Multilingual VERBMOBIL-Dialogs: Experiments, Data Collection and Data Analysis

Susanne J. Jekat[1] and Walther v. Hahn[1*]

University of Hamburg, Computer Science Department, Natural Language Systems Division and SFB 538 Multilingualism, Hamburg, Germany

Abstract. In this article we describe the collection and analysis of multilingual dialogs with a human or machine interpreter within the VERBMOBIL framework. As the dialogs represent very rare speech data with high acoustic quality, analysis is still in progress and further research is ongoing.

1 Some Constraints on Machine Interpretation: The VERBMOBIL Scenario

In human interpretation, three basic forms are distinguished:

1. **Simultaneous Interpretation**: The interpreter translates the source-language input into the target language while the source-language speaker (**sls** in the following) is speaking.
2. **Consecutive Interpretation**: The interpreter translates a longer part of speech (e.g. a lecture) with the help of notes after the sls has finished his talk.
3. **Dialog Interpretation**: The interpreter translates the utterances of one or more sls during a multilingual dialog where the different speakers have to communicate but do not know the language(s) of their dialog partners. The translation is performed for each turn either while the sls is speaking or after he has finished his turn.[1]

Principally, there exists less scientific work on interpretation than on written translation and very little of this research is based upon reliable experimental results (cf. Dillinger (1994)). Within research on interpretation, dialog interpretation, with which we deal here, seems to be more or less ignored by researchers on translation.[2]

* We thank Brigitte Lause and Patricia Nüsslein for their support in the data collection.

[1] The latter constellation had been chosen for the VM dialogs.

[2] One reason for this may be the fact that professional interpreters are trained in simultaneous and consecutive interpretation which are predominant occupations whereas dialog interpretation is not presently considered an important task for them.

VERBMOBIL (VM in the following) as an automatic interpreting system is not designed to replace the human interpreter in highly sophisticated conversations but nevertheless it has been developed for applications in conventional types of communication. So, in order to construct an adequate language model for the system, anticipate possible behavior and possible linguistic input of the user, model interpreting strategies, and train the speech recognizers of the system, it was necessary to establish a special database of interpreted dialogs. The VM application implies several general and specific conditions on the scenario of these dialogs which are described in the following.

The VM system is completely speech driven and can be used either in face-to-face dialogs or via telephone by two or more speakers of different mother tongues whose utterances are then transferred into the language of the respective dialog partner. As the input is processed by different speech recognizers, some conditions are already determined by them:

- No **extralinguistic cues** are processed, therefore extralinguistic cues replacing an utterance (like e.g. pointing) had to be avoided during the recordings.
- **Speaker overlap** which would puzzle the speech recognizers also had to be avoided.
- **Complex prosodic cues**: At the beginning of the VM project, processing of complex prosodic cues, for example, as used for marking an ironic or persuasive meaning of an utterance could not be processed by automatic processing systems. These cues therefore were not elicited in the dialogs (at present, automatic processing of prosody is remarkably improved, see Batliner, *The Prosody Module* and *The Recognition of Emotion*, this volume).

Generally, the limits of automatic speech recognition and automatic processing of spontaneous speech at the beginning of the project have led to three principal features of the multilingual (and as a consequence the monolingual) VM dialogs:

- A **cooperative goal of communication** was defined, all speakers should have the same interest in reaching the goal of communication. None of them has to convince the other to do s.th. that (s)he does not want to.
- **Matter-of-fact negotiation**: all speakers (except the interpreter) should hold the same or a comparable professional position and power within their company so that a realistic negotiation can be developed during the conversation. Question-answer plays or simple orders are then avoided. A comparable professional position also leads to
- **dialog symmetry**. Each speaker has the same time to speak and the same chance to take a turn or make a decision concerning the facts to be negotiated.

2 Multilingual Dialogs with a Human or Machine Interpreter

The collection of multilingual VM dialogs consists of more than 9,000 turns in German-Japanese and German-American English dialogs with human or machine interpreters. We also have run some test recordings with other language pairs: Italian–German, Chinese–German, Finnish–German, Russian–German and French–German. These test dialogs are all translated by a human dialog interpreter.

2.1 Acoustic Quality

The multilingual VM-dialogs are recorded in a language laboratory with a hard disk recorder which delivers digitalized speech data with high acoustic quality and channel separation for different speakers. All speakers are recorded with close microphones but we also use room microphones to generate a variation of the acoustic quality and to take safety copies. The data is transliterated according to the VM-conventions in Burger (1997). With the exception of the hard disk recorder all technical details listed here (as well as the specifications of the scenario described below) are also valid for the monolingual data collection in VM (Japanese-Japanese, American English-American English, German-German without interpreters).

2.2 Experimental Design

The experimental design can be characterized as follows: Two native speakers of different mother tongues (German-Japanese or American English-German) are planning a future business trip and try to find a convenient date, hotel and means of transportation. To meet the conditions mentioned in section 1 (cooperation, symmetry and matter-of-fact negotiation) some details of the speaker's roles had been specified: Both speakers hold comparable professional positions and have the same interest in reaching the communicative goal. A matter-of-fact negotiation is established by certain instructions that the speakers read and discuss before the recordings. These instructions are listed in Jekat et al. (1997b)

- The goal of the communication is a joint business trip.
- The speakers should use prepared calendars in which different dates are marked as occupied. Therefore, the subjects can only choose from a restricted set of dates that are convenient for both of them. Accordingly, in the course of the dialogs several proposals are made and rejected.
- To avoid deviation from the scenario, the visits of sights and events is only broadly discussed. No documents are provided for this topic.

To reach a dialog symmetry, one speaker is equipped with a list of hotels, the other with schedules for planes and trains. Principally, the speech data collected in VM might be characterized as role play. The use of role play was necessary to build a homogeneous database under laboratory conditions. Recordings in different surroundings would have been difficult because of too much background noises. Nevertheless in our opinion our speech data is very realistic because we mainly enrolled subjects whose jobs were similar to that in the role play (businessmen). In most of the cases subjects experienced this kind of communication in their daily professional life and did not play unfamiliar roles. Especially for the Japanese recordings this fact is very important because in Japanese different professional positions trigger different linguistic phenomena. For example politeness formulae that are used in conversations between businessmen from the same company and in the same position differ from those in conversations between professors and students.

Three main types of multilingual dialogs have been collected at the University of Hamburg:

1. Face-to-Face Dialogs,
2. Wizard-of-Oz Dialogs, and
3. End-to-End Dialogs.

Face-to-Face Dialogs As mentioned above, the main scenario for the monolingual and the multilingual data was a face-to-face dialog with travel planning as the topic.[3] During the interpreted dialogs two speakers sit together with one human interpreter who translates in both directions. We also recorded multiparty dialogs with more than two speakers and more than one interpreter. This material turned out to be very useful for dialog analysis but as the dialogs become very long and as the number of overlaps and deviations from the scenario increases the multiparty dialogs cannot be used for acoustic training. The fact that interpreters influence the course of communication is very well documented in Wadensjö (1992). Additionally, our analysis of multiparty dialogs with more than one interpreter showed that the influence on the course of communication was multiplied. The more interpreters took part in a dialog the more misunderstandings and clarification dialogs appeared. The dialogs became very complex as compared to human-machine dialogs within the same scenario. As a consequence in most of the face-to-face dialogs only one human interpreter is brought into action. These recordings represent a sample of natural spontaneous speech which is then compared to the

Wizard-of-Oz Experiments .

In Wizard-of-Oz (WOZ in the following)–Experiments[4] the subjects are unaware of the fact that the translation is performed by a human interpreter because the interpreter is concealed from the subjects in a separate room. Interpreter and speakers are connected respectively via microphone and headphone, the interpreter produces written translations, and the translation output is effected acoustically by a computer voice. The experimental setting which creates the illusion that the translation is indeed carried out by VM leads to a different behaviour on the part of the subjects as compared to the dialogs with an acknowledged human interpreter. For example, subjects accept long pauses within the dialogs and stick more closely to the scenario than in face-to-face dialogs. Therefore, WOZ-dialogs are very suitable for the analysis of human–machine interaction.

[3] Other topics have been appointment scheduling, conference planning, and clarification dialogs.

[4] Some of these experiments are also described in Bade et al. (1994), a more general description stems from Jekat et al. (1997a, p. 6f).

End-to-End Dialogs On the base of the analysis of multilingual interpreted dialogs the end-to-end evaluation of the VM system was also designed and carried out at the University of Hamburg (For methods and results see Jekat and Tessiore (2000) and Tessiore and v. Hahn, *Functional End-to-End Evaluation of an MTSystem: Verbmobil*, this volume). In the analysis, the end-to-end dialogs that represent real applications of the system are compared to the face-to-face dialogs and the Wizard-of-Oz dialogs described above. Each end-to-end test session consists of a dialog between two native speakers of different mother tongues, the communicative goal is the same as for the other types of dialogs (mainly travel planning). VM functions as machine interpreter. As already mentioned, visual cues are not processed by the system and therefore should not influence the ongoing communication, so subjects sit in different rooms, do not see each other and can only communicate with the help of the system to which they are connected via headphone and microphone.

3 Sample of Analysis

As v.Hahn (1995) claims, research in machine translation should study human translation in all necessary details. In our opinion the modelling of some human interpretation strategies might improve machine interpretation. As a sample we describe one strategy of human interpreters which is called **Reduction** here. The modelling of reduction within a machine interpretation system should help to avoid a waste of capacity because redundant input elements can be skipped in an early processing step.

3.1 Reduction as Interpretation Strategy

In Jekat et al. (1997a) **Reduction** is characterised as follows

> When the interpreter uses reduction (s)he concentrates on the central content of the message and ignores redundant information, which (s)he leaves out in the translation process. Redundant parts of speech are e.g. proposals which the speaker immediately withdraws again as well as repetitions, slips of the tongue, false starts, etc..

A first analysis of the VM data (Japanese-German and American English-German dialogs with one human interpreter and some test dialogs with French and German speakers and one human interpreter) reveals three general types of reduction:

- reduction of performance phenomena,
- reduction of redundant information,
- language pair dependant reduction.

Reduction of performance phenomena In speech-to-speech processing of business conversations, reduction (i.e. no further processing) of performance phenomena like slips of the tongue or false starts could save a lot of capacity. In contrast, redundant performance phenomena[5] that are further processed could result in recognition errors and in the worst case trigger wrong translations. Example 1 shows how the human interpreter uses reduction and transforms the German input into a well-formed English utterance. Every non-English utterance in the examples listed below is followed by a transformation into English elements which attempts to reproduce the surface structure of the language in question. The speakers are identified by the first three letters of their native language, the interpreter is always identified as DOL.

Example 1

GER: Ich würde mich freuen, äh, wenn wir mal äh essen gehen könnten. Dann könnte ich Ihnen in aller Ruhe. mal etwas erzählen über unsere Leistungen, über unser Haus. (I WOULD BE HAPPY UHM IF WE ONCE UHM COULD GO OUT TO EAT. THEN I COULD TELL YOU WITH A LOT OF TIME. ONCE S.TH. ABOUT OUR SERVICES, ABOUT OUR COMPANY).

DOL: We could have a meal and I could tell you about our services and our company.

Reduction of redundant information This type of reduction comes into play when the source-language turn contains elements (in our French-German example 2 the time: **ten o'clock**) which then are withdrawn within the same utterance (the final suggestion is **eleven o'clock**).

Example 2

FRE: Hmh . Oui, euh je pense que l'on. pourrai peut-être, c'est oui c'est on disait vers dix, onze heures ... ça. peut-être onze heures (UHM, YES, UHM, I THINK THAT ONE. COULD PERHAPS, IT IS YES IT IS LET'S SAY ABOUT TEN, ELEVEN O'CLOCK, THIS. PERHAPS ELEVEN O'CLOCK).

DOL: Elf Uhr vielleicht besser? (ELEVEN O'CLOCK PERHAPS BETTER?)

In example 3 only the acceptance of the proposal seems to be relevant for the course of the conversation and only this function is translated.

Example 3

FRE: Oui, hm, oui ça serait tout à fait possible . Je n'ai pas de cours non plus le Jeudi . A . Donc pourrai rester. Oui, ça irait (YES, UHM, YES THIS WOULD BE POSSIBLE IN EVERY CASE. I HAVE NO LECTURES ON THURSDAY TOO. UHM SO I COULD STAY. YES THAT WOULD BE POSSIBLE).

DOL: Wunderbar. Kein Problem. Geht genauso (PERFECT. NO PROBLEM. IS POSSIBLE).

In the Japanese-German example 4, some rhetorical elements are omitted:

Example 4

JAP: A sou desu ne, sochira no hou mo hanashite okanai to ikemaseN yo ne, sou

[5] The characteristic features of the dialogs are predictable up to a certain degree. For example, in our domain, the possibility that performance phenomena signal emotional confusion or something similar can be ignored, especially since subtle elements of human communication could not be subjected to machine interpretation so far.

shimashitara juukunichi no hi no yuugata daitai goji gurai no shuppatsu de yoroshii deshou ka? ((TYPICAL JAPANESE RHETORICAL ELEMENT) (WE) ALSO HAVE TO TALK ABOUT THIS. THEN HOW ABOUT LEAVING AROUND FIVE O'CLOCK IN THE EVENING OF THE NINTEENTH ?)
DOL: Ja also am neunundzwanzigsten, wollen wir dann so gegen siebzehn Uhr nach Hause fahren ? (SO ON THE NINTEENTH , SHALL WE GO HOME AT AROUND FIVE O'CLOCK P.M. ?)

Language pair dependant reduction For some elements of the source language there are no equivalents in the target language or a translation of these elements would not fit into the target-language conventions and are therefore left out. In the following examples the German speaker uses many smoothing elements which are typical for a certain speech style in German. These elements are omitted in the English (example 5) as well as in the Japanese translation (example 6).
Example 5
GER: Dann würde ich doch sagen, am Freitag dem sechsten Mai. würde das gut passen. Vielleicht so gegen. neunzehn Uhr? (THEN I WOULD SAY ON FRIDAY THE SIXTH OF MAY THAT WOULD FIT VERY WELL. PERHAPS ABOUT SEVEN?)
DOL: Then on Friday, May sixth around seven?
Example 6
GER: Ich kann das auch noch eben genauer nachsehen , einen Moment. einundzwanzig Uhr fünfzehn würden Sie ankommen in Hamburg. (I CAN JUST LOOK IT UP, ONE MOMENT. YOU WOULD ARRIVE IN HAMBURG NINE FIFTEEN P.M.).
DOL: Ano sou shimashitara , kuji juugofuN ni haNburugu no hou ni touchaku shimasu (THEN (YOU) WILL ARRIVE IN HAMBURG AT NINE FIFTEEN).

In the translation of Japanese into German, sometimes politeness formulae have to be left out because there is no equivalent or a translation would sound strange in the target language (example 7).
Example 7
JAP: Hai arigatou gozaimasu. ano sou shimashitara , juuhachinichi no hi no sou desu ne daitai saNjuppuN gurai mae ni chuuoueki no hou de sou desu ne chuuoueki desu ne omachi shite orimasu (YES THANK YOU VERY MUCH. THEN (I)'LL WAIT (FOR YOU) ON THE EIGHTEENTH, ABOUT THIRTY MINUTES BEFORE (DEPARTURE) AT THE MAINSTATION YES AT THE MAINSTATION).
DOL: Ja gut, dann können wir uns ja am achtzehnten gut eine halbe Stunde vorher treffen am Bahnhof (WELL RIGHT, THEN WE CAN MEET AT THE EIGHTEENTH ABOUT HALF AN HOUR EARLIER AT THE STATION).

4 Conclusion

Our multilingual data serves different functions:

- Acoustic training,

- Identification of scenario-relevant data,
- User modelling,
- Modelling of interpreting strategies,
- End-to-End Evaluation.

As we have shown in our sample analysis, the phenomenon of **Reduction** as a human interpretation strategy appears in the translation process of different source and target languages. This seems to be an indicator for the relevance of the phenomenon and the types of reduction described here. Further research should lead to a complete and systematic description of the different types of reduction and to a theoretical judgement about their possible integration into machine translation systems. In the next step, modelling of reduction will verify empirically whether our proposal to use reduction as an input filter can be realized. Other interpretation strategies which have already been extracted from the data will be subject to future research.

References

U. Bade, S. Heizmann, S.J. Jekat-Rommel, S. Kameyama, D. Krause, I. Maleck, B. Prahl, and W. Preuss. (1994). Wizard-of-Oz -Experimente mit dem Verbmobil-Simulator. VM-Memo 24, Universität Hamburg.

S. Burger. (1997). Transliteration spontansprachlicher Daten - Lexikon der Transliterationskonventionen - VERBMOBIL II. VM-Techdok 56, LMU München.

M. Dillinger. (1994) Comprehension during Interpreting: What do Interpreters know that Bilinguals don't? In S. Lambert and B. Moser-Mercer, eds., *Bridging the Gap. Empirical Research in Simultaneous Interpretation*, John Benjamins, Amsterdam. 155–189.

W. v.Hahn. (1995). Introduction. In W.v. Hahn, S.J. Jekat, and I. Maleck, eds., *Machine Translation and Machine Interpretation. (Proceedings of the VERBMOBIL-Workshop Oct.1995).* Universität Hamburg, Fachbereich Informatik.

S.J. Jekat, H. Tappe, H. Gerlach, and T. Schöllhammer. (1997a). Dialogue Interpreting: Data and Analysis. VM-Report 189, Universität Hamburg.

S.J. Jekat, C. Scheer, and T. Schultz. (1997b). VMII Szenario I: Instruktionen für alle Sprachstellungen - VMII Scenario I: Instructions. VM-Techdoc 62, Universität Hamburg, LMU München, Universität Karlsruhe.

S.J. Jekat, B. Lause, and L. Tessiore. (1999). Offline Evaluation of Verbmobil 0.7 . VM-Memo 143, University of Hamburg.

S.J. Jekat, and L. Tessiore. (2000). End-to-End Evaluation of Machine Interpretation Systems: A Graphical Evaluation Tool. In *Proceedings of LREC 2000, Second International Conference on Language Resources and Evaluation.* Athens (Greece), 31 May - 2 June 2000.

C. Wadensjö. (1992). *Interpreting as Interaction*. Linköping Studies in Arts and Science. Linköping University.