In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence,
Seattle, Washington, August 2001, San Francisco: Morgan Kaufmann, Vol. 2, pp. 1484 - 1493.

Robust Translation of Spontaneous Speech: A Multi-Engine Approach

Wolfgang Wahlster
DFKI
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
wahlster@dfki.de

Abstract

Verbmobil is a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs that can be accessed via GSM mobile phones. It handles dialogs in three business-oriented domains, with context-sensitive translation between four languages (English, German, Japanese, and Chinese). We show that in Verbmobil's multi-blackboard and multi-engine architecture the results of concurrent processing threads can be combined in an incremental fashion. We argue that all results of concurrent processing modules must come with a confidence value, so that statistically trained selection modules can choose the most promising result. 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. 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. One of the main lessons learned from the Verbmobil project is that the problem of speech-to-speech translation can only be cracked by the combined muscle of deep and shallow processing approaches.

1 Introduction

Verbmobil is a software system that provides mobile phone users with simultaneous dialog interpretation services for restricted topics [Wahlster, 1993; 2000b]. As the name Verbmobil suggests, the system supports **verb**al communication with foreign interlocutors in **mobil**e 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. Verbmobil uses an explicit dialog memory and exploits domain knowledge. The dialog context is used to resolve ambiguities and to produce an adequate translation in a particular conversational situation.

Figure 1: Mobile speech-to-speech translation with Verbmobil

Figure 1 illustrates the use of Verbmobil in a travel scenario. Let's suppose that an American business traveller has arrived at Frankfurt airport and wants to call Mrs. Meyer, the secretary of his German business partner. Since he does not speak German and knows that the secretary does not speak English, he activates Verbmobil using the voice dialing mode of his cell phone. After telling Verbmobil the phone number of Mrs. Meyer, the speech translation system initiates a conference call between the American traveller, the German secretary and Verbmobil. Verbmobil translates all input of the American speaker into German and all input of the German speaker into English.

Verbmobil is the first speech-only dialog translation system. Verbmobil users can simply pick up a standard mobile phone and use speech commands in order to initiate a dialog translation session (see Figure 2). 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.

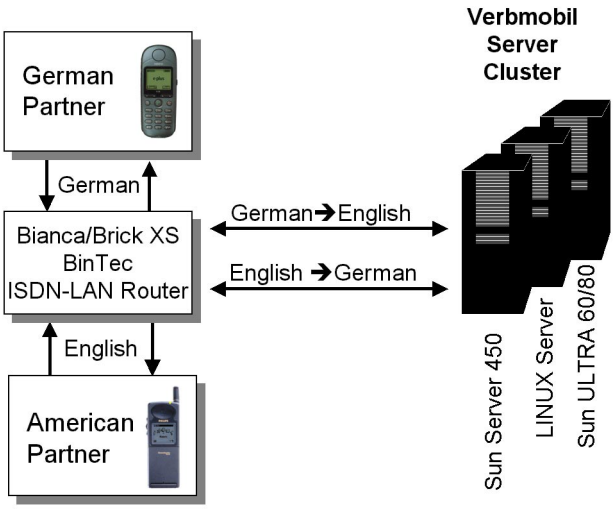

Figure 2: Three-party conference calls with Verbmobil

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 (cell phone) quality. Thus, Verbmobil includes different speech recognizers for 16 kHz and 8 kHz sampling rates. 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 pronounciation variants due to the dialectal diversity of the user community.

2 Understanding Spontaneous Speech

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.

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 [Spilker et al, 2000]. The word lattice is augmented by an additional path that does no

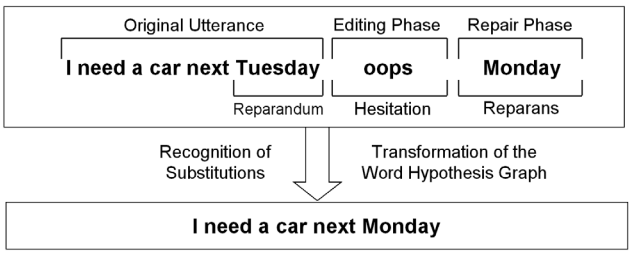

Figure 3: Repairing self-corrections

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 (see Figure 3).

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 [Pinkal et al., 2000]

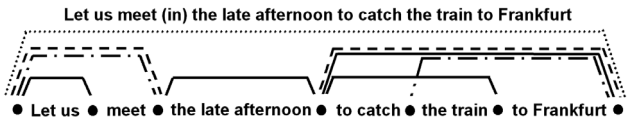

Figure 4: Finding a spanning analysis by type-raising

Due to a speech recognition error or a corrupted input signal, the word hypotheses graph in Figure 4 does not contain any temporal preposition in front of the temporal nominal phrase *the late afternoon*. A type coercion rule maps this phrase to a temporal modifier that expresses an underspecified temporal relation, that is later lexicalized as the default *in* during language generation.

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.

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. The travel planning scenario with a vocabulary of 10,000 words was used for the end-to-end evaluation of the final Verbmobil system. The PC maintenance task had a much larger vocabulary of almost 35,000 words from IT sublanguage lexica. Verbmobil is a hybrid system incorporating both deep and shallow processing schemes [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' handcrafted knowledge sources to achieve an adequate level of robustness and accuracy.

3 Verbmobil's Training Corpora

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. The so-called partitur (German word for musical score) format used for the Verbmobil speech corpora orchestrates fifteen strata of annotations (see Figure 5, [Burger et al., 2000]). Multi-channel recordings of 3,200 spontaneous dialogs with 79,562 turns from 1,658 different speakers were transcribed and distributed on 56 CDs with a total of 21,5 GB of annotated speech corpora (available from BAS, see www.phonetik.uni-muenchen.de/Bas/BasKorporaeng.html).

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. Three treebanks for German, English and Japanese have been developed with 85,000 trees annotated on three strata: morpho-syntax, phrase structure, and predicate-argument structure. The treebanks were used to train the statistical parser and the chunk parser. In addition, machine learning methods were applied to the treebanks to extract semantic construction rules and transfer rules for translation. 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.

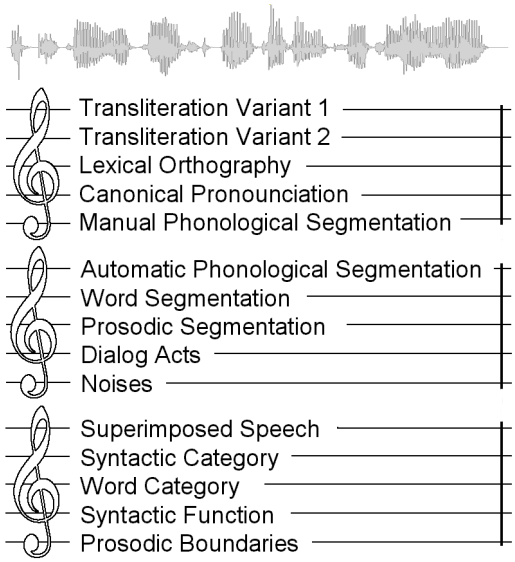

Figure 5: Verbmobil's multi-stratal annotation of speech

4 The Anatomy of Verbmobil

A distinguishing feature of Verbmobil is its multi-engine parsing and translation architecture. The screenshot of Verbmobil's control panel provides an overview of the main components of the system (see Figure 6). 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. Since Verbmobil is a multilingual system it incorporates four speech recognizers and four speech synthesizers for German, English, Japanese, and Chinese.

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. 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

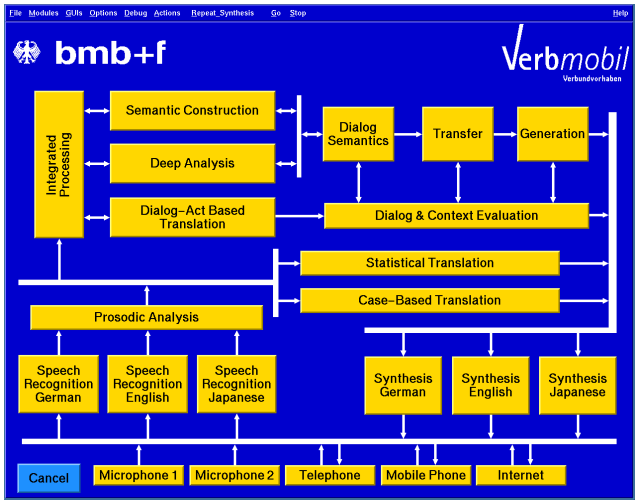

Figure 6: A snapshot of Verbmobil's control panel

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 (UDRS, [Reyle, 1993]. VITs provide a compact representation of lexical and structural ambiguities and scope underspecification of quantifiers, negations and adverbs. The linguistic information is encoded into variable-free sets of non-recursive terms (see Figure 7). These streams of literals serve as flat multi-stratal representations that are very efficient for incremental processing. The various linguistic strata are cross related by a labelling system. Since VIT terms are the central information structure in Verbmobil, they are treated as an abstract data type. VITs are used as a common representation scheme for linguistic information exchange between all components and processing threads of Verbmobil.

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 [Schiehlen 2000]. 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 [Kiefer et al., 2000]

```
Vit (vitID (sid (104,a,en,10,80,1,en,y,semantics), % SegmentID
          [word (he, 1, [26]),                     % WHG String
           word(is, 2, []),
           word(coming, 3, [27]),
           word(at, 4, [36]),
           word(the ,5, [28]),
           word(beginning, 6, [35]),
           word(of, 7, [35]),
           word(``August'', 8, [34])]),
     index (38, 25 ,i35),                          % Index
    [beginning (35, i37),                          % Conditions
     arg3 (35, i37 ,i38),
     come (27, i35),
     arg1 (27, i35, i36),
     decl (37, h43),
     pron (26, i36),
     at (36, i35, i37),
     mofy (34 ,i38, aug),
     def (28, i37, h42, h41),
     udef (31, i38, h45, h44)],
    [in_g (26, 25), in_g (37, 38),                 % Constraints
     in_g (27, 25), in_g (28, 30),
     in_g (31, 33), in_g (34, 32),
     in_g (35, 29), in_g (36, 25),
     leq (25, h41), leq (25, h43),
     leq (29, h42), leq (29, h44),
     leq (30, h43), leq (32, h45),
     leq (33, h43)],
    [s_sort (i35, situation),                      % Sorts
     s_sort (i37, time),
     s_sort (i38, time)],
    [dialog_act (25, inform),                      % Discourse
     dir (36, no),
     prontype (i36, third,std)],
    [cas (i36, nom),                               % Syntax
     gend (i36, masc),
     num (i36, sg), num (i37, sg), num (i38, sg),
     pcase (l135, i38, of)],
    [ta_aspect (i35, progr),                       % Tense and Aspect
     ta_mood (i35, ind),
     ta_perf (i35, nonperf),
     ta_tense (i35, pres)],
    [pros_accent (l135)]                           % Prosody
```

Figure 7: VIT for "He is coming at the beginning of August"

The statistical translation module starts with the single best sentence hypothesis of the speech recognizer [Vogel et al., 2000]. 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 [Block, 2000]. The other component for case-based translation is based on 30,000 translation templates learned from a sentence-aligned corpus. 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.

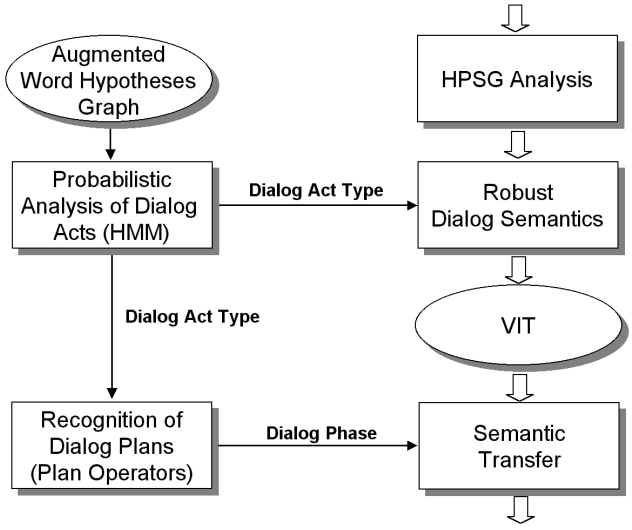

Figure 8: The use of stochastic dialog act and plan recognition

The dialog component includes a plan processor, that structures an ongoing dialog hierarchically in different dialog phases, games and moves. Dialog acts are the terminal nodes of the tree structure that represents the dialog structure. Information about the dialog phase is used e.g. during the semantic-based transfer for disambiguation tasks (see Figure 8). 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. Consider the German sentence *Wir treffen die Kollegen in Berlin* and its English equivalent *We will meet the colleagues in Berlin*. English and German have the same PP-attachment ambiguity in which *in Berlin* is either attached to the noun phrase *the colleagues* or to the verb *meet*.

The transfer component is basically a rewriting system for underspecified semantic representations using Verbmobil's VIT formalism [see Emele et al., 2000]. 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 20,000 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 , 2000]. 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.

The speech synthesizer for German and American English follows a concatenative approach based on a large corpus of annotated speech data. 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.

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.

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 bye-product of the deep processing thread and the dialog processor of Verbmobil. The most specific accepted negotiation results are selected from the dialog memory [Alexandersson et al., 2000]. 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 an 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 value-added service.

5 Exploiting Prosodic Information

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 (see Figure 9). 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 fully operational 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.

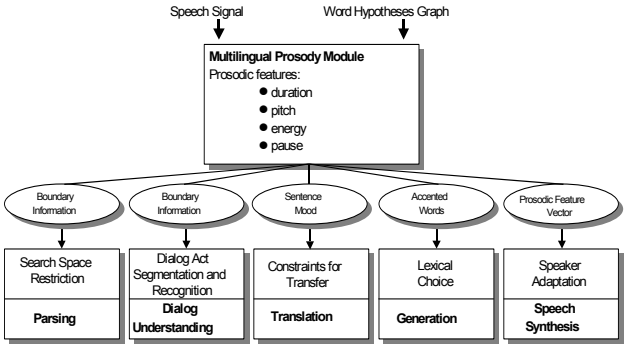

Figure 9: The role of prosodic information in Verbmobil

6 Verbmobil's Multi-Blackboard Architecture

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.

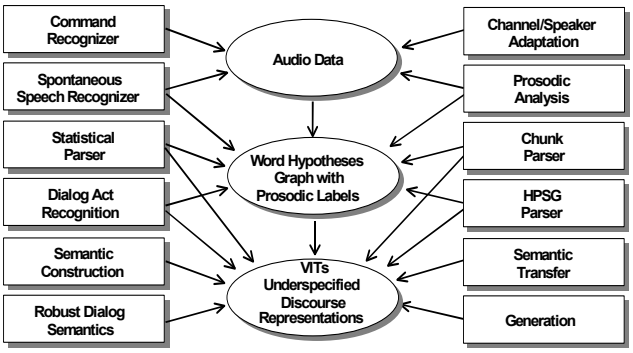

Figure 10**:** Some key blackboards with their subscribing modules

The final Verbmobil system is based on a multi-blackboard architecture that pools processing modules around blackboards representing intermediate results at each processing stage (see Figure 10). 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.

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.

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 [Hinrichs et al., 2000] 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.

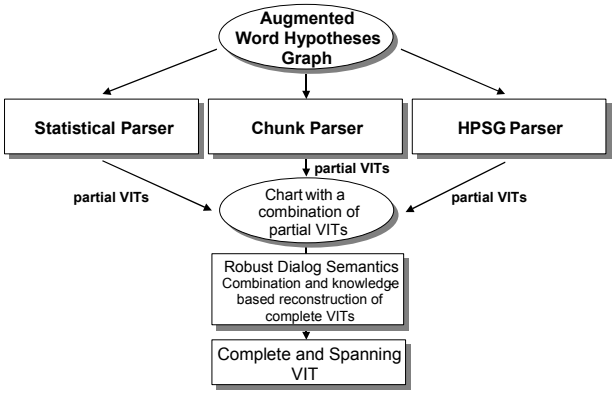

Figure 11: Verbmobil's multi-engine parsing approach

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 (see Figure 11).

The final Verbmobil system includes five translation engines (see Figure 12): 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 translation quality of the final Verbmobil system was rigorously evaluated. 65 evaluators checked 43,180 Verbmobil translations and judged their correctness. We call a translation "approximately correct", if it preserves the intention of the speaker and the main information of his utterance. Table 1 shows clearly that no single translation engine achieves more than 81% approximately correct translations, but that the selection of the appropriate translation result increases the overall performance significantly. In Verbmobil, we used the judgements of the human evaluators (see Table 1, Manual Selection) to construct a training corpus for an instance-based learning algorithm that picks the best translation for a given WHG of a particular turn segment (see Table 1, Automatic Selection).

Translation Thread	Word Accuracy ≥ 50% 5069 Turns	Word Accuracy ≥ 75% 3267 Turns	Word Accuracy ≥ 80% 2723 Turns
Case-based Translation	37%	44%	46%
Statistical Translation	69%	79%	81%
Dialog-Act based Translation	40%	45%	46%
Semantic Transfer	40%	47%	49%
Substring-based Translation	65%	75%	79%
Automatic Selection	**78%**	**83%**	**85%**
Manual Selection	**88%**	**95%**	**97%**

Table 1: Quality of Translations from German to English

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., 2000]. Verbmobil's architecture supports multiple process instances of all components, so that Verbmobil can be used as a translation server for multiparty dialogs (e.g. two Germans, a Japanese and an American planning a joint trip).

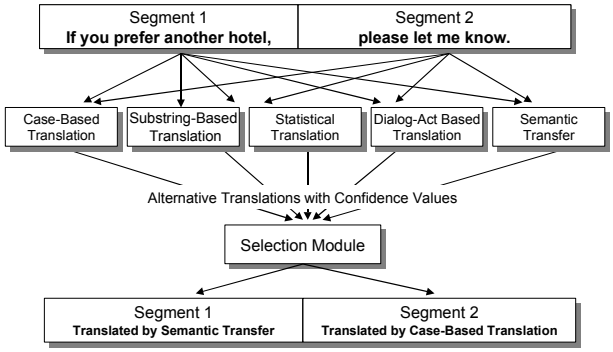

Figure 12. The multi-engine translation approach of Verbmobil

8 Lessons Learned from Verbmobil

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 (www.verbmobil.de) and a comprehensive book [Wahlster, 2000a].

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. In particular, underspecification allows disambiguation requirements to be delayed until later processing stages where better-informed decisions can be made.
– The massive use of underspecification makes the syntax-semantic interface and transfer rules almost deterministic, thereby boosting processing speed.

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 technology 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 Federal Ministry of Education and Research (BMBF) for the Verbmobil Project has been to bring European language technology to the stage of achieving real industrial impact by the turn of the century. Participating companies developing spin-off applications at their own expense have already brought twenty 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. Seven spin-off companies in language technology have been created by former Verbmobil researchers. For example, AixPlain (www.aixplain.de) markets speech translation systems, Sympalog (www.sympalog.de) develops spoken dialog systems, and XtraMind (www.xtramind.com) delivers email response systems based on Verbmobil technology. At present, Verbmobil's large 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.

9 Conclusion and Future Work

Although Verbmobil was a high-risk and long-term project (1993 – 2000), it has successfully met its technical project goals. The Verbmobil consortium brought together 31 partners across three continents. The total amount of public and private funding was about $80 million, resulting in Europe's largest AI project. 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 four language pairs, three domains and a vocabulary size of more than 10,000 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 [Tessiore and v. Hahn, 2000].

SmartKom (1999-2003) is the follow-up project to Verbmobil and reuses some of Verbmobil's components for

the understanding of spontaneous dialogs. SmartKom is a multimodal dialog system that combines speech, gesture, and mimics input and output [Wahlster et al, 2001]. Spontaneous speech understanding is combined with the video-based recognition of natural gestures. One of the major scientific goals of SmartKom is to design new computational methods for the seamless integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level. SmartKom is based on the situated delegation-oriented dialog paradigm (SDDP), in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display. The main contractor of the SmartKom consortium is the German Research Center for Artificial Intelligence (DFKI). The major industrial partners involved in SmartKom are DaimlerChrysler, Philips, Siemens and Sony.

References

[Alexandersson et al., 2000], Jan Aleandersson., Peter Poller, and Michael Kipp, Generating Multilingual Dialog Summaries and Minutes. In: [Wahlster, 2000a] 507-518.

[Becker et al, 2000] Tilman Becker, Anne Kilger, Patrice Lopez, and Peter Poller. The Verbmobil Generation Component VM-GECO. In [Wahlster, 2000a], 481-496.

[Block, 2000] Hans Ulrich Block. Example-Based Incremental Synchronous Interpretation. In [Wahlster, 2000a], 411-417.

[Bub et al., 1997] Thomas Bub, Wolfgang Wahlster, and Alex Waibel, A. 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 *1997*.

[Burger et al, 2000] Susanne Burger, Karl Weilhammer, Florian Schiel, and Hans G. Tillmann. Verbmobil Data Collection and Annotation. In [Wahlster, 2000a], 537-549.

[Emele et al., 2000] Martin Emele, Micheal Dorna, Anke Lüdeling, Heike Zinsmeister, and Christian Rohrer. Semantic-Based Transfer. In [Wahlster, 2000a], 359-376.

[Hinrichs et al., 2000] Erhard Hinrichs, Sandra Kübler, Valia Kordoni, and Frank Müller. Robust Chunk Parsing for Spontaneous Speech. In [Wahlster, 2000a], 163-182.

[Kiefer et al., 2000] Bernd Kiefer, Hans-Ulrich Krieger, and Mark Jan Nederhof. Efficient and Robust Parsing of Word Hypotheses Graphs. In [Wahlster, 2000a], 279-295.

[Pinkal et al., 2000] Manfred Pinkal, C.J. Rupp, and Karsten Worm. Robust Semantic Processing of Spoken Language. In [Wahlster, 2000a], 321-335.

[Reyle, 1993] Uwe Reyle. Dealing with Ambiguities by Under-specification: Construction, Representation and Deduction. Journal of Semantics 10 (2): 123-179, 1993.

[Schiehlen, 2000] Michael Schiehlen. Semantic Construction. In [Wahlster, 2000a], 200-215.

[Spilker et al., 2000] Jörg Spilker, Martin Klarner, and Günther Görz. Processing Self-Corrections in a Speech-to-Speech System. In [Wahlster, 2000a], 131-140.

[Takezawa et al., 2000] Toshiyuki Takezawa, Tsuyoshi Morimoto, Yonishori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. A Japanese-to-English Speech Translation System: ATR-MATRIX. In *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP*, Sydney, 957–960, 1998.

[Tessiore and v. Hahn, 2000] Lorenzo Tessiore and Walther v. Hahn. Functional Validation of a Machine Interpretation System: Verbmobil. In [Wahlster, 2000a], 611-631.

[Vogel et al, 2000] Stepham Vogel, Franz Josef Och, Christoph Tillmann, Sonja Niessen, Hassan Sawaf, and Hermann Ney. Statistical Methods for Machine Translation. In [Wahlster, 2000a], 377-393.

[Waibel et al, 2000] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze. Multilingual Speech Recognition. In [Wahlster, 2000a], 33-45.

[Wahlster, 1993] Wolfgang Wahlster. Verbmobil: Translation of Face-to-Face Dialogs. In *Proceedings of the Fourth Machine Translation Summit*, Kobe, Japan, 128–135.

[Wahlster, 2000a] Wolfgang Wahlster (ed.). Verbmobil: Foundations of Speech-to-Speech Translation. Berlin, New York: Springer 2000.

[Wahlster, 2000b] Wolfgang Wahlster. Mobile Speech-to-Speech Translation: An Overview of the Final Verbmobil System. In: [Wahlster, 2000a], 3-21.

[Wahlster et al., 2001] Wolfgang Wahlster, Norbert Reithinger, and Anselm Blocher. SmartKom: Multimodal Communication with a Life-Like Character. In: *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Eurospeech, 2001.