gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

Some Experiments in Speech Act Prediction

Norbert Reithinger

DFKI GmbH

Vm
Report 49
Dezember 1994

Dezember 1994

Norbert Reithinger

DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken

Tel.: (0681) 302 - 5346
e-mail: reithinger@dfki.uni-sb.de

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 101 K/1 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Some Experiments in Speech Act Prediction

Norbert Reithinger*
DFKI GmbH
Stuhlsatzenhausweg 3
D–66132 Saarbrücken, Germany
e–mail: bert@dfki.uni-sb.de

Abstract

In this paper, we present a statistical approach for speech act prediction in the dialogue component of the speech-to-speech translation system VERBMOBIL. The prediction algorithm is based on work known from language modelling and uses N-gram information computed from a training corpus. We demonstrate the performance of this method with 10 experiments. These experiments vary in two dimensions, namely whether the N-gram information is updated while processing, and whether deviations from the standard dialogue structure are processed. Six of the experiments use complete dialogues, while four process only the speech acts of one dialogue partner. It is shown that the predictions are best when using the update feature and deviations are not processed. Even the processing of incomplete dialogues then yields acceptable results. Another experiment shows that a training corpus size of about 40 dialogues is sufficient for the prediction task, and that the structure of the dialogues of the VERBMOBIL corpus we use differs remarkably with respect to the predictions.

*This paper appeared also in "Proceedings of the AAAI-95 Spring Symposium on Empirical Methods in Discourse".

1 Introduction

Speech processing systems like VERBMOBIL, a speech-to-speech translation system [Wahlster, 1993], require top-down predictions from the dialogue level of processing to narrow down the search space in earlier processing levels. One example is the prediction of the next possible speech act which can be used, amongst others, to select a dialogue dependent language model for the acoustic processing [Niedermair, 1992, Nagata and Morimoto, 1993]. [Niedermair, 1992] shows a perplexity reduction between 19% and 60% when using context dependent language models compared to a general language model.

Speech act predictions from structural knowledge sources like plans or dialogue grammars are difficult to use, because usually one can infer a large number of possible follow-up speech acts that are not ranked [Nagata and Morimoto, 1993]. Therefore, statistic based approaches for speech act prediction are now under development that show encouraging results.

In this paper, we will present an overview of the statistics based dialogue processing within the VERBMOBIL system and show the performance ot the prediction process. Especially, we will address the following topics:

- How good is the prediction process?
- How do deviations in the dialogues influence hit rates?
- How good is the prediction for incomplete dialogues?
- How many training dialogues must be provided?
- How do the dialogues differ in their structure?

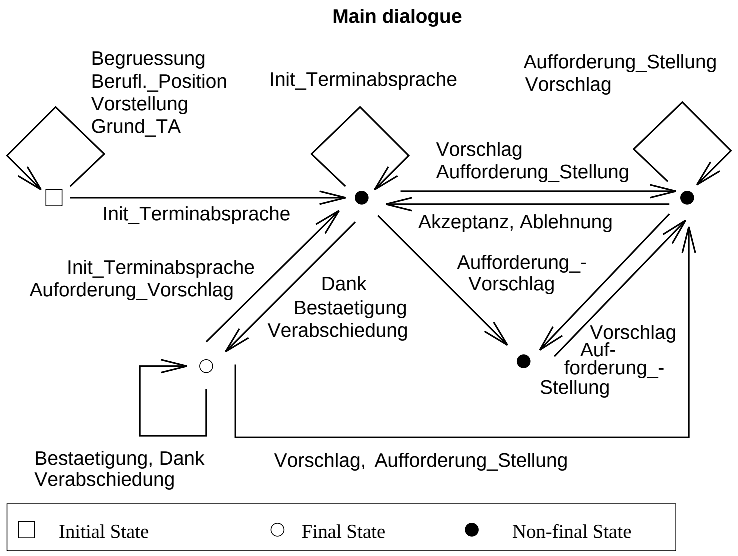
2 An Overview of VERBMOBIL's Dialogue Component

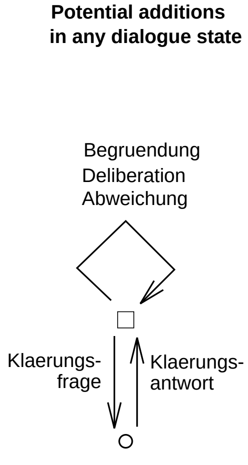
Dialogue processing in VERBMOBIL differs from systems like SUNDIAL [Andry, 1992, Niedermair, 1992] in an important aspect: VERBMOBIL *mediates* the dialogue between two human dialogue participants; the system is not a participant of its own. It does not control the dialogue as it happens in the flight scheduling scenario of SUNDIAL, since it is not a participant in the dialogue. Only when clarifications are needed, VERBMOBIL functions as a full fledged dialogue system.

Another special feature of VERBMOBIL is that it is activated only 'on demand': it is assumed that both dialogue participants have at least a passive knowledge of English. If the owner of the VERBMOBIL system needs the translation of an utterance, she presses a button and lets the system translate it. A consequence of this translation on demand is that the VERBMOBIL system processes maximally 50% of the dialogue, namely if the owner speaks only German. We try to get a hold on the English passages of the dialogue by using a keyword spotter that tracks the ongoing dialogue superficially. It is provided with top-down information of the dialogue component to select the appropriate set of keywords to look for.

Processing is centered around speech acts (see e.g. [Bilange, 1991], [Mast *et al.*, 1992]). For practical purposes within the VERBMOBIL system, speech acts support both the translation module in computing the adequate translation equivalent and also provide the basis for top-down predictions that can reduce the search space in e.g. the speech recognizer or the keyword spotter. We selected a set of 18 speech acts by analyzing the VERBMOBIL corpus of transliterated appointment scheduling dialogues [Maier, 1994]. Using the rules also defined in [Maier, 1994], we annotated more than 200 dialogues with speech act information. This corpus of annotated dialogues serves as training and test material. Figure 1 shows our dialogue model which consists of a network representation of admissible sequences of speech acts. The model for conventional/ expected dialogues is given in the upper network; deviations that can occur everywhere in the dialogue are displayed to the left at the bottom of the figure.

To cope with the requirements mentioned above, we developed and implemented a three layered architecture that combines different processing approaches to get a flexible and robust component (see fig. 2). The statistics module is one of three processing modules within VERBMOBIL's dialogue processing component (see [Alexandersson *et al.*, 1995] for an overview of the dialogue component). The contextual knowledge built up by the processing modules is stored in the dialogue memory which consists of an intentional and a thematic structure, and referential objects for items mentioned during the dialogue.





English Equivalents for German Speech Act Names:

Begruessung	Greeting
Berufl_Position	Position
Vorstellung	Introduction
Grund_TA	Reason_for_Appointment
Init_Terminabsprache	Initialisation
Aufforderung_Stellung	Request_for_Statement
Aufforderung_Vorschlag	Request_for_Suggestion
Akzeptanz	Accept
Ablehnung	Reject
Vorschlag	Suggestion
Bestaetigung	Confirmation
Verabschiedung	Bye
Dank	Thanks
Deliberation	Deliberation
Abweichung	Deviation
Klaerungsfrage	Clarification_Question
Klaerungsantwort	Clarification_Answer
Begruendung	Reason

Figure 1: A dialogue model for the description of appointment scheduling dialogues

Each of the modules has a different task:

- the statistics module provides information about the next possible speech acts
- the finite state machine checks the structure of the dialogue using the transition networks of the dialogue model
- the planner builds up the intentional structure of the dialogue.

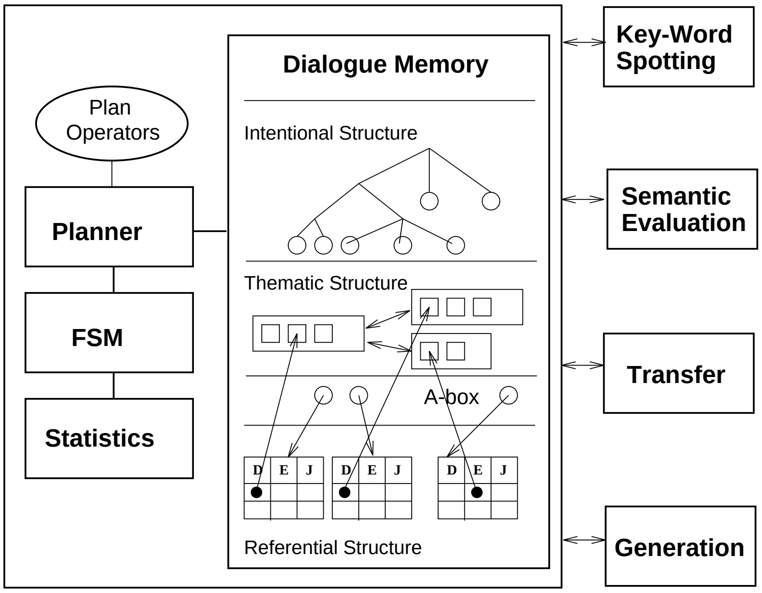


Figure 2: Architecture of the dialogue module

The three modules interact during processing, e.g. statistic knowledge is used when the automaton must be set up again after a speech act has not fit into the dialogue model. In the figure, also the other components of VERBMOBIL are shown to which the dialogue component is connected.

3 The Statistical Method

In speech recognition language models are commonly used to reduce the search space when determining a word that can match a part of the input

[Jellinek, 1990]. In our application, the units to be processed are not words, but a set of speech acts of a text or a dialogue. Also, there is no input signal that is to be interpreted, but we have to predict the most probable speech act(s).

A dialogue S can be seen as a sequence of utterances S_i where each utterance has a corresponding speech act s_i. If $P(S)$ is the statistical model of S, the probability can be approximated by the N-gram probabilities

$$P(S) = \prod_{i=1}^{n} P(s_i | s_{i-N+1}, ..., s_{i-1})$$

Therefore, to predict the nth speech act s_n, we can use the previously uttered speech acts

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

We approximate the conditional probability P with the standard smoothing technique known as deleted interpolation [Jellinek, 1990]. The basis of processing is a training corpus annotated with the speech acts of the utterances. This corpus is used to gain statistical information about the dialogue structure, namely unigram, bigram and trigram frequencies of speech acts. P can then be approximated 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 and $\sum q_i = 1$

4 Overall Predictions Hit Rates

Given this formula and the required N-grams we can determine the k best predictions for the next speech act. In order to evaluate the statistical model, we made various experiments to show the influence of deviations and of continuous retraining to the overall predictions hit rates

For the experiments presented, we took dialogues annotated with speech acts from the VERBMOBIL corpus as training and test data. In all experiments, test and training data are disjunct. We present the prediction hit rate for one, two, and three predictions. A hit for one prediction is counted when a new speech act is predicted with the best score by the predictions algorithm. For two and three predictions a hit is counted, when the new speech act is among the two or three predictions with the highest score. We show the hit rate as precentage of hits in relation to all speech acts processed. The tables below show at the left side the numbers of predictions and in the columns for the experiments the respective hit rate in percent.

4.1 Complete Dialogues

The experiments for complete dialogues use 52 dialogues with 2538 speech acts as training data. The first four experiments, TS1 to TS4 use the same test set of 81 dialogues with 3715 speech acts.

Experiments TS1 and TS2 were made to show the influence of a continuous update of the N-gram frequencies:

Pred.	*TS1*	*TS2*
1	33.78%	44.24 %
2	54.25 %	66.47 %
3	66.94 %	81.46 %

For TS1 the frequencies were not updated during processing, while in TS2 the statistical data were continuously updated. As can be seen, the hit rates go up between 10 and 15 percent, if updating is enabled. This adaption feature is especially important and useful in VERBMOBIL: a user is supposed to 'own' one system which, using the adaption feature of the statistics module, can adapt itself to the language and dialogue behaviour of the respective user.

While in the experiments above, deviations as defined by the dialoge model in the lower left network of figure 1 were not processed, the 720 deviation speech acts of the test corpus that can occur in any dialogue state were included in experiments TS3 and TS4:

Pred.	*TS3*	*TS4*
1	29.13%	38.79 %
2	45.41 %	56.53 %
3	57.25 %	69.23 %

Compared to experiments TS1 and TS2 we see a dramatic decline in the hit rate. If the update feature is diabled (TS3), the quality of the prediction is unsatisfactory. Even with the update, prediction hit rates are 10 percent below the ones of experiments TS2. Since VERBMOBIL is activated only on demand, it can be expected that a user knows what she wants to say, minimizing deviations that have to be processed. Therefore, this decline will not occur too often.

The next two experiments use 15 English dialogues with 363 speech acts as test data. These dialogues have a slightly different setting and were recorded in the USA at Carnegie Mellon University. Update during processing is enabled, but deviations are processed only in experiment TS6:

Pred.	*TS5*	*TS6*
1	43.95%	39.12 %
2	67.11 %	53.44 %
3	81.87 %	70.25 %

Again it can bee seen that prediction hit rates are higher if deviations are left out. Also, the prediction algorithm still delivers pretty good results, even if the overall setting of the test dialogues differs from that of the training material.

Compared to the data from [Nagata and Morimoto, 1993] who report prediction rates of 61.7 %, 77.5 % and 85.1% for one, two or three predictions respectively, the predictions in these six experiments are less reliable. However, their set of speech acts (or the equivalents, called illocutionary force types) does not include speech acts to handle deviations. Also, since the dialogues in our corpus are rather unrestricted besides the limitation to the task, they have a rather big variation in their structure (see below). Therefore, a comparison with [Nagata and Morimoto, 1993] is hardly possible, due to the big differences in the setting and the underlying dialogue model.

4.2 Incomplete Dialogues

Up till now, the experiments were made with complete dialogues, which will not be the case in the standard scenario of VERBMOBIL. As mentioned in the overview, VERBMOBIL usually will process only the contributions of one dialogue participant.

We tested the prediction algorithm for this case, using 52 dialogues as training data, where only the contributions of one speaker are used to build up the N-gram data. In total, 1843 speech acts are processed for the training. As test data we took 177 dialogues, splitting them up according to the speaker identifiers `A` or `B`. For speaker identifier `A`, there are 3561 speech acts, with

580 deviations, for speaker identifier `B` 3413 speech acts with 526 deviations. All experiments are made with the update feature enabled.

TS7 and TS8 show the prediction hit rates when processing the speech acts of **A**, with TS8 excluding deviations.

Pred.	*TS7*	*TS8*
1	34.74%	39.99 %
2	53.10 %	61.96 %
3	65.80 %	74.44 %

The next table is the same as above, but processing the speech acts of speaker identifier `B`

Pred.	*TS9*	*TS10*
1	34.66%	40.49 %
2	52.97 %	62.14 %
3	64.72 %	75.41 %

The figures show that if only one half of a dialogue is processed, the prediction hit rates are not so good as e.g. experiment TS2. But they are only approximately 5% worse and can still deliver predictions to the other components.

5 Size of the Training Corpus

To check whether the size of the training corpus is too small or whether the dialogues differ remarkably in their structure, we did some experiments on the size of the training set and the differences between the dialogues in respect to the prediction hit rate. An example of one of these experiments will be presented in this and the following section.

We randomly took 42 dialogues with 2513 speech acts from our corpus as training material and 47 dialogues with 2283 speech acts as test data. We tested the prediction hit rates of the method, using these 47 dialogues and test corpora that consisted from zero to 42 dialogues from the training material. During the experiments, the update of the statistics during processing was enabled. Speech acts for deviations were not computed, since they distort the results.

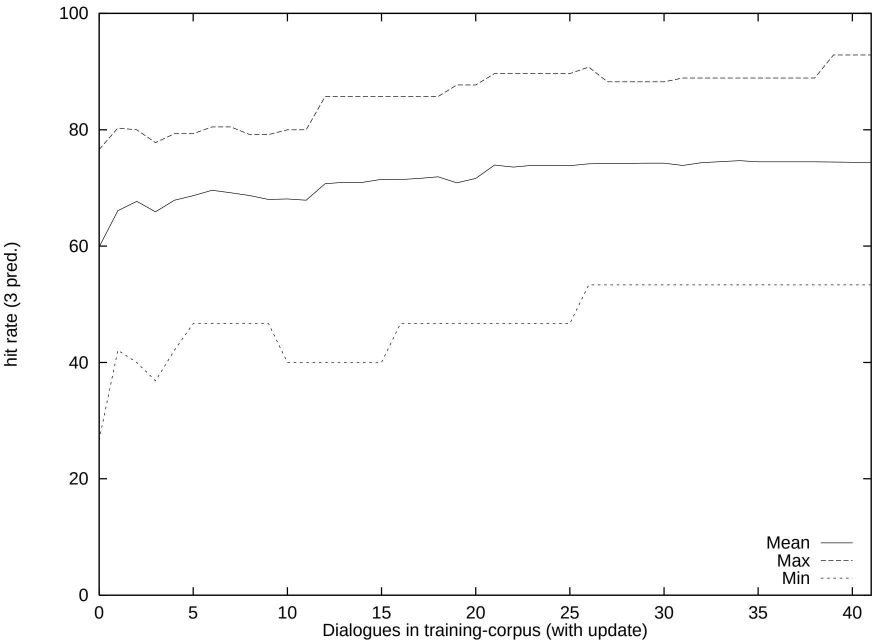


Figure 3: Learning curves

Figure 3 shows the dependencies between the number of dialogues in the training corpus on the x-axis and the prediction hit rate for three predictions on the y-axis. As can be seen, the mean hit rate increases up to the point where 30 - 40 dialogues are in the training corpus. We noticed this behaviour also in a number of other experiments. The hit rate for the best rated dialogues, which is shown with a dashed line at the top, still increases up to approximately 95 %, if we add new training material. The hit rate for the worst dialogues, shown with a dotted line, never raises above approximately 55 %.

6 Differences in the Prediction Hit Rates

The main reason for this big difference in the prediction hit rate is that the dialogues in our corpus frequently do not follow conventional dialogue behaviour, i.e. the dialogue structure differs remarkably from dialogue to

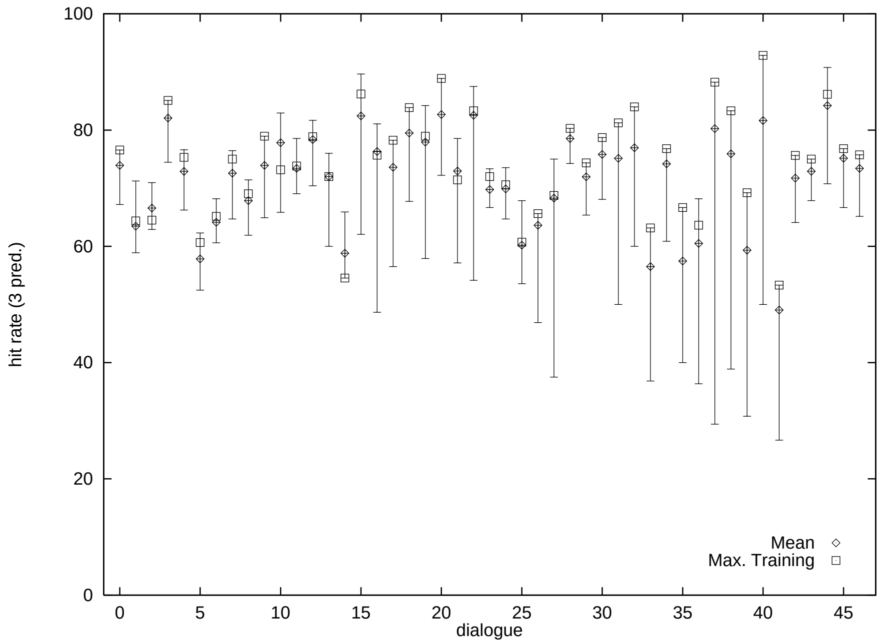


Figure 4: Hit Rate of the 42 dialogues

dialogue. Figure 4 shows the variation of the prediction hit rate for each of the 47 dialogues in the test corpus. The mean value over all training corpora is shown as a diamond, while the prediction hit rate for the biggest training corpus is shown as as square box.

The figure confirms the great variation in the structure of the test dialogues. It is remarkable that only for three dialogues (#3, #11, and #15) the prediction hit rate is worse than the mean value when using the biggest training corpus. But it is better or at least equal to the mean hit rate in the other cases. Also, it can be seen that for the largest training corpus there are large differences between the hit rates. The most extreme difference are the adjacent dialogues #40 and # 41 with hit rates of approximately 93% vs. 53%. While dialogue #40 fits very well in the statistical model acquired from the training corpus, dialogue #41 does not. This figure gives a rather good impression of the wide variety of material the dialogue component has to cope with.

Finally, fig. 5 shows the relationship between the dialogues (x-axis), the number of training dialogues (y-axis), and the prediction hit rate (z-axis).

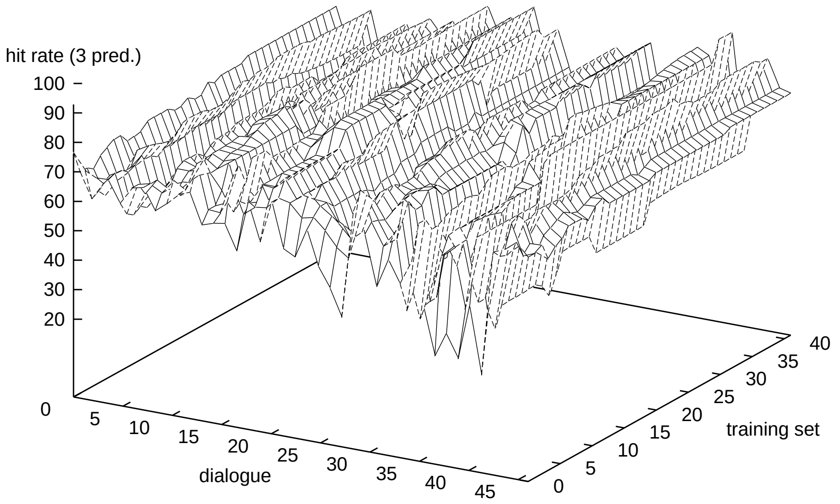


Figure 5: Connection between the number of training dialogues and prediction hit rate

It gives an impression on the dynamic development of the prediction hit rate when increasing the number of dialogues in the training corpus. At the right side, the development of the hit rates for dialog #41 can be seen which remain low as the training set increases, while e.g. dialogue #0 starts with a high hit rate that increases only for about 10% while the training set grows.

7 Conclusion

This paper gives an outline of the statistical processing in the dialogue component of VERBMOBIL. From the analysis of the examples we conclude that

- the presented method usually delivers a prediction hit rate that is good enough to be used in a speech processing system, even if only the contributions of one speaker are processed;

- if the statistic is updated during processing, it can adapt itself to the dialogue patterns of the owner of VERBMOBIL, leading to a higher prediction hit rate;

- a relatively small number of dialogues, namely 30 – 40, is sufficient as training corpus;

- real dialogues differ remarkably in their structure. Therefore predictions have great variations when using a single training corpus.

The predictions computed with this method are currently integrated into the overall VERBMOBIL system. They are e.g. used in the semantic evaluation module, where they support the selection of the speech act for the next utterance. As already mentioned, also the selection of the most probable keywords for the keyword spotter is based on these predictions.

Future work will be concerned, amongst others, with

- identifying clusters of dialogues with similar structure to get less different dialogue structures in the training data and to select the most appropriate training set for a dialogue [Carter, 1994].

- computing the perplexity of the language models defined by the speech acts. Currently, the speech act definition is guided by the needs of the semantic processing and transfer components of VERBMOBIL. We will now look whether the word sequences of the respective speech acts have a lower perplexity than the overall dialogues. Then, we can compute dynamic language models for the speech recognition component which can be selected according to the current dialogue state.

Acknowledgements

Thanks to the other members of the VERBMOBIL dialogue group, especially to Jan Alexandersson and Elisabeth Maier.

References

[Alexandersson *et al.*, 1995] Jan Alexandersson, Elisabeth Maier, and Norbert Reithinger. A Robust and Efficient Three-Layered Dialog Component for a Speech-to-Speech Translation System. In *Proceedings of EACL-95*, 1995.

[Andry, 1992] 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.

[Bilange, 1991] 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, April 1991.

[Carter, 1994] David Carter. Improving language models by clustering training sentences. In *ANLP-94*, 1994.

[Jellinek, 1990] Fred Jellinek. 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.

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

[Mast *et al.*, 1992] Marion Mast, Ralf Kompe, Franz Kummert, Heinrich Niemann, and Elmar Nöth. The Dialogue Modul of the Speech Recognition and Dialog System EVAR. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'92)*, volume 2, pages 1573–1576, 1992.

[Nagata and Morimoto, 1993] 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.

[Niedermair, 1992] 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.

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