Gefördert vom

bmb+f

Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Verbmobil

Verbundvorhaben

Evaluating a Scheme for Dialogue Annotation

Elisabeth Maier

DFKI GmbH

REPORT 193

April 1997

April 1997

Elisabeth Maier

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany

Tel.: (0681) 302 - 5347/5346
Fax: (0681) 302 - 5341
e-mail: {maier}@dfki.un-sb.de

Gehört zum Antragsabschnitt: 2/4/6

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01IV101K/1 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Zusammenfassung

This paper describes the evaluation of a coding scheme for the segmentation and annotation of transliterated spoken dialogues with dialogue act information. Using the kappa method for measuring the reliability of this coding scheme we find that both for segmentation and labeling we receive reliable results. It is shown that this observation holds both for an evaluation of intercoder reliability and for the stability of coding by one coder over time. Our studies also suggest some improvements of our coding scheme which will be incorporated into the scheme for future use.

1 Introduction

This paper presents some results concerning the data collection and preparation that has been carried out in the VERBMOBIL project. It is the task of this project to develop a prototype for the automatic translation of face-to-face dialogues. Since for the training of speech systems large amounts of data are a prerequisite, over 3000 German, English and Japanese dialogues have been collected and transcribed.

Already in a very early stage of the project it became evident that data augmented with discourse information was necessary, e.g. for the (manual and automatic) development of dialogue models. To establish such a corpus we used dialogue acts [Bunt, 1981] as categories. These dialogue acts had already been determined as basic units in dialogue processing [Alexandersson *et al.*, 1995].

The annotation task was carried out at two sites over approximately two years; at each site two persons were working on the task part-time, producing a corpus of over 500 annotated dialogues.

This paper presents the results of a first study to estimate the reliability of the coding scheme. These results will serve as starting point for an improved annotation scheme that will be used (and periodically evaluated) in the second phase of the project. The study is based on a sample of 20 English dialogues, where each set was labeled by two annotators.

The paper is structured as follows: after a brief description of our annotation scheme (section 2) we introduce a measure to estimate the reliability of coding schemes (section 3) and present the results of two studies: in one experiment we measure the agreement of two coders concerning the segmentation and labeling of transcribed dialogues (*reproducibility*); in the second experiment we examine replicability when one coder carries out both tasks twice (*stability*). From these results we derive improvements for our coding scheme (section 4). After a discussion of related work (section 5) we outline future developments (section 6).

2 The annotation scheme

The scheme used for the annotation of the transcribed corpus consists of two parts: (i) guidelines for the segmentation of spontaneous speech into single utterances [Mast *et al.*, 1995] and (ii) a code book describing the dialogue acts which are used as labels [Jekat *et al.*, 1995]. In the following we briefly describe this annotation scheme.

2.1 Segmentation of Spontaneous Speech into Individual Utterances

Unlike in written discourse utterance boundaries in spontaneous speech are often not clearly marked. Where in written language punctuation is used to delimit utterances, in spoken language this task can at best be attributed to pauses, intonation or speaker change. When labeling such signals cannot be recognized reliably, unless it is possible to listen to speech data. Also, spoken language is often fragmentary and incorrect. Common syntactic rules for the determination of well-formed sentences do not apply to spontaneous speech. Therefore a set of criteria for the determination of utterances had to be developed. According to the guidelines an utterance can be:

- a verb and the material that belongs to its frame (**frame rule**);
- a conventionalized phrase, like e.g. a greeting (**convention rule**);
- a particle that has a specific dialogue function, like e.g. *okay, sorry, great* (**particle rule**).

Fragmentary or incomprehensible input occurring between two utterances is also considered a segment.

2.2 Dialogue Acts

For annotation we used a set of 43 dialogue acts that are tailored towards our domain of appointment scheduling. These acts can be grouped into 18 abstract illocutionary classes:

- REQUEST_SUGGEST: the dialogue participant is asked to make a suggestion;
- SUGGEST_SUPPORT and SUGGEST_EXCLUDE: an item is proposed for or explicitly excluded from further consideration;
- ACCEPT and REJECT: a proposed item is accepted or rejected;

- REQUEST_COMMENT: the dialogue participant is asked to express his opinion concerning a topic;
- CLARIFY: clarificatory information is elicited or provided;
- GREET, THANK, BYE: conventionalized dialogue actions are performed;
- GARBAGE: no other dialogue act applies.

The dialogue acts that occur most frequently in our corpus can be derived from the illocutions listed here by specializing them further according to the propositional content they usually convey (i.e. date, duration, location).

Since in our domain an utterance can serve more than one function we have foreseen that utterances can be labeled with *multiple dialogue acts*. When attributing such acts the labels are ordered according to their prominence in the utterance; i.e. the primary dialogue act is positioned first followed by the dialogue acts of decreasing prominence.

2.3 A Sample Dialogue

In the following we give an example for a brief dialogue taken from our corpus.

```
TJD000: hi (GREET AB)
        how you doing (GREET AB)
        I would like to schedule a meeting some time in June
        m(INIT_DATE AB) m(SUGGEST_SUPPORT_DATE AB)
        how does the eleventh sound (SUGGEST_SUPPORT_DATE AB)
DSG001: okay (FEEDBACK_ACKNOWLEDGEMENT)
        the eleventh (DELIBERATE_EXPLICIT)
        any time after twelve is fine with me (SUGGEST_SUPPORT_DATE BA)
TJD002: I will take a one o'clock appointment (SUGGEST_SUPPORT_DATE AB)
        thank you (THANK_INIT AB)
DSG003: alright one o'clock it is (ACCEPT_DATE BA)
```

3 Measuring Reliability

To measure the agreement between feature-attributed data sets the so-called *kappa coefficient* is of outstanding importance. It has been mostly applied in the area of medical and psychological research. In the framework of the MAPTASK project [Anderson *et al.*, 1991] it has been recently used to estimate the reliability of coding schemes for the annotation of speech data with discourse information ([Carletta, 1996], see also section 5).

The kappa coefficient is defined as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) represents the probability that the annotators agree, while P(E) stands for the probability that the coders agree by chance. The per chance agreement is determined as

$$P(E) = \sum_{i=1}^{n} p_i^2$$

where p_i measures the likelihood that a given label occurs in the data.

According to Krippendorff [Krippendorff, 1980] the interpretation of the kappa value depends on the goals followed with the coding. In the field of content analysis a kappa value > 0.8 is considered good replicability for the correlation between two variables, while a kappa of $0.67 < \kappa < 0.8$ still allows tentative conclusions to be drawn.

3.1 Replicability of Dialogue Segmentation

To compute the kappa value for the determination of utterance boundaries segmentation is considered a binary decision: for each word in our sample we examine whether it is followed by an utterance boundary.

The results for our data set, which consists of 10 unsegmented dialogues, are shown in Table 1. The sample data set contains 1058 words and therefore 1058 potential utterance boundaries. We receive a very good boundary agreement of P(A)=98.49 % for the two coders.

For the computation of the kappa coefficient we get the following result:

$$\kappa = \frac{0.9849 - 0.7859}{1 - 0.7859} = 0.9293$$

We can infer that the segmentation of dialogues can be carried out quite reliably using our guidelines. It can be noted that none of the coders puts boundaries significantly more often than the respective other coder: the fact that coder1 labels a boundary while coder2 does not can only be observed in seven cases (0.66 %); the contrary happens in nine cases, which corresponds to a ratio of 0.85%.

The main sources of divergence concern the following issues:

- *segmentation of subordinate clauses*: concerning the separation of subordinate clauses from the main clause the two coders mostly disagreed; for instance, the coders disagreed whether a contrastive subordinate clause

	Boundary (coder1)	Not-Boundary (coder2)
Boundary (coder1)	121	7
Not-Boundary (coder2)	9	921

Tabelle 1: Segmentation of our Sample Data Set by two Coders.

should be separated from the main clause, or whether clauses linked by means of an *if ... then* construction should form one or two segments. The following dialogue fragment was the source of such a disagreement:

```
okay it is going to be a little bit tight towards the
end but I think I can make it
```

Evidently, the code book specification left it unclear to the coders whether the frame rule (see subsection 2.1) applies to subordinate clauses as well.

- *different readings*: in some cases, it is unclear to which of two available verbs a dialogue fragment belongs; many of these cases can be resolved using phonological information; the possibility to listen to the recorded data would have clearly prevented disagreements of this type. This is illustrated by the following example:

```
before noon possibly from ten o'clock to twelve o'clock on
the twenty fourth would that be alright
```

While this fragment was considered one utterance by coder2, coder1 split this fragment into two utterances, defining the boundary after `twenty fourth`.

- *underspecifications of manual*: in various respects the manual is underspecified: it does not, for instance, give any guidelines how to label gerund and participial constructions, like in

```
Looking at my schedule I am free both Tuesday and Wednesday..
```

Also, the guidelines do not include instructions whether or when to separate the repair from the reparandum. Therefore, the coders sometimes disagree in the segmentation of turns including such phenomena.

- *sloppiness*: naturally, some of the disagreements stem from sloppiness of the coders, who did not always follow the guidelines.

3.2 Replicability of Dialogue Act Coding

For this experiment we used 10 presegmented dialogues which altogether consist of 170 utterances. The utterance labels for the two coders coincide in as many as 82.94 % of the cases. The computation of kappa leads to a value that shows that dialogue acts can be coded quite reliably:

$$\kappa = \frac{0.8294 - 0.1575}{1 - 0.1575} = 0.7975$$

Some of the reasons for disagreement can be observed from the confusion matrix shown in Table 2[1]. This table shows a reduced version of the full 43 X 43 dialogue act matrix, which cannot be presented here. The dialogue acts that have not been used for annotation are left out, as well as the dialogue acts on which both coders always agreed or which never participated in a disagreement. This method leaves the table"outbalanced": dialogue acts that have been used by coder1 may have never been used by coder2, since coder2 used a different label for the same phenomenon; therefore the dialogue act sets for both coders as given in Table 2 do not coincide.

coder1 *coder2*	*RjDt*	*AcDt*	*RCLo*	*RCDt*	*SEDt*	*SSDt*	*InDt*	*Bye*	*FBck*	*Cnfm*
RjDt	13	0	0	0	***5**	0	0	0	0	0
DlEx	0	**1**	0	0	0	**1**	0	0	0	0
AcDt	0	17	0	0	0	**1**	0	**2**	0	0
GvRe	**2**	0	0	0	**1**	**1**	0	0	**1**	0
RSLo	0	0	**1**	0	0	0	0	0	0	0
RSDt	0	0	0	**1**	0	0	0	0	0	0
RCDt	0	0	0	14	0	**1**	0	0	0	0
ClSt	0	**1**	0	0	0	0	0	0	0	0
SEDt	***2**	0	0	0	9	**3**	0	0	0	0
SSDt	0	0	0	0	0	50	**1**	**1**	0	**1**
FAck	0	**1**	0	0	0	0	0	0	0	0

Tabelle 2: Two Coders – Part of the Confusion Matrix for Dialogue Act Coding (Highest Priority Acts only); cases where coders disagreed are given in boldface.

The divergencies are related to the following points:

[1] In this paper we abbreviate dialogue act labels as follows: ACCEPT_DATE (AcDt), ACCEPT_LOCATION (AcLc), BYE (Bye), CLARIFY_STATE (ClSt), CONFIRM (CnFm), DIGRESS_SCENARIO (DgSc) DELIBERATE_EXPLICIT (DlEx), FEEDBACK_ACKNOWLEDGEMENT (FAck), FEEDBACK_BACKCHANNELING (FBck), GIVE_REASON (GvRe), GARBAGE (Grbg), INIT_DATE (InDt), REQUEST_COMMENT_DATE (RCDt) REQUEST_COMMENT_LOCATION (RCLo), REJECT_DATE (RjDt), REQUEST_SUGGEST_DATE (RSDt), REQUEST_SUGGEST_LOCATION (RSLo), SUGGEST_EXCLUDE_DATE (SEDt), SUGGEST_SUPPORT_DATE (SSDt), SUGGEST_SUPPORT_LOCATION (SSLo).

- *unclear differentiation of categories*: while SUGGEST_EXCLUDE_DATE should be used in contexts where a dialogue participant mentions a time frame as a non-viable option (e.g. a week, a day, an hour), REJECT_DATE should only be used where a previously proposed time frame gets a negative evaluation. The following dialogue fragment is an example where the two coders disagree concerning these two categories:

```
JDH002: maybe be together by one o'clock or so
        (SUGGEST_SUPPORT_DATE AB)
SMA003: well I have a class starting at two (?)
```

  From table 2 we can see that these classes are frequently confused: the two coders agree upon using REJECT_DATE (RjDt) in 13 cases, they both label nine utterances with SUGGEST_EXCLUDE_DATE (SEDt), while they confuse the two categories in 5 + 2 cases (indicated with an asterisk in Table 2). Obviously, the two classes are not sufficiently differentiated to handle these cases.

- *personal annotation styles*: while in our experiment coder1 labeled five utterances as GIVE_REASON – this act is used to characterize utterances which describe the motivation for a suggestion / acceptance / rejection, etc. – coder2 did not use this category at all.
  A difference in labeling can also be observed from the frequency with which multiple dialogue acts are coded: while coder1 uses them for 12.4 % of the utterances the frequency of multiple dialogue acts for coder2 is only 3.5 %.

The fact that we only took the highest rated of a set of multiple dialogue acts into consideration is responsible for 4 of the 5 abovementioned cases, where one coder uses GIVE_REASON while the other coder uses a different category.

In the following example the two coders both use multiple dialogue act coding, which only differs with respect to the ranking they attribute to the individual dialogue acts:

```
well every day looks great for me next week but not for work
(REJECT_DATE BA)
I am going to be on vacation all of next week and the
following week (??)
```

While coder1 labeled the last utterance with m(GIVE_REASON) m(SUGGEST_EXCLUDE_DATE) coder2 applied the two dialogue acts in the reverse order m(SUGGEST_EXCLUDE_DATE) m(GIVE_REASON). We therefore decided to relax our notion of agreement for the computation of the reliability of dialogue act coding: we count an agreement when the annotations of both labelers for the same utterance overlap in at least one category.

When using this relaxed notion of coding agreement we receive a significantly improved kappa value of

$$\kappa = \frac{0.8529 - 0.1543}{1 - 0.1543} = 0.8261$$

3.3 Stability of Dialogue Segmentation

For this study one of our coders had to resegment dialogues she already worked on ten months before this study was carried out. The sample consists of five dialogues, which altogether include 1335 words. With an actual agreement of 98.65 % we get a kappa value of

$$\kappa = \frac{0.9865 - 0.7879}{1 - 0.7879} = 0.9364$$

which is only slightly better than the kappa value for two coders.

	Boundary (exp2)	Not-Boundary (exp2)
Boundary (exp1)	152	2
Not-Boundary (exp1)	16	1165

Tabelle 3: Resegmentation of our Sample Data Set by one Coder.

As can be seen from Table 3 sixteen utterance boundaries identified in the second annotation round have not been treated as such in the first study. This divergence can be explained by the introduction of the particle rule in the time between the two experiments. With this source of divergency eliminated and the data set updated according to the particle rule the kappa value is

$$\kappa = \frac{0.9903 - 0.7879}{1 - 0.7879} = 0.9543$$

which now significantly exceeds the kappa value for two coders.

3.4 Stability of Dialogue Act Coding

For testing the replicability of dialogue acts we asked our coder to relabel dialogues she had already annotated in the past. We used five presegmented dialogues which altogether contain 191 utterances.

The kappa coefficient for this study is

$$\kappa = \frac{0.8586 - 0.1028}{1 - 0.1028} = 0.8424$$

taking only the highest priority acts into account. Computing kappa using the relaxed notion of agreement results in an insignificant improvement of only 0.0006 and is now 0.8430. The coder maintained her annotation style: she still distributes multiple dialogue acts according to the same priorities as in the first annotation round. From these kappa values we can infer a reliable coding stability over time.

Exp1 – Exp2	*DlEx* †	*DgSc* †	*AcDt*	*ClSt* †	*SEDt*	*SSLc*	*SSDt*	*Bye*	*FAck* †
RjDt	0	0	0	0	**3**	0	0	0	0
DlEx †	3	**1**	0	**3**	0	0	0	0	0
DgSc †	0	30	0	**1**	0	**2**	0	**1**	**1**
AcLc	0	0	0	0	0	0	0	0	**1**
GvRe †	0	0	0	**1**	**2**	0	**1**	0	0
RSDt	0	0	0	0	0	0	**1**	0	0
ClSt †	**2**	**1**	**1**	5	0	0	**2**	0	**1**
FAck †	0	0	0	0	0	0	0	**1**	1
Grbg †	0	0	0	0	0	0	0	0	**1**

Tabelle 4: Stability Test – Part of the Confusion Matrix for Dialogue Act Coding (Highest Priority Acts only).

With multiple speech acts not being a reason for disagreement we identify the following main problem from the confusion matrix (see table 4): of the 54 times that dialogue acts participate in a disagreement, 35 occurrences, i.e. 65, % belong to a dialogue act class which we call *digression* (members of this class are indicated with a † in Table 4): this class consists of dialogue acts that may occur at any point of the dialogue and that do not actively contribute to an advancement of the task. Counting all confusions where both coders use dialogue acts of the digression type eight (29.6 %) out of 27 disagreements can be found.

4 Lessons Learned

From the studies described in section 3 we derive the following suggestions for an improved coding scheme:

- *class merging*: some of the dialogue acts have to be merged into new classes to improve coding reliability:

- SUGGEST_EXCLUDE_DATE *and* REJECT_DATE: in our corpus these two dialogue acts are often disagreed upon; this relates to the fact that both acts have the same function in the dialogue: they express that something is not a viable option. We therefore propose to merge the two dialogue acts.
- *dialogue acts indicating digression*: the dialogue acts of the digression type are often a source of confusion; this confusion can be reduced by creating a new dialogue act DIGRESS that includes all these dialogue acts. This idea is also supported by some experiments in dialogue act prediction, which is also based on the annotated corpus. When digression acts are clustered we get a 5-7 % improvement for the prediction of dialogue acts [Reithinger *et al.*, 1996].

• *improvement of coding manual*: Both coding and segmentation benefit significantly from an improvement of the manual. This concerns two main points:

  - *the guidelines for segmentation* must provide more detail, in particular concerning the treatment of discourse particles. In many cases it is unclear whether particles carry a dialogue function and have to be segmented as utterance. Also, it has to be examined, whether dependent clauses should get a uniform treatment as full segments, or whether they should be further subclassified into clauses that are treated as an utterance and as clauses that have to be grouped together with the sentence they depend on.
  - *the code book for dialogue acts* has to be elaborated further concerning the use of multiple acts; currently, coders were instructed to only label multiple acts where absolutely necessary. Besides more specific guidelines about when to use multiple coding we also need more information concerning the ranking of multiple dialogue acts.

These improvements will be incorporated in the annotation scheme and used for future coding.

5 Related Research

The large-scale annotation of dialogues is a field of growing importance as more and more applications use statistical models which exploit information acquired from large corpora. In the last decade a growing number of dialogues has been collected which cover many different domains and dialogue types; among the data annotated with discourse information are the problem-solving dialogues in TRAINS [Heeman and Allen, 1995], the MAPTASK dialogues [Anderson *et al.*, 1991], information seeking dialogues collected at the University of Delft [van Vark *et al.*, 1996] and many others.

To our knowledge the only reliability studies concerning a coding scheme for the segmentation and annotation of spoken dialogues with discourse information (e.g. with games, transactions and moves, the latter being roughly equivalent to dialogue acts) has been carried out in the framework of the MAPTASK project [Isard and Carletta, 1995]. In this paper the authors address the replicability of transaction coding, i.e. the determination of task boundaries in dialogues also using the kappa coefficient.

Other replicability studies with respect to the labeling of speech data have been carried out in the area of intonation and prosody. Both for the ToBI (**T**ones and **B**reak Indices) standard and for its German counterpart GToBI the replicability of tone and accent labeling has been assessed (see [Pitrelli *et al.*, 1994] and [Grice *et al.*, 1996]).

6 Conclusion and Future Work

The results of our evaluation study show that our annotation scheme can be applied reliably – in nearly all cases the kappa value is significantly higher than 0.8. Our study also suggests further improvements of our scheme which will be incorporated and used in the second project phase.

In the next project phase we will also annotate Japanese dialogues with discourse information; while we applied dialogue acts successfully for both German and English, we expect that additional dialogue acts will be necessary for Japanese. Also, we plan to separate out domain information from our dialogue acts and to code illocutions and propositional content independently. The results of this labling will undergo a regular evaluation.

7 Acknowledgements

We wish to thank Jean Carletta for introducing us to kappa, Paula Sevastre, Marion Mast and Barbara Müller for providing the annotated corpus, Michael Kipp for implementing tools, and Norbert Reithinger for many helpful comments on this paper.

Literatur

[Alexandersson *et al.*, 1995] Jan Alexandersson, Elisabeth Maier, and Norbert Reithinger. A Robust and Efficient Three-Layered Dialog Component for a Speech-to-Speech Translation System. In *Proceedings of the 7th Conference of the European Chapter of the ACL (EACL-95)*, pages 188–193, Dublin, Ireland, 1995. also available as Verbmobil Report Nr. 50, DFKI GmbH, Dezember 1994. available in the cmp-lg electronic archive under no. cmp-lg-9502008.

[Anderson *et al.*, 1991] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowto, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. The HCRC Maptask Corpus. *Language and Speech*, 34(4):351–366, 1991.

[Bunt, 1981] Harry C. Bunt. Rules for the Interpretation, Evaluation and Generation of Dialogue Acts. In *IPO Annual Progress Report 16*, pages 99–107, Tilburg University, 1981.

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

[Grice *et al.*, 1996] Martine Grice, Matthias Reyelt, Ralf Benzmueller, Joerg Mayer, and Anton Batliner. Consistency in Transcription and Labelling of German Intonation with GToBI. In *Proceedings of ICSLP-96*, Philadelphia, PA, October 1996.

[Heeman and Allen, 1995] Peter Heeman and James Allen. The TRAINS 93 Dialogues. Technical Report TRAINS Technical Note 94-2, Computer Science Department, University of Rochester, March 1995.

[Isard and Carletta, 1995] Amy Isard and Jean Carletta. Replicability of Transaction and Action Coding in the Map Task Corpus. In *Working Notes of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 60–66, Stanford, sc ca, March 1995.

[Jekat *et al.*, 1995] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. Dialogue Acts in VERBMOBIL. Verbmobil Report 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.

[Krippendorff, 1980] K. Krippendorff. *Content Analysis: An introduction into its methodology.* Sage publications, 1980.

[Mast *et al.*, 1995] Marion Mast, Elisabeth Maier, and Birte Schmitz. Criteria for the Segmentation of Spoken Input into Individual Utterances. Verbmobil Report 97, Universität Erlangen, DFKI Saarbrücken, TU Berlin, December 1995.

[Pitrelli *et al.*, 1994] J. Pitrelli, M. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of ICSLP-94*, Yokohama, 1994.

[Reithinger *et al.*, 1996] Norbert Reithinger, Ralf Engel, Michael Kipp, and Martin Klesen. Predicting Dialogue Acts for a Speech-To-Speech Translation System. In *Proceedings of International Conference on Spoken Language Processing (ICSLP-96)*, pages 654–657, Philadelphia, PA, October 1996.

[van Vark *et al.*, 1996] R.J. van Vark, J.P.M. de Vreught, and L.J.M. Rothkranz. Classification of Public Transport Information Dialogues Using an Information Based Coding Scheme. In Elisabeth Maier, Marion Mast, and Susann LuperFoy, editors, *Workshop Notes of the ECAI-96 Workshop on Dialogue Processing in Spoken Language Systems*, pages 92–99, Budapest, Hungary, August 1996.