Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech

Andreas Stolcke*
SRI International

Klaus Ries
Carnegie Mellon University and
University of Karlsruhe

Noah Coccaro
University of Colorado at Boulder

Elizabeth Shriberg
SRI International

Rebecca Bates
University of Washington

Daniel Jurafsky
University of Colorado at Boulder

Paul Taylor
University of Edinburgh

Rachel Martin
Johns Hopkins University

Carol Van Ess-Dykema
U.S. Department of Defense

Marie Meteer
BBN Technologies

We describe a statistical approach for modeling dialogue acts in conversational speech, i.e., speech-act-like units such as STATEMENT, QUESTION, BACKCHANNEL, AGREEMENT, DISAGREEMENT, *and* APOLOGY. *Our model detects and predicts dialogue acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialogue act sequence. The dialogue model is based on treating the discourse structure of a conversation as a hidden Markov model and the individual dialogue acts as observations emanating from the model states. Constraints on the likely sequence of dialogue acts are modeled via a dialogue act n-gram. The statistical dialogue grammar is combined with word n-grams, decision trees, and neural networks modeling the idiosyncratic lexical and prosodic manifestations of each dialogue act. We develop a probabilistic integration of speech recognition with dialogue modeling, to improve both speech recognition and dialogue act classification accuracy. Models are trained and evaluated using a large hand-labeled database of 1,155 conversations from the Switchboard corpus of spontaneous human-to-human telephone speech. We achieved good dialogue act labeling accuracy (65% based on errorful, automatically recognized words and prosody, and 71% based on word transcripts, compared to a chance baseline accuracy of 35% and human accuracy of 84%) and a small reduction in word recognition error.*

* Speech Technology and Research Laboratory, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, 1-650-859-2544. E-mail: stolcke@speech.sri.com.

Table 1
Fragment of a labeled conversation (from the Switchboard corpus).

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	So do you go to college right now?
A	ABANDONED	Are yo-,
B	YES-ANSWER	*Yeah,*
B	STATEMENT	*it's my last year [laughter].*
A	DECLARATIVE-QUESTION	You're a, so you're a senior now.
B	YES-ANSWER	*Yeah,*
B	STATEMENT	*I'm working on my projects trying to graduate [laughter].*
A	APPRECIATION	Oh, good for you.
B	BACKCHANNEL	*Yeah.*
A	APPRECIATION	That's great,
A	YES-NO-QUESTION	um, is, is N C University is that, uh, State,
B	STATEMENT	*N C State.*
A	SIGNAL-NON-UNDERSTANDING	What did you say?
B	STATEMENT	*N C State.*

1. Introduction

The ability to model and automatically detect discourse structure is an important step toward understanding spontaneous dialogue. While there is hardly consensus on exactly how discourse structure should be described, some agreement exists that a useful first level of analysis involves the identification of **dialogue acts** (DAs). A DA represents the meaning of an utterance at the level of illocutionary force (Austin 1962). Thus, a DA is approximately the equivalent of the speech act of Searle (1969), the conversational game move of Power (1979), or the adjacency pair part of Schegloff (1968) and Sacks, Schegloff, and Jefferson (1974).

Table 1 shows a sample of the kind of discourse structure in which we are interested. Each utterance is assigned a unique DA label (shown in column 2), drawn from a well-defined set (shown in Table 2). Thus, DAs can be thought of as a tag set that classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria. The computational community has usually defined these DA categories so as to be relevant to a particular application, although efforts are under way to develop DA labeling systems that are domain-independent, such as the Discourse Resource Initiative's DAMSL architecture (Core and Allen 1997).

While not constituting dialogue understanding in any deep sense, DA tagging seems clearly useful to a range of applications. For example, a meeting summarizer needs to keep track of who said what to whom, and a conversational agent needs to know whether it was asked a question or ordered to do something. In related work DAs are used as a first processing step to infer dialogue games (Carlson 1983; Levin and Moore 1977; Levin et al. 1999), a slightly higher level unit that comprises a small number of DAs. Interactional dominance (Linell 1990) might be measured more accurately using DA distributions than with simpler techniques, and could serve as an indicator of the type or genre of discourse at hand. In all these cases, DA labels would enrich the available input for higher-level processing of the spoken words. Another important role of DA information could be feedback to lower-level processing. For example, a speech recognizer could be constrained by expectations of likely DAs in a given context, constraining the potential recognition hypotheses so as to improve accuracy.

The goal of this article is twofold: On the one hand, we aim to present a comprehensive framework for modeling and automatic classification of DAs, founded on well-known statistical methods. In doing so, we will pull together previous approaches

Table 2
The 42 dialogue act labels. DA frequencies are given as percentages of the total number of utterances in the overall corpus.

Tag	Example	%
STATEMENT	*Me, I'm in the legal department.*	36%
BACKCHANNEL/ACKNOWLEDGE	*Uh-huh.*	19%
OPINION	*I think it's great*	13%
ABANDONED/UNINTERPRETABLE	*So, -/*	6%
AGREEMENT/ACCEPT	*That's exactly it.*	5%
APPRECIATION	*I can imagine.*	2%
YES-NO-QUESTION	*Do you have to have any special training?*	2%
NON-VERBAL	*<Laughter>,<Throat_clearing>*	2%
YES ANSWERS	*Yes.*	1%
CONVENTIONAL-CLOSING	*Well, it's been nice talking to you.*	1%
WH-QUESTION	*What did you wear to work today?*	1%
NO ANSWERS	*No.*	1%
RESPONSE ACKNOWLEDGMENT	*Oh, okay.*	1%
HEDGE	*I don't know if I'm making any sense or not.*	1%
DECLARATIVE YES-NO-QUESTION	*So you can afford to get a house?*	1%
OTHER	*Well give me a break, you know.*	1%
BACKCHANNEL-QUESTION	*Is that right?*	1%
QUOTATION	*You can't be pregnant and have cats*	.5%
SUMMARIZE/REFORMULATE	*Oh, you mean you switched schools for the kids.*	.5%
AFFIRMATIVE NON-YES ANSWERS	*It is.*	.4%
ACTION-DIRECTIVE	*Why don't you go first*	.4%
COLLABORATIVE COMPLETION	*Who aren't contributing.*	.4%
REPEAT-PHRASE	*Oh, fajitas*	.3%
OPEN-QUESTION	*How about you?*	.3%
RHETORICAL-QUESTIONS	*Who would steal a newspaper?*	.2%
HOLD BEFORE ANSWER/AGREEMENT	*I'm drawing a blank.*	.3%
REJECT	*Well, no*	.2%
NEGATIVE NON-NO ANSWERS	*Uh, not a whole lot.*	.1%
SIGNAL-NON-UNDERSTANDING	*Excuse me?*	.1%
OTHER ANSWERS	*I don't know*	.1%
CONVENTIONAL-OPENING	*How are you?*	.1%
OR-CLAUSE	*or is it more of a company?*	.1%
DISPREFERRED ANSWERS	*Well, not so much that.*	.1%
3RD-PARTY-TALK	*My goodness, Diane, get down from there.*	.1%
OFFERS, OPTIONS & COMMITS	*I'll have to check that out*	.1%
SELF-TALK	*What's the word I'm looking for*	.1%
DOWNPLAYER	*That's all right.*	.1%
MAYBE/ACCEPT-PART	*Something like that*	<.1%
TAG-QUESTION	*Right?*	<.1%
DECLARATIVE WH-QUESTION	*You are what kind of buff?*	<.1%
APOLOGY	*I'm sorry.*	<.1%
THANKING	*Hey thanks a lot*	<.1%

as well as new ideas. For example, our model draws on the use of DA n-grams and the hidden Markov models of conversation present in earlier work, such as Nagata and Morimoto (1993, 1994) and Woszczyna and Waibel (1994) (see Section 7). However, our framework generalizes earlier models, giving us a clean probabilistic approach for performing DA classification from unreliable words and nonlexical evidence. For the speech recognition task, our framework provides a mathematically principled way to condition the speech recognizer on conversation context through dialogue structure, as well as on nonlexical information correlated with DA identity. We will present methods in a domain-independent framework that for the most part treats DA labels as an arbitrary formal tag set. Throughout the presentation, we will highlight the simplifications

and assumptions made to achieve tractable models, and point out how they might fall short of reality.

Second, we present results obtained with this approach on a large, widely available corpus of spontaneous conversational speech. These results, besides validating the methods described, are of interest for several reasons. For example, unlike in most previous work on DA labeling, the corpus is not task-oriented in nature, and the amount of data used (198,000 utterances) exceeds that in previous studies by at least an order of magnitude (see Table 14).

To keep the presentation interesting and concrete, we will alternate between the description of general methods and empirical results. Section 2 describes the task and our data in detail. Section 3 presents the probabilistic modeling framework; a central component of this framework, the discourse grammar, is further discussed in Section 4. In Section 5 we describe experiments for DA classification. Section 6 shows how DA models can be used to benefit speech recognition. Prior and related work is summarized in Section 7. Further issues and open problems are addressed in Section 8, followed by concluding remarks in Section 9.

2. The Dialogue Act Labeling Task

The domain we chose to model is the Switchboard corpus of human-human conversational telephone speech (Godfrey, Holliman, and McDaniel 1992) distributed by the Linguistic Data Consortium. Each conversation involved two randomly selected strangers who had been charged with talking informally about one of several, self-selected general-interest topics. To train our statistical models on this corpus, we combined an extensive effort in human hand-coding of DAs for each utterance, together with a variety of automatic and semiautomatic tools. Our data consisted of a substantial portion of the Switchboard waveforms and corresponding transcripts, totaling 1,155 conversations.

2.1 Utterance Segmentation

Before hand-labeling each utterance in the corpus with a DA, we needed to choose an utterance segmentation, as the raw Switchboard data is not segmented in a linguistically consistent way. To expedite the DA labeling task and remain consistent with other Switchboard-based research efforts, we made use of a version of the corpus that had been hand-segmented into sentence-level units prior to our own work and independently of our DA labeling system (Meteer et al. 1995). We refer to the units of this segmentation as **utterances**. The relation between utterances and speaker turns is not one-to-one: a single turn can contain multiple utterances, and utterances can span more than one turn (e.g., in the case of backchanneling by the other speaker in mid-utterance). Each utterance unit was identified with one DA, and was annotated with a single DA label. The DA labeling system had special provisions for rare cases where utterances seemed to combine aspects of several DA types.

Automatic segmentation of spontaneous speech is an open research problem in its own right (Mast et al. 1996; Stolcke and Shriberg 1996). A rough idea of the difficulty of the segmentation problem on this corpus and using the same definition of utterance units can be derived from a recent study (Shriberg et al. 2000). In an automatic labeling of word boundaries as either utterance or nonboundaries using a combination of lexical and prosodic cues, we obtained 96% accuracy based on correct word transcripts, and 78% accuracy with automatically recognized words. The fact that the segmentation and labeling tasks are interdependent (Warnke et al. 1997; Finke et al. 1998) further complicates the problem.

Based on these considerations, we decided not to confound the DA classification

task with the additional problems introduced by automatic segmentation and assumed the utterance-level segmentations as given. An important consequence of this decision is that we can expect utterance length and acoustic properties at utterance boundaries to be accurate, both of which turn out to be important features of DAs (Shriberg et al. 1998, see also Section 5.2.1).

2.2 Tag Set

We chose to follow a recent standard for shallow discourse structure annotation, the Dialogue Act Markup in Several Layers (DAMSL) tag set, which was designed by the natural language processing community under the auspices of the Discourse Resource Initiative (Core and Allen 1997). We began with the DAMSL markup system, but modified it in several ways to make it more relevant to our corpus and task. DAMSL aims to provide a domain-independent framework for dialogue annotation, as reflected by the fact that our tag set can be mapped back to DAMSL categories (Jurafsky, Shriberg, and Biasca 1997). However, our labeling effort also showed that content- and task-related distinctions will always play an important role in effective DA labeling.

The Switchboard domain itself is essentially "task-free," thus giving few external constraints on the definition of DA categories. Our primary purpose in adapting the tag set was to enable computational DA modeling for conversational speech, with possible improvements to conversational speech recognition. Because of the lack of a specific task, we decided to label categories that seemed both inherently interesting linguistically and that could be identified reliably. Also, the focus on conversational speech recognition led to a certain bias toward categories that were lexically or syntactically distinct (recognition accuracy is traditionally measured including all lexical elements in an utterance).

While the modeling techniques described in this paper are formally independent of the corpus and the choice of tag set, their success on any particular task will of course crucially depend on these factors. For different tasks not all the techniques used in this study might prove useful and others could be of greater importance. However, we believe that this study represents a fairly comprehensive application of technology in this area and can serve as a point of departure and reference for other work.

The resulting SWBD-DAMSL tag set was multidimensional; approximately 50 basic tags (e.g., QUESTION, STATEMENT) could each be combined with diacritics indicating orthogonal information, for example, about whether or not the dialogue function of the utterance was related to Task-Management and Communication-Management. Approximately 220 of the many possible unique combinations of these codes were used by the coders (Jurafsky, Shriberg, and Biasca 1997). To obtain a system with somewhat higher interlabeler agreement, as well as enough data per class for statistical modeling purposes, a less fine-grained tag set was devised. This tag set distinguishes 42 mutually exclusive utterance types and was used for the experiments reported here. Table 2 shows the 42 categories with examples and relative frequencies.[1] While some of the original infrequent classes were collapsed, the resulting DA type distribution is still highly skewed. This occurs largely because there was no basis for subdividing the dominant DA categories according to task-independent and reliable criteria.

The tag set incorporates both traditional sociolinguistic and discourse-theoretic notions, such as rhetorical relations and adjacency-pairs, as well as some more form-based labels. Furthermore, the tag set is structured so as to allow labelers to annotate a Switchboard conversation from transcripts alone (i.e., without listening) in about 30 minutes.

1 For the study focusing on prosodic modeling of DAs reported elsewhere (Shriberg et al. 1998), the tag set was further reduced to six categories.

Without these constraints the DA labels might have included some finer distinctions, but we felt that this drawback was balanced by the ability to cover a large amount of data.[2]

Labeling was carried out in a three-month period in 1997 by eight linguistics graduate students at CU Boulder. Interlabeler agreement for the 42-label tag set used here was 84%, resulting in a Kappa statistic of 0.80. The Kappa statistic measures agreement normalized for chance (Siegel and Castellan, Jr. 1988). As argued in Carletta (1996), Kappa values of 0.8 or higher are desirable for detecting associations between several coded variables; we were thus satisfied with the level of agreement achieved. (Note that, even though only a single variable, DA type, was coded for the present study, our goal is, among other things, to model associations between several instances of that variable, e.g., between adjacent DAs.)

A total of 1,155 Switchboard conversations were labeled, comprising 205,000 utterances and 1.4 million words. The data was partitioned into a training set of 1,115 conversations (1.4M words, 198K utterances), used for estimating the various components of our model, and a test set of 19 conversations (29K words, 4K utterances). Remaining conversations were set aside for future use (e.g., as a test set uncompromised of tuning effects).

2.3 Major Dialogue Act Types

The more frequent DA types are briefly characterized below. As discussed above, the focus of this paper is not on the nature of DAs, but on the computational framework for their recognition; full details of the DA tag set and numerous motivating examples can be found in a separate report (Jurafsky, Shriberg, and Biasca 1997).

Statements and Opinions. The most common types of utterances were STATEMENTS and OPINIONS. This split distinguishes "descriptive, narrative, or personal" statements (STATEMENT) from "other-directed opinion statements" (OPINION). The distinction was designed to capture the different kinds of responses we saw to opinions (which are often countered or disagreed with via further opinions) and to statements (which more often elicit continuers or backchannels):

Dialogue Act	Example Utterance
STATEMENT	Well, we have a cat, um,
STATEMENT	He's probably, oh, a good two years old, big, old, fat and sassy tabby.
STATEMENT	He's about five months old
OPINION	Well, rabbits are darling.
OPINION	I think it would be kind of stressful.

OPINIONS often include such hedges as *I think*, *I believe*, *it seems*, and *I mean*. We combined the STATEMENT and OPINION classes for other studies on dimensions in which they did not differ (Shriberg et al. 1998).

Questions. Questions were of several types. The YES-NO-QUESTION label includes only utterances having both the pragmatic force of a yes-no-question *and* the syntactic mark-

2 The effect of lacking acoustic information on labeling accuracy was assessed by relabeling a subset of the data *with* listening, and was found to be fairly small (Shriberg et al. 1998). A conservative estimate based on the relabeling study is that for most DA types at most 2% of the labels might have changed based on listening. The only DA types with higher uncertainty were BACKCHANNELS and AGREEMENTS, which are easily confused with each other without acoustic cues; here the rate of change was no more than 10%.

Table 3
Most common realizations of backchannels in Switchboard.

Frequency	Form	Frequency	Form	Frequency	Form
38%	uh-huh	2%	yes	1%	sure
34%	yeah	2%	okay	1%	um
9%	right	2%	oh yeah	1%	huh-uh
3%	oh	1%	huh	1%	uh

ings of a yes-no-question (i.e., subject-inversion or sentence-final tags). DECLARATIVE-QUESTIONS are utterances that function pragmatically as questions but do not have "question form." By this we mean that declarative questions normally have no *wh*-word as the argument of the verb (except in "echo-question" format), and have "declarative" word order in which the subject precedes the verb. See Weber (1993) for a survey of declarative questions and their various realizations.

Dialogue Act	Example Utterance
YES-NO-QUESTION	Do you have to have any special training?
YES-NO-QUESTION	But that doesn't eliminate it, does it?
YES-NO-QUESTION	Uh, I guess a year ago you're probably watching C N N a lot, right?
DECLARATIVE-QUESTION	So you're taking a government course?
WH-QUESTION	Well, how old are you?

Backchannels. A backchannel is a short utterance that plays discourse-structuring roles, e.g., indicating that the speaker should go on talking. These are usually referred to in the conversation analysis literature as "continuers" and have been studied extensively (Jefferson 1984; Schegloff 1982; Yngve 1970). We expect recognition of backchannels to be useful because of their discourse-structuring role (knowing that the hearer expects the speaker to go on talking tells us something about the course of the narrative) and because they seem to occur at certain kinds of syntactic boundaries; detecting a backchannel may thus help in predicting utterance boundaries and surrounding lexical material.

For an intuition about what backchannels look like, Table 3 shows the most common realizations of the approximately 300 types (35,827 tokens) of backchannel in our Switchboard subset. The following table shows examples of backchannels in the context of a Switchboard conversation:

Speaker	Dialogue Act	Utterance
B	STATEMENT	*but, uh, we're to the point now where our financial income is enough that we can consider putting some away –*
A	BACKCHANNEL	Uh-huh. /
B	STATEMENT	*– for college, /*
B	STATEMENT	*so we are going to be starting a regular payroll deduction –*
A	BACKCHANNEL	Um. /
B	STATEMENT	*— in the fall /*
B	STATEMENT	*and then the money that I will be making this summer we'll be putting away for the college fund.*
A	APPRECIATION	Um. Sounds good.

Turn Exits and Abandoned Utterances. Abandoned utterances are those that the speaker breaks off without finishing, and are followed by a restart. Turn exits resemble abandoned utterances in that they are often syntactically broken off, but they are used mainly as a way of passing speakership to the other speaker. Turn exits tend to be single words, often *so* or *or*.

Speaker	Dialogue Act	Utterance
A	STATEMENT	we're from, uh, I'm from Ohio /
A	STATEMENT	and my wife's from Florida /
A	TURN-EXIT	so, -/
B	BACKCHANNEL	*Uh-huh.* /
A	HEDGE	so, I don't know, /
A	ABANDONED	it's <lipsmack>, - /
A	STATEMENT	I'm glad it's not the kind of problem I have to come up with an answer to because it's not –

Answers and Agreements. YES-ANSWERS include *yes*, *yeah*, *yep*, *uh-huh*, and other variations on *yes*, when they are acting as an answer to a YES-NO-QUESTION or DECLARATIVE-QUESTION. Similarly, we also coded NO-ANSWERS. Detecting ANSWERS can help tell us that the previous utterance was a YES-NO-QUESTION. Answers are also semantically significant since they are likely to contain new information.

AGREEMENT / ACCEPT, REJECT, and MAYBE / ACCEPT-PART all mark the degree to which a speaker accepts some previous proposal, plan, opinion, or statement. The most common of these are the AGREEMENT / ACCEPTS. These are very often *yes* or *yeah*, so they look a lot like ANSWERS. But where answers follow questions, agreements often follow opinions or proposals, so distinguishing these can be important for the discourse.

3. Hidden Markov Modeling of Dialogue

We will now describe the mathematical and computational framework used in our study. Our goal is to perform DA classification and other tasks using a probabilistic formulation, giving us a principled approach for combining multiple knowledge sources (using the laws of probability), as well as the ability to derive model parameters automatically from a corpus, using statistical inference techniques.

Given all available evidence E about a conversation, the goal is to find the DA sequence U that has the highest posterior probability $P(U|E)$ given that evidence. Applying Bayes' Rule we get

$$\begin{aligned} U^* &= \operatorname*{argmax}_U P(U|E) \\ &= \operatorname*{argmax}_U \frac{P(U)P(E|U)}{P(E)} \\ &= \operatorname*{argmax}_U P(U)P(E|U) \end{aligned} \tag{1}$$

Here $P(U)$ represents the prior probability of a DA sequence, and $P(E|U)$ is the likelihood of U given the evidence. The likelihood is usually much more straightforward to model than the posterior itself. This has to do with the fact that our models are generative or causal in nature, i.e., they describe how the evidence is produced by the underlying DA sequence U.

Estimating $P(U)$ requires building a probabilistic **discourse grammar**, i.e., a statistical model of DA sequences. This can be done using familiar techniques from language

Table 4
Summary of random variables used in dialogue modeling. (Speaker labels are introduced in Section 4.)

Symbol	Meaning
U	sequence of DA labels
E	evidence (complete speech signal)
F	prosodic evidence
A	acoustic evidence (spectral features used in ASR)
W	sequence of words
T	speakers labels

modeling for speech recognition, although the sequenced objects in this case are DA labels rather than words; discourse grammars will be discussed in detail in Section 4.

3.1 Dialogue Act Likelihoods

The computation of likelihoods $P(E|U)$ depends on the types of evidence used. In our experiments we used the following sources of evidence, either alone or in combination:

Transcribed words: The likelihoods used in Equation 1 are $P(W|U)$, where W refers to the true (hand-transcribed) words spoken in a conversation.

Recognized words: The evidence consists of recognizer acoustics A, and we seek to compute $P(A|U)$. As described later, this involves considering multiple alternative recognized word sequences.

Prosodic features: Evidence is given by the acoustic features F capturing various aspects of pitch, duration, energy, etc., of the speech signal; the associated likelihoods are $P(F|U)$.

For ease of reference, all random variables used here are summarized in Table 4. The same variables are used with subscripts to refer to individual utterances. For example, W_i is the word transcription of the ith utterance within a conversation (*not* the ith word).

To make both the modeling and the search for the best DA sequence feasible, we further require that our likelihood models are **decomposable by utterance**. This means that the likelihood given a complete conversation can be factored into likelihoods given the individual utterances. We use U_i for the ith DA label in the sequence U, i.e., $U = (U_1, \ldots, U_i, \ldots, U_n)$, where n is the number of utterances in a conversation. In addition, we use E_i for that portion of the evidence that corresponds to the ith utterance, e.g., the words or the prosody of the ith utterance. Decomposability of the likelihood means that

$$P(E|U) = P(E_1|U_1) \cdot \ldots \cdot P(E_n|U_n) \tag{2}$$

Applied separately to the three types of evidence A_i, W_i and F_i mentioned above, it is clear that this assumption is not strictly true. For example, speakers tend to reuse words found earlier in the conversation (Fowler and Housum 1987) and an answer might actually be relevant to the question before it, violating the independence of the $P(W_i|U_i)$. Similarly, speakers adjust their pitch or volume over time, e.g., to the conversation partner or because of the structure of the discourse (Menn and Boyce 1982), violating the independence of the $P(F_i|U_i)$. As in other areas of statistical modeling, we count on the fact that these violations are small compared to the properties actually modeled, namely, the dependence of E_i on U_i.

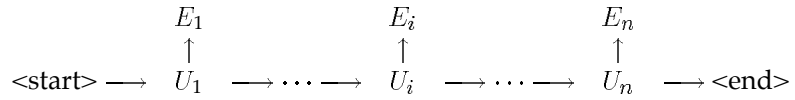

Figure 1
The discourse HMM as Bayes network.

3.2 Markov Modeling

Returning to the prior distribution of DA sequences $P(U)$, it is convenient to make certain independence assumptions here, too. In particular, we assume that the prior distribution of U is Markovian, i.e., that each U_i depends only on a fixed number k of preceding DA labels:

$$P(U_i|U_1, \ldots, U_{i-1}) = P(U_i|U_{i-k}, \ldots, U_{i-1}) \tag{3}$$

(k is the order of the Markov process describing U). The n-gram-based discourse grammars we used have this property. As described later, $k = 1$ is a very good choice, i.e., conditioning on the DA types more than one removed from the current one does not improve the quality of the model by much, at least with the amount of data available in our experiments.

The importance of the Markov assumption for the discourse grammar is that we can now view the whole system of discourse grammar and local utterance-based likelihoods as a kth-order hidden Markov model (HMM) (Rabiner and Juang 1986). The HMM states correspond to DAs, observations correspond to utterances, transition probabilities are given by the discourse grammar (see Section 4), and observation probabilities are given by the local likelihoods $P(E_i|U_i)$.

We can represent the dependency structure (as well as the implied conditional independences) as a special case of Bayesian belief network (Pearl 1988). Figure 1 shows the variables in the resulting HMM with directed edges representing conditional dependence. To keep things simple, a first-order HMM (bigram discourse grammar) is assumed.

3.3 Dialogue Act Decoding

The HMM representation allows us to use efficient dynamic programming algorithms to compute relevant aspects of the model, such as

- the most probable DA sequence (the Viterbi algorithm)
- the posterior probability of various DAs for a given utterance, after considering all the evidence (the forward-backward algorithm)

The Viterbi algorithm for HMMs (Viterbi 1967) finds the globally most probable state sequence. When applied to a discourse model with locally decomposable likelihoods and Markovian discourse grammar, it will therefore find precisely the DA sequence with the highest posterior probability:

$$U^* = \operatorname*{argmax}_U P(U|E) \tag{4}$$

The combination of likelihood and prior modeling, HMMs, and Viterbi decoding is fundamentally the same as the standard probabilistic approaches to speech recognition (Bahl, Jelinek, and Mercer 1983) and tagging (Church 1988). It maximizes the probability

of getting the *entire* DA sequence correct, but it does not necessarily find the DA sequence that has the most DA labels correct (Dermatas and Kokkinakis 1995). To minimize the total number of utterance labeling errors, we need to maximize the probability of getting each DA label correct individually, i.e., we need to maximize $P(U_i|E)$ for each $i = 1, \ldots, n$. We can compute the per-utterance posterior DA probabilities by summing:

$$P(u|E) = \sum_{U:U_i=u} P(U|E) \tag{5}$$

where the summation is over all sequences U whose ith element matches the label in question. The summation is efficiently carried out by the forward-backward algorithm for HMMs (Baum et al. 1970).[3]

For 0th-order (unigram) discourse grammars, Viterbi decoding and forward-backward decoding necessarily yield the same results. However, for higher-order discourse grammars we found that forward-backward decoding consistently gives slightly (up to 1% absolute) better accuracies, as expected. Therefore, we used this method throughout.

The formulation presented here, as well as all our experiments, uses the *entire conversation* as evidence for DA classification. Obviously, this is possible only during offline processing, when the full conversation is available. Our paradigm thus follows historical practice in the Switchboard domain, where the goal is typically the offline processing (e.g., automatic transcription, speaker identification, indexing, archival) of entire previously recorded conversations. However, the HMM formulation used here also supports computing posterior DA probabilities based on partial evidence, e.g., using only the utterances preceding the current one, as would be required for online processing.

4. Discourse Grammars

The statistical discourse grammar models the prior probabilities $P(U)$ of DA sequences. In the case of conversations for which the identities of the speakers are known (as in Switchboard), the discourse grammar should also model turn-taking behavior. A straightforward approach is to model sequences of pairs (U_i, T_i) where U_i is the DA label and T_i represents the speaker. We are not trying to model speaker idiosyncrasies, so conversants are arbitrarily identified as **A** or **B**, and the model is made symmetric with respect to the choice of sides (e.g., by replicating the training sequences with sides switched). Our discourse grammars thus had a vocabulary of $42 \times 2 = 84$ labels, plus tags for the beginning and end of conversations. For example, the second DA tag in Table 1 would be predicted by a trigram discourse grammar using the fact that the *same* speaker previously uttered a YES-NO-QUESTION, which in turn was preceded by the start-of-conversation.

4.1 N-gram Discourse Models

A computationally convenient type of discourse grammar is an n-gram model based on DA tags, as it allows efficient decoding in the HMM framework. We trained standard backoff n-gram models (Katz 1987), using the frequency smoothing approach of Witten and Bell (1991). Models of various orders were compared by their perplexities, i.e., the average number of choices the model predicts for each tag, conditioned on the preceding tags.

3 We note in passing that the Viterbi and Baum algorithms have equivalent formulations in the Bayes network framework (Pearl 1988). The HMM terminology was chosen here mainly for historical reasons.

Table 5
Perplexities of DAs with and without turn information.

Discourse Grammar	$P(U)$	$P(U,T)$	$P(U\|T)$
None	42	84	42
Unigram	11.0	18.5	9.0
Bigram	7.9	10.4	5.1
Trigram	7.5	9.8	4.8

Table 5 shows perplexities for three types of models: $P(U)$, the DAs alone; $P(U,T)$, the combined DA/speaker ID sequence; and $P(U|T)$, the DAs conditioned on known speaker IDs (appropriate for the Switchboard task). As expected, we see an improvement (decreasing perplexities) for increasing n-gram order. However, the incremental gain of a trigram is small, and higher-order models did not prove useful. (This observation, initially based on perplexity, is confirmed by the DA tagging experiments reported in Section 5.) Comparing $P(U)$ and $P(U|T)$, we see that speaker identity adds substantial information, especially for higher-order models.

The relatively small improvements from higher-order models could be a result of lack of training data, or of an inherent independence of DAs from DAs further removed. The near-optimality of the bigram discourse grammar is plausible given conversation analysis accounts of discourse structure in terms of adjacency pairs (Schegloff 1968; Sacks, Schegloff, and Jefferson 1974). Inspection of bigram probabilities estimated from our data revealed that conventional adjacency pairs receive high probabilities, as expected. For example, 30% of YES-NO-QUESTIONS are followed by YES-ANSWERS, 14% by NO-ANSWERS (confirming that the latter are dispreferred). COMMANDS are followed by AGREEMENTS in 23% of the cases, and STATEMENTS elicit BACKCHANNELS in 26% of all cases.

4.2 Other Discourse Models

We also investigated non-n-gram discourse models, based on various language modeling techniques known from speech recognition. One motivation for alternative models is that n-grams enforce a one-dimensional representation on DA sequences, whereas we saw above that the event space is really multidimensional (DA label and speaker labels). Another motivation is that n-grams fail to model long-distance dependencies, such as the fact that speakers may tend to repeat certain DAs or patterns throughout the conversation.

The first alternative approach was a standard cache model (Kuhn and de Mori 1990), which boosts the probabilities of previously observed unigrams and bigrams, on the theory that tokens tend to repeat themselves over longer distances. However, this does not seem to be true for DA sequences in our corpus, as the cache model showed no improvement over the standard n-gram. This result is somewhat surprising since unigram dialogue grammars are able to detect speaker gender with 63% accuracy (over a 50% baseline) on Switchboard (Ries 1999b), indicating that there *are* global variables in the DA distribution that could potentially be exploited by a cache dialogue grammar. Clearly, dialogue grammar adaptation needs further research.

Second, we built a discourse grammar that incorporated constraints on DA sequences in a nonhierarchical way, using maximum entropy (ME) estimation (Berger, Della Pietra, and Della Pietra 1996). The choice of features was informed by similar ones commonly used in statistical language models, as well our general intuitions about potentially information-bearing elements in the discourse context. Thus, the model was designed so that the current DA label was constrained by features such as unigram statistics, the previous DA and the DA once removed, DAs occurring within a window

in the past, and whether the previous utterance was by the same speaker. We found, however, that an ME model using n-gram constraints performed only slightly better than a corresponding backoff n-gram.

Additional constraints such as DA triggers, distance-1 bigrams, separate encoding of speaker change and bigrams to the last DA on the same/other channel did not improve relative to the trigram model. The ME model thus confirms the adequacy of the backoff n-gram approach, and leads us to conclude that DA sequences, at least in the Switchboard domain, are mostly characterized by local interactions, and thus modeled well by low-order n-gram statistics for this task. For more structured tasks this situation might be different. However, we have found no further exploitable structure.

5. Dialogue Act Classification

We now describe in more detail how the knowledge sources of words and prosody are modeled, and what automatic DA labeling results were obtained using each of the knowledge sources in turn. Finally, we present results for a combination of all knowledge sources. DA labeling accuracy results should be compared to a baseline (chance) accuracy of 35%, the relative frequency of the most frequent DA type (STATEMENT) in our test set.[4]

5.1 Dialogue Act Classification Using Words

DA classification using words is based on the observation that different DAs use distinctive word strings. It is known that certain cue words and phrases (Hirschberg and Litman 1993) can serve as explicit indicators of discourse structure. Similarly, we find distinctive correlations between certain phrases and DA types. For example, 92.4% of the *uh-huh*'s occur in BACKCHANNELS, and 88.4% of the trigrams "<start> *do you*" occur in YES-NO-QUESTIONS. To leverage this information source, without hand-coding knowledge about which words are indicative of which DAs, we will use statistical language models that model the full word sequences associated with each DA type.

5.1.1 Classification from True Words. Assuming that the true (hand-transcribed) words of utterances are given as evidence, we can compute word-based likelihoods $P(W|U)$ in a straightforward way, by building a statistical language model for each of the 42 DAs. All DAs of a particular type found in the training corpus were pooled, and a DA-specific trigram model was estimated using standard techniques (Katz-backoff [Katz 1987] with Witten-Bell discounting [Witten and Bell 1991]).

5.1.2 Classification from Recognized Words. For fully automatic DA classification, the above approach is only a partial solution, since we are not yet able to recognize words in spontaneous speech with perfect accuracy. A standard approach is to use the 1-best hypothesis from the speech recognizer in place of the true word transcripts. While conceptually simple and convenient, this method will not make optimal use of all the information in the recognizer, which in fact maintains multiple hypotheses as well as their relative plausibilities.

A more thorough use of recognized speech can be derived as follows. The classification framework is modified such that the recognizer's acoustic information (spectral features) A appear as the evidence. We compute $P(A|U)$ by decomposing it into an acoustic likelihood $P(A|W)$ and a word-based likelihood $P(W|U)$, and summing over

4 The frequency of STATEMENTS across all labeled data was slightly different, cf. Table 2.

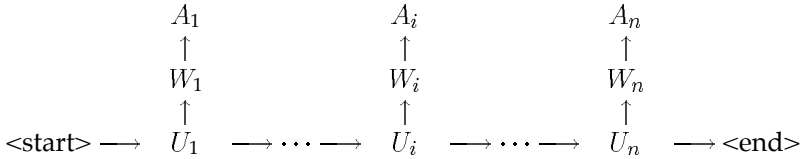

Figure 2
Modified Bayes network including word hypotheses and recognizer acoustics.

all word sequences:

$$\begin{aligned} P(A|U) &= \sum_W P(A|W,U)P(W|U) \\ &= \sum_W P(A|W)P(W|U) \end{aligned} \quad (6)$$

The second line is justified under the assumption that the recognizer acoustics (typically, cepstral coefficients) are invariant to DA type once the words are fixed. Note that this is another approximation in our modeling. For example, different DAs with common words may be realized by different word pronunciations. Figure 2 shows the Bayes network resulting from modeling recognizer acoustics through word hypotheses under this independence assumption; note the added W_i variables (that have to be summed over) in comparison to Figure 1.

The acoustic likelihoods $P(A|W)$ correspond to the acoustic scores the recognizer outputs for every hypothesized word sequence W. The summation over all W must be approximated; in our experiments we summed over the (up to) 2,500 best hypotheses generated by the recognizer for each utterance. Care must be taken to scale the recognizer acoustic scores properly, i.e., to exponentiate the recognizer acoustic scores by $1/\lambda$, where λ is the language model weight of the recognizer.[5]

5.1.3 Results. Table 6 shows DA classification accuracies obtained by combining the word- and recognizer-based likelihoods with the n-gram discourse grammars described earlier. The best accuracy obtained from transcribed words, 71%, is encouraging given a comparable human performance of 84% (the interlabeler agreement, see Section 2.2). We observe about a 21% relative increase in classification error when using recognizer words; this is remarkably small considering that the speech recognizer used had a word error rate of 41% on the test set.

5 In a standard recognizer the total log score of a hypothesis W_i is computed as

$$\log P(A_i|W_i) + \lambda \log P(W_i) - \mu|W_i| \quad ,$$

where $|W_i|$ is the number of words in the hypothesis, and both λ and μ are parameters optimized to minimize the word error rate. The word insertion penalty μ represents a correction to the language model that allows balancing insertion and deletion errors. The language model weight λ compensates for acoustic scores variances that are effectively too large due to severe independence assumptions in the recognizer acoustic model. According to this rationale, it is more appropriate to *divide* all score components by λ. Thus, in all our experiments, we computed a summand in Equation 6 whose logarithm was

$$\frac{1}{\lambda} \log P(A_i|W_i) + \log P(W_i|U_i) - \frac{\mu}{\lambda}|W_i| \quad .$$

We found this approach to give better results than the standard multiplication of $\log P(W)$ by λ. Note that for selecting the best hypothesis in a recognizer only the relative magnitudes of the score weights matter; however, for the summation in Equation 6 the absolute values become important. The parameter values for λ and μ were those used by the standard recognizer; they were not specifically optimized for the DA classification task.

Table 6
DA classification accuracies (in %) from transcribed and recognized words (chance = 35%).

Discourse Grammar	True	Recognized	Relative Error Increase
None	54.3	42.8	25.2%
Unigram	68.2	61.8	20.1%
Bigram	70.6	64.3	21.4%
Trigram	71.0	64.8	21.4%

We also compared the n-best DA classification approach to the more straightforward 1-best approach. In this experiment, only the single best recognizer hypothesis is used, effectively treating it as the true word string. The 1-best method increased classification error by about 7% relative to the n-best algorithm (61.5% accuracy with a bigram discourse grammar).

5.2 Dialogue Act Classification Using Prosody

We also investigated prosodic information, i.e., information independent of the words as well as the standard recognizer acoustics. Prosody is important for DA recognition for two reasons. First, as we saw earlier, word-based classification suffers from recognition errors. Second, some utterances are inherently ambiguous based on words alone. For example, some YES-NO-QUESTIONS have word sequences identical to those of STATEMENTS, but can often be distinguished by their final F0 rise.

A detailed study aimed at automatic prosodic classification of DAs in the Switchboard domain is available in a companion paper (Shriberg et al. 1998). Here we investigate the interaction of prosodic models with the dialogue grammar and the word-based DA models discussed above. We also touch briefly on alternative machine learning models for prosodic features.

5.2.1 Prosodic Features. Prosodic DA classification was based on a large set of features computed automatically from the waveform, without reference to word or phone information. The features can be broadly grouped as referring to duration (e.g., utterance duration, with and without pauses), pauses (e.g., total and mean of nonspeech regions exceeding 100 ms), pitch (e.g., mean and range of F0 over utterance, slope of F0 regression line), energy (e.g., mean and range of RMS energy, same for signal-to-noise ratio [SNR]), speaking rate (based on the "enrate" measure of Morgan, Fosler, and Mirghafori [1997]), and gender (of both speaker and listener). In the case of utterance duration, the measure correlates both with length in words and with overall speaking rate. The gender feature that classified speakers as either male or female was used to test for potential inadequacies in F0 normalizations. Where appropriate, we included both raw features and values normalized by utterance and/or conversation. We also included features that are the output of the pitch accent and boundary tone event detector of Taylor (2000) (e.g., the number of pitch accents in the utterance). A complete description of prosodic features and an analysis of their usage in our models can be found in Shriberg et al. (1998).

5.2.2 Prosodic Decision Trees. For our prosodic classifiers, we used CART-style decision trees (Breiman et al. 1984). Decision trees allow combination of discrete and continuous features, and can be inspected to help in understanding the role of different features and feature combinations.

To illustrate one area in which prosody could aid our classification task, we applied trees to DA classifications known to be ambiguous from words alone. One frequent example in our corpus was the distinction between BACKCHANNELS and AGREEMENTS

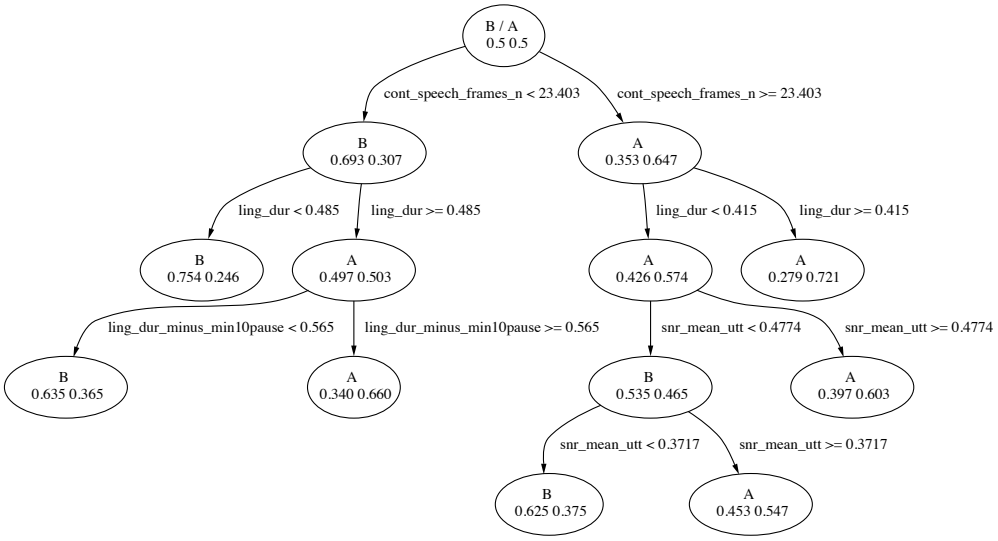

Figure 3
Decision tree for the classification of BACKCHANNELS (B) and AGREEMENTS (A). Each node is labeled with the majority class for that node, as well as the posterior probabilities of the two classes. The following features are queried in the tree: number of frames in continuous (> 1 s) speech regions (*cont_speech_frames*), total utterance duration (*ling_dir*), utterance duration excluding pauses > 100 ms (*ling_dur_minus_min10pause*), and mean signal-to-noise ratio (*snr_mean_utt*).

(see Table 2), which share terms such as *right* and *yeah*. As shown in Figure 3, a prosodic tree trained on this task revealed that agreements have consistently longer durations and greater energy (as reflected by the SNR measure) than do backchannels.

The HMM framework requires that we compute prosodic likelihoods of the form $P(F_i|U_i)$ for each utterance U_i and associated prosodic feature values F_i. We have the apparent difficulty that decision trees (as well as other classifiers, such as neural networks) give estimates for the posterior probabilities, $P(U_i|F_i)$. The problem can be overcome by applying Bayes' Rule locally:

$$P(F_i|U_i) = P(F_i)\frac{P(U_i|F_i)}{P(U_i)} \propto \frac{P(U_i|F_i)}{P(U_i)} \tag{7}$$

Note that $P(F_i)$ does not depend on U_i and can be treated as a constant for the purpose of DA classification. A quantity proportional to the required likelihood can therefore be obtained either by dividing the posterior tree probability by the prior $P(U_i)$,[6] or by training the tree on a uniform prior distribution of DA types. We chose the second approach, downsampling our training data to equate DA proportions. This also counteracts a common problem with tree classifiers trained on very skewed distributions of target classes, i.e., that low-frequency classes are not modeled in sufficient detail because the majority class dominates the tree-growing objective function.

5.2.3 Results with Decision Trees. As a preliminary experiment to test the integration of prosody with other knowledge sources, we trained a single tree to discriminate among

6 Bourlard and Morgan (1993) use this approach to integrate neural network phonetic models in a speech recognizer.

Table 7
DA classification using prosodic decision trees (chance = 35%).

Discourse Grammar	Accuracy (%)
None	38.9
Unigram	48.3
Bigram	49.7

Table 8
Performance of various prosodic neural network classifiers on an equal-priors, six-class DA set (chance = 16.6%).

Network Architecture	Accuracy (%)
Decision tree	45.4
No hidden layer, linear output function	44.6
No hidden layer, softmax output function	46.0
40-unit hidden layer, softmax output function	46.0

the five most frequent DA types (STATEMENT, BACKCHANNEL, OPINION, ABANDONED, and AGREEMENT, totaling 79% of the data) and an Other category comprising all remaining DA types. The decision tree was trained on a downsampled training subset containing equal proportions of these six DA classes. The tree achieved a classification accuracy of 45.4% on an independent test set with the same uniform six-class distribution. The chance accuracy on this set is 16.6%, so the tree clearly extracts useful information from the prosodic features.

We then used the decision tree posteriors as scaled DA likelihoods in the dialogue model HMM, combining it with various n-gram dialogue grammars for testing on our full standard test set. For the purpose of model integration, the likelihoods of the Other class were assigned to all DA types comprised by that class. As shown in Table 7, the tree with dialogue grammar performs significantly better than chance on the raw DA distribution, although not as well as the word-based methods (cf. Table 6).

5.2.4 Neural Network Classifiers. Although we chose to use decision trees as prosodic classifiers for their relative ease of inspection, we might have used any suitable probabilistic classifier, i.e., any model that estimates the posterior probabilities of DAs given the prosodic features. We conducted preliminary experiments to assess how neural networks compare to decision trees for the type of data studied here. Neural networks are worth investigating since they offer potential advantages over decision trees. They can learn decision surfaces that lie at an angle to the axes of the input feature space, unlike standard CART trees, which always split continuous features on one dimension at a time. The response function of neural networks is continuous (smooth) at the decision boundaries, allowing them to avoid hard decisions and the complete fragmentation of data associated with decision tree questions.

Most important, however, related work (Ries 1999a) indicated that similarly structured networks are superior classifiers if the input features are words and are therefore a plug-in replacement for the language model classifiers described in this paper. Neural networks are therefore a good candidate for a jointly optimized classifier of prosodic and word-level information since one can show that they are a generalization of the integration approach used here.

We tested various neural network models on the same six-class downsampled data as used for decision tree training, using a variety of network architectures and output layer functions. The results are summarized in Table 8, along with the baseline result obtained with the decision tree model. Based on these experiments, a softmax network

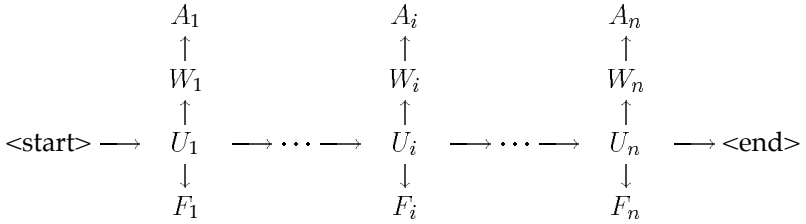

Figure 4
Bayes network for discourse HMM incorporating both word recognition and prosodic features.

(Bridle 1990) without hidden units resulted in only a slight improvement over the decision tree. A network with hidden units did not afford any additional advantage, even after we optimized the number of hidden units, indicating that complex combinations of features (as far as the network could learn them) do not predict DAs better than linear combinations of input features. While we believe alternative classifier architectures should be investigated further as prosodic models, the results so far seem to confirm our choice of decision trees as a model class that gives close to optimal performance for this task.

5.2.5 Intonation Event Likelihoods. An alternative way to compute prosodically based DA likelihoods uses pitch accents and boundary phrases (Taylor et al. 1997). The approach relies on the intuition that different utterance types are characterized by different intonational "tunes" (Kowtko 1996), and has been successfully applied to the classification of move types in the DCIEM Map Task corpus (Wright and Taylor 1997). The system detects sequences of distinctive pitch patterns by training one continuous-density HMM for each DA type. Unfortunately, the event classification accuracy on the Switchboard corpus was considerably poorer than in the Map Task domain, and DA recognition results when coupled with a discourse grammar were substantially worse than with decision trees. The approach could prove valuable in the future, however, if the intonation event detector can be made more robust to corpora like ours.

5.3 Using Multiple Knowledge Sources

As mentioned earlier, we expect improved performance from combining word and prosodic information. Combining these knowledge sources requires estimating a combined likelihood $P(A_i, F_i|U_i)$ for each utterance. The simplest approach is to assume that the two types of acoustic observations (recognizer acoustics and prosodic features) are approximately conditionally independent once U_i is given:

$$\begin{aligned} P(A_i, W_i, F_i|U_i) &= P(A_i, W_i|U_i)P(F_i|A_i, W_i, U_i) \\ &\approx P(A_i, W_i|U_i)P(F_i|U_i) \end{aligned} \quad (8)$$

Since the recognizer acoustics are modeled by way of their dependence on words, it is particularly important to avoid using prosodic features that are directly correlated with word identities, or features that are also modeled by the discourse grammars, such as utterance position relative to turn changes. Figure 4 depicts the Bayes network incorporating evidence from both word recognition and prosodic features.

One important respect in which the independence assumption is violated is in the modeling of utterance length. While utterance length itself is not a prosodic feature, it is an important feature to condition on when examining prosodic characteristics of utterances, and is thus best included in the decision tree. Utterance length is captured

Table 9
Combined utterance classification accuracies (chance = 35%). The first two columns correspond to Tables 7 and 6, respectively.

Discourse Grammar	Accuracy (%)		
	Prosody	Recognizer	Combined
None	38.9	42.8	56.5
Unigram	48.3	61.8	62.4
Bigram	49.7	64.3	65.0

directly by the tree using various duration measures, while the DA-specific LMs encode the average number of words per utterance indirectly through n-gram parameters, but still accurately enough to violate independence in a significant way (Finke et al. 1998). As discussed in Section 8, this problem is best addressed by joint lexical-prosodic models.

We need to allow for the fact that the models combined in Equation 8 give estimates of differing qualities. Therefore, we introduce an exponential weight α on $P(F_i|U_i)$ that controls the contribution of the prosodic likelihood to the overall likelihood. Finally, a second exponential weight β on the combined likelihood controls its dynamic range relative to the discourse grammar scores, partially compensating for any correlation between the two likelihoods. The revised combined likelihood estimate thus becomes

$$P(A_i, W_i, F_i|U_i) \approx \{P(A_i, W_i|U_i)P(F_i|U_i)^\alpha\}^\beta \tag{9}$$

In our experiments, the parameters α and β were optimized using twofold jackknifing. The test data was split roughly in half (without speaker overlap), each half was used to separately optimize the parameters, and the best values were then tested on the respective other half. The reported results are from the aggregate outcome on the two test set halves.

5.3.1 Results. In this experiment we combined the acoustic n-best likelihoods based on recognized words with the Top-5 tree classifier mentioned in Section 5.2.3. Results are summarized in Table 9.

As shown, the combined classifier presents a slight improvement over the recognizer-based classifier. The experiment without discourse grammar indicates that the combined evidence is considerably stronger than either knowledge source alone, yet this improvement seems to be made largely redundant by the use of priors and the discourse grammar. For example, by definition DECLARATIVE-QUESTIONS are not marked by syntax (e.g., by subject-auxiliary inversion) and are thus confusable with STATEMENTS and OPINIONS. While prosody is expected to help disambiguate these cases, the ambiguity can also be removed by examining the context of the utterance, e.g., by noticing that the following utterance is a YES-ANSWER or NO-ANSWER.

5.3.2 Focused Classifications. To gain a better understanding of the potential for prosodic DA classification independent of the effects of discourse grammar and the skewed DA distribution in Switchboard, we examined several binary DA classification tasks. The choice of tasks was motivated by an analysis of confusions committed by a purely word-based DA detector, which tends to mistake QUESTIONS for STATEMENTS, and BACKCHANNELS for AGREEMENTS (and vice versa). We tested a prosodic classifier, a word-based classifier (with both transcribed and recognized words), and a combined classifier on these two tasks, downsampling the DA distribution to equate the class sizes in each case. Chance performance in all experiments is therefore 50%. Results are summarized in Table 10.

Table 10
Accuracy (in %) for individual and combined models for two subtasks, using uniform priors (chance = 50%).

Classification Task Knowledge Source	True Words	Recognized Words
QUESTIONS/STATEMENTS		
prosody only	76.0	76.0
words only	85.9	75.4
words+prosody	87.6	79.8
AGREEMENTS/BACKCHANNELS		
prosody only	72.9	72.9
words only	81.0	78.2
words+prosody	84.7	81.7

As shown, the combined classifier was consistently more accurate than the classifier using words alone. Although the gain in accuracy was not statistically significant for the small recognizer test set because of a lack of power, replication for a larger hand-transcribed test set showed the gain to be highly significant for both subtasks by a Sign test, $p < .001$ and $p < .0001$ (one-tailed), respectively. Across these, as well as additional subtasks, the relative advantage of adding prosody was larger for recognized than for true words, suggesting that prosody is particularly helpful when word information is not perfect.

6. Speech Recognition

We now consider ways to use DA modeling to enhance automatic speech recognition (ASR). The intuition behind this approach is that discourse context constrains the choice of DAs for a given utterance, and the DA type in turn constrains the choice of words. The latter can then be leveraged for more accurate speech recognition.

6.1 Integrating DA Modeling and ASR

Constraints on the word sequences hypothesized by a recognizer are expressed probabilistically in the recognizer language model (LM). It provides the prior distribution $P(W_i)$ for finding the a posteriori most probable hypothesized words for an utterance, given the acoustic evidence A_i (Bahl, Jelinek, and Mercer 1983):[7]

$$\begin{aligned} W_i^* &= \operatorname*{argmax}_{W_i} P(W_i|A_i) \\ &= \operatorname*{argmax}_{W_i} \frac{P(W_i)P(A_i|W_i)}{P(A_i)} \\ &= \operatorname*{argmax}_{W_i} P(W_i)P(A_i|W_i) \end{aligned} \tag{10}$$

The likelihoods $P(A_i|W_i)$ are estimated by the recognizer's acoustic model. In a standard recognizer the language model $P(W_i)$ is the same for all utterances; the idea here is to obtain better-quality LMs by conditioning on the DA type U_i, since presumably the word

7 Note the similarity of Equations 10 and 1. They are identical except for the fact that we are now operating at the level of an individual utterance, the evidence is given by the acoustics, and the targets are word hypotheses instead of DA hypotheses.

distributions differ depending on DA type.

$$\begin{aligned} W_i^* &= \operatorname*{argmax}_{W_i} P(W_i|A_i, U_i) \\ &= \operatorname*{argmax}_{W_i} \frac{P(W_i|U_i)P(A_i|W_i, U_i)}{P(A_i|U_i)} \\ &\approx \operatorname*{argmax}_{W_i} P(W_i|U_i)P(A_i|W_i) \end{aligned} \tag{11}$$

As before in the DA classification model, we tacitly assume that the words W_i depend only on the DA of the current utterance, and also that the acoustics are independent of the DA type if the words are fixed. The DA-conditioned language models $P(W_i|U_i)$ are readily trained from DA-specific training data, much like we did for DA classification from words.[8]

The problem with applying Equation 11, of course, is that the DA type U_i is generally not known (except maybe in applications where the user interface can be engineered to allow only one kind of DA for a given utterance). Therefore, we need to infer the likely DA types for each utterance, using available evidence E from the entire conversation. This leads to the following formulation:

$$\begin{aligned} W_i^* &= \operatorname*{argmax}_{W_i} P(W_i|A_i, E) \\ &= \operatorname*{argmax}_{W_i} \sum_{U_i} P(W_i|A_i, U_i, E)P(U_i|E) \\ &\approx \operatorname*{argmax}_{W_i} \sum_{U_i} P(W_i|A_i, U_i)P(U_i|E) \end{aligned} \tag{12}$$

The last step in Equation 12 is justified because, as shown in Figures 1 and 4, the evidence E (acoustics, prosody, words) pertaining to utterances other than i can affect the current utterance only through its DA type U_i.

We call this the **mixture-of-posteriors** approach, because it amounts to a mixture of the posterior distributions obtained from DA-specific speech recognizers (Equation 11), using the DA posteriors as weights. This approach is quite expensive, however, as it requires multiple full recognizer or rescoring passes of the input, one for each DA type.

A more efficient, though mathematically less accurate, solution can be obtained by combining guesses about the correct DA types directly at the level of the LM. We estimate the distribution of likely DA types for a given utterance using the entire conversation E as evidence, and then use a sentence-level mixture (Iyer, Ostendorf, and Rohlicek 1994) of DA-specific LMs in a *single* recognizer run. In other words, we replace $P(W_i|U_i)$ in Equation 11 with

$$\sum_{U_i} P(W_i|U_i)P(U_i|E) \quad ,$$

a weighted mixture of all DA-specific LMs. We call this the **mixture-of-LMs** approach. In practice, we would first estimate DA posteriors for each utterance, using the forward-backward algorithm and the models described in Section 5, and then rerecognize the conversation or rescore the recognizer output, using the new posterior-weighted mixture LM. Fortunately, as shown in the next section, the mixture-of-LMs approach seems to give results that are almost identical (and as good) the mixture-of-posteriors approach.

8 In Equation 11 and elsewhere in this section we gloss over the issue of proper weighting of model probabilities, which is extremely important in practice. The approach explained in detail in footnote 5 applies here as well.

6.2 Computational Structure of Mixture Modeling

It is instructive to compare the expanded scoring formulae for the two DA mixture modeling approaches for ASR. The mixture-of-posteriors approach yields

$$P(W_i|A_i, E) = \sum_{U_i} \frac{P(W_i|U_i)P(A_i|W_i)}{P(A_i|U_i)} P(U_i|E) \quad , \tag{13}$$

whereas the mixture-of-LMs approach gives

$$P(W_i|A_i, E) \approx \left(\sum_{U_i} P(W_i|U_i)P(U_i|E) \right) \frac{P(A_i|W_i)}{P(A_i)} \quad . \tag{14}$$

We see that the second equation reduces to the first under the crude approximation $P(A_i|U_i) \approx P(A_i)$. In practice, the denominators are computed by summing the numerators over a finite number of word hypotheses W_i, so this difference translates into normalizing either after or before summing over DAs. When the normalization takes place as the final step it can by omitted for score maximization purposes; this shows why the mixture-of-LMs approach is less computationally expensive.

6.3 Experiments and Results

We tested both the mixture-of-posteriors and the mixture-of-LMs approaches on our Switchboard test set of 19 conversations. Instead of decoding the data from scratch using the modified models, we manipulated n-best lists consisting of up to 2,500 best hypotheses for each utterance. This approach is also convenient since both approaches require access to the full word string for hypothesis scoring; the overall model is no longer Markovian, and is therefore inconvenient to use in the first decoding stage, or even in lattice rescoring.

The baseline for our experiments was obtained with a standard backoff trigram language model estimated from all available training data. The DA-specific language models were trained on word transcripts of all the training utterances of a given type, and then smoothed further by interpolating them with the baseline LM. Each DA-specific LM used its own interpolation weight, obtained by minimizing the perplexity of the interpolated model on held-out DA-specific training data. Note that this smoothing step is helpful when using the DA-specific LMs for *word* recognition, but not for DA classification, since it renders the DA-specific LMs less discriminative.[9]

Table 11 summarizes both the word error rates achieved with the various models and the perplexities of the corresponding LMs used in the rescoring (note that perplexity is not meaningful in the mixture-of-posteriors approach). For comparison, we also included two additional models: the "1-best LM" refers to always using the DA-specific LM corresponding to the most probable DA type for each utterance. It is thus an approximation to both mixture approaches where only the top DA is considered. Second, we included an "oracle LM," i.e., always using the LM that corresponds to the *hand-labeled* DA for each utterance. The purpose of this experiment was to give us an upper bound on the effectiveness of the mixture approaches, by assuming perfect DA recognition.

It was somewhat disappointing that the word error rate (WER) improvement in the oracle experiment was small (2.2% relative), even though statistically highly significant ($p < .0001$, one-tailed, according to a Sign test on matched utterance pairs). The

9 Indeed, during our DA classification experiments, we had observed that smoothed DA-specific LMs yield lower classification accuracy.

Table 11
Switchboard word recognition error rates and LM perplexities.

Model	WER (%)	Perplexity
Baseline	41.2	76.8
1-best LM	41.0	69.3
Mixture-of-posteriors	41.0	n/a
Mixture-of-LMs	40.9	66.9
Oracle LM	40.3	66.8

Table 12
Word error reductions through DA oracle, by DA type.

Dialogue Act	Baseline WER	Oracle WER	WER Reduction
NO-ANSWER	29.4	11.8	-17.6
BACKCHANNEL	25.9	18.6	-7.3
BACKCHANNEL-QUESTION	15.2	9.1	-6.1
ABANDONED/UNINTERPRETABLE	48.9	45.2	-3.7
WH-QUESTION	38.4	34.9	-3.5
YES-NO-QUESTION	55.5	52.3	-3.2
STATEMENT	42.0	41.5	-0.5
OPINION	40.8	40.4	-0.4

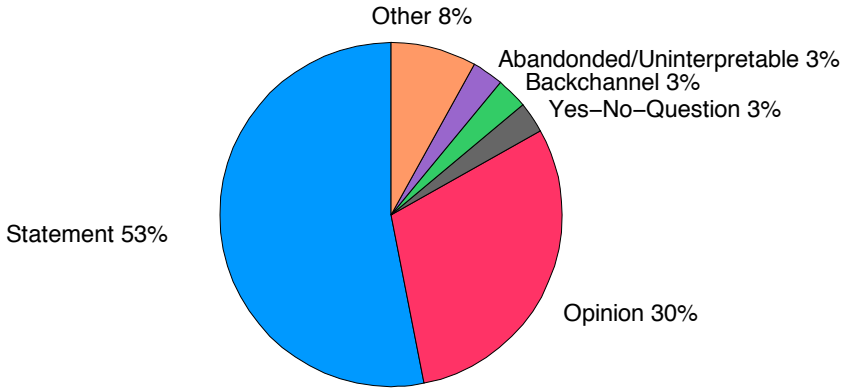

Figure 5
Relative contributions to test set word counts by DA type.

WER reduction achieved with the mixture-of-LMs approach did not achieve statistical significance ($0.25 > p > 0.20$). The 1-best DA and the two mixture models also did not differ significantly on this test set. In interpreting these results one must realize, however, that WER results depend on a complex combination of factors, most notably interaction between language models and the acoustic models. Since the experiments only varied the language models used in rescoring, it is also informative to compare the quality of these models as reflected by perplexity. On this measure, we see a substantial 13% (relative) reduction, which is achieved by both the oracle and the mixture-of-LMs. The perplexity reduction for the 1-best LM is only 9.8%, showing the advantage of the mixture approach.

To better understand the lack of a more substantial reduction in word error, we analyzed the effect of the DA-conditioned rescoring on the individual DAs, i.e., grouping the test utterances by their true DA types. Table 12 shows the WER improvements for a few DA types, ordered by the magnitude of improvement achieved. As shown, all frequent DA types saw improvement, but the highest wins were observed for typically short DAs, such as ANSWERS and BACKCHANNELS. This is to be expected, as such DAs tend to be syntactically and lexically highly constrained. Furthermore, the distribution of

number of words across DA types is very uneven (Figure 5). STATEMENTS and OPINIONS, the DA types dominating in both frequency and number of words (83% of total), see no more than 0.5% absolute improvement, thus explaining the small overall improvement. In hindsight, this is also not surprising, since the bulk of the training data for the baseline LM consists of these DAs, allowing only little improvement in the DA-specific LMs. A more detailed analysis of the effect of DA modeling on speech recognition errors can be found elsewhere (Van Ess-Dykema and Ries 1998).

In summary, our experiments confirmed that DA modeling can improve word recognition accuracy quite substantially in principle, at least for certain DA types, but that the skewed distribution of DAs (especially in terms of number of words per type) limits the usefulness of the approach on the Switchboard corpus. The benefits of DA modeling might therefore be more pronounced on corpora with more even DA distribution, as is typically the case for task-oriented dialogues. Task-oriented dialogues might also feature specific subtypes of general DA categories that might be constrained by discourse. Prior research on task-oriented dialogues summarized in the next section, however, has also found only small reductions in WER (on the order of 1%). This suggests that even in task-oriented domains more research is needed to realize the potential of DA modeling for ASR.

7. Prior and Related Work

As indicated in the introduction, our work builds on a number of previous efforts in computational discourse modeling and automatic discourse processing, most of which occurred over the last half-decade. It is generally not possible to directly compare quantitative results because of vast differences in methodology, tag set, type and amount of training data, and, principally, assumptions made about what information is available for "free" (e.g., hand-transcribed versus automatically recognized words, or segmented versus unsegmented utterances). Thus, we will focus on the conceptual aspects of previous research efforts, and while we do offer a summary of previous quantitative results, these should be interpreted as informative datapoints only, and not as fair comparisons between algorithms.

Previous research on DA modeling has generally focused on task-oriented dialogue, with three tasks in particular garnering much of the research effort. The Map Task corpus (Anderson et al. 1991; Bard et al. 1995) consists of conversations between two speakers with slightly different maps of an imaginary territory. Their task is to help one speaker reproduce a route drawn only on the other speaker's map, all without being able to see each other's maps. Of the DA modeling algorithms described below, Taylor et al. (1998) and Wright (1998) were based on Map Task. The VERBMOBIL corpus consists of two-party scheduling dialogues. A number of the DA modeling algorithms described below were developed for VERBMOBIL, including those of Mast et al. (1996), Warnke et al. (1997), Reithinger et al. (1996), Reithinger and Klesen (1997), and Samuel, Carberry, and Vijay-Shanker (1998). The ATR Conference corpus is a subset of a larger ATR Dialogue database consisting of simulated dialogues between a secretary and a questioner at international conferences. Researchers using this corpus include Nagata (1992), Nagata and Morimoto (1993, 1994), and Kita et al. (1996). Table 13 shows the most commonly used versions of the tag sets from those three tasks.

As discussed earlier, these domains differ from the Switchboard corpus in being task-oriented. Their tag sets are also generally smaller, but some of the same problems of balance occur. For example, in the Map Task domain, 33% of the words occur in 1 of the 12 DAs (INSTRUCT). Table 14 shows the approximate size of the corpora, the tag set, and tag estimation accuracy rates for various recent models of DA prediction. The

results summarized in the table also illustrate the differences in inherent difficulty of the tasks. For example, the task of Warnke et al. (1997) was to simultaneously segment and tag DAs, whereas the other results rely on a prior manual segmentation. Similarly, the task in Wright (1998) and in our study was to determine DA types from speech input, whereas work by others is based on hand-transcribed textual input.

The use of n-grams to model the probabilities of DA sequences, or to predict upcoming DAs online, has been proposed by many authors. It seems to have been first employed by Nagata (1992), and in follow-up papers by Nagata and Morimoto (1993, 1994) on the ATR Dialogue database. The model predicted upcoming DAs by using bigrams and trigrams conditioned on preceding DAs, trained on a corpus of 2,722 DAs. Many others subsequently relied on and enhanced this n-grams-of-DAs approach, often by applying standard techniques from statistical language modeling. Reithinger et al. (1996), for example, used deleted interpolation to smooth the dialogue n-grams. Chu-Carroll (1998) uses knowledge of subdialogue structure to selectively skip previous DAs in choosing conditioning for DA prediction.

Nagata and Morimoto (1993, 1994) may also have been the first to use word n-grams as a miniature grammar for DAs, to be used in improving speech recognition. The idea caught on very quickly: Suhm and Waibel (1994), Mast et al. (1996), Warnke et al. (1997), Reithinger and Klesen (1997), and Taylor et al. (1998) all use variants of backoff, interpolated, or class n-gram language models to estimate DA likelihoods. Any kind of sufficiently powerful, trainable language model could perform this function, of course, and indeed Alexandersson and Reithinger (1997) propose using automatically learned stochastic context-free grammars. Jurafsky, Shriberg, Fox, and Curl (1998) show that the grammar of some DAs, such as appreciations, can be captured by finite-state automata over part-of-speech tags.

N-gram models are *likelihood* models for DAs, i.e., they compute the conditional probabilities of the word sequence given the DA type. Word-based *posterior probability* estimators are also possible, although less common. Mast et al. (1996) propose the use of semantic classification trees, a kind of decision tree conditioned on word patterns as features. Finally, Ries (1999a) shows that neural networks using only unigram features can be superior to higher-order n-gram DA models. Warnke et al. (1999) and Ohler, Harbeck, and Niemann (1999) use related discriminative training algorithms for language models.

Woszczyna and Waibel (1994) and Suhm and Waibel (1994), followed by Chu-Carroll (1998), seem to have been the first to note that such a combination of word and dialogue n-grams could be viewed as a dialogue HMM with word strings as the observations. (Indeed, with the exception of Samuel, Carberry, and Vijay-Shanker (1998), all models listed in Table 14 rely on some version of this HMM metaphor.) Some researchers explicitly used HMM induction techniques to infer dialogue grammars. Woszczyna and Waibel (1994), for example, trained an ergodic HMM using expectation-maximization to model speech act sequencing. Kita et al. (1996) made one of the few attempts at *unsupervised* discovery of dialogue structure, where a finite-state grammar induction algorithm is used to find the topology of the dialogue grammar.

Computational approaches to prosodic modeling of DAs have aimed to automatically extract various prosodic parameters—such as duration, pitch, and energy patterns—from the speech signal (Yoshimura et al. [1996]; Taylor et al. [1997]; Kompe [1997], among others). Some approaches model F0 patterns with techniques such as vector quantization and Gaussian classifiers to help disambiguate utterance types. An extensive comparison of the prosodic DA modeling literature with our work can be found in Shriberg et al. (1998).

DA modeling has mostly been geared toward automatic DA classification, and

much less work has been done on applying DA models to automatic speech recognition. Nagata and Morimoto (1994) suggest conditioning word language models on DAs to lower perplexity. Suhm and Waibel (1994) and Eckert, Gallwitz, and Niemann (1996) each condition a recognizer LM on left-to-right DA predictions and are able to show reductions in word error rate of 1% on task-oriented corpora. Most similar to our own work, but still in a task-oriented domain, the work by Taylor et al. (1998) combines DA likelihoods from prosodic models with those from 1-best recognition output to condition the recognizer LM, again achieving an absolute reduction in word error rate of 1%, similarly disappointing as the 0.3% improvement in our experiments.

Related computational tasks beyond DA classification and speech recognition have received even less attention to date. We already mentioned Warnke et al. (1997) and Finke et al. (1998), who both showed that utterance segmentation and classification can be integrated into a single search process. Fukada et al. (1998) investigate augmenting DA tagging with more detailed semantic "concept" tags, as a preliminary step toward an interlingua-based dialogue translation system. Levin et al. (1999) couple DA classification with dialogue game classification; dialogue games are units above the DA level, i.e., short DA sequences such as question-answer pairs.

All the work mentioned so far uses statistical models of various kinds. As we have shown here, such models offer some fundamental advantages, such as modularity and composability (e.g., of discourse grammars with DA models) and the ability to deal with noisy input (e.g., from a speech recognizer) in a principled way. However, many other classifier architectures are applicable to the tasks discussed, in particular to DA classification. A nonprobabilistic approach for DA labeling proposed by Samuel, Carberry, and Vijay-Shanker (1998) is transformation-based learning (Brill 1993). Finally it should be noted that there are other tasks with a mathematical structure similar to that of DA tagging, such as shallow parsing for natural language processing (Munk 1999) and DNA classification tasks (Ohler, Harbeck, and Niemann 1999), from which further techniques could be borrowed.

How does the approach presented here differ from these various earlier models, particularly those based on HMMs? Apart from corpus and tag set differences, our approach differs primarily in that it generalizes the simple HMM approach to cope with new kinds of problems, based on the Bayes network representations depicted in Figures 2 and 4. For the DA classification task, our framework allows us to do classification given unreliable words (by marginalizing over the possible word strings corresponding to the acoustic input) and given nonlexical (e.g., prosodic) evidence. For the speech recognition task, the generalized model gives a clean probabilistic framework for conditioning word probabilities on the conversation context via the underlying DA structure. Unlike previous models that did not address speech recognition or relied only on an intuitive 1-best approximation, our model allows computation of the optimum word sequence by effectively summing over all possible DA sequences as well as all recognition hypotheses throughout the conversation, using evidence from both past and future.

8. Discussion and Issues for Future Research

Our approach to dialogue modeling has two major components: statistical dialogue grammars modeling the sequencing of DAs, and DA likelihood models expressing the local cues (both lexical and prosodic) for DAs. We made a number of significant simplifications to arrive at a computationally and statistically tractable formulation. In this formulation, DAs serve as the hinges that join the various model components, but also decouple these components through statistical independence assumptions. Conditional

on the DAs, the observations across utterances are assumed to be independent, and evidence of different kinds from the same utterance (e.g., lexical and prosodic) is assumed to be independent. Finally, DA types themselves are assumed independent beyond a short span (corresponding to the order of the dialogue n-gram). Further research within this framework can be characterized by which of these simplifications are addressed.

Dialogue grammars for conversational speech need to be made more aware of the temporal properties of utterances. For example, we are currently not modeling the fact that utterances by the conversants may actually overlap (e.g., backchannels interrupting an ongoing utterance). In addition, we should model more of the nonlocal aspects of discourse structure, despite our negative results so far. For example, a context-free discourse grammar could potentially account for the nested structures proposed in Grosz and Sidner (1986).[10]

The standard n-gram models for DA discrimination with lexical cues are probably suboptimal for this task, simply because they are trained in the maximum likelihood framework, without explicitly optimizing discrimination between DA types. This may be overcome by using discriminative training procedures (Warnke et al. 1999; Ohler, Harbeck, and Niemann 1999). Training neural networks directly with posterior probability (Ries 1999a) seems to be a more principled approach and it also offers much easier integration with other knowledge sources. Prosodic features, for example, can simply be added to the lexical features, allowing the model to capture dependencies and redundancies across knowledge sources. Keyword-based techniques from the field of message classification should also be applicable here (Rose, Chang, and Lippmann 1991). Eventually, it is desirable to integrate dialogue grammar, lexical and prosodic cues into a single model, e.g., one that predicts the next DA based on DA history and all the local evidence.

The study of automatically extracted prosodic features for DA modeling is likewise only in its infancy. Our preliminary experiments with neural networks have shown that small gains are obtainable with improved statistical modeling techniques. However, we believe that more progress can be made by improving the underlying features themselves, in terms of both better understanding of how speakers use them, and ways to reliably extract them from data.

Regarding the data itself, we saw that the distribution of DAs in our corpus limits the benefit of DA modeling for lower-level processing, in particular speech recognition. The reason for the skewed distribution was in the nature of the task (or lack thereof) in Switchboard. It remains to be seen if more fine-grained DA distinctions can be made reliably in this corpus. However, it should be noted that the DA definitions are really arbitrary as far as tasks other than DA labeling are concerned. This suggests using unsupervised, self-organizing learning schemes that choose their own DA definitions in the process of optimizing the primary task, whatever it may be. Hand-labeled DA categories may still serve an important role in initializing such an algorithm.

We believe that dialogue-related tasks have much to benefit from corpus-driven, automatic learning techniques. To enable such research, we need fairly large, standardized corpora that allow comparisons over time and across approaches. Despite its shortcomings, the Switchboard domain could serve this purpose.

10 The inadequacy of n-gram models for nested discourse structures is pointed out by Chu-Carroll (1998), although the suggested solution is a modified n-gram approach.

Table 13
Dialogue act tag sets used in three other extensively studied corpora.

VERBMOBIL. These 18 high-level DAs used in VERBMOBIL-1 are abstracted over a total of 43 more specific DAs; most experiments on VERBMOBIL DAs use the set of 18 rather than 43. Examples are from Jekat et al. (1995).

Tag	Example
THANK	*Thanks*
GREET	*Hello Dan*
INTRODUCE	*It's me again*
BYE	*Alright bye*
REQUEST-COMMENT	*How does that look?*
SUGGEST	*from thirteenth through seventeenth June*
REJECT	*No Friday I'm booked all day*
ACCEPT	*Saturday sounds fine,*
REQUEST-SUGGEST	*What is a good day of the week for you?*
INIT	*I wanted to make an appointment with you*
GIVE_REASON	*Because I have meetings all afternoon*
FEEDBACK	*Okay*
DELIBERATE	*Let me check my calendar here*
CONFIRM	*Okay, that would be wonderful*
CLARIFY	*Okay, do you mean Tuesday the 23rd?*
DIGRESS	*[we could meet for lunch] and eat lots of ice cream*
MOTIVATE	*We should go to visit our subsidiary in Munich*
GARBAGE	*Oops, I-*

Map Task. The 12 DAs or "move types" used in Map Task. Examples are from Taylor et al. (1998).

Tag	Example
INSTRUCT	*Go round, ehm horizontally underneath diamond mine*
EXPLAIN	*I don't have a ravine*
ALIGN	*Okay?*
CHECK	*So going down to Indian Country?*
QUERY-YN	*Have you got the graveyard written down?*
QUERY-W	*In where?*
ACKNOWLEDGE	*Okay*
CLARIFY	{*you want to go... diagonally*} *Diagonally down*
REPLY-Y	*I do.*
REPLY-N	*No, I don't*
REPLY-W	{*And across to?*} *The pyramid.*
READY	*Okay*

ATR. The 9 DAs ("illocutionary force types") used in the ATR Dialogue Database task; some later models used an extended set of 15 DAs. Examples are from the English translations given by Nagata (1992).

Tag	Example
PHATIC	*Hello*
EXPRESSIVE	*Thank you*
RESPONSE	*That's right*
PROMISE	*I will send you a registration form*
REQUEST	*Please go to Kitaooji station by subway*
INFORM	*We are not giving any discount this time*
QUESTIONIF	*Do you have the announcement of the conference?*
QUESTIONREF	*What should I do?*
QUESTIONCONF	*You have already transferred the registration fee, right?*

Table 14
Data on recent DA tagging experiments. The number of DA tokens reflects training set size; accuracy refers to automatic tagging correctness. The error rates should not be compared, since the tasks were quite different. The comment field indicates special difficulties due to the type of input data.

Source	Number of DA Tokens	Number of DA Types/Tag Set	Accuracy	Comments
Woszczyna and Waibel (1994)	150–250(?)	6	74.1%	
Nagata and Morimoto (1994)	2,450	15 / ATR	39.7%	
Reithinger et al. (1996)	6,494	18 / VERBMOBIL	$\approx$ 40%	
Mast et al. (1996)	6,494	18 / VERBMOBIL	59.7%	
Warnke et al. (1997)	6,494	18 / VERBMOBIL	53.4%	unsegmented
Reithinger and Klesen (1997)	2,701	18 / VERBMOBIL	74.7%	
Chu-Carroll (1998)	915	15	49.71%	
Wright (1998)	3,276	12 / Map Task	64%	from speech
Taylor et al. (1998)	9,272	12 / Map Task	47%	
Samuel, Carberry, and Vijay-Shanker (1998)	2,701	18 / VERBMOBIL	75.12%	
Fukada et al. (1998)	3,584	26 / C-Star (Japanese)	81.2%	
Fukada et al. (1998)	1,902	26 / C-Star (English)	56.9%	
Present study	198,000	42 / SWBD-DAMSL	65%	from speech

9. Conclusions

We have developed an integrated probabilistic approach to dialogue act modeling for conversational speech, and tested it on a large speech corpus. The approach combines models for lexical and prosodic realizations of DAs, as well as a statistical discourse grammar. All components of the model are automatically trained, and are thus applicable to other domains for which labeled data is available. Classification accuracies achieved so far are highly encouraging, relative to the inherent difficulty of the task as measured by human labeler performance. We investigated several modeling alternatives for the components of the model (backoff n-grams and maximum entropy models for discourse grammars, decision trees and neural networks for prosodic classification) and found performance largely independent of these choices. Finally, we developed a principled way of incorporating DA modeling into the probability model of a continuous speech recognizer, by constraining word hypotheses using the discourse context. However, the approach gives only a small reduction in word error on our corpus, which can be attributed to a preponderance of a single dialogue act type (statements).

Note

The research described here is based on a project at the 1997 Workshop on Innovative Techniques in LVCSR at the Center for Speech and Language Processing at Johns Hopkins University (Jurafsky et al. 1997; Jurafsky et al. 1998). The DA-labeled Switchboard transcripts as well as other project-related publications are available at http://www.colorado.edu/ling/jurafsky/ws97/.

Acknowledgments

We thank the funders, researchers, and support staff of the 1997 Johns Hopkins Summer Workshop, especially Bill Byrne, Fred Jelinek, Harriet Nock, Joe Picone, Kimberly Shiring, and Chuck Wooters. Additional support came from the NSF via grants IRI-9619921 and IRI-9314967, and from the UK Engineering and Physical Science Research Council (grant GR/J55106). Thanks to Mitch Weintraub, to Susann LuperFoy, Nigel Ward, James Allen, Julia Hirschberg, and Marilyn Walker for advice on the design of the SWBD-DAMSL tag-set, to the discourse labelers at CU Boulder (Debra Biasca, Marion Bond, Traci Curl, Anu Erringer, Michelle Gregory, Lori Heintzelman, Taimi Metzler, and Amma Oduro) and the intonation labelers at the University of Edinburgh (Helen Wright, Kurt Dusterhoff, Rob Clark, Cassie Mayo, and Matthew Bull). We also thank Andy Kehler and the anonymous reviewers for valuable comments on a draft of this paper.

References

Alexandersson, Jan and Norbert Reithinger. 1997. Learning dialogue structures from a corpus. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 4, pages 2231–2234, Rhodes, Greece, September.

Anderson, Anne H., Miles Bader, Ellen G. Bard, Elizabeth H. Boyle, Gwyneth M. Doherty, Simon C. Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

Austin, J. L. 1962. *How to do Things with Words*. Clarendon Press, Oxford.

Bahl, Lalit R., Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, March.

Bard, Ellen G., Catherine Sotillo, Anne H. Anderson, and M. M. Taylor. 1995. The DCIEM Map Task corpus: Spontaneous dialogues under sleep deprivation and drug treatment. In Isabel Trancoso and Roger Moore, editors, *Proceedings of the ESCA-NATO Tutorial and Workshop on Speech under Stress*, pages 25–28, Lisbon, September.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions in Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bourlard, Hervé and Nelson Morgan. 1993. *Connectionist Speech Recognition. A Hybrid Approach*. Kluwer Academic Publishers, Boston, MA.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA.

Bridle, J. S. 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogleman Soulie and J. Herault, editors, *Neurocomputing: Algorithms, Architectures and Applications*. Springer, Berlin, pages 227–236.

Brill, Eric. 1993. Automatic grammar induction and parsing free text: A transformation-based approach. In *Proceedings of the ARPA Workshop on Human Language Technology*, Plainsboro, NJ, March.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carlson, Lari. 1983. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel.

Chu-Carroll, Jennifer. 1998. A statistical model for discourse act recognition in dialogue interactions. In Jennifer Chu-Carroll and Nancy Green, editors, *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium*. Technical Report SS-98-01, pages 12–17. AAAI Press, Menlo Park, CA.

Church, Kenneth Ward. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, TX.

Core, Mark and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA, November.

Dermatas, Evangelos and George Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics*, 21(2):137–163.

Eckert, Wieland, Florian Gallwitz, and Heinrich Niemann. 1996. Combining stochastic and linguistic language models for recognition of spontaneous speech. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 423–426, Atlanta, May.

Finke, Michael, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. 1998. Clarity: Inferring discourse structure from speech. In Jennifer Chu-Carroll and Nancy Green, editors, *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium*. Technical Report SS-98-01, pages 25–32. AAAI Press, Menlo Park, CA.

Fowler, Carol A. and Jonathan Housum. 1987. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26:489–504.

Fukada, Toshiaki, Detlef Koll, Alex Waibel, and Kouichi Tanigaki. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. In Robert H. Mannell and Jordi Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 6, pages 2771–2774, Sydney, December. Australian Speech Science and Technology Association.

Godfrey, J. J., E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, San Francisco, March.

Grosz, B. and C. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hirschberg, Julia B. and Diane J. Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Iyer, Rukmini, Mari Ostendorf, and J. Robin Rohlicek. 1994. Language modeling with sentence-level mixtures. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 82–86, Plainsboro, NJ, March.

Jefferson, Gail. 1984. Notes on a systematic deployment of the acknowledgement tokens 'yeah' and 'mm hm'. *Papers in Linguistics*, 17:197–216.

Jekat, Susanne, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and Joachim Quantz. 1995. Dialogue acts in VERBMOBIL. Verbmobil-Report 65, Universität Hamburg, DFKI GmbH, Universität Erlangen, and TU Berlin, April.

Jurafsky, Dan, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries,

Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings IEEE Workshop on Speech Recognition and Understanding*, pages 88–95, Santa Barbara, CA, December.

Jurafsky, Daniel, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1998. Switchboard discourse language modeling project final report. Research Note 30, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, January.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard-DAMSL Labeling Project Coder's Manual. Technical Report 97-02, University of Colorado, Institute of Cognitive Science, Boulder, CO. `http://www.colorado.edu/ling/-jurafsky/manual.august1.html`.

Jurafsky, Daniel, Elizabeth E. Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120. Association for Computational Linguistics.

Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, March.

Kita, Kenji, Yoshikazu Fukui, Masaaki Nagata, and Tsuyoshi Morimoto. 1996. Automatic acquisition of probabilistic dialogue models. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 1, pages 196–199, Philadelphia, October.

Kompe, Ralf. 1997. *Prosody in speech understanding systems*. Springer, Berlin.

Kowtko, Jacqueline C. 1996. *The Function of Intonation in Task Oriented Dialogue*. Ph.D. thesis, University of Edinburgh, Edinburgh.

Kuhn, Roland and Renato de Mori. 1990. A cache-base natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, June.

Levin, Joan A. and Johanna A. Moore. 1977. Dialogue games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420.

Levin, Lori, Klaus Ries, Ann Thymé-Gobbel, and Alon Lavie. 1999. Tagging of speech acts and dialogue games in Spanish CallHome. In *Towards Standards and Tools for Discourse Tagging* (Proceedings of the Workshop at ACL'99), pages 42–47, College Park, MD, June.

Linell, Per. 1990. The power of dialogue dynamics. In Ivana Marková and Klaus Foppa, editors, *The Dynamics of Dialogue*. Harvester, Wheatsheaf, New York, London, pages 147–177.

Mast, M., R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke. 1996. Dialog act classification with the help of prosody. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 1732–1735, Philadelphia, October.

Menn, Lise and Suzanne E. Boyce. 1982. Fundamental frequency and discourse structure. *Language and Speech*, 25:341–383.

Meteer, Marie, Ann Taylor, Robert MacIntyre, and Rukmini Iyer. 1995. Dysfluency annotation stylebook for the Switchboard corpus. Distributed by LDC, `ftp://ftp.cis.upenn.edu/pub-/treebank/swbd/doc/DFL-book.ps`, February. Revised June 1995 by Ann Taylor.

Morgan, Nelson, Eric Fosler, and Nikki Mirghafori. 1997. Speech recognition using on-line estimation of speaking rate. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 4, pages 2079–2082, Rhodes, Greece, September.

Munk, Marcus. 1999. *Shallow Statistical Parsing for Machine Translation*. Diploma thesis, Carnegie Mellon University.

Nagata, Masaaki. 1992. Using pragmatics to rule out recognition errors in cooperative task-oriented dialogues. In John J. Ohala, Terrance M. Nearey, Bruce L. Derwing, Megan M. Hodge, and Grace E. Wiebe, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 1, pages 647–650, Banff, Canada, October.

Nagata, Masaaki and Tsuyoshi Morimoto. 1993. An experimental statistical dialogue model to predict the speech act type of the next utterance. In Katsuhiko Shirai, Tetsunori Kobayashi, and Yasunari Harada, editors, *Proceedings of the*

International Symposium on Spoken Dialogue, pages 83–86, Tokyo, November.

Nagata, Masaaki and Tsuyoshi Morimoto. 1994. First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203.

Ohler, Uwe, Stefan Harbeck, and Heinrich Niemann. 1999. Discriminative training of language model classifiers. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 4, pages 1607–1610, Budapest, September.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Power, Richard J. D. 1979. The organisation of purposeful dialogues. *Linguistics*, 17:107–152.

Rabiner, L. R. and B. H. Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, January.

Reithinger, Norbert, Ralf Engel, Michael Kipp, and Martin Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 654–657, Philadelphia, October.

Reithinger, Norbert and Martin Klesen. 1997. Dialogue act classification using language models. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 4, pages 2235–2238, Rhodes, Greece, September.

Ries, Klaus. 1999a. HMM and neural network based speech act classification. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 497–500, Phoenix, AZ, March.

Ries, Klaus. 1999b. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 3, pages 1415–1418, Budapest, September.

Rose, R. C., E. I. Chang, and R. P. Lippmann. 1991. Techniques for information retrieval from voice messages. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 317–320, Toronto, May.

Sacks, H., E. A. Schegloff, and G. Jefferson. 1974. A simplest semantics for the organization of turn-taking in conversation. *Language*, 50(4):696–735.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 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*, volume 2, pages 1150–1156, Montreal.

Schegloff, E. A. 1968. Sequencing in conversational openings. *American Anthropologist*, 70:1075–1095.

Schegloff, Emanuel A. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, *Analyzing Discourse: Text and Talk*. Georgetown University Press, Washington, D.C., pages 71–93.

Searle, J. R. 1969. *Speech Acts*. Cambridge University Press, London-New York.

Shriberg, Elizabeth, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):439–487.

Shriberg, Elizabeth, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):to appear, September. Special Issue on Accessing Information in Spoken Audio.

Siegel, Sidney and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition.

Stolcke, Andreas and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 1005–1008, Philadelphia, October.

Suhm, B. and A. Waibel. 1994. Toward better language models for spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 831–834, Yokohama, September.

Taylor, Paul, Simon King, Stephen Isard, and Helen Wright. 1998. Intonation and dialog context as constraints for speech recognition. *Language and Speech*, 41(3-4):489–508.

Taylor, Paul, Simon King, Stephen Isard, Helen Wright, and Jacqueline Kowtko.

1997. Using intonation to constrain language models in speech recognition. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 5, pages 2763–2766, Rhodes, Greece, September.

Taylor, Paul A. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3):1697–1714.

Van Ess-Dykema, Carol and Klaus Ries. 1998. Linguistically engineered tools for speech recognition error analysis. In Robert H. Mannell and Jordi Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 5, pages 2091–2094, Sydney, December. Australian Speech Science and Technology Association.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

Warnke, V., R. Kompe, H. Niemann, and E. Nöth. 1997. Integrated dialog act segmentation and classification using prosodic features and language models. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 1, pages 207–210, Rhodes, Greece, September.

Warnke, Volker, Stefan Harbeck, Elmar Nöth, Heinrich Niemann, and Michael Levit. 1999. Discriminative estimation of interpolation parameters for language model classifiers. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 525–528, Phoenix, AZ, March.

Weber, Elizabeth G. 1993. *Varieties of Questions in English Conversation*. John Benjamins, Amsterdam.

Witten, Ian H. and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, July.

Woszczyna, M. and A. Waibel. 1994. Inferring linguistic structure in spoken language. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 847–850, Yokohama, September.

Wright, Helen. 1998. Automatic utterance type detection using suprasegmental features. In Robert H. Mannell and Jordi Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 1403–1406, Sydney, December. Australian Speech Science and Technology Association.

Wright, Helen and Paul A. Taylor. 1997. Modelling intonational structure using hidden Markov models. In *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, pages 333–336, Athens, September.

Yngve, Victor H. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–577, Chicago, April. University of Chicago.

Yoshimura, Takashi, Satoru Hayamizu, Hiroshi Ohmura, and Kazuyo Tanaka. 1996. Pitch pattern clustering of user utterances in human-machine dialogue. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 837–840, Philadelphia, October.