STANDARDISATION EFFORTS ON THE LEVEL OF DIALOGUE ACT IN THE MATE PROJECT

Marion Klein
DFKI German Research Center for Artificial Intelligence GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken
Marion.Klein@dfki.de
http://www.dfki.de/~mklein

Abstract

This paper describes the state of the art of coding schemes for dialogue acts and the efforts to establish a standard in this field. We present a review and comparison of currently available schemes and outline the comparison problems we had due to domain, task, and language dependencies of schemes. We discuss solution strategies which have in mind the reusability of corpora. Reusability is a crucial point because production and annotation of corpora is very time and cost consuming but the current broad variety of schemes makes reusability of annotated corpora very hard. The work of this paper takes place in the framework of the European Union funded MATE project. MATE aims to develop general methodological guidelines for the creation, annotation, retrieval and analysis of annotated corpora.

INTRODUCTION

Over the last years, corpus based approaches have gained significant importance in the field of natural language processing (NLP). Large corpora for many different languages are currently being collected all over the world, like

- [British National Corpus],
- [The TRAINS Spoken Dialogue Corpus], or
- [The Child Language Data].

In order to use this amount of data for training and testing purposes of NLP systems, corpora have to be annotated in various ways by adding, for example, prosodic, syntactic, or dialogue act information. This annotation assumes an underlying coding scheme. The way such schemes are designed depends on the task, the domain, and the linguistic phenomena on which developers focus. The author's own style and scientific background also has its effects on the scheme. So far, standardisation in the field of dialogue acts is missing and reusability of annotated corpora in various projects is complicated. On the other hand reusability is needed to reduce the costs of corpus production and annotation time.

The participating sites of the EU sponsored project MATE (Multi level Annotation Tools Engineering) reviewed the world-wide approaches, available schemes [Klein et al.1998], and tools on spoken dialogue annotation [Isard et al.1998]. The project builds its own workbench of integrated tools to support annotation, evaluation, statistical analysis and mapping between different formats. MATE also aims to develop a preliminary form of standard concerning annotation schemes on various levels to support the reusability of corpora and schemes.

In this paper we focus on the level of dialogue acts. We outline the results of the comparison of the reviewed coding schemes based on [Klein et al.1998] and discuss best practice techniques for annotation of mass data on the level of dialogue acts. These techniques are considered as a first step towards a standard on the level of dialogue acts.

COMPARISON OF CODING SCHEMES

Plenty of research has been done in the field of annotation schemes and many schemes for different purposes exist. Not all of these schemes can be annotated reliably and are suitable for reuse. In the following we state guidelines we have developed for selecting most appropriate schemes and represent the results of our scheme comparison according to these guidelines.

Firstly, it is important for us that there is a coding book provided for a scheme. Without definition of a tag set, decision trees, and annotation examples, a scheme is hard to apply. Also the scheme has to show that it is easy to handle which means it should have been successfully used by a reasonable number of annotators on different levels of expertise. For reusability reasons, language, task, and domain independence is required. Additionally, it is crucial that the scheme has been applied to large corpora. The annotation of mass data is the best indicator for the usability of a scheme. Finally, it was judged positive if schemes directly proved their reliability by providing a numerical evaluation of inter-coder agreement, e.g. the κ-value [Carletta1996].

Information about schemes was collected from the world wide web, from recent proceedings and through personal contact. We compared 16 different schemes, developed in the UK, Sweden, the US, Japan, the Netherlands, and Germany. Most of these schemes were applied to English language data. Only three of the reviewed schemes were annotated in corpora of more than one language, and thus, indicate some language independence.

A drawback in reusing schemes for different purposes is tailoring them to a certain domain or task. Nevertheless, most of the ongoing projects in corpus annotation look at two-agent, task-oriented dialogues, in which the participants collaborate to solve some problem. These facts are also reflected in the observed schemes which were all designed for a certain task and/or used in a specific domain.

With regard to the evaluation guidelines stated above we can positively mention that all schemes provide coding books. Also, all schemes were applied to corpora of reasonable size (10 K - 16 MB data). In 14 cases expertised annotators were employed to apply the schemes which leads to the assumption that these schemes are rather difficult to use. The inter-coder agreement, given by 10 of the schemes, shows intermediate to good results.

The comparison of tag sets was performed differently with regard to higher and lower order categories. The definition of higher order categories was mainly driven by the linguistic, e.g. [Sacks and Schegloff1973], and/or philosophical theories, e.g. [Searle1969], the schemes were based on. Whereas definitions and descriptions of lower order categories were influenced by the underlying task the scheme was designed for, e.g. information retrieval, and the domain of the corpus the scheme was applied to, e.g. conversation between children.

The only higher order aspect that was implicitly or explicitly covered in all schemes was forward and backward looking functionality. This means that a certain dialogue segment is related to a previous dialogue part, like a "RESPONSE", or to the following dialogue part, like a "CLAIM" that forces a reaction from the dialogue partner.

On the level of lower order tags we could see tags

- with nearly equivalent definitions, e.g.
  - the dialogue act “REQUEST” definition in D. Traum’s scheme:

    “The speaker aims to get the hearer to perform some action.”
    [Traum1996]

    compared to
  - the dialogue act “RA” definition in S. Condon & C. Cech’s scheme:

    “Requests for action function to indicate that the speaker would like the hearer(s) to do something [...]”
    [Condon and Cech1995]

    compared to
  - the dialogue act “REQUEST” definition in the VERBMOBIL scheme:

    “If you realise that the speaker requests some action from the hearer [...] you use the dialogue act REQUEST”
    [Alexandersson et al.1998];

- which broadly seem to cover the same feature with slightly different description facettes, e.g.
  - the dialogue act “OPEN-OPTION” definition in the DAMSL scheme:

    “It suggests a course of action but puts no obligation to the listener.”
    [Allen and Core1997]

    compared with the examples above;

- which differ completely from the rest, e.g.
  - the dialogue act “UPDATE” definition in the LINLIN scheme:

    “where users provide information to the system”
    [Dahlbäck and Jönsson1998]

    — addressed to human-machine dialogues.

Especially the last group can be interpreted as highly task or domain dependent.

HOW TO ACHIEVE A STANDARD

There are several possibilities, how standardisation on the level of dialogue acts can be achieved. One possibility is to develop a single, very general scheme. Our impression is, that such a new scheme which has not proven usability is not going to be accepted by researchers who want to look at a certain phenomena of interest. The CHAT scheme used in the CHILDES system [MacWhinney], for example, distinguishes 67 different dialogue acts — it is very unlikely that a general scheme would fit all of their requirements concerning children's conversation.

Another possibility is to provide a set of coding schemes for several purposes. These already existing coding schemes must hold the condition that they have proven reliability in mass data annotation. As there cannot exist a scheme for every purpose, this approach only serves developers of new schemes who want to get an idea how to proceed. With regard to the problem of standardisation this solution is very unsatisfiable as mapping between schemes is often impossible, if schemes do not have a common ground, like the SLSA scheme that models feedback and own communication management [Nivre et al.1998], and the ALPARON scheme [van Vark and de Vreught1996] with the primary objective to analyse the previously mentioned dialogues to model information transfer.

The Discourse Resource Initiative (DRI) group provided input on a third possibility: Developing best practice methods for scheme design, documentation and annotation.

Scheme Design

We can classify the existing schemes in two categories: multi-dimensional and single-dimensional schemes.

Multi-dimensional schemes are based on the assumption that an utterance covers several different orthogonal aspects, called dimensions. Each dimension can be labeled. DAMSL [Allen and Core1997], for instance, is a scheme that implements a four dimensional hierarchy. These dimensions are tailored to two-agent, task-oriented, problem-solving dialogues. Suggested dimensions are

- *Communicative Status* which records whether an utterance is intelligible and whether it was successfully completed,
- *Information Level* which represents the semantic content of an utterance on an abstract level,
- *Forward Looking Function* which describes how an utterance constraints the future beliefs and actions of the participants, and affects the discourse, and
- *Backward Looking Function* which characterizes how an utterance relates to the previous discourse.

Single-dimensional schemes consist of one single list of possible labels. Their labels belong basically to what is called Forward and Backward Looking Functions in DAMSL. Apart from DAMSL all observed schemes belong to this category.

Comparing both categories, the multi-dimensional approach is more linguistically motivated presenting a clear modeling of theoretical distinctions; but annotation experiments have shown that it takes more effort to apply them than it is for single-dimensional schemes. On the other hand, although a single-dimensional scheme is easier to annotate it is hard to judge from outside what kind of phenomena such a scheme tries to model as dimensions are merged — a major disadvantage if reusability is considered. An example for a dialogue act that merges DAMSL's backward and forward looking function is the "CHECK" move in the Map Task scheme. "A CHECK move requests the partner to confirm information that the speaker has some reason to believe, but is not entirely sure about"[Carletta et al.1996]. This reflects the forward looking aspect of such a dialogue act. "However, CHECK moves are almost always about some information which the speaker has been told" [Carletta et al.1996] — a description that models the backward looking functionality of a dialogue act.

Our suggestion to tackle the problem of what kind of scheme is most appropriate, is to use single- and multi-dimensional schemes in parallel. The developer of a new scheme is asked to think precisely what kind of phenomena will be explored and what kind of tags are needed for this purpose. These tags have to be classified with regard to the dimension they belong to. The theoretical multi-dimensional scheme will then be applied to some test corpora. The example annotation shows which tags are less used than others and which tag combinations often occur together. Based on this information the scheme designer can derive a flattened single-dimensional version of the multi-dimensional scheme. The flattened or merged scheme is used for mass data-annotation. A mapping mechanism has to be provided to convert a corpus from its *surface structure*, annotated using the single-dimensional scheme to the *internal structure*, annotated using the multi-dimensional scheme. The multi-dimensional scheme can easily be reused and extended by adding a new dimension. Furthermore, the corpus annotated with the multi-dimensional scheme is not any longer task dependent.

Scheme Documentation

Each coding scheme should provide a coding book to be applicable. Such a document is needed to help other researchers to understand why a tag set was designed in the way it is. Therefore the introduction part of a coding book should state the purpose, i.e. task and domain, the scheme is designed for, the kind of information that has been labeled with regard to the scheme's purpose, and the theory the scheme is based on.

For detailed information about a tag, a tag set definition is required. Following [Carletta1998], such a definition should be mutual exclusive and unambiguous so that the annotator finds it easy to classify a dialogue segment as a certain dialogue act. Also the definition should be intention-based and hence easy to understand and to remember, so that the annotator does not have to consult the coding book permanently even after using the scheme for quite a while.

We suggest, that a coding book should contain a decision tree that aims to give an overview of all possible tags and how they are related to each other. Additionally, the decision tree has to be supplemented by rules that help to navigate through the tree. For each node in the tree there should be a question which states the condition that has to be fulfilled in order to go to a lower layer in the tree. If no further condition holds, the current node (or leaf) in the tree represents the most appropriate tag. As an example of a subtree plus decision rules see Figure 1, taken from [Alexandersson et al.1998].

Decision CONTROL_DIALOGUE:

if the segment is used to open a dialogue by greeting a dialogue partner
then label with GREET

else if the segment is used to close a dialogue by saying good-bye to a dialogue partner then label with BYE.

else if the segment contains the introduction of the speaker, i.e. name, title, associated company etc. label with INTRODUCE

else if the segment is used to perform an action of politeness like asking about the partner's good health or formulating compliments label with POLITENESS_FORMULA

else if the segment is used to express gratitude towards the dialogue partner label with THANK

else if the segment is used to gain dialogue time by thinking aloud or using certain formulas label with DELIBERATE

else if the segment is used to signal understanding (i.e. acknowledging intact communication) label with BACKCHANNEL

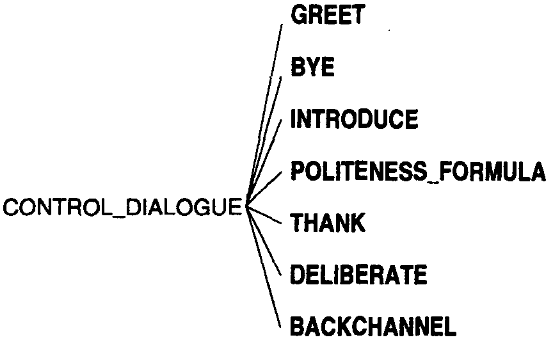

Figure 1

Examples should complement the scheme's description. These examples should present ordinary but also problematic and more difficult cases of annotation. The difficulties should be briefly explained.

Experiences have shown that for new coders tag set definitions are most important to get an understanding of schemes. Annotation examples serve as a starting point to get a feeling for annotation but to manage the annotation task, decision trees are used until coders are experienced enough to perform annotation without using a coding book. This shows how important these three components of a coding book are in order to give new annotators or other scientists best support to understand and apply a coding scheme.

To interpret the evaluation results of intercoder agreement in the right way, the coding procedure that was used for annotation should be mentioned. Such a coding procedure covers, for example, how segmentation of a corpus is performed, if multiple tagging is allowed and if so, is it unlimited or are there just certain combinations of tags not allowed, is look ahead permitted, etc..

For further information on coding procedures we want to refer to [Dybkjær et al.1998] and for good examples of coding books see, for example, [Carletta et al.1996], [Alexandersson et al.1998], or [Thymé-Gobbel and Levin1998].

Annotation Support

Another criterion which is important to increase the effectiveness of annotation is using a user-friendly annotation tool. Such a tool also guarantees consistency, as typing errors are avoided and, hence, improves the quality of annotated corpora. This issue is addressed by the MATE workbench. Other, already existing tools are the ALEMBIC Workbench by [The Mitre Corporation], NB by [Flammia], or FRINGE used in the FESTIVAL system by [The Centre for Speech Technology Research].

DIALOGUE ACT LEVEL REALISATION IN MATE

The approach in MATE is to reuse the DAMSL scheme as an example for an internal multidimensional scheme and a variant of the SWBD-DAMSL scheme [Jurafsky et al.1997] as its example flattened surface counterpart. SWBD-DAMSL was derived from the original DAMSL scheme using the techniques described above. Unfortunately some additional tags were added so that an exact mapping from one scheme to the other is not possible any more. For this reason the MATE SWBD-DAMSL variant omits these additional tags.

MATE uses XML [The W3Ca], a widely accepted interchange and storage format for structured textual data, to represent the schemes and the annotated corpora.

Stylesheets (a subset of XSL [The W3Cb]) are used as a mapping mechanism between corpora annotated with the surface scheme and corpora annotated with the internal scheme.

The choice in MATE to use the W3C (World Wide Web Consortium) proposals is because XML is the latest, most flexible data exchange format currently available and strongly supported by industry. XSL supplements XML insofar that it realises the formatting of an XML document.

The facilities for dialogue act annotation are embedded in the MATE workbench. The workbench is currently being implemented in Java 1.2 as a platform independent approach. This makes the distribution process of the workbench easier and supports wide-spreading MATE's ideas of best practice in annotation.

RELATED WORK

Projects which are related to MATE's aim to develop a preliminary form of standard concerning annotation schemes are the DRI which was started as an effort to assemble discourse resources to support discourse research and application. The goal of this initiative is to develop a standard for semantic/pragmatic and discourse features of annotated corpora [Carletta et al.1997]. Another project, LE-EAGLES, also has the goal to provide preliminary guidelines for the representation or annotation of dialogue resources for language engineering [Leech et al.1998]. These guidelines cover the areas of orthographic transcription, morpho-syntactic, syntactic, prosodic, and pragmatic annotation. LE-EAGLES describes most used schemes, markup languages, and systems for annotation rather than proposing standards.

CONCLUSION AND FUTURE WORK

Having reviewed a large amount of currently available coding schemes for dialogue acts we presented a methodology how to tackle the standardisation problem. We outlined best practice for scheme and coding book design which hopefully will lead to a better understanding and reusability of schemes and corpora annotated using the proposed method.

Our approach is currently being implemented in the MATE workbench and will be tested and enhanced in the remaining time of the project. It will be applied to the CSELT, Danish Dialogue System, CHILDES, MAPTASK and VERBMOBIL corpus to help making it as task, domain and language independent as possible. Inadequacies during the testing phase which are related to the internal scheme we use will be discussed with the members of the DRI to further improve the scheme or its flattened variant.

ACKNOWLEDGMENT

I would like to thank the members of the DRI for stimulating and fruitful discussions and suggestions, and Norbert Reithinger, Jan Alexandersson, and Michael Kipp who gave valuable feedback on this paper.

The work described here is part of the European Union funded MATE LE Telematics Project LE4-8370.

References

[Alexandersson et al.1998] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. 1998. Dialogue acts in verbmobil-2, second edition. Verbmobil Report 226.

[Allen and Core1997] J. Allen and M. Core. 1997. Draft of damsl: Dialogue act markup in several layers. http://www.cs.rochester.edu:80/research/trains/annotation.

[British National Corpus] British National Corpus. http://info.ox.ac.uk/bnc.

[Carletta et al.1996] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1996. Hcrc dialogue structure coding manual. http://www.hcrc.ed.ac.uk/ jeanc/.

[Carletta et al.1997] J. Carletta, N. Dahlbäck, N. Reithinger, and M. A. Walker. 1997. Standards for dialogue coding in natural language processing. Dagstuhl Seminar Report. (editors).

[Carletta1996] J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. In *Computational Linguistics*, volume 22(2), pages 249—254.

[Carletta1998] J. Carletta. 1998. The history of discourse representation. 2nd DRI Meeting, Japan. Slides.

[Condon and Cech1995] S. Condon and C. Cech. 1995. Manual for coding decision-making interactions. ftp://sls-ftp.lcs.mit.edu/pub/multiparty/coding_schemes/condon.

[Dahlbäck and Jönsson1998] N. Dahlbäck and A. Jönsson. 1998. A coding manual for the linkœping dialogue model. http://www.cs.umd.edu/users/traum/DSD/arne2.ps.

[Dybkjær et al.1998] L. Dybkjær, N.O. Bernsen, H. Dybkjær, D. McKelvie, and A. Mengel. 1998. The MATE Markup Framework. MATE Deliverable D1.2.

[Flammia] G. Flammia. The nb annotation tool. http://www.sls.lcs.mit.edu/flammia/Nb.html.

[Isard et al.1998] A. Isard, D. McKelvie, B. Cappelli, L. Dybkjær, S. Evert, A. Fitschen, U. Heid, M. Kipp, M. Klein, A. Mengel, and N. Reithinger M. Baun Mø ller. 1998. Specification of coding workbench. http://www.cogsci.ed.ac.uk/~amyi/mate/report.html.

[Jurafsky et al.1997] D. Jurafsky, L. Shriberg, and D. Biasca. 1997. Switchboard swbd-damsl, shallow-discourse-function annotation. http://stripe.Colorado.EDU/~jurafsky/manual˙.august1.html.

[Klein et al.1998] M. Klein, N. O. Bernsen, S. Davies, L. Dybkjær, J. Garrido, H. Kasch, A. Mengel, V. Pirrelli, M. Poesio, S. Quazza, and C. Soria. 1998. Supported coding schemes. http://www.dfki.de/mate/d11.

[Leech et al.1998] G. Leech, M. Weisser, and A. Wilson. 1998. Draft chapter: Survey and guidelines for the represention and annotation of dialogue. LE-EAGLES-WP4-3.1. Integrated Resources Working Group.

[MacWhinney] B. MacWhinney. The childes project: Tools for analysing talk. http://poppy.psy.cmu.edu/childes/index.html.

[Nivre et al.1998] J. Nivre, J. Allwood, and E. Ahlsén. 1998. Interactive communication management: Coding manual.

[Sacks and Schegloff1973] H. Sacks and E. Schegloff. 1973. Opening up closing. In *Semiotica*, volume 8, pages 289—327.

[Searle1969] J. R. Searle. 1969. *Speech Acts.* Cambridge University Press.

[The Centre for Speech Technology Research] The Centre for Speech Technology Research. The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival.html.

[The Child Language Data] The Child Language Data. http://sunger2.uia.ac.be/childes.

[The Mitre Corporation] The Mitre Corporation. The alembic system. http://www.mitre.org/re sources/centers/it/g063/workbech.html.

[The TRAINS Spoken Dialogue Corpus] The TRAINS Spoken Dialogue Corpus. http://www .cs.rochester.edu/research/speech/cdrom.html.

[The W3Ca] The W3C. a. Extensible markup language. http://www.w3.org/TR/REC-xml.

[The W3Cb] The W3C. b. Extensible stylesheet language. http://www.w3.org/TR/WD-xsl.

[Thymé-Gobbel and Levin1998] A. Thymé-Gobbel and L. Levin. 1998. Speech act, dialogue game, and dialogue activity tagging manual for spanish conversational speech. http://www.cnbc.cmu.edu/~gobbel/clarity/ma nualintro.html.

[Traum1996] D. Traum. 1996. Coding schemes for spoken dialogue structure. ftp://sls-ftp.Lcs.m it.edu/pub/multiparty/coding_schemes/traum.

[van Vark and de Vreught1996] R.J. van Vark and L.J.M. Rothkrantz J.P.M. de Vreught. 1996. Analysing ovr dialogue coding scheme 1.0. ftp://ftp.twi.tudelft.nl/TWI/publications/tech-reports/1996/DUT-TWI-96-137.ps.gz.