Dialogue Act Classification Based on Intra-Utterance Features*

Nick Webb and **Mark Hepple** and **Yorick Wilks**
Natural Language Processing Group
Department of Computer Science
University of Sheffield, UK
{n.webb,m.hepple,y.wilks}@dcs.shef.ac.uk

Abstract

We present recent work in the area of Dialogue Act (DA) tagging. Identifying the dialogue acts of utterances is recognised as an important step towards understanding the content and nature of what speakers say. Our experiments investigate the use of a simple dialogue act classifier based on purely *intra-utterance* features — principally involving word n-gram cue phrases. Such a classifier performs surprisingly well, rivalling scores obtained using far more sophisticated language modelling techniques for the corpus we address. We also discuss the potential utility of classifiers that identify the n most likely dialogue acts for each utterance, leaving it to some later process to choose amongst these alternatives.

Introduction

In the area of spoken language dialogue systems, the ability to assign user input with a functional tag which represents the communicative intentions behind each utterance — the utterance's *dialogue act* — is acknowledged to be a useful first step in dialogue processing. Such tagging can assist the semantic interpretation of user utterances, and can help an automated system in producing an appropriate response.

Researchers, for example Hirschberg & Litman; Grosz & Sidner (1993; 1986), speak of cue phrases in utterances which can serve as useful indicators of dialogue acts. In common with the work of Samuel, Carberry, & Vijay-Shanker (1999), we wanted to detect automatically word n-grams in a corpus that might serve as potentially useful cue phrases. The method we chose for selecting such phrases is based on their *predictivity*. The predictivity of cue phrases can be exploited directly in a simple model of dialogue act classification that employs only intra-utterance features. We report here the results of experiments evaluating this simple approach on the SWITCHBOARD corpus. Surprisingly, the results we obtain rival the best results achieved on that corpus, in work by Stolcke *et al.* (2000), who use a far more complex approach involving Hidden Markov modelling (HMM), that addresses both the sequencing of words *within* utterances and the sequencing of dialogue acts *over* utterances.

This simple classification approach can as well be used to produce a (possibly ranked) list of the n most likely alternative classifications for each utterance, which might feed into some subsequent process, such as a dialogue manager, that could select amongst the restricted set of alternatives offered on the basis of higher-level dialogue information. The subsequent process might alternatively be a machine-learning based component trained to make the final choice of DA based on inter-utterance context, with the possible benefit of having a much reduced feature space from the elimination of word n-gram based features, which would have already been exploited in the simple classifier component.

The work described in this paper forms part of the AMITIES project (Hardy *et al.* 2004), which aims to build automated service counters allowing users to access information (e.g. such as banking information) in a more natural and flexible way. The models we use to achieve this will make use of dialogue act sequencing information.

This paper presents our work on dialogue act classification using intra-utterance information. Previous work with dialogue act modelling is outlined in Section 2. An overview of the available corpora for this task is given in Section 3. Our experiments evaluating the simple cue-based dialogue act classifier approach to assign a single DA to each utterance are described in Section 4. Our initial explorations around classifiers assigning n-best lists of DAs is described in Section 5. We end with some discussion and an outline of intended further work.

Related Work

There has been an increasing interest in using machine learning techniques on problems in spoken dialogue. One thread of this work has addressed dialogue act modelling, i.e. the task of assigning an appropriate dialogue act tag to each utterance in a dialogue. It is only recently, with the availability of annotated dialogue corpora, that research in this area has become possible.

One approach that has been tried for dialogue act tagging is the use of n-gram language modelling, exploiting principally ideas drawn from the area of speech recognition. For

*This paper is based on work supported in part by the European Commission under the 5th Framework IST/HLT Program, and by the U.S. Defense Advanced Research Projects Agency.

Corpus	*Availability*	*Utterance count*	*Dialogue count*	*Word count*	*Distinct words*	*Dialogue type*
SWITCHBOARD	public	223606	1155	1431725	21715	Conversational
VERBMOBIL	public	3117	168	24980	959	Task-oriented
MAPTASK	public	26621	128	152705	2502	Task-oriented
AMITIES GE	restricted	30206	1000	228165	7841	Task-oriented
AMITIES IBM	restricted	122080	5000	1132663	11586	Task-oriented

Figure 1: Summary data for the dialogue corpora

example, Reithinger & Klesen (1997) have applied such an approach to the VERBMOBIL corpus, which provides only a rather limited amount of training data, and report a tagging accuracy of 74.7%. Stolcke *et al.* (2000) apply a somewhat more complicated HMM method to the SWITCHBOARD corpus, one that exploits both the order of words *within* utterances and the order of dialogue acts *over* utterances. They use a single split of the data for their experiments, with 198k utterances for training and 4k utterances for testing, achieving a DA tagging accuracy of 71.0% on word transcripts. These performance differences, with a higher tagging accuracy score for the VERBMOBIL corpus despite significantly less training data, can be seen to reflect the differential difficulty of tagging for the two corpora.

A second approach that has been applied to dialogue act modelling, by Samuel, Carberry, & Vijay-Shanker (1998), uses transformation-based learning over a number of utterance features, including utterance length, speaker turn and the dialogue act tags of adjacent utterances. They achieved an average score of 75.12% tagging accuracy over the VERBMOBIL corpus. A significant aspect of this work, that is of particular relevance here, has addressed the automatic identification of word sequences that might serve as useful dialogue act cues. A number of statistical criteria are applied to identify potentially useful word n-grams which are then supplied to the transformation-based learning method to be treated as 'features'.

Corpora

Publicly available corpora

Three key corpora have been used in most work on DA modelling. First, the VERBMOBIL project, on speech-to-speech translation, produced a corpus of 168 English annotated task-oriented dialogues, whose 'task' is meeting arrangement. The VERBMOBIL corpus is tagged used a total of 46 tags, which are then further clustered into 26 top-level tags. Secondly, the SWITCHBOARD corpus (Jurafsky *et al.* 1998) comprises 1155 annotated conversations of an unstructured, non-directed character, which have a high greater variability of topics, hence exhibit much greater semantic variability than VERBMOBIL and have therefore been thought to present a more difficult problem for accurate DA modelling. The corpus is annotated using an elaboration of the DAMSL tag set (Core & Allen 1997), involving 50 major classes, together with a number of diacritic marks, which combine to generate 220 distinct labels. Jurafsky *et al.* (1998) propose a clustering of the 220 tags into 42 larger classes, listed in Figure 2, and it is this clustered set that was used in the experiments of Stolcke *et al.* (2000). The third corpus is MAPTASK, comprising 128 task-oriented dialogues in which two people negotiate an agreed route on separate (and slightly different) maps. The DA annotation uses 12 distinct DA labels, and part of the corpus is annotated for dialogue games. The dialogues in this corpus are more collaborative in nature.

Restricted Corpora

Although not used in the experiments reported here, the nature of the AMITIES corpora was a motivating factor in the selection of the SWITCHBOARD corpus for our work. Previous work on the automatic selection of cue phrases from a corpus was done on VERBMOBIL, which has a very small vocabulary size and utterance count. We concentrated on SWITCHBOARD in part to study the effect of scale on this task.

During the AMITIES project, we collected 1000 English and 1000 French human-human dialogues from GE call centres. The calls are of an information seeking or transactional type, in which customers interact with their financial accounts by phone to check balances, make payments and report lost credit cards. The AMITIES GE corpus is annotated with DAs (using DAMSL) and additional domain specific semantic information such as account numbers and credit card details (Hardy *et al.* 2002). Later in AMITIES, we acquired 10000 transcribed calls from an IBM call centre (half French, half English), which involve call-routing dialogues, of limited length, for a product and hardware support service. This data is currently being annotated, using the same formalism as AMITIES GE. Both corpora are large (equal to SWITCHBOARD) but are task oriented and so have greater regularity in semantic content that we hope can be exploited to help DA tagging.

Simple DA Classification

In this section we describe our simple approach to DA classification, based on intra-utterance features, together with our experiments to evaluate it. A key aspect of the approach is the selection of the word n-grams to use as cue phrases in tagging. Samuel, Carberry, & Vijay-Shanker (1999) investigate a series of different statistical criteria for use in automatically selecting cue phrases. We use a criterion of *predictivity*, described below, which is one that Samuel, Carberry, & Vijay-Shanker (1999) do not consider. Predictivity values are straightforward to compute, so the approach can feasibly

Dialogue Act	*% of corpus*	*Dialogue Act*	*% of corpus*
statement-non-opinion	36%	action-directive	0.4%
acknowledge	19%	collaborative completion	0.4%
statement-opinion	13%	repeat-phrase	0.3%
agreeaccept	5%	open-question	0.3%
abandoned	5%	rhetorical-questions	0.2%
appreciation	2%	hold before answer	0.2%
yes-no-question	2%	reject	0.2%
non-verbal	2%	negative non-no answers	0.1%
yes answers	1%	signal-non-understanding	0.1%
conventional-closing	1%	other answers	0.1%
uninterpretable	1%	conventional-opening	0.1%
wh-question	1%	or-clause	0.1%
no answers	1%	dispreferred answers	0.1%
response acknowledgement	1%	3rd-party-talk	0.1%
hedge	1%	offers, options commits	0.1%
declarative yes-no-question	1%	self-talk	0.1%
other	1%	downplayer	0.1%
backchannel in question form	1%	maybeaccept-par	$< 0.1\%$
quotation	0.5%	tag-question	$< 0.1\%$
summarisereformulate	0.5%	declarative wh-question	$< 0.1\%$
affirmative non-yes answers	0.4%	apology	$< 0.1\%$

Figure 2: SWITCHBOARD dialogue acts

be applied to very large corpora. As we shall see, predictivity scores are used not only in selecting cue phrases, but also directly as part of the classification method.

Experimental corpus

For our experiments, we used the SWITCHBOARD data set of 1155 annotated conversations. The dialogue act types for this set can be seen in (Jurafsky *et al.* 1997). Altogether these 1155 conversations comprise in the region of 205k utterances. We created three different experimental data sets from this corpus. The first mirrored the size of the experiments performed on the VERBMOBIL corpus, described above, i.e. 3k utterances. To investigate the effects of using a greater amount of data, a 50k utterance data set was used, with 45k used for training and 5k for testing in each experiment. The final split used the same data size as that of Stolcke *et al.* (2000), with 198k utterances for training and 4k for testing. We hoped to show that a significant increase in the amount of training data would translate to a much improved tagging accuracy. Our experiments address both the initial 220 element tag set of the corpus, and the clustered set of 42 tags discussed earlier, and listed in Figure 2. The corpus was pre-processed to remove all punctuation and case information. Some of the corpus mark-up, such as filler information described in Meteer (1995), was also removed.

Our experiments used a cross-validation approach, with results being averaged over 10 runs. For the first two data sets, this was a standard ten-fold approach, i.e. with the data being split into ten approximately equal partitions, each being used in turn for testing, with the remainder combined for training. Cross-validation of this kind is recognised as the standard way to estimate predictive accuracy. For the third data set, created for comparability with Stolcke *et al.* (2000), the test set is much less than a tenth of the overall set, so a standard ten-fold approach does not apply. Instead, we randomly selected dialogues out of the overall data to create ten disjoint subsets of around 4k utterances for use as test sets, which were re-used across the different experimental runs. In each case, the corresponding training set was the overall data minus that subset. In addition to cross-validated results, we also report the single highest score from the ten runs performed for each experimental case. We have done this to facilitate comparison with the results of Stolcke *et al.* (2000).

Cue Phrase Selection

For our experiments, the word n-grams used as cue phrases during classification are computed from the training data. All word n-grams of length 1–4 within the data are considered as candidates. The phrases chosen as cue phrases are selected principally using a criterion of *predictivity*, which is the extent to which the presence of a certain n-gram in an utterance is predictive of it having a certain dialogue act category. For an n-gram n and dialogue act d, this corresponds to the conditional probability: $P(d \mid n)$, a value which can be straightforwardly computed. Specifically, we compute all n-grams in the training data of length 1–4, counting their occurrences in the utterances of each DA category and in total, from which the above conditional probability for each n-gram and dialogue act can be computed. For each n-gram, we are interested in its *maximal* predictivity, i.e. the highest predictivity value found for it with any DA category. This set of n-grams is then reduced by applying thresholds of predictivity and occurrence, i.e. eliminating any n-gram whose maximal predictivity is below some minimum requirement, or whose maximal number of occurrences with any category falls below a threshold value. The n-grams that remain are

Data Set	*Cross Validated Score*	*Single Best Score*
4k, unclustered	51.83%	57.06%
4k, clustered 42 tags	56.47%	67.34%
as above, plus utt. length models	59.13%	66.87%
as above, plus <*start*>,<*end*> tags	61.01%	66.12%
as above, plus interrupted utterances	62.69%	69.47%

Figure 3: Experiments with 4k data set

Data Set	*Cross Validated Score*	*Single Best Score*
50k, unclustered	56.35%	60.67%
50k, clustered 42 tags	61.29%	65.80%
as above, plus utt. length models	65.71%	68.78%
as above, plus <*start*>,<*end*> tags	66.41%	69.53%
as above, plus interrupted utterances	68.42%	**71.98%**

Figure 4: Experiments with 50k data set

used as cue phrases. The threshold values that were used in our experiments were arrived at by conducting a series of experiments at varying levels of threshold and frequency. We recognise that there is an arbitrary nature to this, and we have performed subsequent experiments using a validation set to automatically set the threshold levels independently of the test data (Webb, Hepple, & Wilks 2005).

Using Cue Phrases in Classification

The selected cue phrases are used directly in classifying further utterances in the following manner. To classify an utterance, we identify all the cue phrases it contains, and determine which has the highest predictivity of some dialogue act category, and then that category is assigned. If multiple cue phrases share the same maximal predictivity, but predict different categories, we select the DA for the phrase which has the higher number of occurrences. If the combination of predicitivity and occurrence count is insufficient to determine a single DA, then a random choice is made amongst the remaining candidate DAs. If no cue phrases are present, then a default tag is assigned, corresponding to the most frequent tag within the training corpus.

Experimental cases

For each of the three data sets, we performed five different experiments, whose results are reported in Figures 3–5. The five different experimental cases are described following.

Case 1: unclustered tag set

For these experiments, the classification approach just described was applied using the full 220 element tag set from the SWITCHBOARD corpus. Applied to the 4k data set, the approach yields an average tagging accuracy of 51.83%, which compares against a baseline accuracy of 36.5% from assigning the most frequently occurring tag in the SWITCHBOARD data set (which is **sd** — statement). Applied to the medium data set, the approach yields an average tagging accuracy of 54.5%, which compares to 33.4% from using the most frequent tag. Finally, applied to the large data set, we produced an average tagging accuracy of 55.82%, compared to a baseline of 36%. These baseline remain constant across the following experimental cases for each of the three data sets.

Case 2: clustered tag set

For these experiments, we used the clustering of labels proposed by Jurafsky *et al.* (1998), which maps the full 220 DA labels in the 42 larger classes shown in Figure 2. This move produced a significant improvement in performance, around 5% in all cases. For the 4k data set, average tagging accuracy rose to 56.47% (an improvement of 4.64%). For the 50k data set, the score was 61.29% (an improvement of 4.94%), and for the 202k data set we achieved 60.73% (4.91%).

Case 3: utterance length models

For this case, we trained models sensitive to utterance length. In particular, we grouped training utterances into those of length 1, those with lengths 2–4, and those of length 5+, and produced separate models for each group. We hoped that this move would provide better classification for dialogue acts whose realisation was skewed over, for instance, short utterances like 'okay'. On the whole, the introduction of such models lead to an increase in tagging accuracy of around 4%, except in the case of the 4k set, where data sparsity was more of an issue.

Case 4: position specific cues

Further experiments suggest that we can improve this score. We introduced <*start*> and <*finish*> tags to each utterance - to capture position specific information for particular cues. For example '<*start*> *okay*' effectively identifies occurrence of word 'okay' as the first word in the utterance. The effects of these additions can be seen in the tables, but in summary, the position specific cues added a further percentage point.

Data Set	*Cross Validated Score*	*Single Best Score*
202k, unclustered	55.82%	58.92%
202k, clustered 42 tags	60.73%	65.14%
as above, plus utt. length models	64.76%	69.71%
as above, plus <*start*>,<*end*> tags	65.89%	71.51%
as above, plus interrupted utterances	**69.09%**	71.29%

Figure 5: Experiments with 202k data set

Case 5: interrupted utterances

In addition to the dialogue act mark-up of the corpus, there were several annotations relating directly to the dysfluencies encountered in the data. These are outlined in Meteer (1995). The most important of these is the dialogue act '+', which indicates an utterance which was interrupted by the other speaker, an example of which can be seen in Figure 6.

We saw that a lot of potentially useful word data was being ignored. The '+' tag occurs around 16,000 times in the 202k corpus, around 8% of total annotations. One approach to utilise this data would be to 'reconnect' the divided utterances, i.e. appending any utterance assigned tag '+' to the last utterance by the same speaker. Clearly this approach has it's limits — and such a corpus would lose important sequence information (such as the effect of back-channels on the conversation). However, as a pre-processing step, it is worth exploring. Doing so gave us both our highest cross-validated score of 69.09% and our highest single score of 71.98%.

N-Best Dialogue Act Classification

The experiments described so far have all tried to select the single best-fit candidate dialogue act tag for an utterance. As DA tagging could be seen as a first step before possible refinement by some higher level process, we wanted to investigate the possibility of selecting some list, possibly ranked, of potential dialogue acts. The higher level process, perhaps a dialogue manager or machine learning-based selection component, could choose among some limited selection of possible acts based on additional information outside the utterance itself, such as dialogue context.

Such an approach would address the problem of being unable to resolve some ambiguity on the basis of surface realisation. For example, the utterance 'okay' can be either a back-channel or an accept/confirm, it depends entirely on the context. If we can represent such an ambiguity to a higher level process, a restricted choice can be made based on contributory factors, such as prosody, as indicated in Stolcke *et al.*; Mast *et al.* (2000; 1996).

Our most recent experiment has yielded some promising results. We built a classifier using the medium-sized data set, i.e. with 45k utterance training 5k utterance test sets. However, rather than attempting to find the single best match from the classifier, we tagged each utterance with the top 5 possible dialogue acts, as indicated by the classifier on the basis of the predictivity of the n-grams the utterance contained. All possible DAs suggested by the presence of cue phrases are considered, where the top 5 ordered by predicitivy are used. Duplicate DAs are deleted from the candidate set, so the 2nd ranked DA could be represented by the 5th ranked cue phrase, for example.

On a cross-validation of the corpus, we calculated that 86.74% of the time the correct dialogue act was contained in the 5-best output of the classifier. This score would define some theoretically attainable upper limit of performance attainable by some higher level process that selected amongst the n-best DAs.

In order to create a baseline measure for this task, we computed comparable scores for utterances assigned a default set of tags, consisting of the 5 most frequently occurring tags in the corpus. The number of times the correct dialogue act occurred in the top 5 was 71.09%.

Of course, other DA classification approaches would be able to generate n-best lists of DAs for utterances, so this idea of combining n-best classification together with a higher-level selection component has more general applicability than with just our own classification approach.

Discussion

Combining all features for simple dialogue act tagging, we obtain a cross validated score of 69.09% over the larger, 202k data set. Our highest single run score was 71.98%, using the 50k data set. It is difficult to compare our results directly with those of Stolcke *et al.* (2000), given that they did not use a cross-validation approach, but even so it is striking that our cross-validated score comes so close to their result given their use of a much more complex language modelling approach, that exploits also *inter*-utterance information. It is furthermore possible that their choice of test data was a lucky one, i.e. one giving higher scores than would arise with results averaged in cross-validation.

We have shown that a simple dialogue act tagger can be created that uses just intra-utterance cues for classification. This approach performs surprisingly well given its simplicity. One of the prime motivators for using this approach was to remove a large number of word n-grams from the feature set of machine learning algorithms. By doing so we are hopeful that we can use a wider range of machine learning approaches for this task than has presently been tried. Finally, by analysing the n-best approach to tagging, we have demonstrated that a naive classifier can present a list of ranked possible alternatives, which could be used by some later, higher level process, such as a dialogue manager, to make informed choices in the evaluation of utterances.

`Speaker A: DA="sv":` **probably the biggest thing we're got going right now is the robberies and theft and probably murder –**

`Speaker B: DA="b":` **uh-huh**

`Speaker A: DA="+":` **– are the two top ones we have.**

Figure 6: An utterance interrupted by a back-channel

Future Work

Clearly one next step is to pass the output of our classifier to some machine learning algorithm, to exploit inter-utterance relationships. Transformation-Based Learning (TBL) has been used for this task by previous researchers (Samuel, Carberry, & Vijay-Shanker 1998; Lager & Zinovjeva 1999) and we shall examine the effects of using a single-best, and n-best pre-classification approach.

An interesting area of investigation is to what extent models trained on one set of data can be used to tag data from a different domain and conversational style. This would indicate to what extent our models of DAs were general in nature - whether questions are realised in similar ways across domains, for example. We will try to tag the VERBMOBIL corpus data to determine cue phrase generality.

Finally, we aim to apply these techniques to a new corpus collected for the AMITIES project, consisting of human-human conversations recorded in the call centre domain. We hope that the techniques outlined here will prove a useful first step in creating automatic service counters for call centre applications.

References

Core, M. G., and Allen, J. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Grosz, B., and Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 19(3).

Hardy, H.; Baker, K.; Devillers, L.; Lamel, L.; Rosset, S.; Strzalkowski, T.; Ursu, C.; and Webb, N. 2002. Multi-layered dialogue annotation for automated multilingual customer service. In *Proceedings of the ISLE workshop on Dialogue Tagging for Multimodal Human Computer Interaction, Edinburgh*.

Hardy, H.; Biermann, A.; Inouye, R. B.; Mckenzie, A.; Strzalkowski, T.; Ursu, C.; Webb, N.; and Wu, M. 2004. Data driven strategies for an automated dialogue system. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona*.

Hirschberg, J., and Litman, D. 1993. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics* 19(3):501–530.

Jurafsky, D.; Bates, R.; Coccaro, N.; Martin, R.; Meteer, M.; Ries, K.; Shriberg, E.; Stolcke, A.; Taylor, P.; and Ess-Dykema, C. V. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara.*

Jurafsky, D.; Bates, R.; Coccaro, N.; Martin, R.; Meteer, M.; Ries, K.; Shriberg, E.; Stolcke, A.; Taylor, P.; and Ess-Dykema, C. V. 1998. Switchboad discourse language modeling project final report. Research Note 30, Center for Language and Speech Processing, Johns Hopkins University, Baltimore.

Lager, T., and Zinovjeva, N. 1999. Training a dialogue act tagger with the μ-TBL system. In *Proceedings of the Third Swedish Symposium on Mulitmodal Communication, Linköping University Natural Language Processing Laboratory*.

Mast, M.; Kompe, R.; Harbeck, S.; Kiessling, A.; and Warnke, V. 1996. Dialog act classification with the help of prosody. In *Proceedings of the International Conference on Speech and Language Processing ICSLP '96*, volume 3, 1732–1735.

Meteer, M. 1995. Dysfluency annotation stylebook for the switchboard corpus. Working paper, Linguistic Data Consortium.

Reithinger, N., and Klesen, M. 1997. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*.

Samuel, K.; Carberry, S.; and Vijay-Shanker, K. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Samuel, K.; Carberry, S.; and Vijay-Shanker, K. 1999. Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, Canada*.

Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Ess-Dykema, C. V.; and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Computational Linguistics 26(3), 339–373.*

Webb, N.; Hepple, M.; and Wilks, Y. 2005. Empirical determination of thresholds for optimal dialogue act classification. In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*.