Dialogue Management in Verbmobil VRP1

DISC partner: MIP
Authors: Niels Ole Bernsen and Laila Dybkjær

1. Introduction

This paper present an analysis of dialogue management in the Verbmobil First Phase Research Prototype 1.0 demonstrator (VRP1). The analysis is presented in the form of (a) a 'grid' which describes the system's properties with particular emphasis on dialogue management and evaluation results, (b) a life-cycle model which provides a structured description of the system's development and evaluation process, and (c) supporting material, such as system architecture, example screen shots, dialogues and scenarios.

The presented information will be cross-checked with the developers of Verbmobil as well as with the complementary descriptions of other aspects of Verbmobil provided by the DISC partners. These other descriptions address **speech recognition, done by LIMSI, speech generation, done by KTH, language understanding and generation, done by IMS, dialogue management, done by IMS, human factors, done by Vocalis, system integration, done by Vocalis.**

Demonstrator: Available in November 1997. Phone demo planned for 10-97.
Developer: Verbmobil consortium (4 companies, 16 universities, 3 subcontractors outside Germany).
Contact: Reinhard Karger M.A., DFKI GmbH, Stuhlsatzenhausweg 3, D 66 123 Saarbrücken, Germany. Email: karger@dfki.de. URL: http://www.dfki.de/verbmobil.

Development phases:

Phase 1 (1993-1996):

1. Verbmobil Demonstrator, CeBIT 1995. 1292 words. Translates German appointment scheduling input into spoken English output.

2. Verbmobil Research Prototype 1.0 (VRP1), CeBIT 1997. 2461 words. Translates German and Japanese appointment scheduling input into spoken English output. Performs German-German clarification dialogues with users. See Figure 1.

Phase 2 (1997-2000): Verbmobil-2 (V2). Not discussed here.

Verbmobil grid

The interaction model of Verbmobil, First Phase

The Verbmobil First Phase Research Prototype 1.0 demonstrator (VRP1) is a research prototype of a stand-alone appointment scheduling spoken dialogue translation support system which performs uni-directional German-to-English and Japanese-to-English translation. The Japanese-to-English part of the system will be disregarded in what follows. **OK? The Japanese part seems much smaller and we cannot really do anything with it. Its dialogue manager probably is the same as for German-English.**

URL: **http://www.dfki.de/verbmobil**

System performance	
Cooperativity	The issue of cooperative design of system utterances only applies to the meta-communication in German between VRP1 and the user. **During phase I of Verbmobil, there were working packets that dealt with the questions of user acceptance of VERBMOBIL. An overview of the work is given in [Krause 1997]. Since Verbmobil doesn't interact with the user during normal operation, the main point to ensure was that it (almost) always translates an utterance. That is the cooperative behaviour users expect from a translation system: users simply don't want interactions with Verbmobil as a third dialogue partner.** **In the case of clarification dialogues - which is the only mode of operation where Verbmobil interrupts the dialogue of the human dialogue partners - the dialogues were based on the results of the simulations: no long clarification dialogues; allow only simple yes/no answers to system questions; if you can't immediately resolve, ask the user to repeat his/her utterance to be translated.**
Initiative	Domain communication is fully natural (conversational) mixed initiative human-human communication. Meta-communication is done in German between VRP1 and user. **Can we say: Both the system and the user may initiate meta-communication?**
Influencing users	**The system is intended to be a walk-up-and-use system. Users are advised to remain in the domain of the system. Is there a system's introduction? Is it optional? Any important limitations here (e.g. lack of control of aspects of user behaviour which should be controlled)? The user is not "controlled" since Verbmobil is not a dialogue partner. Evaluated? How? Results? Was the evaluation procedure appropriate? Note that the sysem cannot be walk-up-and-use because it is speaker adaptive.**
Real-time	The system responds in six times real-time. **This is a strong limitation on the present version of the system as close-to-real-time is desirable.**
Transaction success	Defined as speaker-intended contents/dialogue acts which are transferred into understandably equivalent contents in the target language. VRP1 produced +70% approximately correct translations in the domain of appointment scheduling. Evaluation was done by interpreters. **+70% is insufficient for a realistic application. What is the target percentage for realistic applications?** **The notion of transaction success proper is hard to apply to VRP1, e.g. as the proportion of successful, fully translated real-life dialogues in the domain. However, transaction success as defined for VRP1 was measured on isolated corpus sentences rather than on sentences occurring in real-life human-human dialogues. This makes the reported results non-transferable to real-life dialogues.**
General evaluation	**Has any ISO standards or other well-known methods been used?**

	Which? How? Results? Which evaluation methods have been applied to the system/component? Overall results?
Speech input	
Nature	Continuous; spontaneous; speaker-adaptive with speaker-independent core; German, English. **One dialogue partner is American English speaking**. **Any important limitations here (e.g. on the appropriateness of speaker adaptation to the task(s))? Evaluated? How? Results? Was the evaluation procedure appropriate?**
Device	Close microphone. **Telephone and mobile phone present additional challenges.**
Phone server	**N/A**
Acoustic models	**HMM and Neural Network based. See SR analysis?**
Search	**Viterbi search and A*.**
Vocabulary	2461 words. **It's sufficient for the limited domain. In phase II the domain has been extended and the vocabulary size of the data collected didn't exceed approx. 6000 words (except city and street names). Measured? How? Results? Was the measurement procedure appropriate?** Word **recognition** accuracy 73.3% on random samples taken from previously unencountered input. ~~The Japanese system only has a vocabulary of about 400 words~~. **Any important limitations here (e.g. too low for application)? I found another piece of information: VRP1 has a recognition word error rate of 14%. How does that relate to the above? Check with SR analysis.**
Barge-in	**The system currently does not listen when it speaks. If not, what are the problems caused by that? Were they measured? How? Results? Was the measurement procedure appropriate?**
Word hypotheses	Word hypothesis graph with probabilities. **1-best or n-best? Where do the probabilities come from? Recogniser score values used? Is the approach satisfactory/sufficient? Check with SR analysis.**
Ist a graph with acoustic scores. The word chain to be processed is extracted by the parser modules.	
Grammar	**No grammar in the speech recogniser. Staticstical language models are used.**
Prosody	The prosody module recognises breaks, intonation, duration and energy of the input signal. Use of prosodic information for long-utterance segmentation, grammatical processing (speed-up of syntactical analysis, reduction of candidate interpretations), sense disambiguation, translation **[does this mean: use of prosodic information over and above sense disambiguation?]** and dialogue management **[use of prosodic information for what]. In phase 1 the main information was boundary info, prosodic mood, and accent information. Without boundary info, the system doesn't work. The other information leads to different transfer results.**
Speech output	

Device	**Loudspeaker.**
Language(s)	English; German paraphrases **are not generated in Phase 1. What you can hear is the best chain in the word lattice from the acoustic scores.**
Coded/parametric	Parametric speech. **A concatenative approach is used.** Hesitations etc., in input are just left out in output. True-talk **[??]** is used for English synthesis. **Distinguish between German and English synthesis when necessary. Quality measured? How? Results? Was the measurement procedure appropriate? Is the approach satisfactory/sufficient? Check with SS analysis.**
Prosody	No prosody is included in the German synthesis. **Prosody included in the English synthesis. See SS analysis?**
Voice character	**VRP1 reproduces the voice character of the present speaker (e.g. male or female)? See SS analysis?**

User utterances	
Lexicon	2461 words for German-English translation. The 2461 are for German, the English recognition list in phase 1 is approx 900 words. The linguistic coverage in English is approx. 4000 full forms. For the coverage see above. **Evaluated? How? Results? Was the evaluation procedure appropriate? See NLUG analysis.** **Describe the lexical semantics. Is the approach satisfactory/sufficient? Evaluated? How? Results? Was the evaluation procedure appropriate? See NLUG analysis.**
Grammar	German basic grammar for spontaneous speech. **A bit more? Grammatical coverage is sufficient. See [Bubetal97] for the evaluation.**
Parsing	Applies syntactic and semantic constraints simultaneously. Combines deep and shallow semantic analysis: deep analysis through linguistically-based compositional syntactic-semantic analysis and semantics-based transfer of VITs (classes of underspecified DRSs); shallow, approximate analysis through schematic translation, dialogue act-based translation and statistical translation. Information extraction techniques are used to select the best result for semantic transfer German-to-English ~~and Japanese-to-English~~. **The TRUG parser searches through the lattice and finds syntactically correct paths. It's pretty robust. Measured? How? Results? Was the measurement procedure appropriate? Is the approach satisfactory/sufficient (e.g. wrt. possible domain inferences)? Check with NLUG analysis.**
Style	Free in domain communication. Extremely long sentences possible. **No limitations on the length. Meta-communication is avoided if possible. Is the approach satisfactory/ sufficient (e.g. wrt. the load it imposes on recogniser and grammar, or the restrictions it imposes on the users' utterances)? Does meta-communication require a special style? If so, is the approach satisfactory/sufficient (e.g. wrt. the load it imposes on the users' utterances)?**
System utterances	
Lexicon	**Size of English lexicon? What about the German paraphrases? Is the approach satisfactory/sufficient? Check with NLUG analysis. Approx 4000 words are in the generator lexicon. In phase 1 there is no paraphrase the output at all.**
Grammar	Reversible HPSG grammar for English. **The grammatical coverage is sufficient. Measured? How? Results? Was the measurement procedure appropriate? Meta-communication: ?? Check with NLUG analysis.**
Semantics	Domain communication: expression of the results of semantic transfer. **No meta-communication. Check with NLUG analysis.**
Style	Translation of domain utterances: reduced to expression of core messages. **In case of clarification dialogues, utterances are short, the questions can be answered with yes/no. How characterise the style? Meta-communication style? Is the approach satisfactory/sufficient?**

Multimodal aspects	
None	**The system is unimodal (speech-only).**
Attentional state	
Focus, prior	There is contextual knowledge that is used in the translation process. It covers the domain and the usual dialogue focus stuff. There's no separate evaluation possible.
Sub-task id.	**The system does sub-task identification by means of dialogue acts and dialogue phases. See [AlexanderssonReithinger97, ReithingerKlesen97] Measured? How? Results? Was the measurement procedure appropriate? Is the approach satisfactory/sufficient? Does the system continuously monitor and process the input from the dialogue participants (or is that 2nd phase only?)**
Expectations	Are predictions being used? **[This entry may be redundant.]**
Intentional structure	
Task(s)	Appointment scheduling. Two human dialogue partners have to agree on a date to meet at a certain location. The system translates selective portions of their exchanges from German to English when requested by the users through pressing the Verbmobil button. **The system only handles utterances that are strictly related to the appointment scheduling task and not, for instance, the reasons people are used to giving for their (non-) availability. This unnatural restriction constitutes a difficulty for application to real-life dialogues.**
Task complexity	Ill-structured task. Being negotiation-based, VRP1 cannot be characterised as having to fill a specific number of slots.
Communication	Domain communication: unconstrained within the limitation mentioned under 'task(s) above: VRP1 accepts very long sentences, any order of topics, any number of topics per sentence. ****LD: Is it really unconstrained (people know that they may need Verbmobil to help them). NOB: the problem is that we need theory to make such distinctions. Looking at the transcribed dialogues, I would say that they are unconstrained according to the theoretical metrics we have now: very long sentences, any order of topics, any number of topics per sentence.** **System-initiated meta-communication:** The system performs German-German clarification dialogues with users in case of speech recognition or understanding problems. **Repairs are done by clarification dialogues. Initiation is done by modules that detect an error.** Clarification is initiated if two words are phonetically similar in the recognition or in case of inconsistent time expressions like 18 o'clock in the morning. **How?** **User-initiated meta-communication: The user has to wait for the translation, but can optionally get the best chain. If the speaker doesn't like the translation, he can repeat with other words.** **Problems: The system does not have specific problems as a result of how it communicates about the domain and about the**

	communication.
Interaction level	**Only applicable for meta-communication. Describe the levels involved in the system initiated meta-communication, if any. Does the system have specific problems as a result of its level(s) of communication? What have the developers done to analyse the problems? Was their approach appropriate? [See Krause97].**
Dialogue structure	**How represented? How implemented?**
	What may be parameterised, i.e. how are the intentional and linguistic structures indicated? Which kinds of model are there (e.g. task structure, turn-taking structure)? Is the control model separated from the rest of the system? Is it solely based on semantic information so that it is language independent? Which sub-components does the dialogue manager include? Describe the interfaces between the dialogue manager and the other system components. Describe the interfaces between the sub-components of the dialogue manager. Describe the flow internally in the dialogue manager and externally between the dialogue manager and the other system components. What is produced as output from the dialogue manager? Are dialogue patterns used?

Linguistic structure	
Speech acts	The system identifies speech (or dialogue) acts in the users' input, such as 'suggest_date', which model intended utterance interpretations in abstraction from the performance phenomena of spontaneous speech**. How? How many speech acts does the system recognise? Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? Is there any difference between the system's use of speech acts and its ability to do topic spotting (sub-task identification)?** By means of a) knowlege based methods and b) statistically. See ReithingerKlesen97
Discourse particles	**Particles are partially identified and used for translation. How? How many discourse particles does the system recognise? Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? If the system does not identify discourse particles - does this produce any particular problems?**
Co-reference	**The system does partial co-reference resolution if needed for translation. How? No seperate evaluation was done.**
Ellipses	**The system does partial processing of ellipses if needed for translation. How? No seperate evaluation was done.**
Segmentation	**The system does user turn segmentation.** It's done partly by prosody and syntax. If not done, the system does not run. **Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? If the system does not do any user turn segmentation - does this produce any particular problems?**
Interaction history	
Linguistic	The language of the user, and indirectly the words uttered are being recorded. **What is the record being used for? Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? If the system does not maintain a surface language record - does this produce any particular problems?**
Topic	**The system maintains a record of the order in which topics have been addressed through the interaction. It's used for focus determination, which is used for reference resolution. No seperate evaluation was done.**
Task	**The system maintains a record of the task-relevant information which has been exchanged. It's used for focus determination, which is used for reference resolution. No seperate evaluation was done.**
Performance	**The system does not maintain a record of the user's performance during interaction. This is currently not needed.**
Domain model	
Data	**The system has a world model, consisting of the domain, i.e. calendar knowledge. The data are pretty realistic.**
Rules	**Describe the rules operating on the domain data, such as completions and constraints. Have the rules been evaluated? How? Results? Was the evaluation procedure appropriate? No rules, but all actions you**

	usually do with calendars.
User model	
Goals	The user goal is assumed to be appointment scheduling, i.e. fixing time and location for meetings, without going into the user's reasons for their proposals and decisions made. **This is an artificial curtailment of the user's goals in the domain.**
Beliefs	Only applicable to meta-communication. **Describe what the system does to handle the user's beliefs during interaction. Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? N/A.**
Preferences	Only applicable to meta-communication. **Describe what the system does to handle the user's preferences during interaction. Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? Not applicable for Verbmobil. What's being said should be translated.**
User group	**Does the system assume any distinctions among user groups, such as between domain novices and experts, novices and experts in using the system, other? Has the approach been evaluated? How? Results? Was the evaluation procedure appropriate? If no distinction between user groups is being made - does this cause specific problems? Not applicable for Verbmobil. What's being said should be translated**.
Cognition	**Has anything been done to take into account the specific cognitive characteristics of users, such as task load, limited memory, natural "response packages" or limited attention span? Are such characteristics not being considered relevant to the interaction? If not, is this justified, or is it possible to characterise specific problems the system has because, e.g., cognitive load issues were not considered? Not applicable for Verbmobil. What's being said should be translated.**
Architecture	
Platform	Standard hardware: Processor: SPARC. Memory: 500 MB. OS: UNIX (Solaris), other UNIX versions on demand; Linux version planned. Disk System: 500 MB, Swap: 2-3 Gb. **Is the platform adequate according to today's standards? Yes.**
Tools and methods	**Describe the tools and methods used.**
Generic	Multi-agent, object-oriented. **Is the generic software architecture adequate according to today's standards? Obviously. See [BubSchwinn96].**
No. components	43 **(cf. the architecture diagram).**
Flow	**Describe the process flow among the system components (cf. the architecture diagram). See BubSchwinn96 or Bubetal97.**
Processing times	On average: 38% speech recognition, 17% prosody, 25% syntax and semantics, 14% semantic evaluation and dialogue, 3% transfer, 3% generation. **Are the proportions satisfactory? Has this been evaluated? How? Results? Was the evaluation procedure**

	appropriate? We have an adjustable real-time factor. See Bubetal97.

Figure 1. High-level description of the Verbmobil First Phase Research Prototype 1.0. **Boldface has been used for comments, new text and entries, and unresolved questions. Underlining has been used for evaluative comments.**

Verbmobil architecture

Maybe we should put the architecture diagram here for comments in an architecture section?

Verbmobil dialogue(s)

Proposal: For each DM analysis, DISC should provide one or more complete and annotated/commented example dialogues with the system being analysed. This will contribute to providing a concrete “feel” for the system in question. The example dialogues should be selected by the analysers, not by the developers.

Do we have any transcribed dialogues with Verbmobil which we can put here?

One is described in [Alexanderssonetal97].

Verbmobil screen shot(s)

Any?
See Bubetal97, BubSchwinn96 or Alexanderssonetal97. If you need a PS-File (large), please contact Reinhard.