioned systems like SUNDIAL, VERBMOBIL does not control the dialogue. Therefore the dialogue component cannot take over the part it plays in these other systems, namely guiding the user so that the information he gives can finally

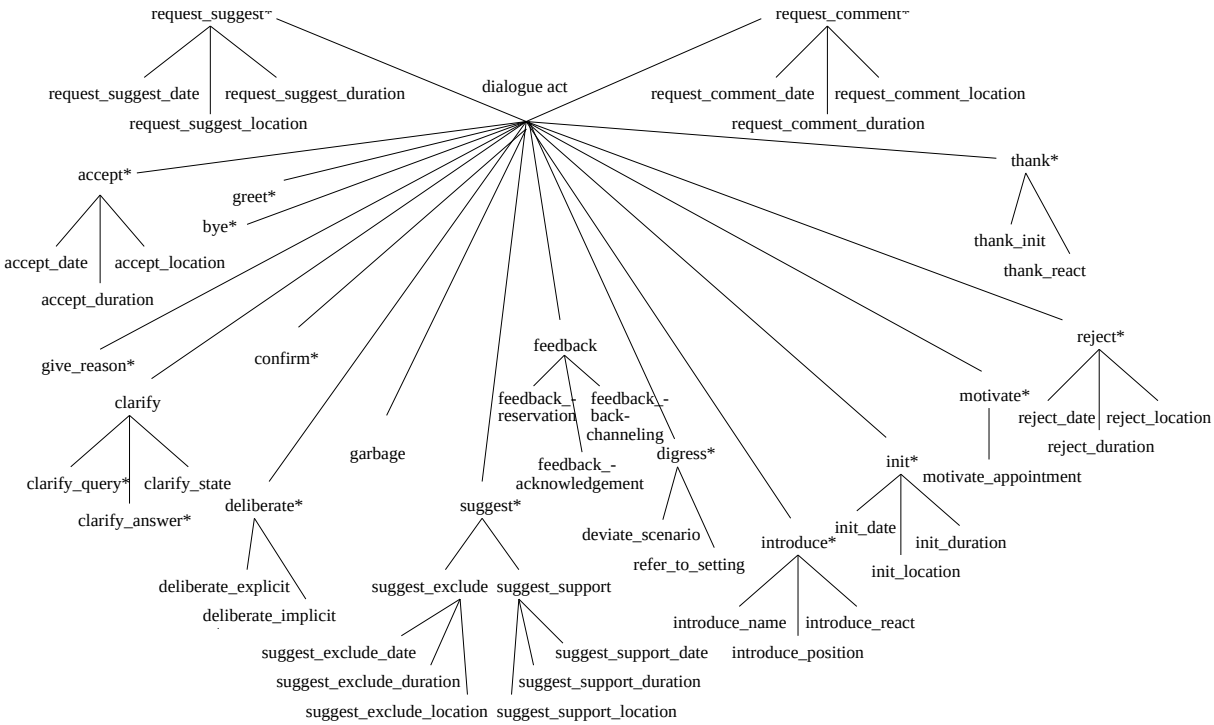

Figure 1: The taxonomy of dialogue acts

be used (e.g. to return a scheduling information). In the VERBMOBIL scenario the dialogue is between the two humans, and VERBMOBIL is only a tool for one of them.

The Internal Structure

To store the contextual information, we follow the approach of [4] for modeling the context, and describe it by three interconnected knowledge sources

- an intentional structure, a tree-like structure which contains information about the intentions for parts of the dialogue. This information is used amongst others to determine the dialogue act of the actual utterance.
- a thematic structure which represents local and global focus and the development of the different topics mentioned in the dialogue. It is for instance used by the transfer component.
- a referential structure which links the conceptual and language-related information for the objects mentioned. One example of the application of this knowledge is the generation of noun phrases in the target language.

To build and maintain these structures and to provide predictions, we had to select the appropriate processing mechanisms.

The first and obvious step for the implementation is to put the dialogue model into software, i.e. to implement it as an automaton. This is a technique frequently to be found in speech processing systems. It is a simple way to follow and control the flow of discourse. However, as VERBMOBIL does not actively participate in the dialogue, it has no control over the dialogue steps, and cannot rely on a reasonable sequence of dialogue acts, as it is e.g. the case in travel information systems. Also, the first two of the abovementioned requirements are hardly met with such a simple model.

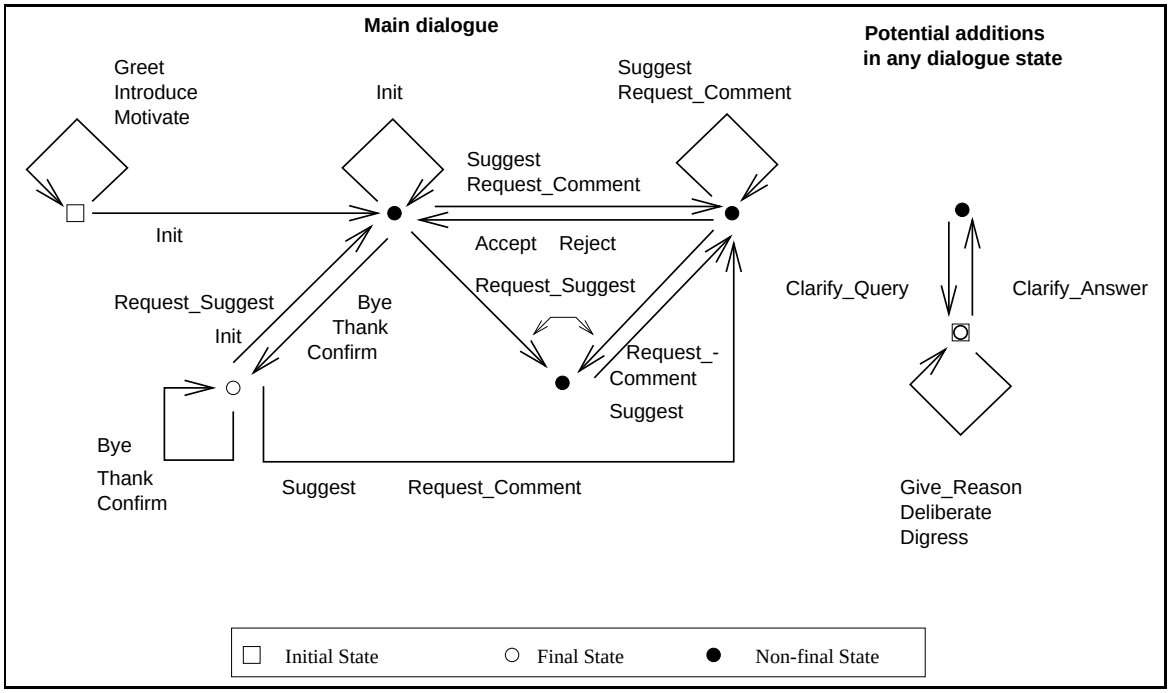

Figure 2: The Dialogue Model

We therefore divided processing up into two parts. One is responsible for building up the intentional and thematic structure, and one for the prediction process.

The development of the knowledge structures is the task of a plan recognizer. Its input consists of the dialogue acts and the propositional content expressed by the utterance when deep processing takes place. The leaves of the intentional structure are the acts, while the intermediate nodes represent subsections of the dialogue like the greeting or negotiating phase. This tree is built up incrementally, as the dialogue acts are provided by the other components of VERBMOBIL.

The source of the dialogue acts can either be deep linguistic processing in cases where one dialogue partner presses the activation button and demands a translation, or either the keyword spotter which tracks the English parts superficially. This spotter is one component provided with predictions for follow-up dialogue acts.

Structural knowledge sources are usually useless for prediction purposes since they provide too many, unscored predictions. To compute weighted dialogue act predictions we evaluated two methods: The first method is to attribute probabilities to the arcs of our network by training it with annotated dialogues from our corpus. The second method adopts information theoretic methods from speech recognition. We implemented and tested both methods and currently favor the second one because it is insensitive to digressions from the dialogue structure as described by the dialogue model and generally yields better prediction rates (see below).

In the next section, we describe, how these two processing approaches can be combined to form a synergetic processing environment.

3 ARCHITECTURE

Figure 3 shows an overview of the internal structure of the dialogue component. In the middle the three processing modules *plan rec-*

ognizer, *finite state machine (FSM)* and *statistics* are given. On the left and the right hand side the different components we interact with and the type of interaction we perform are sketched. Below we first present the implementation of the three different components, and in the next section we present their evaluation.

The Finite State Machine

The finite state machine uses the dialogue model (see figure 2) as knowledge base. It parses the incoming dialogue acts and checks whether the ongoing dialogue is in accordance with the dialogue structure as defined in the networks. If the incoming dialogue act is not within the language generated by the model, the FSM uses a fall-back strategy, based on the information in the statistics component, to find the most probable follow-up state.

We have also tried to use the FSM to predict the next dialogue act to come but as we will see in section 4 the results are quite poor.

Currently, we use this module only to drive the graphical user interface as one means to visualize the ongoing dialogue.

The Statistics Component

To provide weighted predictions for the most probable follow-up speech acts, we use a statistical approach. A knowledge source like the network model cannot be used for this task since the average number of predictions in any state of the main network is 10 when including the speech acts that can occur everywhere in the dialogue. Given a total number of 17 dialogue acts this is not sufficiently restrictive, especially when there is no ranking information attached.

The statistical method is based on n-gram speech act probabilities, a method adapted from speech recognition (see [6, 14, 13], and [15] for an evaluation of the method used in VERBMOBIL).

The task of this module is to compute those dialogue acts that will most probably follow a given dialogue history, i.e. to compute

$$s_n := \max_s P(s|s_{n-1}, s_{n-2}, s_{n-3}, ...)$$

To approximate the conditional probability $P(.|.)$ the standard smoothing technique known as deleted interpolation is used [6] with

$$P(s_n|s_{n-1}, s_{n-2}) = q_1 f(s_n) + q_2 f(s_n|s_{n-1}) + q_3 f(s_n|s_{n-1}, s_{n-2})$$

where f are the relative frequencies of dialogue act uni-, bi-, and trigrams computed from a training corpus and $\sum q_i = 1$. The weights q_i are determined with the HMM-based algorithm described in [6].

In [15] we show that a training corpus size of about 50 dialogues is sufficient to get an approximation that is based mainly on bigrams. I.e. q_2 is large compared to the uni- and trigramm coefficients. The annotated corpus is currently too small to allow for stable frequencies of higher n-grams.

The Plan Recognizer

The plan recognizer explores the connection between plan recognition and parsing as pointed out by Vilain [19]. In his paper he shows how a plan hierarchy can be compiled into a context free grammar. This approach is convenient since parsing is a well understood technique with a lot of both fast and robust approaches. Our first version is basically a simple top down parser with backtracking where the plan operators are processed strictly left-to-right (mimicking the behaviour of a prolog interpreter).

The plan recognizer operates on a set of *plan operators*. In our perspective, however, they are rules forming a grammar. The rules are used to encode both the dialogue model and methods for recovery from erroneous dialogue states. The latter is especially important: even when the dialogue partners deviate from a well-formed dialogue as defined in

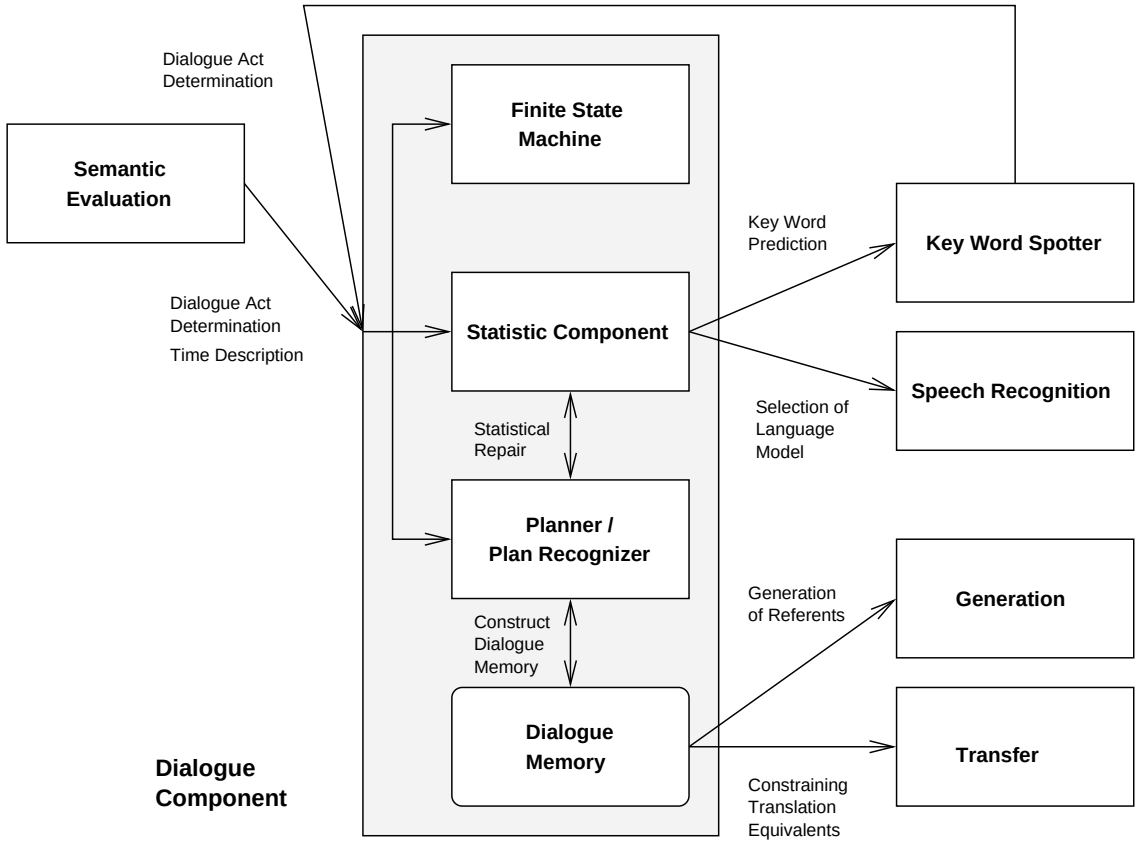

Figure 3: The Architecture of the Dialogue Module

the dialogue model, the planner has to continue to construe the intentional structure of the dialogue.

Each rule represents a specific `goal` which it is able to fulfill in case specific `constraints` hold. These constraints mostly address the context, but they can also be used to check pragmatic features, like e.g. whether the dialogue participants know each other. Also, every plan operator can trigger follow-up `actions`. To be able to fulfill a goal a plan operator can define `subgoals` which have to be achieved in a pre-specified order (see for instance [11, 12] for comparable approaches).

One of the big problems with parsers is that they are recognizers – they either accept or reject the input. We cannot allow the recognizer to fail since it would cause the whole module to fail. To prevent this we use two techniques when our input deviates from what the grammar allows for. The first method relies on the information of the statistical component allowing for reinterpreting our input, while the second uses a set of so-called *repair operators* for "repairing" the parse tree.

Below we first present the methods and then we give two examples taken from our corpus of annotated dialogues. The English translation, however, is not produced by VERBMOBIL.

Statistical Repair

This method for error recovery is based on the hypothesis that the attribution of only one speech act to a given utterance is insufficient and that an utterance has more than one speech act reading.

...
mhw3_1_07: ja Montag der vierzehnte Juni paßt mir ausgezeichnet (ACCEPT)
(*Yes Monday the 14th suits me well*)

wir können jede Zeit nehmen die Ihnen gefällt (SUGGEST)
(*We could pick any time you want*)

mps1_1_08: ja morgens um halb elf hab' ich Zeit (SUGGEST)
(*I have time at half past ten in the morning*)

mhw3_1_09: also gut treffen wir uns am Montag den vierzehnten Juni um zehn Uhr dreißig zu unserem Termin (CONFIRM)
(*Ok let us meet on Monday the 14th of June at half past ten*)

...

Figure 4: Example turns – statistical repair

If a dialogue act not compatible with our dialogue model is encountered, the statistical component is looked up in order to find out whether any statistically relevant dialogue acts exists which are able to bridge the previous and the current (incompatible) dialogue act. If such a speech act can be found and if the insertion of this speech act renders the dialogue compatible, a multiple reading is proposed for either the current, or one of the former turns.

Plan Based Repair

The second mechanism uses a set of special *repair operators* which are used when the plan recognizer does not succeed in parsing the next token using the normal plan operators. The simplest case covers the dialogue acts in the subnet in figure 2. The problem with these acts is that they can appear anywhere in the dialogues. One could handle this by adding these dialogue acts to each state in the dialogue model. However, this method is costly in performance and grammar size. We instead process these dialogue acts using the repair operators.

Also, when an input is not admissible by the grammar, and our statistical repair technique has not been able to adjust the input, we repair the tree with this technique.

An Example of Statistical Repair

In this example (see fig. 4) a confirmation (CONFIRM) follows a suggestion (SUGGEST) – a sequence not admissible for the plan recognizer. The trace in fig 5 shows how the recognizer discovers that it can not process the sequence. It consults the statistical component for suggestions to bridge the two dialogue acts. The only suggestion from the statistical component in this example (ACCEPT[1]) is then checked with the surrounding dialogue acts to see which reading to modify. Here the CONFIRM gets an additional reading of ACCEPT.

```
...
Planner: -- Processing ACCEPT
Planner: -- Processing SUGGEST
Planner: -- Processing SUGGEST
Planner: -- Processing CONFIRM
Warning -- Repairing...

Trying to find a dialogue act to
bridge SUGGEST and CONFIRM ...

Possible insertion(s) and
its (their) score(s):
((ACCEPT 98256))

Testing ACCEPT for compatibility
with surrounding dialogue acts...

The current dialogue act CONFIRM has
an additional reading of ACCEPT:

CONFIRM -> ACCEPT CONFIRM !

Planner: -- Processing
...
```

Figure 5: Trace of statistical repair

An Example of Plan Based Repair

We here show a typical example of a clarification dialogue and how the recognizer inserts a clarification dialogue using the repair

[1]The score is the product of the transition probabilities times 1000 between the previous dialogue act, the potential insertion and the current dialogue act.

technique. In the example (see figure 6) the ongoing sub-dialogue is interrupted by a clarification dialogue between the dialogue participants.

turn_2_speaker_b_MW1001': wie wär's denn am Dienstag den dreizehnten April vormittags (SUGGEST)
(*How about Tuesday the 13th of April in the morning*)

turn_3_speaker_a_PS1002: tut mir leid ,am dreizehnten April bin ich noch im Urlaub . genauso wie am zwölften April Montag (REJECT)
(*I'm sorry, but I'm on vacation the 13:th. The same with Monday the 12:th*)

turn_3_speaker_a_PS1002: ich habe erst wieder ab dem vierzehnten April Zeit. (SUGGEST)
(*I'm free from the 14th of April*)

turn_4_speaker_b_MW1003: der vierzehnte is' ein Mittwoch , richtig (CLARIFICATION_QUESTION)
(*The 14th is a Wednesday, isn't it*)

turn_5_speaker_a_PS1004: ja genau (CLARIFICATION_ANSWER)
(*Yes, exactly*)

turn_5_speaker_a_PS1004: allerdings hab' ich da von neun bis zehn Uhr schon einen Arzt-Termin (REJECT)
(*I have to see my doctor at ten*)

turn_5_speaker_a_PS1004: deshalb würde ich vielleicht den Donnerstag vorschlagen (SUGGEST)
(*Maybe I could propose Thursday*)

Figure 6: Example turns – plan based repair

Figure 7 shows a screen snapshot from the plan recognizer. It is taken from a system version with the "old" 17 core dialogue acts named in German. Also, the names of the plan operators are truncated to 20 characters[2]. In the figure we see the difference between how the plan recognizer construes the intentional structure for a normal sub-negotiation dialogue SUGGEST (vorschlag) – REJECT (ablehnung) to the left, and the repaired SUGGEST (vorschlag) – CLARIFICATION_QUESTION (klaerungsfrage) – CLARIFICATION_ANSWER (klaerungsantwort) – REJECT (ablehnung) to the right.

The repair operator repair-operator is inserted and allows for the insertion of the clarification sub-dialogue.

[2]...when possible

4 TESTING THE MODULE WITH THE CORPUS

In this section we describe the results on testing our component on the corpus. For evaluating the dialogue model and plan recognizer we used 177 hand-annotated dialogues containing 7469 dialogue acts.

Evaluating the dialogue model

To test the coverage of the dialogue model we parsed the above mentioned dialogues with the FSM. In 6633 (91.1 %) cases admissible state changes were encountered. In 836 (8.9 %) cases a non valid sequence of dialogue acts was encountered. However, when trying to use the model to predict the next dialogue act to come, the results is not as good as when using the statistical method (see below).

Evaluating the plan recognizer

We also tested the plan recognizer with the same 177 dialogues. We got 1249 repairs. 795 of them (63.65 %) are concerned with digressions. Of the remaining 454, 95 could be repaired using statistics, i.e. 7.61 % of all repairs and 20.92 % of the repairs without digressions. It shows, that the repair mechanism plays an important role in the plan processing module. The role of statistical repair covering one fifth of all "real" repairs is important but has to be investigated further.

Evaluating The Prediction Process

To evaluate the prediction process, we took 52 dialogues with 2538 dialogue acts and trained both the FSM and the statistical

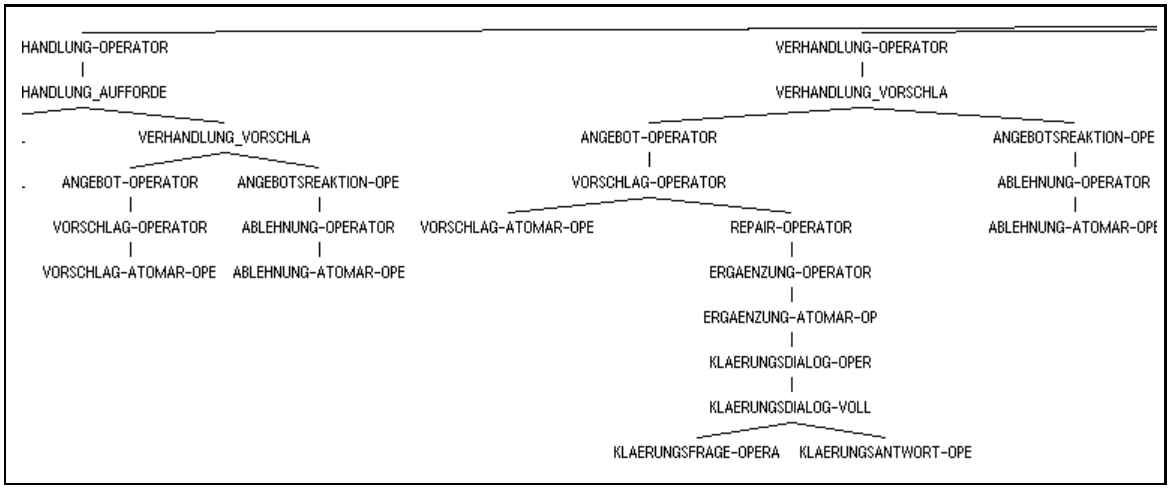

Figure 7: Screen dump – Plan Recognizer

Pred.	Statistics	FSM
3	2127 (57.25 %)	2069 (55.69 %)
2	1687 (56.53 %)	1427 (38.41 %)
1	1082 (38.79 %)	970 (30.07 %)

Table 1: Predictions without update

Pred.	Statistics	FSM
3	2572 (69.23 %)	2073 (55.80 %)
2	2100 (56.53 %)	1764 (47.48 %)
1	1441 (38.79 %)	1642 (30.07 %)

Table 2: Predictions with update

component. We then took 81 dialogues containing 3715 dialogue acts and tried to predict the next dialogue act to come. Two methods where used. In the first, the predictions where made without update, and in the second with update, i.e. processed dialogue acts are added to the training data. The results are shown in tables 1 and 2. We tested the hit rates for one to three predictions. It can bee seen that the n-gram based statistical method performs better than the FSM, because it is trained on real data and not handcrafted, and because it is possible to integrate longer histories of dialogue acts by using trigrams. The difference is even more obvious when the two components are allowed to adjust for the new dialogues. For more information about the prediction method and its evaluation, see [15].

5 DISCUSSION AND FUTURE WORK

In this paper we gave an overview of the design process and the inner structure of the dialogue component of VERBMOBIL. One point we want to stress is the importance of a careful analysis of the application environment. It was not possible to simply take the approaches of dialogue processing as used in earlier speech processing systems. Due to the passive, non-controlling character of VERBMOBIL in the scenario, the dialogue structures to be processed can vary unforeseeable. Yet, they have to be processed by the dialogue component.

Our design process for the dialogue component of VERBMOBIL consists of the following steps

1. annotate a corpus
2. extract a "standard" dialogue model from the annotations
3. check the requirements from the other

components in the system and identify information needed from the dialogue component

4. select appropriate processing methods, in our case a plan based and a statistical approach which are combined for robustness reasons

5. evaluate the system with real data

6. tune the system, again using real data

Evaluation shows certain deficits in e.g. prediction. We are currently in the process of replacing the prediction module with a reimplementation that delivers up to five percent better prediction results. Still, the prediction process is far from optimal. Since the structure of the dialogue varies a lot [16], we are now testing whether dialogues with similar dialogue structure can be automatically clustered together in different training sets. The idea is to switch between the training data to find the best one for a dialogue.

Our current dialogue model is handcrafted, which may explain its poor results in the prediction process. To automate the extraction of a dialogue model given an annotated corpus is also topic for further research.

For the plan recognizer we have two main challenging tasks to work on. As mentioned above the input of the dialogue component contains gaps. Extending the plan recognizer to cope with this is a big challenge. The current version also processes the plan operators strictly left to right. In future versions the plan operators will be selected on basis of statistical information collected from a corpus.

References

[1] Francois Andry. Static and Dynamic Predictions : A Method to Improve Speech Understanding in Cooperative Dialogues. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'92)*, volume 1, pages 639–642, 1992.

[2] John Austin. *How to do things with words.* Oxford: Clarendon Press, 1962.

[3] Eric Bilange. A Task Independent Oral Dialogue Model. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics (EACL-91)*, pages 83–88, Berlin, Germany, 9-11 April 1991.

[4] B. Grosz and C. Sidner. Attention, Intentions and the Structure of Discourse. *Journal of Computational Linguistics*, 12(3), 1986.

[5] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. Dialogue Acts in VERBMOBIL. Verbmobil-Report 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.

[6] Fred Jelinek. Self-Organized Language Modeling for Speech Recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann, 1990.

[7] Martin Kay, Jean Mark Gawron, and Peter Norvig. *Verbmobil. A Translation System for Face-to-Face Dialog.* Chicago University Press, 1994. CSLI Lecture Notes, Vol. 33.

[8] Klaus Kohler, Gloria Lex, Matthias Paetzold, Michael Scheffers, Adrian Simpson, and Werner Thon. Handbuch zur Datenerhebung und Transliteration in TP14 von VERBMOBIL - 3.0. Verbmobil Technisches Dokument 11, Universitaet Kiel, 1994.

[9] Elisabeth Maier. Dialogmodellierung in VERBMOBIL – Festlegung der Sprechhandlungen für den Demonstrator. Technical Report Verbmobil–Memo 31, DFKI Saarbrücken, Juli 1994.

[10] M. Mast, R. Kompe, F. Kummert, H. Niemann, and E. Nöth. The dialog module of the speech recognition and dialog system EVAR. In *Proceedings of the International Conference on Spoken Language Processing, Banff, Canada*, pages 1573–1576, 1992.

[11] Mark T. Maybury. *Planning Multisentential English Text Using Communicative Acts.* PhD thesis, University of Cambridge, Cambridge, GB, 1991.

[12] Johanna Moore. *Participating in Explanatory Dialogues.* The MIT Press, 1994.

[13] Masaaki Nagata and Tsuyoshi Morimoto. An Experimental Statistical Dialogue Model to Predict the Speech Act Type of the Next Utterance. In *Proceedings of International Symposium on Spoken Dialogue (ISSD'93), Nov. 10-12, Waseda University, Tokyo*, pages 83–86, 1993.

[14] Gerhard Th. Niedermair. Linguistic Modelling in the Context of Oral Dialogue. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'92)*, volume 1, pages 635–638, Banff, Canada, 1992.

[15] Norbert Reithinger. Some Experiments in Speech Act Prediction. In *AAAI 95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[16] Norbert Reithinger and Elisabeth Maier. Utilizing Statistical Speech Act Processing in VERBMOBIL. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, Cambridge, MA, June 1995.

[17] Bärbel Ripplinger and Folker Caroli. Konzept-basierte Übersetzung in Verbmobil. Technical report, IAI Saarbrücken, May 1994.

[18] John R. Searle. *Speech Acts.* Cambridge/GB: University Press, 1969.

[19] Marc B. Vilain. Getting Serious about Parsing Plans: a Grammatical Analysis of Plan Recognition. In *Proceedings of American Association for Artificial Intelligence*, pages 190–197, 1990.

[20] Wolfgang Wahlster. Verbmobil–Translation of Face-to-Face Dialogs. Technical report, German Research Centre for Artificial Intelligence (DFKI), 1993. In Proceedings of MT Summit IV, Kobe, Japan, 1993.