Verbmobil VRP1 Dialogue Engineering Life-Cycle

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

Overall design goal(s): *What is the general purpose(s) of the design process?*

To give Germany research and industry world leadership in speech and language technology through the collaborative building of a system which can translate spoken appointment negotiation dialogues. **Comparable only to the CMU Janus system, this is a path-breaking project. LD: There may be other translation systems. Check DMDialogue.**

Hardware constraints: *Were there any a priori constraints on the hardware to be used in the design process?*

The project takes a pragmatic stance on hardware: to include the hardware necessary for meeting the software objectives.

Software constraints: *Were there any a priori constraints on the software to be used in the design process?*

The first plan was to use keyword spotting to allow Verbmobil to follow a dialogue. However, the keyword spotter used did not work acceptably so the idea was dropped. In the November 1996 demonstrator one has to press a button before the utterance is spoken to which Verbmobil has to listen. After the utterance is spoken the button is released. **Is it correct that this means that VRP1 does not "shadow" the dialogue at all but only translates utterances when ordered to do so, thus not benefitting from contextual knowledge about what happened when it was not asked to translate? In that case, VRP1 is essentially a single-sentence translator. You are not correct. Verbmobil translates everything.**

Each partner site in the project was supposed to contribute the best version(s) of the relevant software it might have at project start. **The lack of common software constraints at the beginning of the project is unusual.**

Customer constraints: *Which constraints does the customer (if any) impose on the system/component? Note that customer constraints may overlap with some of the other constraints. In that case, they should only be inserted once, i.e. under one type of constraint.*

No customers are involved. No hypothetical customer constraints were assumed. **The lack of even hypothetical customer constraints marks Verbmobil as a long-term research endeavour.**

Other constraints: *Were there any other constraints on the design process?*

The Verbmobil project is divided into two phases: Phase 1 from 1993 to 1996 and Phase 2 from 1997 to 2000. Phase 1 should, and did, deliver a convincing 1st demonstrator system. Funding comes from the German government plus 60% **[less??** Reinhard has these figures.**]** contribution from the involved industries. Total 1st Phase funding was close to 100 mio. DM. **No standards conformation requirements at all - not even internal ones? Comment, if needed, on the constraints and their effects from the pov. of best practice. I don't know what you mean with conformation standards. There are lot's of design, interaction, and software standards.**

Design ideas: *Did the designers have any particular design ideas which they would try to realise in the design process?*

E.g. innovative product features, innovative experimental features, other? Describe the effects, if any, of the ideas. Comment, if needed, on the ideas and their effects from the pov. of best practice.

Yes of course. There are some things new where patents are pending. So, no information about that :-) And there were some accepted papers (>200 from all Verbmobilists) to various conferences. Perhaps there are some new ideas described.

Designer preferences: *Did the designers impose any constraints on the design which were not dictated from elsewhere?*

E.g. programming language preferences, development methodology. Describe the effects, if any, of the preferences. Comment, if needed, on the preferences and their effects from the pov. of best practice.

Astonishingly few things where dictated. Basically that: build a working system. One thing we decided upon is the system's architecture, where we agreed on the whiteboard idea, which proved to be pretty good (flexible system, interchangeable modules,...)

Design process type: *What is the nature of the design process?*

Exploratory research. **A peculiarity of the design process is the large number of partners, academic as well as industrial, and the very heterogeneous nature of the initial software contributions.**

Development process type: *How was the system/component developed?*

E.g. through Wizard of Oz, using development methodology X (describe it). Comment on any peculiarities of the development process from the pov. of best practice. The Hamburg site has developed a small WOZ corpus.

We did WOZ dialogues (see Krause97), corpus analysis and evaluation,...

Requirements and design specification documentation: *Is one or both of these specifications documented?*

Describe the specifications. Comment on any peculiarities of the specifications from the pov. of best practice.

No, I don't think it's publically available. Please ask Reinhard.

Development process representation: *Has the development process itself been explicitly represented in some way? How?*

E.g. bits and pieces can be found in scientific papers, the entire process was carefully documented in semi-formal notation, most of the process has been systematically represented in reports or meeting protocols, other. Comment on any peculiarities from the pov. of best practice.

We have a project plan, technical docs about the interfaces, protocols etc.

Realism criteria: *Will the system/component meet real user needs, will it meet them better, in some sense to be explained, than known alternatives, is the system/component "just" meant for exploring specific possibilities (explain), other (explain)?*

It is not clear that the system can meet real user needs given that is covers an artificially limited domain: it does not allow users to state the reasons for their positions during appointment scheduling negotiations.

Functionality criteria: *Which functionalities should the system/component have (this entry expands the overall design goals)?*

E.g. "allow users to do tasks X and Y", "include barge-in", "real-time". Note that this entry is more general than, but may partially overlap with, the "grid" properties. Comment on any peculiarities from the pov. of best practice.

Provide for a proper translation, be robust, meet defined real-time requirements, ...

Customer(s): *Who is the customer for the system/component (if any)?*

Verbmobil has no customer. However, industries involved in Verbmobil have produced spin-off products which do have customers. **True? Which? There are some. Please ask Reinhard about public available data.**

Users: *Who are the intended users of the system/component?*

Users are people who are going to meet for one reason or another and who have to agree on a date and a time for the meeting. Users are supposed to negotiate in English but to speak German or Japanese to Verbmobil. The system is walk-up-and-use. **[Check.] Yes.**

Usability criteria: *What are the aims in terms of usability?*

Adequate translation, for walk-up-and-use users, of all possible German dialogue contributions in the domain. Is this too strong? Check. Yes, but also the English contributions are processed.

Organisational aspects: *Will the system/component have to fit into some organisation or other, how?*

N/A.

Evaluation criteria: *Which quantitative and qualitative performance measures should the system/component satisfy?*

E.g. word error rate, naturalness (explain) transaction success, synthesis quality, robustness (explain), user satisfaction, other. Comment on any peculiarities from the pow. of best practice. Were there no performance targets at all?

Provide for a proper translation, be robust, meet defined real-time requirements....

Evaluation: *At which stages during design and development was the system/component subjected to testing/evaluation? How?*

Testing is done all the time (test data is exchanged between module developers in an early stage, later on you integrate other's modules and test in the Verbmobil testbed), evaluation was done at a large scale at the end of phase 1.

Transaction success: End-to-end evaluation: 20.000 turns, approximately 30.000 sentences; test sets were randomly selected and included no unknown words. The test corpus contains human-human negotiation dialogues in German. The dialogues are scenario-based. Justifications are left out on purpose since Verbmobil cannot handle justifications. **German->English? Results? At which stage was this done on VRP1? Can we add an example scenario below? One scenario, time-scheduling. Has nothing else been evaluated at any stage (wrt. dialogue management)? <u>This evaluation of transaction (translation) success must be characterised as early: it is corpus-based rather than based on real-time user interaction; unknown words were excluded; justifications were excluded.</u>**

Requirements and design specification evaluation: *Were the requirements and/or design specifications themselves subjected to evaluation in some way, prior to system/component implementation? If so, how?*

Comment on any peculiarities from the pov. of best practice. I don't know. Perhaps Reinhard does.

Development time: *When was the system developed? What was the actual development time for the system/component (estimated in person/months)? Was that more or less than planned? Why?*

Comment on any peculiarities from the pov. of best practice. Look at the Verbmobil hompage about the duration. During that time people developed modules. Reinhard has the person/months numbers, but let's see: let's assume 120 Persons for 4 years gives 480 person years.

Developers: *How many people took significant part in the development? Did that cause any significant problems (time delays, loss of information, other (explain))? Characterise each person who took part in terms of novice/intermediate/expert wrt. developing the system/component in question and in terms of relevant background (e.g., novice phonetician, skilled human factors specialist, intermediate electrical engineer).*

Comment on any peculiarities from the pov. of best practice. All persons were involved. I simply can't tell you details about such a large work force. We had all levels of expertise.

Mastery of the development and evaluation process: *Of which parts of the process did the team have sufficient mastery in advance? Of which parts didn't it have such mastery?*

Comment on any peculiarities from the pov. of best practice. Well, we had mastery of all aspects, I think.

Problems during development and evaluation: *Were there any major problems during development and evaluation? Describe these.*

E.g. problems of collaboration in the team, major delays caused by ?, difficulties in satisfying specification requirement X, developer Y left the team, lack of quality of what was delivered by some in the team. Comment on any peculiarities from the pov. of best practice. Well, we did a good job in a grassroots way and came up with a good self organisation. There were of some edges, but we worked it out.

Maintenance: *How easy is the system to maintain, cost estimates, etc.*

Comment on any peculiarities from the pov. of best practice. Astonishingly easy for such a big and diverse system. No estimates, sorry.

Portability: *How easily can the system/component be ported?*

E.g. OS dependencies, machine dependencies. Comment on any peculiarities from the pov. of best practice.

We are in the process to port it to LINUX, and first modules work. AS long as a reasonable OS (POSIX compliant, gcc, lisp, prolog running) is used, there should be no major problems.

Modifications: *What is required if the system is to be modified?*

Comment on any peculiarities from the pov. of best practice. In which hindsight? New domains, new languages?

Robustness: *How robust is the system/component? How has this been measured? What has been done to ensure robustness?*

Comment on any peculiarities from the pov. of best practice. See Bubetal97. Robustness is enhanced by parallel tracks.

Platform and architecture: *On which OS(s) will the system/component run? Machine requirements? Was a particular development platform used? Description of the architecture of the system/component.*

Sun Ultra 2; Solaris; 1 GB memory; 8 GB internal and external disk space; gradien **[??]** analogue-to-digital converter (October 1996 demonstrator). **We need more here - architecture. Comment on any peculiarities from the pov. of best practice. See Bubetal97, BubSchwinn96.**

Component selection/design: *Describe the components and their origins.*

The Verbmobil system is a hybrid system which means that it takes from different modules what they produce and concatenates the results. The Phase II Verbmobil system is expected to use a more intelligent selection.

Programming languages: Several programming languages were used, including Fortran, C and C++.

Speech recognition: Two sites have developed recognisers: D-B and Karlruhe. Karlsruhe both have a German and a Japanese recogniser. The D-B recogniser was used in the November 1996 demonstrator.

Speech recognition: No prosody is included in the German synthesis. True-talk **[??]** is used for English synthesis.

Syntax and semantics: Two sites have developed syntax-semantics modules. **Which? IBM Heidelberg, Siemens AG Munich, Univ. des Saarlandes, Saarbrücken.**

Dialogue management: Dialogue management was developed at DFKI Saarbrücken.

Inter-process communication: For inter-process communication ICE was used. ICE was developed by Verbmobil as an add-on to PVM which in itself was not powerful enough. **These are the basic communication mechanisms. PVM is the well-known parallel virtual machine, ICE is the Intarc Communication Evnironment, which provides an easy to use interface-layer on top of PVM. See BubSchwinn96, Bubetal97, Alexanderssonetal97.**

We need more here. Comment on any peculiarities from the pov. of best practice.

Development and evaluation process sketch: *Please summarise in a couple of pages key points of development and evaluation of the system/component. To be done by the developers.*

Management of Verbmobil is centralised (DFKI Saarbrücken). System integration is done at DFKI Kaiserslautern. **We need more here. Comment on any peculiarities from the pov. of best practice. See BubSchwinn96, Bubetal97. For the dialogue module see [Alexanderssonetal97].**

Property rights:

Seen from outside software belongs to the group which developed it. However, it is part of the contract that the involved industries have the right to use university partners' software and to obtain the source code. Universities may get source code from other universities but not from industry. For example, DFKI get binaries from D-B for system integration.

References

1. URLs:

http://www.dfki.de/verbmobil.
http://www.dfki.de/verbmobil/Vm.Info.Phase2.html.
http://www.dfki.de/verbmobil/VM.English.Mail.30.10.96.html
http://www.dfki.de/verbmobil/tp2/tp02.html
http://www.dfki.de/verbmobil/tp2/tp04.html

http://www.dfki.de/verbmobil/tp2/tp05.html
http://www.dfki.de/~finkler/vm-2-tp5/
http://www.dfki.de/verbmobil/tp2/tp06.html
http://www.dfki.de/cgi-bin/verbmobil/htbin/doc-access.cgi

2. Papers and reports:

General:
Thomas Bub, Wolfgang Wahlster, Alex Waibel. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. DFKI GmbH Kaiserslautern. In Proceedings of ICASSP'97.
Burgard, R. and Karger, R.: Foliendokumentation der Vorträge zur 5. Projektlenkungssitzung für Verbmobil. DFKI. Technisches Dokumentation Nr. 58. May 1997. **Available. Read NOB. Difficult to use.**
Karger, R. and Wahlster, W.: Verbmobil: Multilinguale Verarbeitung von Spontansprache. **Available. Read, used NOB. Good report. GET DATE!**

General and Evaluation:
Thomas Bub, Johannes Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. DFKI GmbH Kaiserslautern. In Proceedings of ICSLP'96, Philadelphia.

Dialogue Management: Scenario.
D. Krause Nutzer-Verbmobil-Klärungsdialoge und Szenarienwahl im Experiment Verbmobil. Memo 60, 1995.
Dialogue Management: Dialogue Acts.
Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Maier, E., Reithinger, N., Schmitz, B. and Siegel, M.: Dialogue Acts in Verbmobil -2. Report 204, May 1997. **Available. Read, used NOB. Good report. Details on all dialogue acts and discourse particles.**
J. Alexandersson, N. Reithinger, E. Maier. Insights into the Dialogue Processing of Verbmobil. Verbmobil Report 191, 1997. 14pp.
B. Schmitz, J. Quantz. Dialogue Acts in Automatic Dialogue Interpreting. Verbmobil Report 173, 1997. 50 pp.
S. Wermter, M. Löchel. Learning Dialog Act Processing. Verbmobil Report 139, July 1996. 14pp. **Available.**
N. Reitinger, E. Maier. Utilizing Statistical Dialogue Act Processing in Verbmobil. Verbmobil Report 80, 1997. 13pp.
An up-to-date report on dialogue acts in Verbmobil is in preparation (available late summer 1997).
Maier, E.: Evaluating a Scheme for Dialogue Annotation. Report 193, April 1997. **Available.**

Language Understanding, Syntax:
The grammars are so far largely undocumented. Work on grammar engeneering is underway (contact Klaus Netter,DFKI, netter@dfki.uni-sb.de).
Grammars:
German: DFKI (Stefan Mueller, Christine Thielen).
English: CSLI Stanford (Dan Flickinger).
Japanese: DFKI/Uni Saarbruecken (C.J. Rupp).

Language Understanding, Semantics:
Johan Bos, Michael Schiehlen, Markus Egg. Definition of the Abstract Semantic Classes for the Verbmobil Forschungsprototyp 1.0. Universität des Saarlandes, IBM Heidelberg, Universitaet Stuttgart. August 1996. 7 Seiten. Verbmobil-Report 165.
A detailed draft on lexical semantics in VM (Kasper, Bos, Schiehlen, Thielen) is upcoming.

Language Generation:
T. Becker, W. Finkler, A. Kilger. Generation in Dialog Translation: Requirements, Techniques, and their Realization in Verbmobil. Draft Verbmobil Report, 1997. 28pp. DRAFT, May 5, 1997.

Testing and Evaluating NL parts:
M. Auerswald, Thomas Bub, H. Kirchmann, H.J. Kroner, J. Schwinn. Verbmobil Integrations- und Testumgebung -- Benutzerhandbuch. Verbmobil-Report 178, DFKI GmbH Kaiserslautern, Oktober 1996.

Evaluation work::
Is only starting: end-to-end module evaluation, input/output tests; schemata are being defined summer 97. No reports so far; contact Uwe Jost (Hamburg; jost@informatik.uni-hamburg.de).
Jost, U.: System- und Modulevaluation. Memo 125, July 1997.

```
@incollection{Krause97,
      author = "Detlev Krause",
      title = "{Using an Interpretation System - Some Observations in Hidden
            Operator Simulations of VERBMOBIL}",
      booktitle = "{Dialogue Processing in Spoken Language Systems}",
      editor = "Elisabeth Maier and Marion Mast and Susann LuperFoy",
      publisher = "Lecture Notes in Artificial Intelligence,
            Springer-Verlag",
      year = 1997,
      address = "Heidelberg"  }

@inproceedings{Bubetal97,
      author = "Thomas Bub and Wolfgang Wahlster and Alex Waibel",
      title = "VERBMOBIL:  The Combination  of Deep and Shallow
            Processing for Spontaneous  Speech Translation.",
      booktitle = "Proceedings of ICASSP-97",
      address =  "Munich",
      PAGES = {71--74},
      year = 1997     }

@INPROCEEDINGS{AlexanderssonReithinger97,
      AUTHOR = {Jan Alexandersson and Norbert Reithinger},
      TITLE = {Learning Dialogue Structures from a
Corpus},
      BOOKTITLE = {Proceedings of EuroSpeech-97},
      ADDRESS = {Rhodes},
      PAGES = {2231--2235},
      YEAR = {1997}
```

```
}
@INPROCEEDINGS{ReithingerKlesen97,
      AUTHOR = {Norbert Reithinger and Martin Klesen},
      TITLE = {Dialogue Act Classification Using Language Models},
      BOOKTITLE = {Proceedings of EuroSpeech-97},
      ADDRESS = {Rhodes},
      PAGES = {2235--2238},
      YEAR = {1997}
}

@INPROCEEDINGS{BubSchwinn96,
      AUTHOR = {Thomas Bub and Johannes Schwinn},
      TITLE = {VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech
Translation System},
      BOOKTITLE = {Proceedings of ICSLP-96},
      YEAR = {1996},
      PAGES = {2371--2374},
      ADDRESS = {Philadelphia, PA.}
}

@inproceedings{Alexanderssonetal97,
      author = "Jan Alexandersson and Norbert
            Reithinger and Elisabeth Maier",
      year = 1997,
      pages=  "33-40",
      title = "{Insights into the Dialogue Processing of {\sc Verbmobil}}",
      booktitle = "Proceedings of the Fifth Conference on Applied
           Natural Language Processing, ANLP '97",
      address = "Washington, DC"
      }
```

2. Dialogue management

1. *Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Maier, E., Reithinger, N., Schmitz, B. and Siegel, M.: Dialogue Acts in Verbmobil -2. Report 204, May 1997.*
See also: http://www.dfki.de/cgi-bin/verbmobil/htbin/doc-access.cgi
The paper gives a draft description of dialogue acts to be used in 2nd phase of Verbmobil.

Dialogue acts (DAs) express the primary communicative intention behind an utterance. *Dialogue turns* are sub-divided into *utterances* each of which reflect one dialogue act. A complex definition of 'utterance' is provided (71-73): an utterance may contain one finite verb, possibly with additional complement (finite) verbs, or no finite verb - if it is an entire turn with no finite verb or consists of fixed lexemes of phrases characteristic of particular DAs, or be a nominal phrase (for certain DAs). Segmentation rules are defined based of the above definition of 'utterance' - for segmenting turns into DAs, false starts, and discourse particles (see below).

Use of DAs: DAs help identify the best translation where several possibilities exist, e.g. of discourse particles such as 'ja', 'bitte' or 'vielleicht'; in shallow processing, DAs help select the templates used to generate target language expressions; DAs form the basis for protocol generation, identifying the core dialogue steps taken in the dialogue.

Coding reliability (between two coders or for the same coder at different times) is measured through (a) confusion matrices which show the DAs that have been confused most frequently, and (b) kappa values. The VRP1 DA coding scheme yielded a kappa value of 0.83 for 10 pre-segmented dialogues labelled by two coders with equal experience. VRP1 had a DA recognition rate between 65% and 75%. There are significant differences in how easily different DAs are identified. These differences are due to varying clarity of DA definitions and to how easily the DAs can be distinguished through their surface language expressions. The authors argue that all task-relevant DAs can be recognised with satisfactory accuracy. All other DAs could be mapped onto a single 'garbage' DA. It is not clear from the paper how this will be done or which DAs will be involved. Verbmobil should ultimately use statistical training-cum-a-priori-success-rates + rule-based heuristics (weighted default rules) to identify DAs.

V2 will distinguish these five *dialogue phases* of negotiation dialogues:

- Hello.
- Opening.
- Negotiation.
- Closing.
- Good_bye.

One turn may cover several phases. Some DAs only seem to occur in certain phases as shown in the hierarchical model of negotiation dialogue DAs (Figure 2).

[Insert Figure 2 here: the V2 DA hierarchy including decision tree.]

All DAs are described in terms of name, occurrence in which dialogue phases, related propositional content, definition, examples of occurrence in context in German, English and Japanese. Some of the definitions reveal the conceptual difficulties involved (se, e.g., the definition of FEEDBACK_POSITIVE).

In VRP1, only the leaves in Figure 2 were used for DA classification. V2 will use the hierarchy in Figure 2 as a decision tree with decision procedures (described in the paper) attached at the numbered branchings in the hierarchy. This will give the possibility of using more abstract DA classifications than those provided by the leaves of the tree.

Discourse particles: In addition to the DAs, the hierarchical model in Figure 2 also shows the category discourse particles, i.e. discourse markers which are not DAs or utterances but nevertheless has a discourse function. The authors distinguish between four categories of discourse particles (VRP1), some of which has sub-categories:

- structuring, including uptake, check, repair marking;
- speaker-attitude signalling;
- smoothing;

- coherence marking.

Occasionally, it may be difficult to distinguish between discourse particles and DAs. The authors exemplify heuristics to support this task.

Chapter 6 shows some fully annotated dialogues.

Information parked here after parts of it has been used above or when it concerns Verbmobil-2

Verbmobil Phase 2 (1997-2000): Verbmobil-2 (V2).

Platform: Central server accessible through ISDN, ATM-based telecooperation services or GSM mobile networks.

Real time behaviour.

Bidirectional translation: Speaker-independent DE/EN (10.000 words), DE/JAP (2.500 words). The system continuously monitors and processes the input from the dialogue participants.

Domains: Multi-functionality: combined appointment scheduling, travel planning, hotel reservation from English to German. 1999/2000: Multi-party travel planning teleco-operation scenario involving parallel German-English and German-Japanese translation. Information: http://www.dfki.de/verbmobil/Vm.Info.Phase2.html.

Speech recognition: More robustness. Coping with errors in the input dealing with long utterances (>25 words). From close microphone to telephone and mobile phone. Automatic detection of user utterance start and finish.

NL analysis: Integration of deep and flat analysis. Depending on the needs wrt. efficiency, robustness, coverage etc., the system will select the appropriate method. Disambiguation.

Dialogue: Identification of dialogue context, under the constraints of multi-domain, multi-party exchanges, possibly multilingually (http://www.dfki.de/verbmobil/tp2/tp04.html). Domain and topic switching without user prompting to be investigated.

Generation: Integration with speech synthesis (concept-to-speech) generation based on morpho-syntactic and prosodic 'concepts' (http://www.dfki.de/verbmobil/tp2/tp05.html, http://www.dfki.de/~finkler/vm-2-tp5/). The generator produces annotated information for prosody-based synthesis. Speaker emphasis will be communication through syntactic means, prosodic means, or their combination. Production of spoken and written text dialogue protocols (summaries) in the relevant languages.

System integration: Integration of knowledge sources: lexicons, syntax, semantics, prosody, dialogue module (http://www.dfki.de/verbmobil/tp2/tp02.html). Architectures, standards, implementation of integrated system.

Tools: Tools development.

2nd phase: The idea is also to implement software for portable devices. But the devices themselves will not be developed in the project.

DISC QUESTIONS AND COMMENTS

1. Apparently no independent human factors group involved.