Large Scale Dialogue Annotation in VERBMOBIL

Norbert Reithinger and Michael Kipp (or vice versa)
DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrücken
e-mail: {reithinger, kipp}@dfki.de

Abstract

In this paper we present an overview of the approach to dialogue related annotations in VERBMOBIL. We introduce the information annotated in dialogues of the VERBMOBIL corpus and present the rationale behind the use of the open partitur format for annotations. A tool to facilitate the task of the annotators was developed that supports two dialogue related annotation levels. Finally, we show our approach to measure the intercoder reliability.

1 Introduction to the Environment and Tasks

Large scale language processing systems like VERBMOBIL, a speech-to-speech translation system in the domain of time-scheduling (Bub et al., 1997), heavily rely on corpus data that can be used for the training and test of knowledge sources and algorithms. For VERBMOBIL, currently about 20 CDROMs with German, English, and Japanese dialogues are available[1], containing both speech signals and transliterations. The transliterations of the signal are more than mere orthographic transcriptions. They also cover phenomena like dialectal peculiarities, a symbolic transliteration of all sorts of noises contained in the audio signal, and word breaks.

One of the most important units of processing for the dialogue module are so-called *dialogue acts* like ACCEPT, REJECT, or SUGGEST that characterise the core intention of an utterance.[2] Currently, we use 30 acts that are structured hierarchically. Early during the development of the dialogue module of VERBMOBIL (Alexandersson et al., 1997b), we noticed that the determination and processing of acts has to be based on real data. Therefore, we started to tag the transliterated dialogues of the corpus with dialogue acts.

In this paper, we first describe the format of the dialogues and the annotation level. We then present the tool we developed for annotation and finally show how we take care of the coder's reliability. We close with a look into the future.

2 Annotations and the Partitur Format

When we started annotating the dialogues, we edited the original transliteration files. We added dialogue act tags directly after the utterance using an ad-hoc annotation markup, supported by some EMACS macros. The transliterations looked like this:

```
ja , guten <!1 gu'n> Tag , Herr
Metze .  <Ger"ausch> <#Klopfen>
@(GREET AB) <"ahm> <#> <Ger"ausch>
wir sollen hier einen <!1 ein'>
Termin vereinbaren .  <Ger"ausch>
<A> @(INIT AB)
```

where annotation consisted of a special character, the dialogue act and the speaker direction in brackets. This format was OK as long as there was just one level of markup (i.e.

[1] http://www.phonetik.uni-muenchen.de/Bas/BasKorporadeu.html

[2] We currently use the third version of the dialogue acts as presented in (Alexandersson et al., 1997a).

dialogue acts) and one annotator involved. However, as soon as we added another level of markup[3], a pandora's box of problems was opened. E.g. version problems appeared between different files annotated by different people. Version control tools alone cannot cope with this problems easily, because both the transliterations and the annotations can be changed by different persons and at different institutions and the tools available tell you where there are changes, but not how to integrate these changes.

As a remedy, we first considered the use of available markup-tools like ALEMBIC (Day, 1996) which provides for a structured markup and supports annotation. But it couldn't solve one of the major problems: the data collection agency, Bavarian Archive for Speech Signals (BAS), updates the transliterations in the master files regularly to correct errors, and these changes couldn't be merged easily with our annotated files since we changed the original text with the markup text.

The solution was to move to an extensible format BAS provides, the so-called *Partitur format*. In this format all levels of description are independent but time aligned like the single parts of a score.

It is an open format that contains independent descriptions of as many different levels of the speech signal as necessary, for instance orthography, canonical transcript, phonology, phonetics, prosody, dialog acts, or POS-tags. Symbolic links between the independent levels allow logical assignments related to the linear flow of language. These links are based on the word units of the utterance and are realized as numbers.[4]

A part of the partitur for the above mentioned sentence is show in figure 1. At the beginning it contains some bookkeeping information, followed by the transliteration, the orthographic transcription, the canonic phoneme representation, and finally by the dialogue act.

```
LHD: Partitur 1.2.3
REP: Muenchen
SNB: 2
SAM: 16000
SBF: 01
SSB: 16
NCH: 1
SPN: AAJ
LBD:
TR2: 0 ja ,
TR2: 1 guten <!1 gu'n>
TR2: 2 Tag ,
TR2: 3 Herr
TR2: 4 ~Metze . <Ger"ausch> <#Klopfen>
TR2: 5 <"ahm> <#> <Ger"ausch>
TR2: 6 wir
TR2: 7 sollen
TR2: 8 hier
TR2: 9 einen <!1 ein'>
TR2: 10 Termin
TR2: 11 vereinbaren . <Ger"ausch> <A>
ORT: 0 ja
ORT: 1 guten
ORT: 2 Tag
ORT: 3 Herr
ORT: 4 Metze
ORT: 5 <"ahm>
ORT: 6 wir
ORT: 7 sollen
ORT: 8 hier
ORT: 9 einen
ORT: 10 Termin
ORT: 11 vereinbaren
KAN: 0 j'a:
KAN: 1 g'u:t@n
KAN: 2 t'a:k
KAN: 3 h'E6
KAN: 4 m"Ets@
KAN: 5 QE:m
KAN: 6 vi:6+
KAN: 7 zOl@n+
KAN: 8 h'i:6
KAN: 9 QaIn@n+
KAN: 10 tE6m'i:n
KAN: 11 f6Q'aInba:r@n
DAS: 0,1,2,3,4 @(GREET AB)
DAS: 5,6,7,8,9,10,11 @(INIT AB)
```

Figure 1: The partitur-file for the example

As can be seen, each level of description is marked by a key, e.g. `TR2` for transliteration, `ORT` for orthographic transcription or `DAS` for dialogue acts, followed by the information for this level. The `DAS` level shows

[3]We needed to mark whole turns with a special turn-related set of tags

[4]See `http://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html#Partitur` for a detailed description.

the link between the dialogue act with the basic unit indices, that are the canonic KAN units.

Since the KAN track is guaranteed to always stay the same and our own annotations do refer to the KAN layer, changes in the TR2 or ORT layer (by BAS) do not affect our annotation. On the other hand, our annotation never in any way modifies other layers (like TR2 or ORT) so that partitur files annotated at different institutes with other layers of information can be compared and merged with other files easily and consistently. Furthermore, new levels needed, e.g. for turn information, could be added swiftly due to the open specification.

In contrast to the human readable textual transliterations, the partitur is not intended to be worked on with text editors. Since the annotators are only moderately computer literate, we developed a tool to ease the pain of these poor guys.

3 Colourful Tools for Annotation

Human annotation of the training data is a tedious task. Since human annotators have to read and tag thousands of utterances, mistakes of syntactical and semantic nature are commonplace. Syntactic errors occur even when using tools, since annotators tend to sidestep the tools sometimes, e.g. they use other means to just quickly correct or add a tag with their favourite editor, even if they are told not to do so. Semantic mistakes stem from documentation of the current dialogue act definitions that cover most, but not all phenomena, due to the fact that language is – as we all know – very flexible. Also, over the time of the annotation the guidelines dynamically change through the feedback of annotators. This source of problems can only be reduced by reliability checks on a regular basis where disagreement in annotation is analysed in detail (see next section). The other important source of mistakes is the lack of technical support which should provide the annotator with sufficient means to concentrate on the semantic aspect of the annotation task.

What we learnt from these errors is that an annotation tool should draw a clear line between annotator and data. In one direction data should visualised in a way that suits the annotator. In the other direction manipulation of data must only be allowed within very limited bounds. The original base document should not be accessible to the annotators directly.

The obvious way of doing this is by visualising only task-relevant parts of the tool's internal dialogue representation and open one level to manipulation. Writing back to the original partitur file then changes not the transliteration but only the level under consideration, e.g. the DAS level when annotating dialogue acts.

Our tool, named ANNOTAG, reads partitur files of one dialogue, visualises the transliteration of the turns and presents the annotator buttons for the various dialogue acts (see fig. 2). Human annotators are now to partition these turns into utterances by labelling a part of the text with one or more dialogue act(s). There maybe more than one illocutionary act performed in one utterance which forces us to either define a new dialogue act whose definition covers the phenomenon or to label two or more of the old ones (in the latter case we speak of a multiple dialogue act). The definition of our domain-specific dialogue acts should keep the number of multiple dialogue acts low.

Segmentation and annotation are done in a single step. We are convinced that a separation of these two steps is both unnecessary and impractical. The decision that there is a dialogue act boundary and the decision which dialogue act to annotate can be made in parallel (and *should* be made in parallel since our manual actually makes use of the dialogue act definition for segmentation (Alexandersson et al., 1997a)). Splitting it simply requires the annotator to look at the same data twice and artificially split the reasoning. Also, currently some people (most of them do not annotate large corpora) argue that you must listen to the speech data in order to be able to segment properly. We

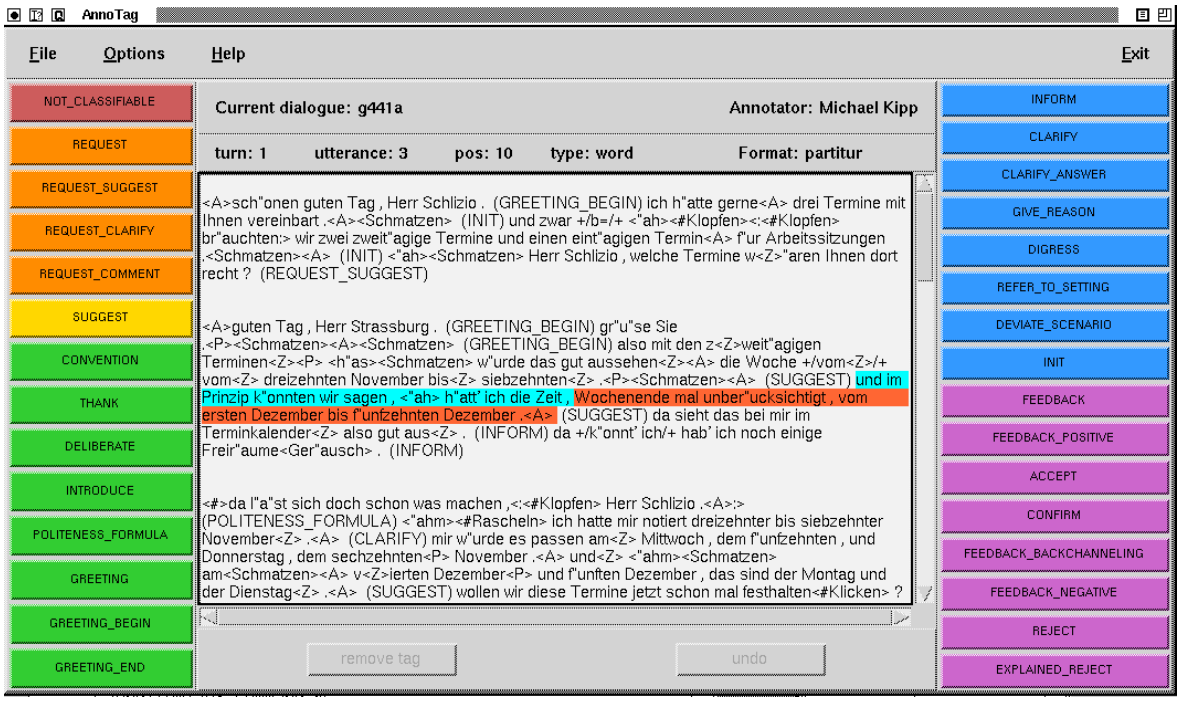

Figure 2: ANNOTAG annotation tool in dialogue act mode

also think this might help in some rare cases, but in general it is not necessary. For example, pauses are transliterated, so the annotators have access to this information, too. But pauses are no criterion for segmentation, since in spoken language pauses do not tell you that much about proper segmentation. Personal communication with other annotation project like Switchboard–DAMSL and MAPTASK shows that all other large scale annotation projects also do not separate the two steps of segmentation and annotation.

Annotation with ANNOTAG works like this: Mark the text by clicking on a word of a turn. ANNOTAG then highlights all the text beginning at the preceding segment border and ending in this word. A click on one of the dialogue act buttons inserts the corresponding tag into the text. Furthermore, we can remove tags (the segments left and right of the tag are merged to one segment), undo the last action and add more tags to an existing tag (making it a multiple dialogue act). To improve readability, the text can be filtered before being displayed. Furthermore, dialogue act buttons are colour-coded according to their position in the hierarchy and unknown dialogue acts in the text are marked red.

So far we have made true our twofold wishes: first, to provide the annotator with a comfortable, easy to use surface that filters out all technical details and discomforts; second, to keep the human user in safe distance from the valuable original data. What remains are questions of extendibility. The tool is wholly written in Tcl/Tk (plus the Tix[5] extension) and can easily be manipulated. E.g. to change the set of dialogue acts we only need to change a list of strings. Recently, we added a feature enabling the user to annotate turns with turn classes which are written to a particular level in the partitur file format.

Additionally, we developed tools that were integrated to serve our special needs. For example, when we revised the set of dialogue acts we used a facility which maps all dialogue acts from one one version to an-

[5] `http://www.xpi.com/tix/`

other. There is also a conversion tool that allows to use annotated files in old plain text transliteration format and maps them to the partitur format. The fact that both formats do not match perfectly (there are words omitted, different turn segmentation etc.) called for a semi-automatic mechanism, showing conversion suggestions for approval/dismissal. We successfully converted more than 400 dialogues this way. Further tools allow inspection of original files, extraction of special information, or conversion of dialogues to LaTeX format.

4 κ or not: Comparing Annotators

To be useful for training and test purposes in a speech processing system our hand-annotated dialogues have to fulfil very high quality standards.

In order to assess this we carried out a number of reliability studies for the segmentation and annotation of the data.

To measure the agreement between feature-attributed data sets the κ *coefficient* is of outstanding importance (for details see (Carletta, 1996)). In the field of content analysis a κ value > 0.8 is considered good reliability for the correlation between two variables, while a κ of $0.67 < \kappa < 0.8$ still allows tentative conclusions to be drawn.

For measuring inter-coder replicability of dialogue act coding we used 10 dialogues which altogether consist of 170 utterances and had two human coders annotate this data. The utterance labels for the two coders coincide in as many as 85.30 % of the cases. The κ value of 0.8261 shows that dialogue acts can be coded quite reliably.

For testing the stability of dialogue acts we asked one coder to relabel five dialogues which altogether contain 191 utterances. The κ coefficient for this study is 0.8430 with an overall agreement of 85.86%.

Currently, we annotate German, Japanese, and English dialogues. The annotators are mostly students of these nationalities, with or without linguistic background. They are trained on a set of training dialogues, where the training consist, amongst others, in annotating a set of test dialogues that are compared to the annotation of the other annotators who also annotated this set. Proceeding this way, we can identify the overall agreement of the annotators and can identify classes where an annotator seems to misunderstand the definitions of the manual.

One major use of the dialogue act annotation is to train a statistical dialogue act recogniser (Reithinger and Klesen, 1997). To test the quality of the system, we also measure the κ coefficient between the annotations from humans and the program. For a test set of 51 German dialogues, we achieved agreement in 63.48% of the cases, and a κ value of 0.58, using 215 dialogues for training, which of course did not contain the tested dialogues. For 18 Japanese dialogues the agreement was 71.65% with a κ value of 0.68, using 81 dialogues for training.

5 Future Work

At the time we write these lines, and using the approach and the tools presented above. we have about 830 German, English, and Japanese dialogues annotated with dialogue acts, and some 100 with turn classes. The partitur format demonstrates its power in daily use, since it can easily be processed for different purposes with common UNIX tools like `sed` or `perl`.

In the future, besides annotating more dialogues with dialogue act and turn class information, we will extend our reliability studies to detect and repair possible weaknesses in the definition and description of the tags. Also, we will annotate the propositional content of the utterances using a domain description language.

As we now have a pool of annotated dialogues we will use them to bootstrap an automatic annotator which will propose the human annotator the most likely tag for a unit to be annotated. We hope that we finally will get a (semi-)automatic annotator.

References

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997a. Dialogue Acts in VERBMOBIL-2. Technical Report 204, DFKI Saarbrücken, Universität Stuttgart, Technische Universität Berlin, Universität des Saarlandes.

Jan Alexandersson, Norbert Reithinger, and Elisabeth Maier. 1997b. Insights into the Dialogue Processing of VERBMOBIL. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLP '97*, pages 33–40, Washington, DC.

Thomas Bub, Wolfgang Wahlster, and Alex Waibel. 1997. Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In *Proceedings of ICASSP-97*, pages 71–74, Munich.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistics. *Computational Linguistics*, 22(2):249–254, June.

David S. Day, 1996. *Alembic Workbench User's Guide*. The MITRE Corporation, Bedford, MA, December.

Norbert Reithinger and Martin Klesen. 1997. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, Rhodes.