

Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System

Wolfgang Wahlster

DFKI GmbH, Saarbrücken, Germany

Abstract. Verbmobil is a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations. It recognizes spoken input, analyses and translates it, and finally utters the translation. The multilingual system handles dialogs in three business-oriented domains, with context-sensitive translation between three languages (German, English, and Japanese). Since Verbmobil emphasizes the robust processing of spontaneous dialogs, it poses difficult challenges to human language technology, that we discuss in this paper. We present Verbmobil as a hybrid system incorporating both deep and shallow processing schemes. We describe the anatomy of Verbmobil and the functionality of its main components. We discuss Verbmobil's multi-blackboard architecture that is based on packed representations at all processing stages. These packed representations together with formalisms for underspecification capture the non-determinism in each processing phase, so that the remaining uncertainties can be reduced by linguistic, discourse and domain constraints as soon as they become applicable. We present Verbmobil's multi-engine approach, eg. its use of five concurrent translation engines: statistical translation, case-based translation, substring-based translation, dialog-act based translation, and semantic transfer. Distinguishing features like the multilingual prosody module and the generation of dialog summaries are highlighted. We conclude that Verbmobil has successfully met the project goals with more than 80% of approximately correct translations and a 90% success rate for dialog tasks.

1 Introduction

Verbmobil is a software system that provides mobile phone users with simultaneous dialog interpretation services for restricted topics (see Wahlster, 1993, and Kay et al., 1994). As the name Verbmobil suggests, the system supports **verbal** communication with foreign interlocutors in **mobile** situations. It recognizes spoken input, analyses and translates it, and finally utters the translation. The multilingual system handles dialogs in three business-oriented domains, with bidirectional translation between three languages (German, English, and Japanese).

In contrast to previous dialog translation systems that translate sentence-by-sentence, Verbmobil provides context-sensitive translations. Since Verbmobil must "hear between the words"—things that were communicated earlier and things about the topic being discussed—it uses an explicit dialog memory and exploits domain knowledge. Figure 1 illustrates the use of Verbmobil in the travel planning scenario.

It shows that the Verbmobil server translates the German word “nächste” into English either as “next” or “nearest” depending on a temporal or spatial question context. In Verbmobil, the dialog context is used to resolve ambiguities and to produce an adequate translation in a particular conversational situation.

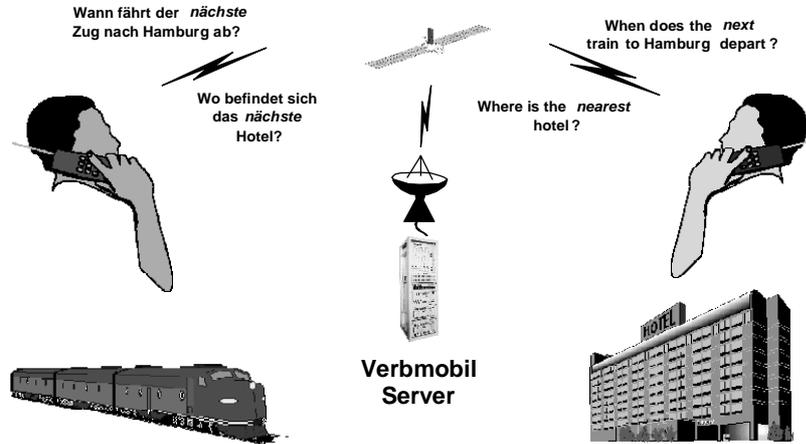


Fig. 1. Context-sensitive speech-to-speech translation

Verbmobil is the first speech-only dialog translation system. Verbmobil users can simply pick up a standard mobile phone and use voice dialing and speech commands in order to initiate a dialog translation session (see Figure 2). In contrast to previous versions of Verbmobil and other systems in the C-STAR consortium (see Woszczyzna, 1999), the operation of the final Verbmobil system is completely hands-free without any push-to-talk button. Since the Verbmobil speech translation server can be accessed by GSM mobile telephones, the system can be used anywhere and anytime. No PC, notebook or PDA must be available to access the Verbmobil translation service, just a phone for each dialog participant. In addition, no waiting time for booting computers and keyboard or mouse input to start the Verbmobil system is needed—dialog translation can begin instantaneously. Although the primary goal of Verbmobil is to support face-to-face conversations, in which all participants use their mobile phones as translation devices, Verbmobil can also be used for conversations in which the participants cannot see one another.

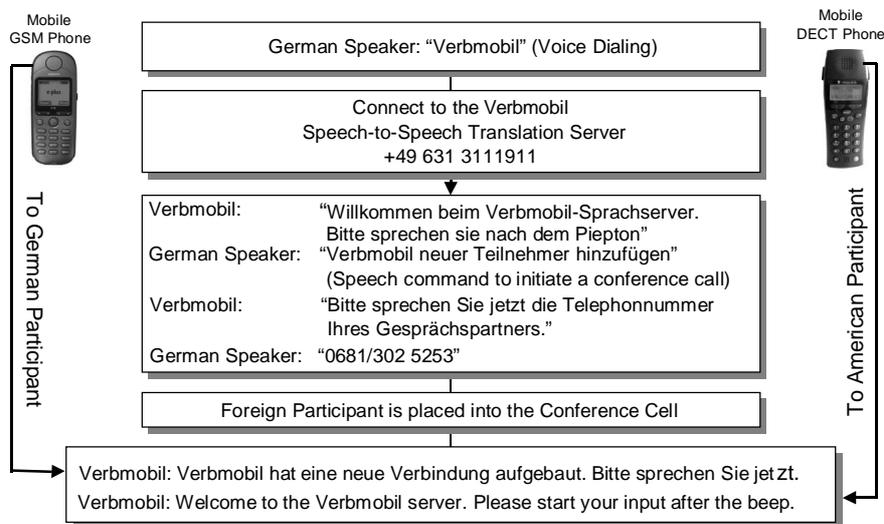


Fig. 2. Setting up a Verbmobil session with speech commands

Verbmobil emphasizes the robust processing of spontaneous dialogs posing difficult challenges to human language technology, that are summarized in Figure 3 and discussed in more detail below.

2 Some Grand Challenges Tackled by Verbmobil

Verbmobil is the only dialog translation system to date based on an open microphone condition. It is not a "push-to-talk" system which has to be told which chunks of the sound signal represent coherent contributions by individual speakers: Verbmobil works that out for itself from the raw input signal. The signal may be of different qualities—not necessarily from a lab-quality close-speaking microphone, for instance it can be GSM (mobile phone) quality. Thus, Verbmobil includes different speech recognizers for 16 kHz and 8 kHz sampling rates.

Verbmobil deals with spontaneous speech. This does not just mean continuous speech like in current dictation systems, but speech which includes realistic disfluencies and repair phenomena, such as changes of tack in mid-sentence (or mid-word), ums and ers, and cases where short words are accidentally left out in rapid speech. For example, in the Verbmobil corpus about 20% of all dialog turns contain at least one self-correction and 3% include false starts. Verbmobil uses a combination of shallow and deep analysis methods to recognize a speaker's slips and translate what he tried to say rather than what he actually said.

	Input Conditions	Naturalness	Adaptability	Dialog Capabilities
Increasing Complexity	Close-Speaking Microphone/Headset Push-to-talk	Isolated Words	Speaker Dependent	Monolog Dictation
	Telephone, Pause-based Segmentation	Read Continuous Speech	Speaker Independent	Information- seeking Dialog
	Open Microphone, GSM Quality	Spontaneous Speech	Speaker adaptive	Multiparty Negotiation

Verbmobil

Fig. 3. Some challenges for speech and language technology

At an early processing stage prosodic cues are used to detect self-corrections. A stochastic model is used to segment the repair into the “wrong” part (the so-called reparandum) and the correction. Then the corrected input is inserted as a new hypothesis into the word hypotheses graph. Thus, Verbmobil’s repair processing is a filter between speech recognition and syntactic analysis (see Spilker et al., in this volume). The word lattice is augmented by an additional path that does no longer contain those parts of the utterances that the speaker tried to correct. This transformation of the word lattice is used in addition to simple disfluency filtering, that eliminates sounds like “ahh” that users often make while speaking.

In addition to this shallow statistical approach, other forms of self-corrections are also processed at a later stage on the semantic level. A rule-based repair approach is applied during robust semantic processing to a chart containing possible semantic interpretations of the input (the so-called VIT Hypotheses Graph (VHG)). Verbmobil applies various hand-crafted rules to detect repairs in semantic representations and to delete parts of the representation that corresponds to slips of the speaker (see Pinkal et al., in this volume).

Verbmobil is a speaker-adaptive system, i.e. for a new speaker it starts in a speaker-independent mode and after a few words have been uttered it improves the recognition results by adaptation. A cascade of unsupervised methods, ranging from very fast adaptation during the processing of a single utterance to complex adaptation methods that analyze a longer sequence of dialog turns, is used to adjust to the acoustic characteristics of the speaker’s voice, the speaking rate and

pronunciation variants due to the dialectal diversity of the user community (see Haiber et al., in this volume).

Verbmobil deals with mixed-initiative dialogs between human participants. Each partner has a clear interaction goal in a negotiation task like appointment scheduling or travel planning. Although these tasks encourage cooperative interaction, the participants have often conflicting goals and preferences that lead to argumentative dialogs. Therefore Verbmobil has to deal with a much richer set of dialog acts than previous systems that focused on information-seeking dialogs (see Kipp et al., in this volume).

In order to ensure domain independence and scalability, Verbmobil was developed for three domains of discourse (appointment scheduling, travel planning, remote PC maintenance) with increasing size of vocabularies and ontologies (see Figure 4). Scenario 2 with a vocabulary of 10 000 words was used for the end-to-end evaluation of the final Verbmobil system.

Scenario 1 Appointment Scheduling	Scenario 2 Travel Planning	Scenario 3 Remote PC Maintenance
<p style="text-align: center;">When?</p> <p style="text-align: center;">Focus on temporal expressions</p> <p style="text-align: center;">Vocabulary Size: 2500/6000</p>	<p style="text-align: center;">When? Where? How?</p> <p style="text-align: center;">Focus on temporal and spatial expressions</p> <p style="text-align: center;">Vocabulary Size: 7000/10000</p>	<p style="text-align: center;">What? When? Where? How?</p> <p style="text-align: center;">Integration of special sublanguage lexica</p> <p style="text-align: center;">Vocabulary Size: 15000/30000</p>

Fig. 4. Three domains of discourse for Verbmobil

Verbmobil is a hybrid system incorporating both deep and shallow processing schemes (see Bub et al., 1997). It integrates a broad spectrum of corpus-based and rule-based methods. Verbmobil combines the results of machine learning from large corpora with linguists' hand-crafted knowledge sources to achieve an adequate level of robustness and accuracy.

3 Verbmobil's Massive Data Collection Effort

A significant programme of data collection was performed during the Verbmobil project to extract statistical properties from large corpora of spontaneous speech. A distinguishing feature of the Verbmobil speech corpus is the multi-channel recording. The voice of each speaker was recorded in parallel using a close-speaking microphone, a room microphone, and various telephones (GSM phone, wireless DECT phone and regular phone), so that the speech recognizers could be trained on data sets with various audio signal qualities.

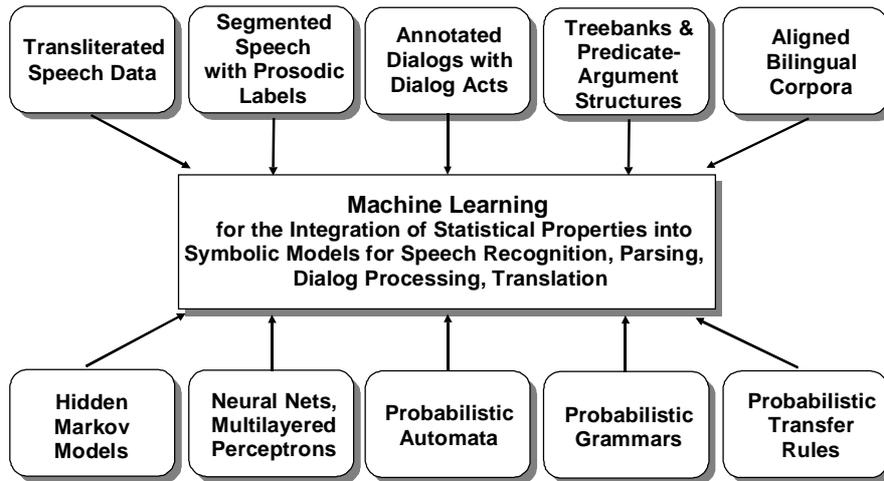


Fig. 5. Extracting statistical properties from large corpora

The so-called partitur (German word for musical score) format used for the Verbmobil speech corpora orchestrates fifteen strata of annotations: two transliteration variants, lexical orthography, canonical pronunciation, manual phonological segmentation, automatic phonological segmentation, word segmentation, prosodic segmentation, dialog acts, noises, superimposed speech, syntactic category, word category, syntactic function, and prosodic boundaries (see Burger et al., in this volume). In addition to the monolingual data, the multilingual Verbmobil corpus includes bilingual dialogs (from Wizard-of-OZ experiments, face-to-face dialogs with human interpreters, or dialogs interpreted by various versions of Verbmobil) and aligned bilingual transliterations (see Jekat and v. Hahn, in this volume). Three treebanks for German, English and Japanese have been developed with annotations on three strata: morpho-syntax, phrase structure, and predicate-argument structure (see Hinrichs et al. (b), in this volume).

Various machine learning methods have been used to train Hidden Markov Models, neural nets, probabilistic automata, parsers, rule systems, translation methods and plan recognizers. The end-to-end evaluations of the various Verbmobil prototypes have shown clearly, that the robustness, coverage and accuracy of a speech-to-speech translation system for spontaneous dialogs depends critically on the quantity and quality of the training corpora.

4 The Main Components of Verbmobil

The screenshot of Verbmobil's control panel provides an overview of the main components of the system. The overall control and data flow is indicated by arrows pointing upwards on the left side of the screenshot, from left to right in the middle and downwards on the right side. On the bottom various input devices can be selected (Europeans call a cell phone a "handy"). Since Verbmobil is a multilingual system it incorporates three speech recognizers and three speech synthesizers for German, English and Japanese.

A distinguishing feature of Verbmobil is its multi-engine parsing architecture. Three parsers based on different syntactic knowledge sources are used to process the word hypotheses graphs (WHG) that are augmented by prosodic information extracted by the prosody module (see Section 5 below). All parsers use the multi-stratal VIT representation as an output format. Since in most cases the parsers produce only fragmentary analyses, their results are combined in a chart of VIT structures. A chart parser and a statistical LR parser are combined in a package that is visualized in the screenshot as "integrated processing". These shallow parsers produce trees that are transformed into VIT structures by a module called semantic construction (see Figure 6). This syntax-semantics interface is primarily lexically driven (see Schiehlen, in this volume). The module with the label "deep analysis" is based on a HPSG parser for deep linguistic processing in the Verbmobil system. Verbmobil is the only completely operational speech-to-speech translation system that is based on a wide-coverage unification grammar and tries to preserve the theoretical clarity and elegance of linguistic analyses in a very efficient implementation. The parser for the HPSG grammars processes the n best paths produced by the integrated processing module. It is implemented as a bidirectional bottom-up active chart parser (see Kiefer et al., in this volume).

The statistical translation module starts with the single best sentence hypothesis of the speech recognizer (see Vogel et al. in this volume). Prosodic information about phrase boundaries and sentence mode are utilized by the statistical translation module. The output of this module is a sequence of words in the target language together with a confidence measure that is used by the selection module (not shown in the control panel) for the final choice of a translation result.

Verbmobil includes two components for case-based translation. Substring-based translation is a method for incremental synchronous interpretation, that is based on machine learning methods applied to a sentence-aligned bilingual corpus. Substrings of the input for which a contiguous piece of translation can be found in the corpus are the basic processing units. Substring pairs are combined with patterns for word

order switching and word cluster information in an incremental translation algorithm for a sequence of input segments (see Block in this volume). The other component for case-based translation is based on 30000 translation templates learned from a sentence-aligned corpus (see Auerswald, in this volume). Date, time and naming expressions are recognized by definite clause grammars (DCGs) and marked in the WHG. An A* search explores the cross-product graph of the WHG with the subphrase tags and the template graph. A DCG-based generator is used to produce target language output from the interlingual representation of the recognized date, time and naming expressions. These subphrases are used to instantiate the target language parts of translation templates.

Dialog-act based translation includes the statistical classification of 19 dialog acts and a cascade of more than 300 finite-state transducers that extract the main propositional content of an utterance. The statistical dialog classifier is based on n-grams and takes the previous dialog history into account. The recognized dialog act, the topic and propositional content are represented by a simplistic frame notation including 49 nested objects with 95 possible attributes covering the appointment scheduling and travel planning tasks. A template-based approach to generation is used to transform these interlingual terms into the corresponding target language. The shallow interlingual representation of an utterance is stored together with topic and focus information as well as a deep semantic representation encoded as a VIT in the dialog memory for further processing by the dialog and context evaluation component.

The dialog component includes a plan processor, that structures an ongoing dialog hierarchically in different dialog phases, games and moves (see Kipp et al in this volume). Dialog acts are the terminal nodes of the tree structure that represents the dialog structure. Information about the dialog phase is used eg. during the semantic-based transfer for disambiguation tasks. In addition, inference services are provided by the dialog and context component eg. for the completion of underspecified temporal expressions and the resolution of anaphora or ellipsis. Temporal reasoning is used for example to transform expressions like “two hours later” or “next week” into fully specified times and dates stored in the dialog memory for summarizing the results of a negotiation. The transfer module triggers contextual reasoning process only in cases where a disambiguation or resolution is necessary for a given translation task. For example the German noun “Essen” can be translated into “lunch” or “dinner” depending on the time of day, which can be derived by contextual reasoning. Disambiguation and resolution on demand is typical for Verbmobil’s approach to translation, since various forms of underspecification and ambiguity can be carried over into target language, so that the hearer can resolve them.

The transfer component is basically a rewriting system for underspecified semantic representations using Verbmobil’s VIT formalism (see Emele et al., in this volume). Semantic-based transfer receives a VIT of a source language utterance and transforms it into a VIT for the target language synthesis. This means that the transfer module abstracts away from morphological and syntactic analysis results. The final Verbmobil system includes more than 20000 transfer rules. These rules

include conditions that can trigger inferences in the dialog and context evaluation module to resolve ambiguities and deal with translation mismatches, whenever necessary. The transfer component uses cascaded rule systems, first for the phrasal transfer of idioms and other non-compositional expressions and then for the lexical transfer. The translation of spatial and temporal prepositions is based on an interlingual representation in order to cut down the number of specific transfer rules. Semantic-based transfer is extremely fast and consumes on the average less than 1% of the overall processing time for an utterance.

Verbmobil's multilingual generator includes a constraint-based microplanning component and a syntactic realization module that is based on the formalism of lexicalized tree-adjoining grammars (see Becker et al in this volume). The input to the microplanning component are VITs produced by the transfer module. A sentence plan is generated that consists basically of lexical items and semantic roles linking them together. The microplanner decides about subordination, aggregation, focus and theme control as well as anaphora generation. The syntactic realization component can either use LTAG grammars that are compiled from the HPSG grammars used for deep analysis or a hand-written LTAG generation grammar. For English and Japanese the grammars that were designed for analysis are usable for generation after an offline-compilation step (see Becker and Lopez, in this volume).

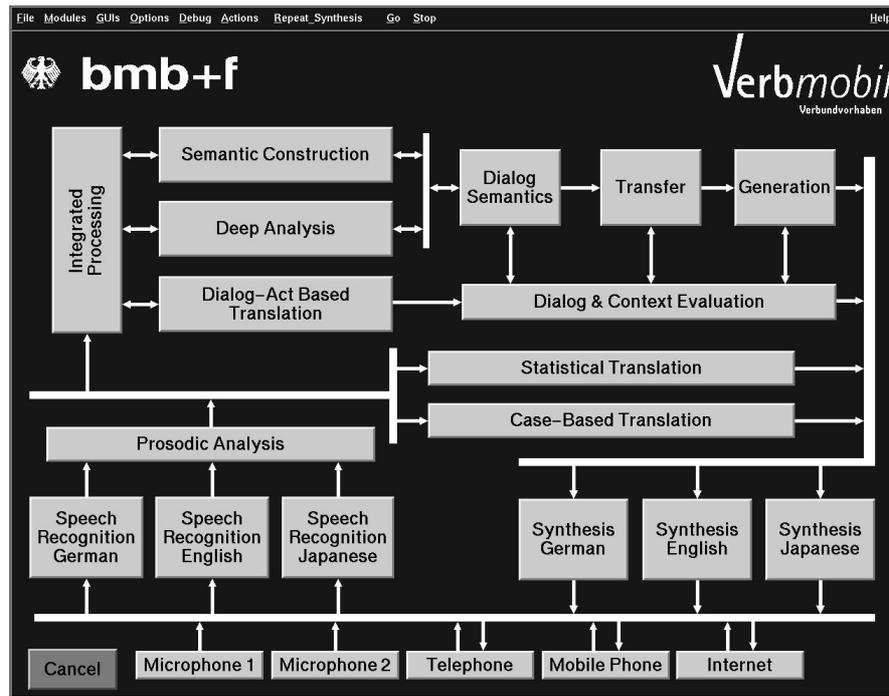


Fig. 6. A snapshot of Verbmobil's control panel

The speech synthesizer for German and American English follows a concatenative approach based on a large corpus of annotated speech data (see Stöber et al., in this volume). The word is the basic unit of concatenation, so that subword units are only used if a word is not available in the database.

The synthesizer applies a graph-based unit selection procedure to choose the best available synthesis segments matching the segmental and prosodic constraints of the input. Whenever possible the synthesizer exploits the syntactic, prosodic and discourse information provided by previous processing stages. Thus for the deep processing stream it provides concept-to-speech synthesis, whereas for the shallow translation threads it operates more like a traditional text-to-speech system resulting in a lower quality of its output.

5 Using Prosodic Information at all Processing Stages

Verbmobil is the first spoken-dialog interpretation system that uses prosodic information systematically at all processing stages. The results of Verbmobil's multilingual prosody module are used for parsing, dialog understanding, translation, generation and speech synthesis. This means that prosodic information in the source utterance is passed even through the translation process to improve the generation and synthesis of the target utterance. Prosodic differences in one language can correspond to lexical or syntactic differences in another; for instance, a German utterance beginning *wir haben noch ...* may be translated by Verbmobil into English either as *we still have ...* or as *we have another ...* depending whether *noch* is stressed. Although prosody is used in some other recent speech recognition systems, the exploitation of prosodic information is extremely limited in these approaches. For example, the ATR Matrix system (see Takezawa et al., 1998) uses prosody only to identify sentence mood (declarative vs. question). We believe that Verbmobil is the first system to make significant use of prosodic aspects of speech.

The prosody module of Verbmobil uses the speech signal and the word hypotheses graph (WHG) produced by the speech recognizer as an input and outputs an annotated WHG with prosodic information for each recognized word. The system extracts duration, pitch, energy, and pause features and uses them to classify phrase and clause boundaries, accented words and sentence mood. A combination of a multilayer perceptron and a polygram-based statistical language model annotate the WHG with probabilities for the classified prosodic events.

Verbmobil uses the probabilistic prosodic information about clause boundaries to reduce the search space for syntactic analysis dramatically. During parsing, the clause boundary marks that are inserted into the WHG by the prosody module play the role of punctuation marks in written language. Dialog act segmentation and recognition is also based on the boundary information provided by the prosody module. Prosodic cues about sentence mood is often used in Verbmobil's translation modules to constrain transfer results, if there is not enough syntactic or semantic evidence for a certain mood (e.g. question). The information about word accent is used to guide lexical choice in the generation process. Finally, during speech synthesis the extracted prosodic features are used for speaker-adaptation.

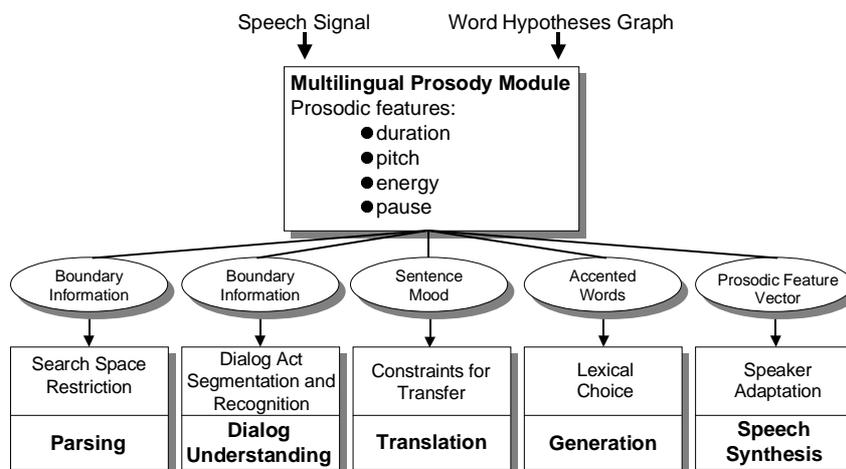


Fig. 7. The role of prosodic information in Verbmobil

6 The Multi-Blackboard Architecture of Verbmobil

The final Verbmobil system consists of 69 highly interactive modules. The transformation of speech input in a source language into speech output in a target language requires a tremendous amount of communication between all these modules. Since Verbmobil has to translate under real-time conditions it exploits parallel processing schemes whenever possible. The non-sequential nature of the Verbmobil architecture implies that not only inputs and results are exchanged between modules but also top-down expectations, constraints, backtracking signals, alternate hypotheses, additional parameters, probabilities, and confidence values.

198 blackboards are used for the necessary information exchange between modules. A module typically subscribes to various blackboards. Modules can have several instances, e.g. in a multiparty conversation there may be two German speakers, so that two instances of the German speech recognition module are needed.

The final Verbmobil system is based on a multi-blackboard architecture that pools processing modules around blackboards representing intermediate results at each processing stage. It turned out that such a multi-blackboard approach is much more efficient than the more general multi-agent architecture used in the first Verbmobil prototype. Due to the huge amount of interaction between modules a multi-agent architecture with direct communication among module agents would imply 2380 different interfaces for message exchanges between the 69 agents (see Klüter et al., in this volume for more details). Figure 8 summarizes the advantages of

the multi-blackboard approach vs. the multi-agent approach for the Verbmobil architecture.

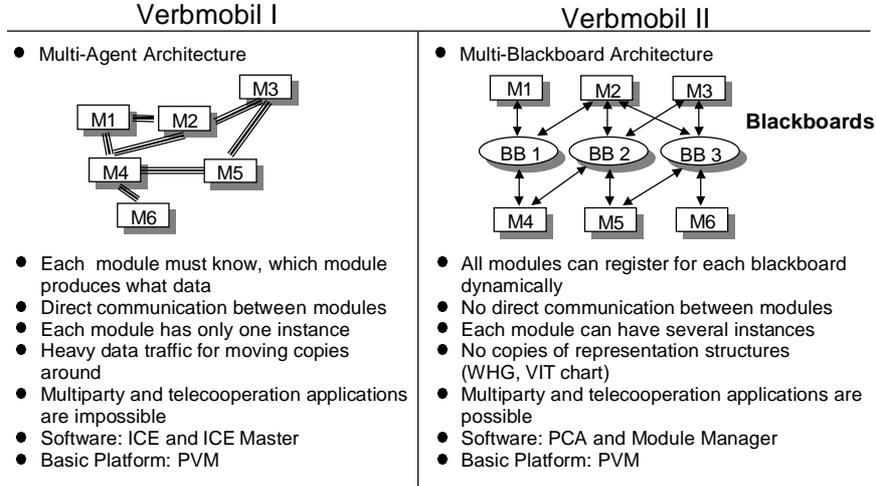


Fig. 8. A comparison of the architecture of Verbmobil I and II

In a multi-blackboard architecture based on packed representations at all processing stages (speech recognition, parsing, semantic processing, translation, generation, speech synthesis) using charts with underspecified representations the results of concurrent processing threads can be combined in an incremental fashion. All results of concurrent processing modules come with a confidence value, so that selection modules can choose the most promising results at each processing stage or delay the decision until more information becomes available. Packed representations such as the WHG (Word Hypotheses Graph) and VHG (VIT Hypotheses Graph) together with formalisms for underspecification capture the non-determinism in each processing phase, so that the remaining uncertainties can be reduced by linguistic, discourse and domain constraints as soon as they become applicable.

VITs (Verbmobil Interface Terms) are used as a multi-stratal semantic representation by the central blackboards for the deep processing threads in Verbmobil. The semantic representation in a VIT is augmented by various features concerning morpho-syntax, tense, aspect, prosody, sortal restrictions and discourse information. VITs form the input and output of the modules for robust semantic processing and semantic -based transfer. The initial design of the VIT representation language was inspired by underspecified discourse representation structures. VITs provide a compact representation of lexical and structural ambiguities and scope underspecification of quantifiers, negations and adverbs. Figure 9 illustrates the role of VITs as a common semantic representation language for the blackboards of Verbmobil.

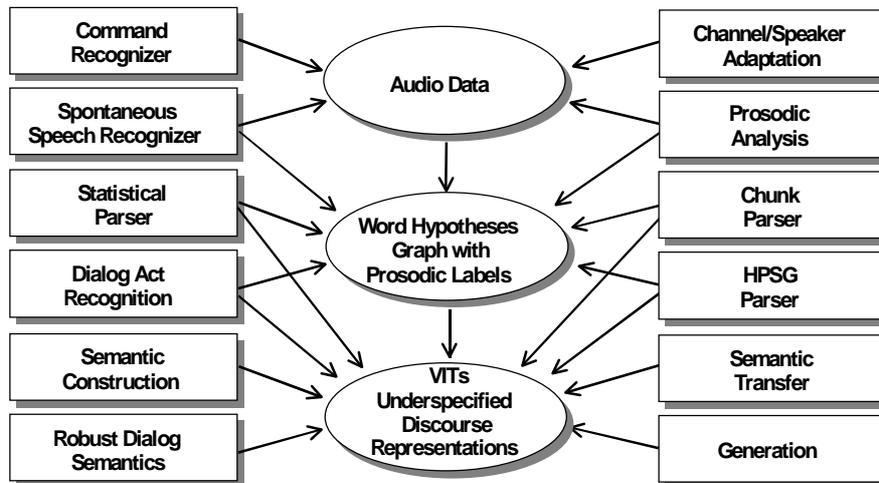


Fig. 9. Some key blackboards with their subscribing modules

7 Verbmobil's Multi-Engine Approach

Verbmobil performs language identification, parsing and translation with several engines simultaneously. Whereas the multi-engine parsing results are combined and merged into a single chart, a statistical selection module chooses between the alternate results of the concurrent translation threads, so that only a single translation is used for generating the system's output.

Verbmobil uses three parallel parsing threads: an incremental chunk parser, a probabilistic LR parser and a HPSG parser. These parsers cover a broad spectrum with regard to their robustness and accuracy. The chunk parser (see Hinrichs et al. (a), in this volume) produces the most robust but least accurate results, whereas the HPSG parser delivers the most accurate but least robust analysis. All parsers process the same word hypotheses graph with its prosodic annotations. The search for the best scored path (according to the acoustic score and the language model) is controlled by a central A* algorithm that guides the three parsers through the word hypotheses graph. The HPSG parser may return more than one analysis for ambiguous inputs, whereas the chunk parser and statistical parser return always only one result. Each parser uses a semantic construction component to transform its analysis results into a semantic representation term. Even partial results of the different parsing engines are integrated into a chart of VITs, that is further analyzed by the robust semantic processing component.

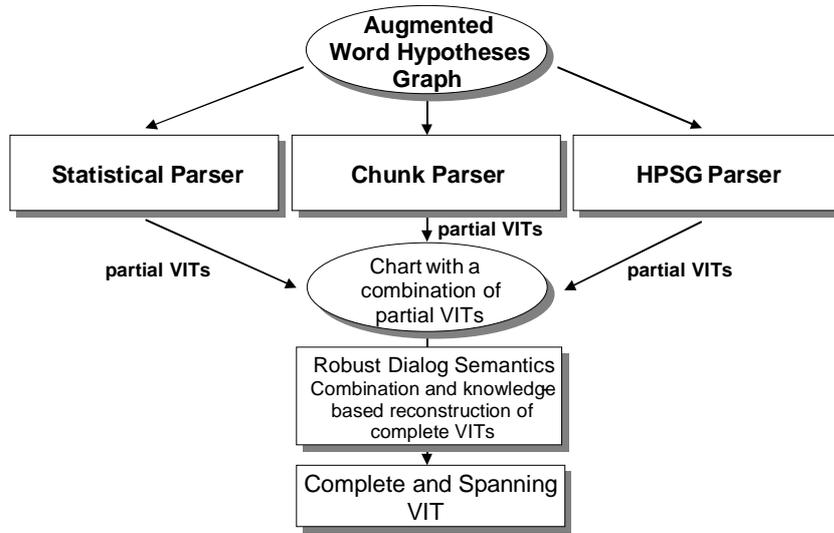


Fig. 10. Verbmobil's multi-engine parsing approach

The final Verbmobil system includes five translation engines (see Figure 11): statistical translation, case-based translation, substring-based translation, dialog-act based translation, and semantic transfer. These engines cover a wide spectrum of translation methods. While statistical translation is very robust against speech recognition problems and produces quick-and-dirty results, semantic transfer is computationally more expensive and less robust but produces higher quality translations. However, it is one of the fundamental insights gained from the Verbmobil project, that the problem of robust, efficient and reliable speech-to-speech translation can only be cracked by the combined muscle of deep and shallow processing approaches.

The language identification component of Verbmobil uses also a multi-engine approach to identify each user's input language. The three instances of the multilingual speech recognizer for German, English, and Japanese run concurrently for the three first seconds of speech input. A confidence measure is used to decide which language is spoken by a particular dialog participant. The language identification component switches to the selected recognizer that produces a word hypotheses graph for the full utterance. Verbmobil's error rate for this type of language identification task is only 7.3% (see Waibel et al., in this volume).

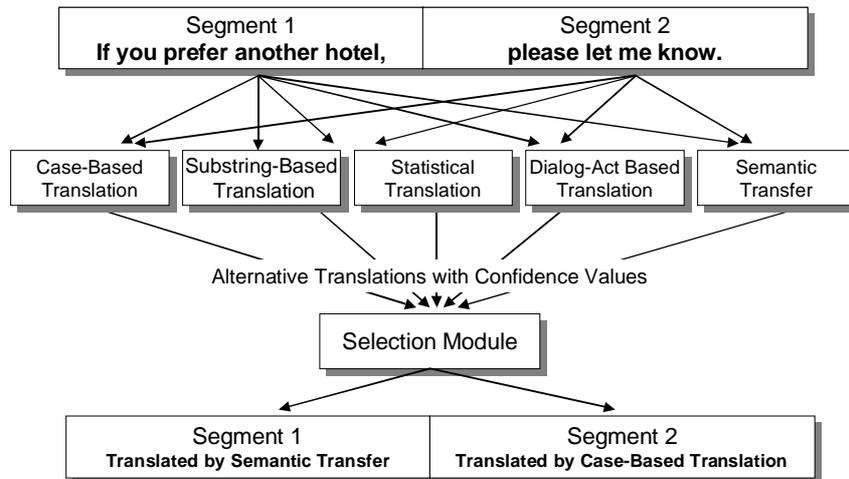


Fig. 11. The multi-engine translation approach of Verbmobil

8 Summarizing Dialogs

Another novel functional feature of Verbmobil is the ability to generate dialog summaries. Suppose that two speakers negotiate a travel plan: one can ask the system either to specify the final agreement, omitting the negotiating steps, or to summarize the steps of argument while leaving out irrelevant details of wording. A dialog summary can be produced on demand after the end of a conversation (see Figure 12).

The summaries are based on the semantic representation of all dialog turns stored in the dialog memory of Verbmobil. It is interesting to note that dialog summaries are mainly a by-product of the deep processing thread and the dialog processor of Verbmobil. The most specific accepted negotiation results are selected from the dialog memory (see Alexandersson et al., in this volume). The semantic-based transfer component and the natural language generators for German and English are used for the production of multilingual summaries. This means that after a conversation over a cell phone the participants can ask for a written summary of the dialog in their own language. The dialog summary can be sent as a HTML document using email. In the context of business negotiations Verbmobil's ability to produce written dialog summaries of a phone conversation is an important valued-added service.

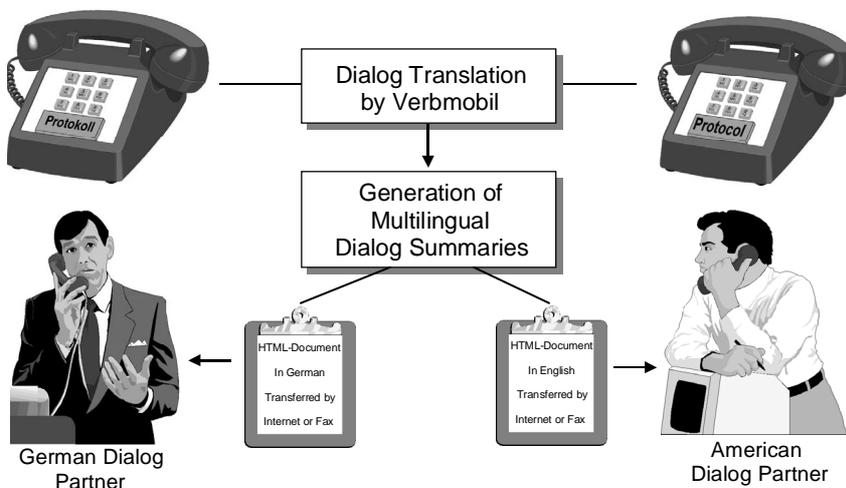


Fig. 12. The generation of a multilingual dialog summaries

9 Conclusion: Some Lessons Learned from the Verbmobil Project

Verbmobil has successfully met the project goals. The technical challenge was to design and implement

- a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations
- that works in an open microphone mode and can cope with speech over GSM mobile phones
- for three language pairs, three domains and a vocabulary size of more than 10000 word forms
- with an average processing time of four times of the input signal duration
- with a word recognition rate of more than 75% for spontaneous speech
- with more than 80% of approximately correct translations that preserve the speaker's intended effect on the recipient in a large-scale translation experiment
- a 90% success rate for dialog tasks in end-to-end evaluations with real users

Various benchmark tests and large-scale end-to-end evaluation experiments with unseen test data have convincingly shown that all these objectives have been met by the final Verbmobil system and some goals have been surpassed (see Tessitore and v. Hahn, and Malenke et al., in this volume).

The broad range of scientific discoveries in the areas of speech, language and discourse processing, dialog translation, language generation and speech synthesis

that resulted from the Verbmobil project are documented in more than 800 publications and this book.

One of the main lessons learned from the Verbmobil project is that the problem of speech-to-speech translation of spontaneous dialogs can only be cracked by the combined muscle of deep and shallow processing approaches:

- deep processing can be used for merging, completing and repairing the results of shallow processing strategies
- shallow methods can be used to guide the search in deep processing
- statistical methods must be augmented by symbolic models to achieve higher accuracy and broader coverage
- statistical methods can be used to learn operators or selection strategies for symbolic processes

The final Verbmobil architecture supports large and robust dialog systems and maximizes the necessary interaction between processing modules:

- in Verbmobil's multi-blackboard and multi-engine architecture, that is based on packed representations on all processing levels and uses charts with underspecified multi-stratal representations, the results of concurrent processing threads can be combined in an incremental fashion
- all results of concurrent and competing processing modules come with a confidence value, so that statistically trained selection modules can choose the most promising result at each stage, if demanded by a following processing step
- packed representations together with formalisms for underspecification capture the uncertainties in each processing phase, so that the uncertainties can be reduced by linguistic, discourse and domain constraints as soon as they become applicable.

Verbmobil has shown the need to take software engineering considerations in language technology projects seriously. Verbmobil's system integration group included professional software engineers with no particular language or speech background; they were responsible for ensuring that the software is robust and maintainable, and that modules developed in different programming languages by a distributed team fit together properly. These issues are too important to leave to subject specialists who see them as a side issue. An important achievement of the Verbmobil project is the consistent integration of a very large number of modules created by diverse groups of researchers from disparate disciplines and to produce a set of capabilities which have not been demonstrated in an integrated speech-to-speech translation system before.

Organizationally, Verbmobil underlines the importance of competition among research teams, with frequent objective evaluations. Competition was fostered naturally within the Verbmobil framework, because the processing model itself is a competitive one. Crucial to the success of Verbmobil was the fact that various teams within the project developed rival solutions to particular tasks, with formal evaluations being used to winnow out the most successful or to combine it with the next best solutions to improve the overall performance of the system.

The objective of the public funding provided by the German Ministry of Education and Research (BMBF) for the Verbmobil Project has been to move German language technology into a position of leadership. BMBF aimed to bring German language technology to the stage of achieving real industrial impact by the turn of the century. Participating companies developing spin-offs at their own expense have already brought eighteen products to market that are all based on results from Verbmobil. There have been various patents and inventions resulting from Verbmobil, in areas such as speech processing, parsing, dialog, machine translation and generation. Six spin-off companies in language technology have been created by former Verbmobil researchers. At present, the Verbmobil industrial partners (DaimlerChrysler, Philips, Siemens, Temic) are among the top European companies using language technology in the marketplace.

The sharable language resources collected and distributed during the Verbmobil project will be useful beyond the project lifespan, since these transliterated and richly annotated corpora of spontaneously spoken dialogs can be used for building, improving or evaluating natural language and speech algorithms or systems in coming years.

Along the way, the Verbmobil project has done a great deal to bring researchers in Germany together across the language/speech and the academic/industrial divides. This is an important contribution from the point of view of a long-range research policy for the field of human language technology. More than 900 young researchers (among them 238 master students, 164 PhD students, and 16 habilitation postdocs) gained experience in advanced speech and language technology through their work on Verbmobil during the project lifespan from 1993 – 2000.

References

- Alexandersson, J., Poller P., and Kipp, M. Generating Multilingual Dialog Summaries. In *this volume*.
- Auerswald, M. Example-Based Machine Translation with Translation Templates. In *this volume*.
- Becker, T., and Lopez, P. The Application of HPSG-to-TAG Compilation Techniques. In *this volume*.
- Becker, T., Kilger, A., Lopez, P., and Poller, P. The Verbmobil Generation Component VM-GECO. In *this volume*.
- Block, H. Example-Based Incremental Synchronous Interpretation. In *this volume*.
- Bub, T., Wahlster, W., and Waibel, A. (1997). Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, München, Germany*, 71–74, IEEE.
- Burger, S., Weilhammer, K., Tillmann H., and Schiel, F. Verbmobil Data Collection and Annotation. In *this volume*.
- Emele, M., Dorna, M., Lüdeling, A., Zinsmeister, H., and Rohrer C. Semantic-Based Transfer. In *this volume*.
- Haiber, U., Mangold H., Regel-Brietzmann P, Ruske, G., and Schleiß, V. Robust Recognition of Spontaneous Speech. In *this volume*.

- Hinrichs, E.W., Kübler, S., Kordoni, V., and Müller, F. (a) Robust Chunk Parsing for Spontaneous Speech. In *this volume*.
- Hinrichs, E.W., Bartels, J., Kawata Y., Kordoni, V., and Telljohann, H. (b) The Tübingen Treebanks for Spoken German, English, and Japanese. In *this volume*.
- Jekat, S., and v. Hahn, W. Multilingual Verbmobil-Dialogs: Experiments, Data Collection and Data Analysis. In *this volume*.
- Kay, M., Gawron, J.M., and P. Norvig. (1994) Verbmobil: A Translation System for Face-to-Face Dialog. Stanford, CSLI Lecture Notes No. 33 CA: CSLI Publications
- Kiefer, B., Krieger, H.-U., and Nederhof, M.-J. Efficient and Robust HPSG Parsing of Word Graphs. In *this volume*.
- Kipp, M., Alexandersson, J., Reithinger, N., and Engel, R. Dialog Processing. In *this volume*.
- Malenke, M., Bäumler, M., and Paulus, E. Speech Recognition Performance Assessment. In *this volume*.
- Klüter, A., Ndiaye, A., and Kirchmann, H. Verbmobil from a Software Engineering Point of View: System Design and Software Integration. In *this volume*.
- Pinkal, M., Rupp C.J., and Worm, K. Robust Semantic Processing of Spoken Language. In *this volume*.
- Schiehlen, M. Semantic Construction. In *this volume*.
- Spilker J., Klarner, M., and Görz, G. Processing Self-Corrections in a Speech-to-Speech System. In *this volume*.
- Stöber, K., Wagner, P., Helbig, J., Köster, S., Stall, D., Thomae, M., Blauert, J., Hess, W., Hoffmann, R., and Mangold, H. Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora. In *this volume*.
- Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A., and Yamamoto, S. (1998): A Japanese-to-English Speech Translation System: ATR-MATRIX. In *Proceedings of the ICSLP*, 957–960.
- Tessitore, L., and v. Hahn, W. Functional Validation of a Machine Interpretation System: Verbmobil. In *this volume*.
- Vogel, S., Och, F.J., Tillmann, C., Niessen, S., Sawaf, H., and Ney, H. Statistical Methods for Machine Translation. In *this volume*.
- Waibel, A., Soltau, H., Schultz, T., Schaaf, Th., and Metze, F. Multilingual Speech Recognition. In *this volume*.
- Wahlster, W (1993): Verbmobil: Translation of Face-to-Face Dialogs. In *Proceedings of the Fourth Machine Translation Summit*, Kobe, Japan, 128–135.
- Woszczyna, M., ed., (1999) Proceedings of the C-STAR Workshop. Schwetzingen, Germany, September 1999.