The Many Functions of Discourse Particles: A computational model of pragmatic interpretation

Gabriele Scheler (SCHELER@ICSI.BERKELEY.EDU)
International Computer Science Institute
1947 Center Street, Berkeley 94704, USA[1]
Kerstin Fischer (FISCHER@NOV1.LILI.UNI-BIELEFELD.DE)
Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft
Postfach 100131, D 33501 Bielefeld

Abstract

We present a connectionist model for the interpretation of discourse particles in real dialogues that is based on neuronal principles of categorization (categorical perception, prototype formation, contextual interpretation). It can be shown that discourse particles operate just like other morphological and lexical items with respect to interpretation processes. The description proposed locates discourse particles in an elaborate model of communication which incorporates many different aspects of the communicative situation. We therefore also attempt to explore the *content* of the category *discourse particle*. We present a detailed analysis of the meaning assignment problem and show that 80% – 90% correctness for unseen discourse particles can be reached with the feature analysis provided. Furthermore, we show that 'analogical transfer' from one discourse particle to another is facilitated if prototypes are computed and used as the basis for generalization. We conclude that the interpretation processes which are a part of the human cognitive system are very similar with respect to different linguistic items. However, the analysis of discourse particles shows clearly that any explanatory theory of language needs to incorporate a theory of communication processes.

Discourse Particles, Meaning Assignment, and the Communication System

In a number of papers [11, 13], it has been argued that feature-based categorization is an effective model for morphological and lexical meaning analysis. This applies to the issue of generating a morphological category or a lexeme from a given feature representation, but it is even more effective in extracting meanings from the context of a written text or dialogue. The central idea is that linguistic elements are used in communication (whether with self or others) via reference to a conceptual level, which is closely tied to cognitive categories (event and temporal structure, spatial cognition, knowledge about objects, motion, changes etc.).

In this paper, it will be argued that pragmatic meanings basically operate in the same way: a limited number of discourse elements in a language serves to communicate a greater number of pragmatic meanings in a predictable way. This model of categorization as meaning assignment will be instantiated for **discourse particles**, characteristic items of spoken language discourse, such as *well*, *yes*, *oh*, *ah*, *okay*, *uh* and *um*. These elements fulfill many different functions with respect to the turn-taking system, speech management, discourse structure, and the interactive level between the communication partners. For example, they segment utterances; they indicate new topics and mark important information; they establish a harmonious atmosphere between speakers and hearers; they help taking, yielding or holding the turn, and they signal speaker-attitude [8, 14, 7].

Most analyses of these particles are characterized by restricted perspectives on their function, which is mirrored in a large number of terms for the phenomena under consideration, for instance **segmentation marker, cue phrase, connector, interjection** [4]. Additionally, so far no automatic methods for assigning discourse functions to discourse particles have been proposed. Using a specific model of the communication situation [3], a method for assigning discourse functions to particles in natural dialogues will be presented which is based on neuronal categorization principles as embodied within the connectionist framework of NEUROSEM [1].

The aims of the paper are therefore

- to show that feature-based categorization is an effective model for pragmatic interpretation processes;
- to automatically assign meanings to occurrences of discourse particles in context;
- to show how knowledge about the functions of one discourse particle can support another by means of generalization from prototypes.

A computational model of discourse function

The linguistic model developed for discourse particles was motivated by the goal of finding consistent patterns of the pragmatic function of these particles, which seem to occur almost randomly at first sight. However, our hypothesis was that if we apply a computational model that is capable of accounting for a high degree of *context-dependency* and *multifunctionality*, the different meanings of discourse particles can be computed with a considerable degree of accuracy. In the following we shall present briefly the linguistic model and the computational model applied to it. We refer the reader to [3] and [10] for a fuller account of these approaches.

The Linguistic Model In contrast to studies on discourse particles so far, the model proposed here is not restricted to a particular function discourse particles fulfill in spoken language discourse but it considers the whole range within a model of communication.

The model proposed treats discourse particles as lexemes with an invariable cognitive content which is employed on different communicative levels. The variablility of aspects of the communicative situation a particle can refer to causes

[1] This work was funded in part by the DFG under contract number Br 609/7-1

the apparent multifunctionality of the class. In particular, the content can refer to the following communicative domains: the speaker's mental state, the hearer's supposed mental state, the propositional level, the speaker-hearer interaction level, as well as the action level [4].

It is assumed that the occurrence of each discourse particle has a specific reading in a natural dialogue. A number of such *discourse functions* has been identified in hypothesis-test cycles on four German corpora. The basic idea is that although many different features are involved on many different communicative levels, there is categorical perception involved in the interpretation of discourse particles. I.e. discourse functions are proposed to be more or less stable feature bundles of surface and pragmatic properties [3, 4]. The inventory of these discourse functions is supposed to be valid for all discourse particles, defining these linguistic elements from a functional perspective.

As an example for different, yet related readings of a discourse particle consider the following two occurrences of English *yes* : Both are reactions to a proposal. In the first case, the semantic content of *yes* "you and I think the same" refers to the propositional level, signalling agreement on the subject in question as an answer particle. In the second example, *yes* is used to signal basic agreement to the communication partner ALTHOUGH the speaker has to reject the proposal. Here the semantic content refers to the interaction level.

Example 1 *yes, that would suit me.*

Example 2 *yes, it is problematic because of the holidays.*

The meanings of discourse particles are consequently modeled with a fixed lexical core and additional context-dependent features which refer to different levels of communication. This renders the postulation of several polysemous items superfluous, and exploits context-dependent *systematic ambiguity* instead.

Occurrences of discourse particles can be characterized further according to the specific functions they fulfill with respect to the turn-taking system, the speech formulation and planning process, the discourse structure, as well as concerning their surface-level properties in utterances. For instance, *yes* can be used to take the turn as in example 3, to introduce a new topic (or even to open a conversation as in example 4) , and in utterance repairs (example 5).

Example 3 *yes, what would you suggest?*

Example 4 *yes, hi hello my name is Quell.*

Example 5 *oh yes, sorry, June.*

The individual features of description used concern both surface features (such as *turn position: initial*) and functions located in different pragmatic domains, for instance turn-taking or speech management functions.

The corpus under consideration was recorded in a setting where one speaker had to teach another one to build a toy-airplane [9]. A simple example of a surface feature concerns the speaker's role in uttering a discourse particle: for instance, it is more likely that a feedback signal is uttered by the constructor than the instructor. Other features, for example, concern the position of the particle in an utterance, as well as its combinations with other particles.

Besides the functions a particle displays, an example of a pragmatic feature is constituted by the speech act the particle occurs in. It was found in distributional and functional analyses that there are significant correlations between particular dialogue acts, domain-specific speech acts of the preceding and the current utterance [15], and the respective discourse particle [5]. Consequently, the preceding and current dialogue acts were coded for the feature-based description as pragmatic features. A catalogue of features and discourse functions is given in the appendix.

To sum up, the linguistic features which are employed concern the many different functions of discourse particles in spoken language discourse, the different communicative domains the cognitive content of discourse particles refers to, as well as the surface realization of an occurrence.

Principles of Linguistic Categorization Feature-based analysis has a long tradition in linguistics. Feature representations play a prominent role in most phonological theories, and they are also of considerable impact in theories of lexical meaning and grammatical categories. Looking at linguistic categorization from a cognitive perspective, we may emphasize the symbolic, i.e. binary (or n-ary) nature of representational features and the transient, context-dependent nature of category assignment.

An exciting possibility from a neurocognitive point of view concerns the interpretation of linguistic units as truly perceptual categories on a par with visual image formation or auditory percepts [10]. Main analogies concern the existence of *categorical perception*, i.e. classification to stable higher-level units from lower-level descriptive input features, *prototype formation*, which concerns the abstraction from a large set of input patterns to a few central reference patterns, and *contextual interpretation*, which refers to the human brain's ability to switch between various classifications of the same item depending on the perceived context. The system NEUROSEM has been developed with the goal of providing the specific tools necessary to perform connectionist semantic analysis for a wide range of applications. A precursor of NEUROSEM was used for machine translation of aspectual categories and text correction of definite and indefinite articles [13, 12]. The main parts of NEUROSEM concern a flexible binary encoding scheme VGEN, an optional input tagger for surface categories [2], a data analysis tool DATMAP, and a set of statistical and connectionist learning procedures.

Clustering and Data Analysis

Representation Issues and Sample Size According to the feature catalogue, a feature-based analysis was carried out on 150 randomly chosen occurrences of *ja* in a large German corpus (cf. [9]) and an additional set of 20 occurrences of *oh*. We arrive at a database of atomic features describing various aspects of the linguistic signal and its communicative setting. The goal of the computational model is (a) to provide an analysis of the data with respect to their stochastic and classificational properties and (b) to effectively realize meaning assignment for discourse particles in context given the features involved.

Our emphasis in this paper is on how speakers perform meaning analysis, looking at this question from a theoretical perspective which opens up the possibility of using this ap-

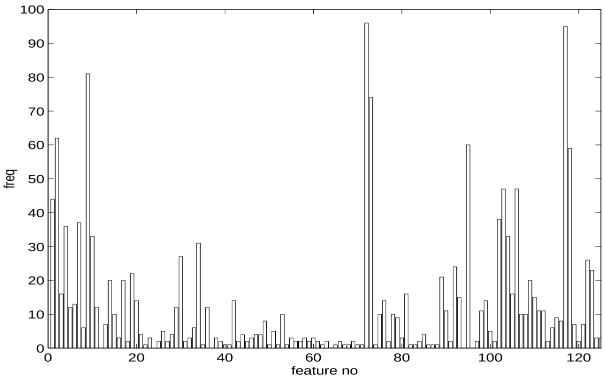

Figure 1: Frequency of feature values in the attested sample

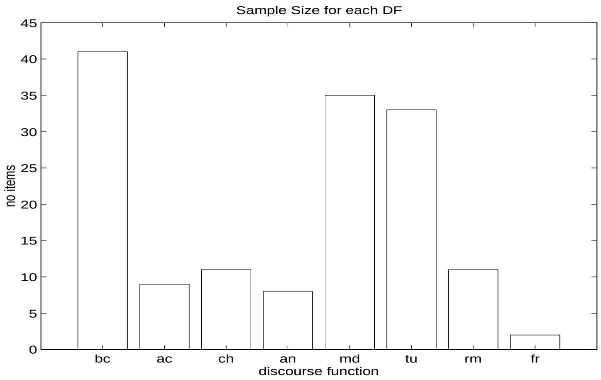

Figure 2: Distribution of Patterns according to Discourse Function

proach in practical language engineering tasks as well. This is in contrast to statistical natural language processing where it has become customary to conduct analyses on the scale of 10's of millions of words. It seems, however, that small-scale, intensive analyses such as the one presented here are more realistic with respect to the language learner than analyses conducted on large corpora. It is doubtful whether large-scale statistical analysis will be applicable to the situation of a person taking in one case at a time, but the method proposed here is amenable to such on-line learning.

Similarity measures, clustering and frequency counts The data we are using are the result of a specific representational method, namely using attribute-value descriptions, where several values attach to each descriptional dimension (or attribute). They are further pre-processed by being converted into binary data vectors, for which a number of options are available (optimal binary coding, linear 1-of-n coding, error-correcting coding etc.), and where optimal binary coding was chosen. (The data used are sparse in the sense that for each pattern several attributes were not used, so binary coding produced good results.)

We performed some initial analyses of the data such as frequency counts, clustering based on Hamming distance and computing within-class and between-class distances of the

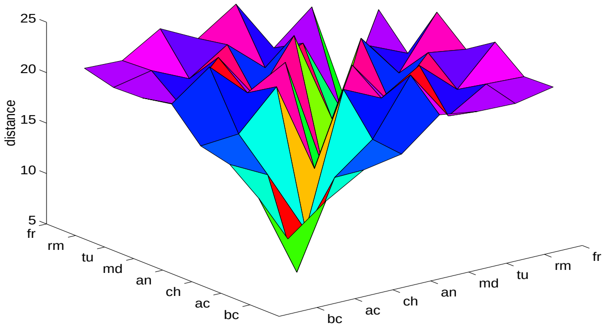

Figure 3: Mean Distances for Patterns according to Discourse Function

data labeled according to discourse function.

There are 127 different input features (values of attributes) which are arranged in 17 dimensions. These features occur with varying frequency in the dataset (cf. Fig. 1). (Mean frequency of a feature in the total set is 13.8).

Class-labeled data In Fig. 2 we show the number of patterns in each class defined by discourse functions. We also compute the mean overall Hamming distance of pairs of patterns both taken from the same class and from different classes (cf. Fig. 3).

These data show that within-class distances (data points on the diagonal) are generally smaller than those between different classes. They present evidence for the linguistic classification to be supported by the descriptive features in attested patterns.

Finally, for each class we may select the pattern with the smallest overall mean distance to other patterns in the same class. These are the most central attested patterns, which can be regarded as the *prototypical* member of that category. We will be using these prototypes in meaning assignment tasks.

Meaning assignment

Interpretation with Full and Reduced Feature Sets In the first set of experiments, it was determined in how far it is possible to automatically assign discourse functions to *ja* and *oh* using the paradigm of supervised learning. We show the influence of using different training and testing samples and experiment with reduced feature sets to determine the specific influence of surface features (1-7), pragmatic features (8-17) and discourse function (18) respectively. (The numbers in parentheses here and below refer to the feature numbering in the appendix.)

Full Feature Set Meaning Assignment The total set of examples contains 150 *ja*-patterns and 20 *oh*-patterns. Binary coding produced 50 input and 4 output nodes (= 8 classes of discourse functions), and a fully-connected feedforward 50-10-4 net was used for training. The training procedure used in all cases was standard backpropagation, which allowed to create a large number of easily comparable generalization results.

The basic experiment concerned the task of learning a func-

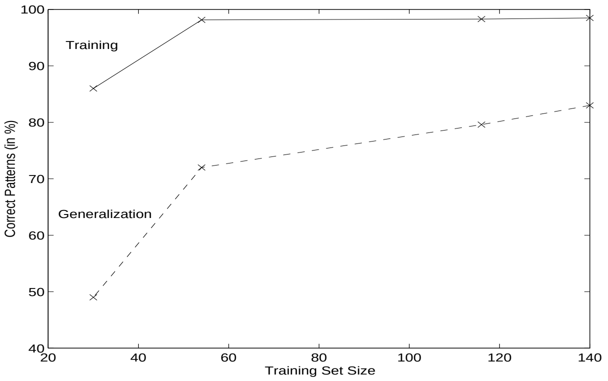

Figure 4: Influence of the Training Set Size on Training and Generalization

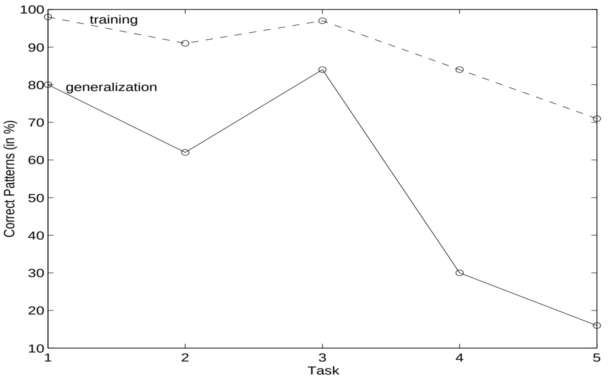

Figure 5: Results for Learning with Reduced Feature Sets: 1: input features → discourse function(DF) 2: surface features (sfs) → DF, 3: pragmatic features → DF, 4: sfs → dialogue act, 5: sfs → pragmatic features

tional mapping between input features and discourse functions, using a training set and generalizing to a test set from the 170 coded examples. We experimented with different sizes of training sets and different feature sets.

In Fig. 4 we show the influence of the training set size on generalization performance. We can see that performance starts to level off when we use about 50 patterns for training. This boundary condition on the training size has interesting implications for lexical learning. It may underscore the notion that lexical learning requires only a comparatively small set of examples to realize a functional mapping from input to meaning for a great number of other cases (cf. [13]). Below we show that analogical transfer to a new particle may be successful for an even smaller, *prototypical* training set.

Reduced feature sets The surface features used are easily extractible from text corpora using preprocessing and statistical corpus analysis tools (cf. [2]). Therefore it is of great practical interest to perform meaning assignment given only the surface features. In that case there would be no need for human intervention in the training process, and a fully automated discourse function assignment system would result. Accordingly, we have experimented with various reduced feature sets. We used a 100 training set, 70 test set scenario in all cases. The results are shown in Fig. 5. Task 1 refers to the basic experiment with the full feature set (1-17 input, 18 output). In task 2 we restricted the input to the surface features (1-7) and learned the assignment of the discourse function (18). We see that both training and generalization performance drop from 98% to 91% and 80% to 62% respectively. In Task 3 we used only the pragmatic features (8-17) to predict discourse function (18) and find that the performance matches or exceeds that of the full feature set (98%, 84%). Task 4 and 5 confirm the view that the contribution of pragmatic features concerns information that cannot be extracted from the surface features. In each case, we tried to predict pragmatic features from surface features alone, and find only weak dependencies between feature sets. In task 4 the output was a single pragmatic feature, the dialogue act of the current utterance (9) (which has 14 different feature values), and in task 5 the output was the whole range of pragmatic features (8-17). These results show that discourse functions may be predicted from *pragmatic* features alone as well as from surface and pragmatic features combined. The pragmatic features used here have been manually encoded using operationalized guidelines for intersubjective agreement. We may assume that they are in themselves complex features, which incorporate some surface cues, i.e. they are not statistically independent of the surface features. This is evidenced in task 3 and task 4. Therefore, given the complex, pragmatic features, the surface features are not really needed to perform the meaning assignment. However, pragmatic features are difficult to derive automatically, because they incorporate additional properties of a wider context and shared-background understanding. In that light, the level of performance achieved using only surface features is rather encouraging. It seems that a considerable part of the discourse function assignment can be performed by looking at the immediate surface utterance context of a discourse particle.

We conclude that we do get significant generalization results for the meaning assignment of discourse functions given a full feature analysis of the linguistic and pragmatic context. All of these features contribute towards the determination of a function of a particular lexeme - for fully automated analysis to be feasible there must be a way to extract pragmatic features from discourse as well.

Generalization and Prototype effects An interesting question from a cognitive point of view is the mechanism of ‘analogical transfer’ or acquiring a new particle of the same type. In this study, we use a comparatively large sample of *ja*-occurrences to generalize to a smaller one of *oh*-occurences. Generally, analogical transfer should be possible only to a limited degree, because the specific influence of the lexeme will be disregarded (i.e. there is at least one untrained feature in the set). We performed several experiments on generalizing from one discourse particle to the other, using the computed prototypes for *ja* to speed up the learning process. We found that performance on generalization was 50% correctness with the unedited sample, it was slightly higher (55%), when the frequency of prototypes was significantly increased (10x), but the best results were achieved with a training set of only prototypes (i.e. 8 patterns) (65%).

Obviously it is easier to classify a set of descriptive patterns if a net has been trained on a small set of most salient, central patterns than if a lot of spurious feature patterns are reflected in the weights of the network. This is a general property of network learning which should occur when we have stable, consistent pattern-class mappings. These results underline the usefulness of using neurally inspired classification methods for linguistic tasks: We can make the notion of 'analogical transfer' more explicit and improve performance for language engineering tasks as well.

Uniqueness of meaning assignment A question not adressed in this paper is the uniqueness of discourse function. The examples used have all been analysed for their dominant reading only. It is a general feature of natural languages that disambiguation procedures are not always completely realized (e.g. pp-attachment, systematic lexical ambiguity, pronominal reference) (cf. e.g. [6]) and subjective judgments on meaning assigment problems vary. We must expect a certain level of dubious cases even with a perfect meaning assignment model. In our attempt to characterize the individual's capacity for understanding, the role of interactive clarification processes and unresolved mis-assignments in everyday communication should not be underestimated. It is possible that even humans may perform only in the 80%-90% range (of correctly understood discourse meanings) in real settings.

Conclusion

Discourse particles offer a fascinating view on linguistic cognitive abilities because of their simultaneous reference to the communicative setting and their expression of semantic content. Viewing language as a cognitive ability automatically puts spoken language discourse at the center of attention, rather than the derived ability of producing written text according to the norms of a standard language. We need to test our theories of linguistic ability against the empirical data of real dialogues as embodied in spoken language corpora. In this paper we have tried to move a step in that direction.

The experiments reported above show that discourse particles follow the general pattern of categorical meaning assignment and that the contribution of different types of features from the communicative situation can be explored in considerable detail.

Prototype abstraction was shown to be a significant factor in learning a new discourse particle on the basis of the contextual distribution and functional properties of another. Meaning assignment for discourse particles may be regarded as an exemplification of general lexical interpretation processes - where the influence of the communicative situation is highly apparent. We may conjecture that cognitive models of other types of lexical or morphological items may similarly have to be constrained by discourse factors, at least when analysed in the context of spoken language.

References

[1] URL:http://www7.informatik.tu-muenchen.de/projects/NEUROSEM/

[2] S. Bauer. Entwicklung eines Eingabe-Taggers für lexikalisch-syntaktische Information. Master's thesis, Technische Universität München, November 1995.

[3] K. Fischer. A construction-based approach to the lexicalization of interjections. In M. Gellerstam, S. Malmberg, K. Noeren, L. Rogstrom, and C. Rojder Papmehl, editors, *Proceedings of EuraLex'96*, 1996.

[4] K. Fischer. Distributed representation formalisms for discourse particles. In D. Gibbon, editor, *Natural Language Processing and Speech Technology*. Mouton-de Gruyter, 1996.

[5] K. Fischer and M. Johanntokrax. Ein linguistisches Merkmalsmodell für die Lexikalisierung von diskurssteuernden Einheiten. Technical Report SFB-360-18, Universität Bielefeld, 1995.

[6] G. M. Green. Ambiguity resolution and discourse interpretation. In K. van Deemter and S. Peters, editors, *Semantic Ambiguity and Underspecification*, Cambridge University Press, 1996.

[7] B.J. Grosz, M.E. Pollack, and C.L.Sidner. Discourse. In M. Posner, editor, *Foundations of Cognitive Science*. MIT Press, 1989.

[8] W.J.M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[9] G. Sagerer, H.J. Eikmeyer, and G. Rickheit. "Wir bauen jetzt ein Flugzeug": Konstruieren im Dialog. Technical Report SFB 360-1, University of Bielefeld, 1994.

[10] G. Scheler. Feature-based perception of semantic concepts. Freksa, C., editor, *Cognition and Computation*. Springer, 1997 (to appear).

[11] G. Scheler. Feature selection with exception handling—an example from phonology. In Robert Trappl, editor, *Proceedings of the European Meeting on Systems and Cybernetics*, 1994.

[12] G. Scheler. Generating English plural determiners from semantic representations. In S. Wermter, E. Riloff, and G. Scheler, editors, *Learning for natural language processing: Statistical, connectionist and symbolic approaches*. Springer, 1996.

[13] G. Scheler. Learning the semantics of aspect. In H. Somers, editor, *New Methods in Language Processing*. University College London Press, 1996.

[14] D. Schiffrin. *Discourse Markers*. Cambridge University Press, 1987.

[15] B. Schmitz and J.J.Quantz. Dialogue acts in automatic dialogue processing. In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation*, 1997.

Inventory of Descriptive Features

Features are numbered in parentheses and rendered in **boldface**, feature-values are given in *italics*. (Not all feature values are reported here.)

Surface Features

lexeme (1) (here: *ja* and *oh*);

turn (2) and **utterance** (3) **position** of the discourse particle, i.e. *own, init, 2nd, medial, final*;

right (4) and **left** (5) **context**, i.e. the preceding and following syntactic constituents, for instance, *NP, VP, PP*, but also *pause, adverb*, etc.;

combination with other discourse particles (6), such as *so, ah, gut*;

the **role** (7) the speaker fulfills in the discourse situation, i.e. *instructor* or *constructor*.

Pragmatic Features

pragmatic functions, with respect to the:

- **turn-taking system** (10), i.e. *taking, holding, yielding* and *supporting* a turn;
- **information management** domain (17), e.g. signalling the beginning of a *new topic* or *highlighting* important information, *segmenting* utterances or indicating that the current utterance is *relevant* to the preceding one;
- **speech management** level (11), e.g. concerning the *time* for speech planning activities;

domain-specific dialogue acts, of the:

- **preceding utterance** (8), e.g. *request, acknowledgment*;
- **current utterance** (9), with the same set of values;

cognitive content of *oh* and *ja*, refering to:

- the **mental state** of the **speaker** (12);
- the supposed **mental state** of the **hearer** (13);
- the **action** level (14);
- the **propositional** level (15);
- and the **interactional** level (16).

Discourse Functions

(18)

take-up(tu): gives feedback to the other speaker and signals that one intends to take the turn to say something relevant;

backchannel(bc): gives feedback and supports the other's turn;

frame(fr): introduces a new topic or concludes the previous one;

repair marker(rm): signals problems in the formulation process;

answer(an): signals agreement on the same proposition;

action(ac): refers to the task the speaker fulfills in the situation; for instance in the toy-airplane construction dialogues, German *ja* can be used to indicate that the action is completed;

check(ch): signals the hearer that the speaker would like to get positive feedback;

modal(md): refers to the hearer's supposed mental state.