Integration of Multi-modal Data and Annotations into a Simple Extendable Form: the Extension of the BAS Partitur Format

Florian Schiel*, Silke Steininger†, Nicole Beringer†, Ulrich Türk†, Susen Rabold†

*Bavarian Archive for Speech Signals (BAS)
†Institut für Phonetik und Sprachliche Kommunikation

University of Munich, Schellingstr. 3, 80799 München, Germany
{schiel,kstein,beringer,tuerk,rabold}@phonetik.uni-muenchen.de

Abstract

Multi-modal resources typically consist of very different data in terms of content and format. This paper discusses a practical solution for the integration of different physical signals as well as associated symbolic data into a common framework. There are ongoing efforts like for instance the ISLE project to develop guidelines and best-of-practice for the standardized representation of such data collections. Since these efforts have not yet converged into a widely accepted concept, we suggest as a starting point to use two different already existing frameworks that can be easily combined for this purpose: The QuickTime format for the handling of synchronized multi-modal signals and the (extended) BAS Partitur Format for the handling of all symbolic data. We can show that with this simple approach it is already possible to integrate the rather complex data streams of the SmartKom Corpus into an easy-to-use format that will be distributed via the Bavarian Archive for Speech Signals (BAS) starting in July 2002.

1. Introduction

The last years have seen quite a number of projects starting to work on the processing / recognition / output of multi-modal data in man-machine-interaction systems. However, a quick survey in the Web sites of LDC[1], ELDA[2], CSLU[3] as well as in general search engines shows that such data are not widely available to the scientific community outside of dedicated project groups[4]. On the other hand projects like ISLE[5] started with the aim to extend the EAGLES initiative with guidelines and standards for multi-modal data, but has not produced any recommendations yet. Although standards and role models do not exist, in most scientific projects people had to get started collecting data for their special needs, in most cases gathering material for training and evaluation of multi-modal input devices. Almost like twenty years ago when the creation of language resources started to get going the concerned scientists nowadays collect and annotate data to their needs and with the tools and standards available.

So did we when we started to collect data for the German SmartKom project[6] beginning of 2000. Unfortunately, this MO will very likely aggravate the future use of these corpora, which is a shame considering the very high efforts (and costs) that are invested into these resources.

Meanwhile the SmartKom group at BAS has collected a vast amount of multi-modal data (about 1500 GByte) and has solved most of the technical problems that come with such a task. As reported elsewhere (Tuerk, 2001) the SmartKom data collection consists of 9 different audio channels, two high resolution video streams, one infrared video stream (black and white) and a screen capture (very low frame rate), a HID input and a pen input. Within the last year we were faced with the problem to integrate all these different modalities (signals) together with the various annotation of data streams into a common framework that may be used for the final distribution of the corpus (starting in July 2002 with the first release of SK Public). The two main problems here are that on the one hand different modalities are recorded by different non-synchronized capture devices, on the other hand annotations to different modalities are produced with the use of different – sometimes even self-written – software tools. All this results in a huge variety of resolutions, time bases, file formats that will hinder the easy usage of the corpus by others.

2. Practical solution

In this contribution we would like to give a proposal (to be precise: two independent proposals) how to handle these problems with existing frameworks. We do not claim that our proposal will be the ultimate and best solution. However, it could act as an intermediate step that allows the immediate work with multi-modal data and might make the conversion of multi-modal resources into a future standard (whatever it might be) less painful.

Let us first list a few basic requirements denoting the intended characteristics of the framework for multi-modal resources (FMMR). Our intended FMMR

- should be extensible and flexible.
  In almost all cases a fixed format for data resources is bad news for the scientist or developer, because he then uses a lot of unnecessary time to solve data format problems. Although this has been true for mono-modal resources as well, the problem multiplies when

[1] http://www.ldc.upenn.edu/

[2] http://www.elda.fr/

[3] http://cslu.cse.ogi.edu/corpora/

[4] The only exception being the M2VTS biometrical corpus available at ELDA

[5] http://isle.nis.sdu.dk/

[6] http://smartkom.dfki.de/

it comes to multi-modal data. Therefore the framework should not be a fixed definition for different kinds of modalities and how to treat them but rather an extensible framework that can be easily adapted to upcoming needs.

- should be easy to process.
  The reason for this key point is obvious. The conclusion is that we may use a well developed format for which tools are available (for instance XML) or that we use such a simple format that it may be processed with standard tools on the operation system level.

- should not integrate signals and annotations in one file format.
  According to our experience in many cases users of a data resources do not need to access all signals or all annotations at the same time. To simplify handling and distribution we therefore strongly recommend that signal and annotation data are separated in storage but linked together via the time base (like it was done in the SAM and BAS Partitur File (BPF) standards).

With these basic requirements in mind our proposed method can be summarized as follows:

1. To integrate the raw data we use QuickTime (QT)[7] for all data that are measured signals or events.
2. To integrate annotations we use BPF or a similar flexible framework (e.g. annotation graphs (Bird, 2001)).
3. We link both representations through the physical time base only.
4. We use what ever necessary relational/hierarchical linking only between the annotation layers.

Note that although we use the BPF in the following examples, this is exchangeable to any other equally qualified format. The point we want to stress here is not the format but that the symbolic (annotation) data should be kept seperate from the signals, but be grouped into a single framework for easier analysis.

We will discuss the pro and cons of our approach in the following section using the SmartKom corpus as an example.

3. Example SmartKom

To demonstrate that our proposal does actually work we show as an example the integration of a complex data collection in the SmartKom project where a wide range of signals and annotations are currently used.

3.1. Integration of signals in QT

Let us first look at the integration of signals into a QT frame. QT allows the integration of several kinds of media into a single multi-media file. Theoretically every signal format that describes physical measurements (signals or events) may be incorporated, if you provide the necessary interface to QT. Fortunately, interfaces for most of the common file formats do already exist. Therefore, it is possible to integrate for instance video, audio, images, vector graphic and even text into a QT frame without the need to transform the single modalities from their original format; since they remain in their original files, it is also possible to access to the data via other tools than the QT player, if necessary. The only problem is the synchronisation of different time bases, e.g. the synchronisation of a video stream with 25 frames per sec on one computer with an audiostream captured at 48 kHz on another system. We have not found yet an elegant solution to synchronize automatically. At the moment we use a technique quite similar as in movie productions: we synchronize manually with regard to a significant acoustical and visual event at the beginning of each recording. Even more difficult is the synchronization of 2D spatial data with the video signals. In the Smartkom corpus the output of the gesture analyzer consists of a stream of coordinates in the working area indicating pointing gestures of the user. We solved this problem by converting the two-dimensional data into so-called sprites – that are little bit maps that move in the visual plane – and then overlap both pictures to synchronize the infrared picture of the hand with the sprite. Please refer to (Tuerk, 2001) for a detailed discussion of the synchronization problem.

In Smartkom a typical session file contains the following tracks:

- video of the face, frontal, DV format.
- video of upper body, from left, DV format.
- video of infrared camera directed on display to capture hand gestures, from top, DV format.
- audio in 10 channels (microphone array (4), directed mic, headset (2), backround noise (2), system output) captured by a 10-channel audio card with 48 kHz
- graphical system output captured by a screen capture application at 4fps, AVI format.
- combined video frame with face, upper body, system output and infrared, AVI format.
- coordinate logfiles: output of either the gesture recognition system (finger tip) or the output of the graphic tableau (pen tip)

For performance reasons all streams are captured on different computers. Coordinate logfiles are transformed into a sprite track to make coordinates visible in the video signals. Then all raw signals are synchronised and integrated into a QT frame.

3.2. Pros and Cons of QT

As mentioned above QT is an open format that serves some of our intended purposes: it is quite easy to use, it is extensible to new, yet unknown formats, and data are accessible via the QT standard library. The synchronization is still a problem but solvable. The alternative would be a fully synchronized capturing hardware, but that was far out of our budged range. The original formats of the data are still accessible on the distribution media which makes the

[7] http://developer.apple.com/techpubs/quicktime/quicktime.html

access easy for people that do not want to use QT. Furthermore, parts of the synchronized stream may be used across different data collections.

When the SmartKom project started we also discussed other possible formats than QT. The Java Media Framework (JMF) was already out at that time and would have had the advantange to run completely in JAVA. However, this also caused a very low performance compared to QT which is coded in C++ (encapsulated in a JAVA class library). Also, we could not get necessary drivers in JMF for our intended platforms, for instance no recording drivers for Mac and no DV codec.
The other alternative would have been the Microsoft Media Format (MMF, nowadays mostly replaced by AVI). MMF was only available for MS platforms and – being a mere format definition and no consistent system like JMF or QT – was not flexible enough for our needs.

One major drawback of QT is the still missing QT library and QT player for Linux OS (we managed to get a QT player running in a Win emulation environment, but the performance is very bad). We hope that with the further spreading of QT this will be solved in the near future.
Depending on how many video streams are integrated into the QT frame it is sometimes necessary to spread the frame over more than one DVD-5 which makes working with the data difficult. Also the time deviation between the time bases of the capturing devices is getting significant in longer recording sessions. We avoid this by restricting the length of one recording session to 300 sec.

Figure 1 shows four data streams of a SmartKom recording within a single flattened video frame. In the upper left quadrant the video signal of the face camera is shown; in the upper right quadrant the video signal of the body from the left; in the lower left quadrant the displayed output of the system, in the lower right quadrant the output of the system and as an overlay the video signal of the infrared camera that captures the user's gestures. The shown frame is actually from a video stream that was calculated from the original QT frame; the QT Player Pro is principally capable to show many video streams simultaneously, however the performance on a standard Intel platform is still unsatisfying.

3.3. Integration of Annotations into BPF

During the last 5 years we have shown that the BAS Partitur Format (BPF) developed at the Bavarian Archive for Speech Signals in 1995 is very successful to integrate so called 'symbolic information' (that is in most cases some kind of annotation) of speech recordings into a simple text based format (see for instance (Schiel et al., 1998)). A BPF is a simple text file very similar to the first SAM label file standard, but has no fixed format concerning the syntax and semantics of the contained tier information blocks. Therefore it is quite easy to extend the format to new needs as long as the meta structure is followed to. Based on the UNIX filter concepts it is possible to add new tier information blocks to a BPF without the need to re-write existing application software (as long as this software does not need to access to the new tier information, of course). A simple chaining mechanism within the different tiers allows the integration of annotations without any direct link to the physical time base; by following the chaining to such a tier all remaining tiers are automatically projected to their right position within the signal.

Let us have a closer look at the structure of the BPF[8]:
A BPF file is a simple ASCII file in which each line has a three character key followed by a colon at the beginning that defines the syntax and semantic of this particular line. A BPF consists of a mandatory header structure (compatible to SAM) that must contain a minimum of descriptors, for instance:

```
LHD: Partitur 1.2.11
REP: Muenchen
SNB: 2
SAM: 16000
SBF: 01
SSB: 16
NCH: 1
SPN: ABZ
LBD:
```

Most important entry in this context is 'SAM' which denotes the sampling frequency for all time references in the following annotation tiers.
After this header block an arbitrary number of tier blocks may follow marked by their respective line key. Registered BPF tiers together with their syntax and semantics can be found on the BAS Web pages. For instance the tier block

```
ORT:    0       all
ORT:    1       right
ORT:    2       Mister
ORT:    3       Durante
ORT:    4       <uh>
```

transcribes the pure lexical words of a short utterance. The numbers in the second column are 'links' between different tiers. In principle there may any sort of links units defined (for instance chunks, words, syllables, events etc.). At the moment the BPF standard uses only one type of link that is the word unit counted from the beginning of the recording. Therefore BPF tiers come in only 5 basic types:

1. Events attched to a word, a group of words or the time slot between two words.
2. Events that denote a segment of time without a relation to the word structure.
3. Events that denote a singular time point without a relation to the word structure.
4. Events that denote a segment of time associated with a word, a group of words or the time slot between two words.
5. Events that denote a singular time point associated with a word, a group of words or the time slot between two words.

[8] http://www.bas.uni-muenchen.de/Bas/BasFormatseng.html

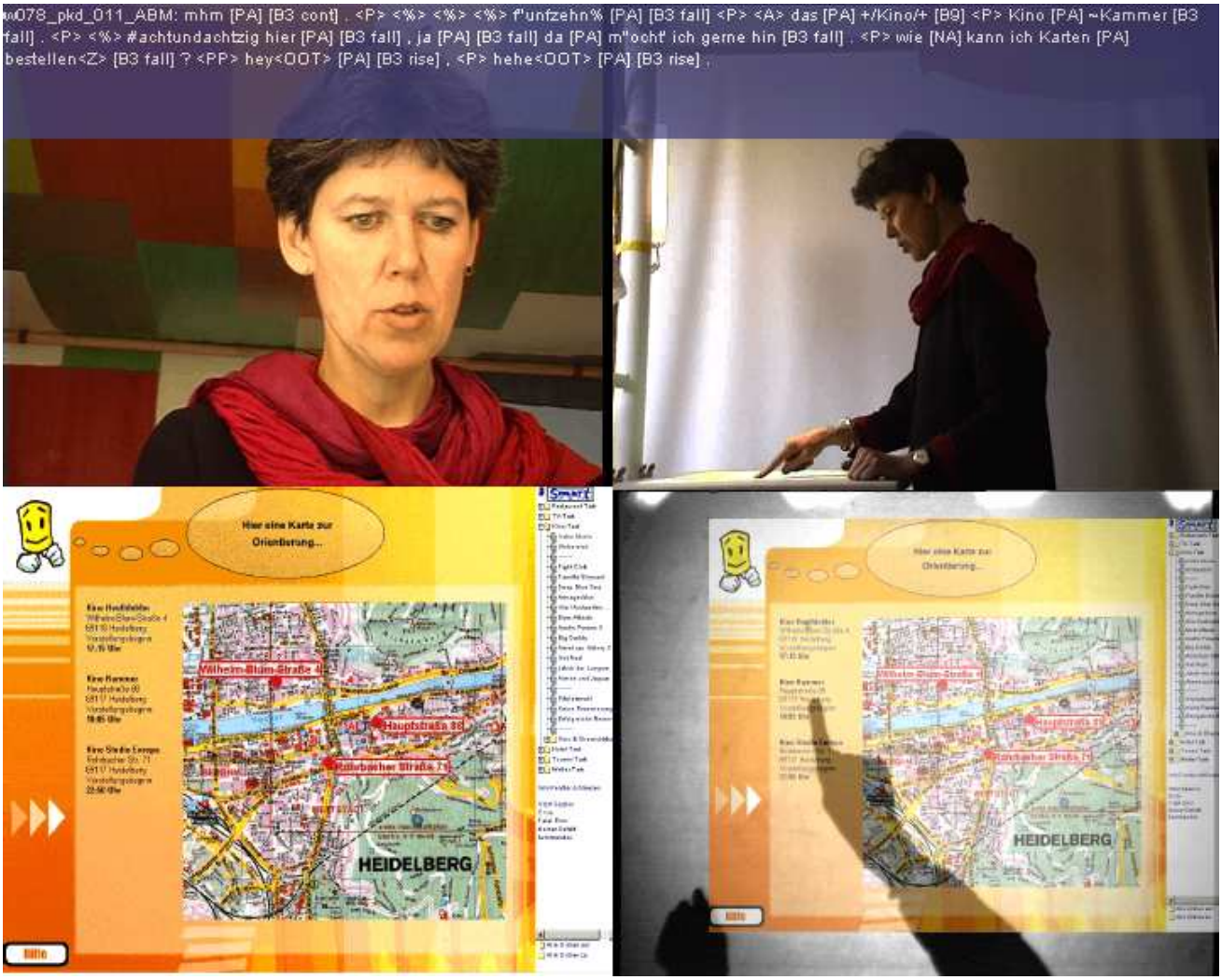

Figure 1: Four synchronized video streams extracted from a SmartKom QT file (see text)

The tier blocks have no preference in order[9] nor hierarchical structure. It is therefore quite easy to cut and paste BPF tiers with standard UNIX tools.

We have shown that the BPF is capable to integrate a variety of symbolic information that was produced within the German Verbmobil project corpus. These data range from simple word alignment over complex syntactic-prosodic tagging up to syntax tree structures. A total of 21 different tiers to the speech signal were used in the Verbmobil corpus (Weilhammer et al., 2002).

Encouraged by this success we started to think about the possibility of integrating symbolic information of multimodal data as well. Surprisingly enough we managed without changing the meta structure of BPF to integrate the following tier information into an BPF (in brackets the corresponding BPF tier keys):

- SmartKom Transliteration of audio channels (TRS,SUP,NOI,ORT,KAN)
- Turnsegmentation (TRN)
- Segmentation and labeling of gestures in the 2D plane (GES)
- Segmentation and labeling of user state (facial and speech) (USH)
- Segmentation and labeling of user state from facial expression only (USM)
- Segmentation and labeling of complex prosodic features to recognize 'emotions' (USP)

Please note that the above annotations are produced with a variety of different software tools (eg. USS, CLAN, Interact). Simple Perl scripts are used to transform the label and segmentation information into the BPF tier information block and add them by concatenation to the existing BPF.

The following example shows an extract from a SmartKom BPF. For better readability the file is abbreviated to the first 12 words of the dialogue and the header block is omitted.

```
TRS:    0       <"ah> [NA] [B2]
TRS:    1       hallo [PA] [B3 fall] . <A> <P>
TRS:    2       kennst [NA]
TRS:    3       du
```

[9] not even within one tier, although the readability is better if the entries follow the time flow

```
TRS:    4       den [B2]
TRS:    5       Wetterbericht [PA]
TRS:    6       f"ur
TRS:    7       heute
TRS:    8       abend [B3 fall] ? <P>
TRS:    9       <:<#> na:> [NA] [B2] ,
TRS:    10      vergi"s [PA]
TRS:    11      es [B3 fall] . <#>
...
SUP:    42,43   w104_mt_SMA.par @1m"ochtest @1du
SUP:    55      w104_mt_SMA.par Pl"atze . <P>2@>
SUP:    56      w104_mt_SMA.par <:<#> hier3@:>
SUP:    61      w104_mt_SMA.par bitte . <P>4@>
ORT:    0       <"ah>
ORT:    1       hallo
ORT:    2       kennst
ORT:    3       du
ORT:    4       den
ORT:    5       Wetterbericht
ORT:    6       f"ur
ORT:    7       heute
ORT:    8       abend
ORT:    9       na
ORT:    10      vergi"s
ORT:    11      es
...
KAN:    0       QE:
KAN:    1       hal'o:
KAN:    2       k'Enst
KAN:    3       d'u:+
KAN:    4       d'e:n+
KAN:    5       v'Et6#b@r"ICt
KAN:    6       f'y:6+
KAN:    7       h'OYt@
KAN:    8       Q'a:b@nt
KAN:    9       n'a+
KAN:    10      f6g'Is
KAN:    11      Q'Es+
...
TRN:    66560   197888  0,1,2,3,4,5,6,7,8,9,10,11 002
TRN:    377984  43776   12,13,14,15 004
...
NOI:    1;2     <A>
NOI:    9       <#>
NOI:    11;12   <#>
...
USH:    0       244480  Neutral
USH:    244480  519040  "Uberlegen/Nachdenken
USH:    517760  25600   Hand im Gesicht
...
USM:    0       515840  Neutral
USM:    515840  216960  "Uberlegen/Nachdenken
USM:    517760  25600   Hand im Gesicht
...
USP:    1364144 3936    27      CLEAR_ART
USP:    1377776 3536    30      CLEAR_ART
USP:    3437728 5856    63      EMPHASIS
USP:    3983392 14992   73      PAUSE_SYLL
...
GES:    265600  32000   U-Geste U - "uberleg - \
        p re Stift nicht erkennbar     640
GES:    376320  30080   I-Geste I - tipp +     \
        re Stift nicht erkennbar
GES:    515200  29440   R-Geste R - emot -     \
        re Hand 393600  8320 "Uberlegung/Nachdenken
...
```

In this example the following tier blocks are contained (see references for details about labeling systems and conventions):

- TRS : SmartKom transliteration (Oppermann et al., 2000)
- SUP : Labeling of cross talk between user and system
- ORT : Lexical entity
- KAN : Citation form in SAM-PA
- TRN : Turn segmentation
- NOI : Noise labeling
- USH : User state labeling using video and audio (Steininger et al., 2002b)
- USM : User state labeling using video only (Steininger et al., 2002a)
- USP : Prosodic labeling of features for user state detection
- GES : Labeling of 2D gestures (Steininger et al., 2001)

3.4. Pros and Cons of BPF

BPFs of Smartkom are fully compatible to BPFs of mono-modal resources. For instance we can easily train a speech recognizer with the data of Smartkom as well as the data of Verbmobil together, since the BPFs tier information blocks for this purpose are identical.
Since the BPF is an open format it is very simple to extend it, for instance by a new tier that contains the time synchronized coordinates of the finger tip delivered by an early stage of the gesture recognizer.
As defined in the BPF format the link to the actual physical signals is solely achieved by reference to the physical time base. It is clear that by doing this the format of the individual signals is arbitrary. It may be the QT format that we use; it may be another format or it may be even just an extraction of a certain modality, as long as the time synchrony is maintained.
Software tools that read only a specific tier information do not need to be adapted when the BPF is extended to a new tier (except of course that the tool needs to process the new tier blocks).
Since the BPF is a simple ASCII file it is usable across platforms.

The BPF does not allow free hierarchical structuring as for instance in the EMU system.
There is no provision in BPF to use UNICODE for special languages or for IPA.
There is no general purpose viewer available for BPF. Up to now we use Praat[10] or SFS[11] to view traditional mono-modal BPFs resources. For the SmartKom corpus we use the QT library that allows to blend in time-aligned text labels as can be seen in figure 1.
There is no dedicated databank system for the BPF. Although we have developed a PROLOG based databank system for the Web that allows simple and complex queries, this is not a general purpose tool. However, it is quite easy to import BPF files into any data bank system.
Last but not least: BPF is not XML. We have started to use parsers that convert BPF tiers into XML. However, it turns out that BPF is easier to read by humans than the XML version.

4. Conclusion

Our approach to use two existing data frameworks, QickTime (QT) and BAS Partitur Format (BPF) for multi-modal data collections was borne out of the need to get started without having any role models and/or applicable

[10] http://www.praat.org/
[11] http://www.phon.ucl.ac.uk/resource/sfs/

standards. We recognize that our current mode of operation is a compromise with some drawbacks. On the other hand it is quite surprising that the integration of multi-modal signal data together with their annotations went rather smoothly. We hope that our experiences will help other researchers that face similar logistic problems as well as researchers that are in the process of defining best-of-practice procedures in the field of multi-modal speech resources.
The SmartKom corpus will be made accessible for the public beginning July 2002. Following our policies with monomodal speech resources we will provide a free access to the symbolic data of the corpus via simple FTP download from the BAS server[12]. To obtain the QT files on DVD-5 media please contact bas@bas.uni-muenchen.de or consult the general BAS Web documentation[13].

5. References

St. Bird. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

D. Oppermann, S. Burger, S. Rabold, and N. Beringer. 2000. Transliteration spontanprachlicher Daten - Lexikon der Transliterationskonventionen. TechDok 02-V4, The SmartKom Project.

F. Schiel, S. Burger, A. Geumann, and K. Weilhammer. 1998. The Partitur Format at BAS. *Proc. of the 1st Int. Conf. on Language Resources and Evaluation, Granada, Spain*, pages 1295–1301.

S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in SmartKom - The coding system. *Springer "Gesture Workshop 2001", London*, page to appear.

S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. 2002a. Development of the user-state conventions for the multimodal corpus in SmartKom. *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*, page to appear.

S. Steininger, F. Schiel, and A. Glesner. 2002b. Labeling procedures for the multimodal data collection of SmartKom. *Proceedings of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Spain*, page to appear.

U. Tuerk. 2001. The technical processing in the SmartKom data collection: A case study. *Proceedings of EUROSPEECH Scandinavia*, pages 1541–1544.

K. Weilhammer, F. Schiel, and U. Reichel. 2002. Multitier annotations in the Verbmobil corpus. *Proc. of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Spain*, page to appear.

[12] ftp://ftp.bas.uni-muenchen.de/pub/BAS

[13] http://www.bas.uni-muenchen.de/Bas