To appear in Laurence R. Horn and Gregory Ward (eds.)
Handbook of Pragmatics. Oxford: Blackwell

Pragmatics and Computational Linguistics

Dan Jurafsky

1 Introduction

These days there's a computational version of everything. Computational biology, computational musicology, computational archaeology, and so on, ad infinitum. Even movies are going digital. This chapter, as you might have guessed by now, thus explores the computational side of pragmatics. Computational pragmatics might be defined as the computational study of the relation between utterances and context. Like other kinds of pragmatics, this means that computational pragmatics is concerned with indexicality, with the relation between utterances and action, with the relation between utterances and discourse, and with the relationship between utterances and the place, time, and environmental context of their being uttered.

As Bunt and Black (2000) point out, computational pragmatics, like pragmatics in general, is especially concerned with INFERENCE. Four core inferential problems in pragmatics have received the most attention in the computational community: REFERENCE RESOLUTION, the interpretation and generation of SPEECH ACTS, the interpretation and generation of DISCOURSE STRUCTURE AND COHERENCE RELATIONS, and ABDUCTION. Each of these four problems can be cast as an inference task, one of somehow filling in information that isn't actually present in the utterance at hand. Two of these tasks are addressed in other chapters of this volume; abduction in Hobbs (this volume) , and discourse structure and coherence in Kehler (this volume). Reference resolution is covered in Kehler (2000). I have therefore chosen the interpretation of speech acts as the topic of this chapter.

Speech act interpretation, a classic pragmatic problem, is a good choice for this overview chapter for many reasons. First, the early computational work drew very strongly from the linguistic literature of the period. This enables us to closely compare the ways that computational linguistic and non-computational linguistic approaches differ in their methodology Second, there are two distinct computational paradigms in speech act interpretation: a logic-based approach and a probabilistic approach. I see these two approaches as good vehicles for motivating the two dominant paradigms in computational linguistics; one based on logic, logical inference, feature-structures, and unification, and the other based on probabilistic approaches. Third, speech act interpretation provides a good example of pragmatic inference: inferring a kind of linguistic structure which is not directly present in the input utterance. Finally, speech act interpretation is a problem that applies very naturally both to written and spoken genres. This allows us to discuss the computational processing of speech input, and in general talk about the way that computational linguistics has dealt with the differences between spoken and written inputs.

I like to think of the role of computational models in linguistics as a kind of musical conversation among three melodic voices. The base melody is the role of computational linguistics as a core of what we sometimes call 'mathematical foundations' of linguistics, the study of the formal underpinnings of models such as rules or trees, features or unification, indices or optimality. The middle note is the attempt to do what we sometimes call language engineering. One futuristic goal of this research is the attempt to build artificial agents that can carry on conversations with humans in order to perform tasks like answering questions, keeping schedules, or giving directions. The third strain is what is usually called 'computational psycholinguistics': the use of computational techniques to build processing models of human psycholinguistic performance. All of these melodic lines appear in computational pragmatics, although in this overview chapter we will focus more on the first two roles; linguistic foundations and language engineering.

The problem with focusing on speech act interpretation, of course, is that we will not be able to address the breadth of work in computational pragmatics. As suggested above above, the interested reader should turn to other chapters in this volume (especially Kehler (this volume) and Hobbs (this volume)) and also to Jurafsky and Martin (2000), which covers a number of computational pragmatic issues from a pedagogical perspective. Indeed, this chapter itself began as an expansion of, and meditation on, the section on dialogue act interpretation in Jurafsky and Martin (2000).

2 Speech Act Interpretation: The problem, and a quick historical overview

The problem of speech act interpretation is to determine, given an utterance, which speech act it realizes. Of course, some speech acts have surface cues to their form; some questions, for example, begin with *wh*-words or with aux-inversion. The Literal Meaning Hypothesis (Gazdar 1981), also called the Literal Force Hypothesis (Levinson 1983), is a strong version of this hypothesis, suggesting that every utterance has an illocutionary force which is built into its surface form. According to this hypothesis, aux-inverted sentences in English have QUESTION force; subjectdeleted sentences have IMPERATIVE force, and so on.

But it has long been known that many or even most sentences do not seem to have the speech act type associated with their syntactic form. Consider two kinds of examples of this phenomenon. One example is INDIRECT REQUESTS, in which what looks on the surface like a question is actually a polite form of a directive or a request to perform an action. The sentence

(1) Can you pass the salt?

looks on the surface like a *yes-no* question asking about the hearer's ability to pass the salt, but functions actually as a polite directive to pass the salt.

There are other examples where the surface form of an utterance doesn't match its speech act form. For example, what looks on the surface like a statement can really be a question. A very common kind of question, called a CHECK question (Carletta *et al.* 1997b; Labov and Fanshel 1977) is used to ask the other participant to confirm something that this other participant has privileged knowledge about. These checks are questions, but they have declarative word order, as in the bold-faced utterance in the following snippet from a travel agent conversation:

(2)

A	I was wanting to make some arrangements for a trip that I'm going to be taking uh to LA uh beginning of the week after next.
B	OK uh let me pull up your profile and I'll be right with you here. [pause]
B	**And you said you wanted to travel next week?**
A	Uh yes.

There are two computational models of the interpretation of speech acts. The first class of models was originally motivated by indirect requests of the "pass the salt" type. Gordon and Lakoff (1971), and then Searle (1975), proposed the seeds of this INFERENTIAL approach. Their intuition was that a sentence like *Can you pass the salt?* is unambiguous, having the literal meaning of a question: *Do you have the ability to pass me the salt?* The request speech act *Pass me the salt* is inferred by the hearer in a later step of understanding after processing the literal question. Computational implementations of this idea focus on using belief logics to model this inference chain.

The second class of models has been called CUE-BASED or PROBABILISTIC (Jurafsky and Martin 2000). The name CUE-BASED draws on the key role of cues in such psychological models as the Competition Model of Bates and MacWhinney (MacWhinney 1987; MacWhinney *et al.* 1984). These models are motivated more by indirect requests like CHECK questions. Here the problem is to figure out that what looks on the surface like a statement is really a question. Cue-based models think of the surface form of the sentence as a set of CUES to the speaker's intentions. Figuring out these intentions does require inference, but not of the type that chains through literal meanings.

These two models also differ in another important way. The inferential models are based on belief logics and use logical inference to reason about the speaker's intentions. The cue-based models tend to be probabilistic machine learning models. They see interpretation as a classification task, and solve it by training statistical classifiers on labeled examples of speech acts.

Despite their differences, these models have in common the use of a kind of abductive inference. In each case, the hearer infers something that was not contained directly in the semantics of the input utterance. That makes them an excellent pair of examples of these two different ways of looking at computational linguistics. The next section introduces a version of the inferential model called the PLAN INFERENCE or BDI model, and the following section the CUE-BASED model.

3 The Plan Inference (or BDI) Model of Speech Act Interpretation

The first approach to speech act interpretation we will consider is generally called the BDI (belief, desire, and intention) or PLAN-BASED model, proposed by Allen, Cohen, and Perrault and their colleagues (e.g., Allen 1995). Bunt and Black (2000) define this line of inquiry as follows:

> to apply the principles of rational agenthood to the modeling of a (computer-based) dialogue participant, where a rational communicative agent is endowed not only with certain private knowledge and the logic of belief, but is considered to also assume a great deal of common knowledge/beliefs with an interlocutor, and to be able to update beliefs about the interlocutor's intentions and beliefs as a dialogue progresses.

The earliest papers, such as Cohen and Perrault (1979), offered an AI planning model for how speech acts are GENERATED. One agent, seeking to find out some information, could use standard planning techniques to come up with the plan of asking the hearer to tell the speaker the information. Perrault and Allen (1980) and Allen and Perrault (1980) also applied this BDI approach to COMPREHENSION, specifically the comprehension of indirect speech effects.

Their application of the BDI model to comprehension draws on the plan-inference approach to dialogue act interpretation, first proposed by Gordon and Lakoff (1971) and Searle (1975). Gordon, Lakoff, and Searle noticed that there was a structure to what kind of things a speaker could do to make an indirect request. In particular, they noticed that a speaker could mention or question various quite specific properties of the desired activity to make an indirect request. For example, the air travel request "*Give me certain flight information*" can be realized as many different kinds of indirect requests. Here is a partial list from Jurafsky and Martin (2000) with examples from the ATIS[1] corpus of sentences spoken to a computerized speech understanding system for planning air travel:

1. The speaker can question the hearer's ability to perform the activity
   - Can you give me a list of the flights from Atlanta to Boston?
   - Could you tell me if Delta has a hub in Boston?
   - Would you be able to, uh, put me on a flight with Delta?
2. The speaker can mention speaker's wish or desire about the activity
   - I want to fly from Boston to San Francisco.
   - I would like to stop somewhere else in between.
   - I'm looking for one way flights from Tampa to Saint Louis.
   - I need that for Tuesday.
   - I wonder if there are any flights from Boston to Dallas.
3. The speaker can mention the hearer's doing the action
   - Would you please repeat that information?
   - Will you tell me the departure time and arrival time on this American flight?
4. The speaker can question the speaker's having permission to receive results of the action
   - May I get a lunch on flight UA 21 instead of breakfast?
   - Could I have a listing of flights leaving Boston?

Based on the realization that there were certain systemic ways of making indirect requests, Searle (1975:73) proposed that the hearer's chain of reasoning upon hearing *Can you give me a list of the flights from Atlanta to Boston?* might be something like the following (Searle's sentence was actually different; I've modified it to this ATIS example):

1. X has asked me a question about whether I have the ability to give a list of flights.

2. I assume that X is being cooperative in the conversation (in the Gricean sense) and that his utterance therefore has some aim.

3. X knows I have the ability to give such a list, and there is no alternative reason why X should have a purely theoretical interest in my list-giving ability.

4. Therefore X's utterance probably has some ulterior illocutionary point. What can it be?

5. A preparatory condition for a directive is that the hearer have the ability to perform the directed action.

6. Therefore X has asked me a question about my preparedness for the action of giving X a list of flights.

7. Furthermore, X and I are in a conversational situation in which giving lists of flights is a common and expected activity.

8. Therefore, in the absence of any other plausible illocutionary act, X is probably requesting me to give him a list of flights.

The inferential approach thus explains why *Can you give me a list of flights from Boston?* is a reasonable way of making an indirect request in a way that *Boston is in New England* is not: the former mentions a precondition for the desired activity, and there is a reasonable inferential chain from the precondition to the activity itself.

As we suggested above, Perrault and Allen (1980) and Allen and Perrault (1980) applied this BDI approach to the comprehension of indirect speech effects, essentially cashing out Searle's (1975) promissory note in a computational formalism.

I'll begin by summarizing Perrault and Allen's formal definitions of belief and desire in the predicate calculus. I'll represent "S believes the proposition P" as the two-place predicate $B(S,P)$. Reasoning about belief is done with a number of axiom schemas inspired by Hintikka (1969) (such as $B(A,P) \wedge B(A,Q) \Rightarrow B(A,P \wedge Q)$; see Perrault and Allen (1980) for details). Knowledge is defined as "true belief"; *S knows that P* will be represented as $KNOW(S,P)$, defined as follows:

$$\text{KNOW}(S,P) \equiv P \wedge B(S,P)$$

In addition to *knowing that*, we need to define *knowing whether*. *S knows whether* (KNOWIF) a proposition P is true if S KNOWs that P or S KNOWs that $\neg P$:

$$\text{KNOWIF}(S,P) \equiv \text{KNOW}(S,P) \vee \text{KNOW}(S,\neg P)$$

The theory of desire relies on the predicate WANT. If an agent S wants P to be true, we say $WANT(S,P)$, or $W(S,P)$ for short. P can be a state or the execution of some action. Thus if ACT is the name of an action, $W(S,\text{ACT}(H))$ means that S wants H to do ACT. The logic of WANT relies on its own set of axiom schemas just like the logic of belief.

The BDI models also require an axiomatization of actions and planning; the simplest of these is based on a set of ACTION SCHEMAS similar to the AI planning model STRIPS (Fikes and Nilsson 1971). Each action schema has a set of parameters with CONSTRAINTS about the type of each variable, and three parts:

- PRECONDITIONS: Conditions that must already be true in order to successfully perform the action.
- EFFECTS: Conditions that become true as a result of successfully performing the action.
- BODY: A set of partially ordered goal states that must be achieved in performing the action.

In the travel domain, for example, the action of agent A booking flight F for client C might have the following simplified definition:

BOOK-FLIGHT(A,C,F):

Constraints:	Agent(A) ∧ Flight(F) ∧ Client(C)
Precondition:	Know(A,departure-date(F)) ∧ Know(A,departure-time(F)) ∧ Know(A,origin-city(F)) ∧ Know(A,destination-city(F)) ∧ Know(A,flight-type(F)) ∧ Has-Seats(F) ∧ W(C,(Book(A,C,F))) ∧ . . .
Effect:	Flight-Booked(A,C,F)
Body:	Make-Reservation(A,F,C)

Cohen and Perrault (1979) and Perrault and Allen (1980) use this kind of action specification for speech acts. For example, here is Perrault and Allen's definition for three speech acts relevant to indirect requests. INFORM is the speech act of informing the hearer of some proposition (the Austin/Searle ASSERTIVE) The definition of INFORM is based on Grice's 1957 idea that a speaker informs the hearer of something merely by causing the hearer to believe that the speaker wants them to know something:

INFORM(S,H,P):

Constraints:	Speaker(S) $\wedge$ Hearer(H) $\wedge$ Proposition(P)
Precondition:	Know(S,P) $\wedge$ W(S, INFORM(S, H, P))
Effect:	Know(H,P)
Body:	B(H,W(S,Know(H,P)))

INFORMIF is the act used to inform the hearer whether a proposition is true or not; like INFORM, the speaker INFORMIFs the hearer by causing the hearer to believe the speaker wants them to KNOWIF something:

INFORMIF(S,H,P):

Constraints:	Speaker(S) $\wedge$ Hearer(H) $\wedge$ Proposition(P)
Precondition:	KnowIf(S, P) $\wedge$ W(S, INFORMIF(S, H, P))
Effect:	KnowIf(H, P)
Body:	B(H, W(S, KnowIf(H, P)))

REQUEST is the directive speech act for requesting the hearer to perform some action:

REQUEST(S,H,ACT):

Constraints:	Speaker(S) $\wedge$ Hearer(H) $\wedge$ ACT(A) $\wedge$ H is agent of ACT
Precondition:	W(S,ACT(H))
Effect:	W(H,ACT(H))
Body:	B(H,W(S,ACT(H)))

Perrault and Allen's theory also requires what are called "surface-level acts". These correspond to the "literal meanings" of the imperative, interrogative, and declarative structures. For example the "surface-level" act S.REQUEST produces imperative utterances:

S.REQUEST (S, H, ACT):

Effect: B(H, W(S,ACT(H)))

The effects of S.REQUEST match the body of a regular REQUEST, since this is the default or standard way of doing a request (but not the only way). This "default" or "literal" meaning is the start of the hearer's inference chain. The hearer will be given an input which indicates that the speaker is requesting the hearer to inform the speaker whether the hearer is capable of giving the speaker a list:

S.REQUEST(S,H,InformIf(H,S,CanDo(H,Give(H,S,LIST))))

The hearer must figure out that the speaker is actually making a request:

REQUEST(H,S,Give(H,S,LIST))

The inference chain from the request-to-inform-if-cando to the request-to-give is based on a chain of *plausible inference*, based on heuristics called **plan inference** (**PI**) rules. We will use the following subset of the rules that Perrault and Allen (1980) propose:

- **(PI.AE) Action-Effect Rule:** For all agents S and H, if Y is an effect of action X and if H believes that S wants X to be done, then it is plausible that H believes that S wants Y to obtain.
- **(PI.PA) Precondition-Action Rule:** For all agents S and H, if X is a precondition of action Y and if H believes S wants X to obtain, then it is plausible that H believes that S wants Y to be done.
- **(PI.BA) Body-Action Rule:** For all agents S and H, if X is part of the body of Y and if H believes that S wants X done, then it is plausible that H believes that S wants Y done.
- **(PI.KD) Know-Desire Rule:** For all agents S and H, if H believes S wants to KNOWIF(P), then H believes S wants P to be true:

$$B(H,W(S,\text{KNOWIF}(S,P))) \overset{\text{plausible}}{\Longrightarrow} B(H,W(S,P))$$

- **(EI.1) Extended Inference Rule:** if $B(H,W(S,X)) \overset{\text{plausible}}{\Longrightarrow} B(H,W(S,Y))$ is a PI rule, then

$$B(H,W(S,B(H,(W(S,X))))) \overset{\text{plausible}}{\Longrightarrow} B(H,W(S,B(H,W(S,Y))))$$

  is a PI rule (i.e., you can prefix $B(H,W(S))$ to any plan inference rule).

Let's see how to use these rules to interpret the indirect speech act in *Can you give me a list of flights from Atlanta?* Step 0 in the table below shows the speaker's initial speech act, which the hearer initially interprets literally as a question. Step 1 then uses Plan Inference rule *Action-Effect*, which suggests that if the speaker asked for something (in this case information), they probably want it. Step 2 again uses the *Action-Effect* rule, here suggesting that if the Speaker wants an INFORMIF, and KNOWIF is an effect of INFORMIF, then the speaker probably also wants KNOWIF.

Rule	Step	Result
	0	S.REQUEST(S,H,InformIf(H,S,CanDo(H,Give(H,S,LIST))))
PI.AE	1	B(H,W(S,InformIf(H,S,CanDo(H,Give(H,S,LIST)))))
PI.AE/EI	2	B(H,W(S,KnowIf(H,S,CanDo(H,Give(H,S,LIST)))))
PI.KP/EI	3	B(H,W(S,CanDo(H,Give(H,S,LIST))))
PI.PA/EI	4	B(H,W(S,Give(H,S,LIST)))
PI.BA	5	REQUEST(H,S,Give(H,S,LIST))

Step 3 adds the crucial inference that people don't usually ask about things they aren't interested in; thus if the speaker asks whether something is true (in this case CanDo), the speaker probably wants it (CanDo) to be true. Step 4 makes use of the fact that CanDo(ACT) is a precondition for (ACT), making the inference that if the speaker wants a precondition (CanDo) for an action (Give), the speaker probably also wants the action (Give). Finally, step 5 relies on the definition of REQUEST to suggest that if the speaker wants someone to know that the speaker wants them to do something, then the speaker is probably REQUESTing them to do it.

In summary, the BDI model of speech act interpretation is based on three components:

1. an axiomatization of belief, of desire, of action and of planning inspired originally by the work of Hintikka (1969)
2. a set of plan inference rules, which codify the abductive heuristics of the understanding system
3. a theorem prover

Given these three components and an input sentence, a plan-inference system can interpret the correct speech act to assign to the utterance by simulating the inference chain suggested by Searle (1975).

The BDI model has many advantages. It is an explanatory model, in that its plan-inference rules explain why people make certain inferences rather than others. It is a rich and deep model of the knowledge that humans use in interpretation; thus in addition to its basis for building a conversational agent, the BDI model might be used as a formalization of a cognitive model of human interpretation. The BDI model also shows how linguistic knowledge can be integrated with non-linguistic knowledge in building a model of cognition. Finally, the BDI model is a clear example of the role of computational linguistics as a foundational tool in formalizing linguistic models.

In giving this summary of the plan-inference approach to indirect speech act comprehension, I have left out many details, including many necessary axioms, as well as mechanisms for deciding which inference rule to apply. The interested reader should consult Perrault and Allen (1980).

4 The cue-based model of speech act interpretation

The plan-inference approach to dialogue act comprehension is extremely powerful; by using rich knowledge structures and powerful planning techniques the algorithm is designed to address even subtle indirect uses of dialogue acts. Furthermore, the BDI model incorporates knowledge about speaker and hearer intentions, actions, knowledge, and belief that is essential for any complete model of dialogue. But although the BDI model itself has crucial advantages, there are a number of disadvantages to the way the BDI model attempts to solve the speech act interpretation problem.

Perhaps the largest drawback is that the BDI model of speech act interpretation requires that each utterance have a single literal meaning, which is operated on by plan inference rules to produce a final non-literal interpretation. Much recent work has argued against this literal-first non-literal-second model of interpretation. As Levinson (1983) suggests, for example, the speech act force of *most* utterances does not match their surface form. Levinson points out, for example, that the imperative is very rarely used to issue requests in English. He also notes another problem: that indirect speech acts often manifest surface syntactic reflexes associated with their indirect force as well as their putative 'literal force'.

The psycholinguistic literature, similarly, has not found evidence for the temporal primacy of literal interpretation. Swinney and Cutler (1979), just to give one example, found that literal and figurative meanings of idioms are accessed in parallel by the human sentence processor.

Finally, for many speech act types that are less well studied than the "big three" (question, statement, request), it's not clear what the "literal" force would be. Consider, for example, utterances like "*yeah*" which can function as YES-ANSWERS, AGREEMENTS, and BACKCHANNELS. It's not clear why any one of these should necessarily be the literal speech act and the others be the inferred act.

An alternative way of looking at disambiguation is to downplay the role of a "literal meaning". In this alternate **cue** model, we think of the listener as using different cues in the input to help decide how to build an interpretation. Thus the surface input to the interpretive algorithm provides clues to structure-building, rather than providing a literal meaning which must be modified by purely inferential processes. What characterizes a cue-based model is the use of different sources of knowledge (cues) for detecting a speech act, such as lexical, collocational, syntactic, prosodic, or conversational-structure cues.

The cue-based approach is based on metaphors from a different set of linguistic literature than the plan-inference approach. Where the plan-inference approach relies on Searle-like intuitions about logical inference from literal meaning, the cue-based approach draws from the conversational analytic tradition. In particular, it draws from intuitions about what Goodwin (1996) called **microgrammar** (specific lexical, collocation, and prosodic features which are characteristic of particular conversational moves), as well as from the British pragmatic tradition on conversational games and moves (Power 1979). In addition, where the plan-inference model draws most heavily from analysis of written text, the cue-based literature is grounded much more in the analysis of spoken language. Thus, for example, a cue-based approach might use cues from many domains to recognize a true question, including lexical and syntactic knowledge like aux-inversion, prosodic cues like rising intonation, and conversational structure clues, like the neighboring discourse structure, turn boundaries, etc.

4.1 Speech Acts and Dialogue Acts

Before I give the cue-based algorithm for speech act interpretation, I need to digress a bit to give some examples of the kind of speech acts that these algorithms will be addressing. This section summarizes a number of computational 'tag sets' of possible speech acts. The next section chooses one such act, CHECK, to discuss in more detail.

While speech acts provide a useful characterization of one kind of pragmatic force, more recent work, especially computational work in building dialogue systems, has significantly expanded this core notion, modeling more kinds of conversational functions that an utterance can perform. The resulting enriched acts are often called **dialogue acts** (Bunt 1994) or **conversational moves** (Power 1979; Carletta *et al.* 1997b).

The phrase 'dialogue act' is unfortunately ambiguous. As Bunt and Black (2000) point out, it has been variously used to loosely mean 'speech act, in the context of a dialogue' (Bunt 1994), to mean a combination of the speech act and semantic force of an utterance (Bunt 1995), or to mean an act with internal structure related specifically to its dialogue function (Allen and Core 1997). The third usage is perhaps the most common in the cue-based literature, and I will rely on it here.

In the remainder of this section, I discuss various examples of dialogue acts and dialogue act structures. A recent ongoing effort to develop dialogue act tagging schemes is the DAMSL (Dialogue Act Markup in Several Layers) architecture (Allen and Core 1997; Walker *et al.* 1996; Carletta *et al.* 1997a; Core *et al.* 1999), which codes various kinds of dialogue information about utterances. As we suggested above, DAMSL and other such computational efforts to build practical descriptions of dialogue acts, like cue-based models in general, all draw on a number of research areas outside of the philosophical traditions that first defined speech acts. Perhaps the most important source has been work in conversation analysis and related fields. These include work on **repair** (Schegloff *et al.* 1977) work on **grounding** (Clark and Schaefer 1989), and work on the relation of utterances to the preceding and succeeding discourse (Allwood *et al.* 1992; Allwood 1995; Schegloff 1968; Schegloff 1988).

For example, drawing on Allwood's work, the DAMSL tag set distinguishes between the **forward looking** and **backward looking** function of an utterance. The forward looking function of an utterance corresponds to something like the Searle/Austin speech act. The DAMSL tag set is more complex in having a hierarchically structured representation that I won't discuss here and differs also from the Searle/Austin speech act in being focused somewhat on the kind of dialogue acts that tend to occur in task-oriented dialogue:

STATEMENT	a claim made by the speaker
INFO-REQUEST	a question by the speaker
CHECK	a question for confirming information
INFLUENCE-ON-ADDRESSEE	(=Searle's directives)
OPEN-OPTION	a weak suggestion or listing of options
ACTION-DIRECTIVE	an actual command
INFLUENCE-ON-SPEAKER	(=Austin's commissives)
OFFER	speaker offers to do something, (subject to confirmation)
COMMIT	speaker is committed to doing something
CONVENTIONAL	other
OPENING	greetings
CLOSING	farewells
THANKING	thanking and responding to thanks

The backward looking function of DAMSL focuses on the relationship of an utterance to previous utterances by the other speaker. These include accepting and rejecting proposals (since DAMSL is focused on task-oriented dialogue), as well as acts involved in grounding and repair:

AGREEMENT	speaker's response to previous proposal
ACCEPT	accepting the proposal
ACCEPT-PART	accepting some part of the proposal
MAYBE	neither accepting nor rejecting the proposal
REJECT-PART	rejecting some part of the proposal
REJECT	rejecting the proposal
HOLD	putting off response, usually via subdialogue
ANSWER	answering a question
UNDERSTANDING	whether speaker understood previous
SIGNAL-NON-UNDER.	speaker didn't understand
SIGNAL-UNDER.	speaker did understand
ACK	demonstrated via backchannel or assessment
REPEAT-REPHRASE	demonstrated via repetition or reformulation
COMPLETION	demonstrated via collaborative completion

DAMSL and DAMSL-like sets of dialogue acts have been applied both to task-oriented dialogue and to non-task-oriented casual conversational speech. We give examples of two dialogue act tagsets designed for task-oriented dialogue and one for casual speech.

The task-oriented corpora are the Map Task and Verbmobil corpora. The Map Task corpus (Anderson *et al.* 1991) 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 speakers map, all without being able to see each other's maps. The Verbmobil corpus consists of two-party scheduling dialogues, in which the speakers were asked to plan a meeting at some future date. Tables 1 and 2 show the most commonly-used versions of the tagsets from those two tasks.

Switchboard is a large collection of 2400 6-minute telephone conversations between strangers who were asked to informally chat about certain topics (cars, children, crime). The SBWB-DAMSL tagset (Jurafsky *et al.* 1997b) was developed from the DAMSL tagset in an attempt to label the kind of non-task-oriented dialogues that occur in Switchboard. The tagset was multidimensional, with approximately 50 basic tags (QUESTION, STATEMENT, etc.) and various diacritics. A labeling project described in Jurafsky *et al.* (1997b) labeled every utterance in about 1200 of the Switchboard conversations; approximately 200,000 utterances were labeled. Approximately 220 of the many possible unique combinations of the SWBD-DAMSL codes were used by the coders. To obtain a system with somewhat higher inter-labeler agreement, as well as enough data per class for statistical modeling purposes, a less fine-grained tag set was devised, distinguishing 42 mutually exclusive utterance types (Jurafsky *et al.* 1998a; Stolcke *et al.* 2000). Table 3

Table 1: The 18 high-level dialogue acts used in Verbmobil-1, abstracted over a total of 43 more specific dialogue acts. Examples are from Jekat *et al.* (1995).

Tag	**Example**
THANK	*Thanks*
GREET	*Hello Dan*
INTRODUCE	*It's me again*
BYE	*Allright 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-*

Table 2: The 12 move types used in the Map Task. Examples are from Taylor *et al.* (1998).

Tag	**Example**
INSTRUCT	*Go round, ehm horizontally underneath diamond mind*
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*

shows the 42 categories with examples and relative frequencies.

None of these various sets of dialogue acts are meant to be an exhaustive list. Each was designed with some particular computational task in mind, and hence will have domain-specific inclusions or absences. I have included them here mainly to show the kind of delimited task that the computational modeling community has set for themselves. As is clear from the examples above, the various tag sets do include commonly studied speech acts like QUESTION and REQUEST. They also, however, include acts that have not been studied in the speech act literature. In the next section, I summarize one of these dialogue acts, the CHECK, in order to give the reader a more in-depth view of at least one of these various 'minor' acts.

4.2 The Dialogue Act CHECK

We saw in previous sections that the motivating example for the plan-based approach was based on indirect requests (surface questions with the illocutionary force of a REQUEST). In this section we'll look at a different kind of indirect speech act, one that has motivated some of the cue-based literature. The speech act we will look at, introduced very briefly above, is often called a CHECK or a CHECK QUESTION (Carletta *et al.* 1997b; Labov and Fanshel 1977). A CHECK is a subtype of question which requests the interlocutor to confirm some information; the information may have been mentioned explicitly in the preceding dialogue (as in the example below), or it may have been inferred from what the interlocutor said:

(3)	A	I was wanting to make some arrangements for a trip that I'm going to be taking uh to LA uh beginning of the week after next.
	B	OK uh let me pull up your profile and I'll be right with you here. [pause]
	B	**And you said you wanted to travel next week?**
	A	Uh yes.

Here are some sample realizations of CHECKs in English from various corpora, showing their various surface forms:

(4) As tag questions (example from the Trains corpus; Allen and Core 1997):

U **and it's gonna take us also an hour to load boxcars right?**

S right

(5) As declarative questions, usually with rising intonation (Quirk *et al.* 1985:p. 814) (example from the Switchboard corpus; Godfrey *et al.* 1992)

Table 3: The 42 dialogue act labels, from Stolcke *et al.* (2000). Dialogue act frequencies are given as percentages of the total number of utterances in the 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 ACKNOWLEDGEMENT	*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 kid s.*	.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%

A and we have a powerful computer down at work.

B Oh (laughter)

B **so, you don't need a personal one (laughter)?**

A No

(6) As fragment questions (subsentential units; words, noun-phrases, clauses) (Weber 1993) (example from the Map Task corpus; Carletta *et al.* 1997b)

G Ehm, curve round slightly to your right.

F **To my right?**

G Yes.

The next section will discuss the kind of cues that are used to detect CHECKS and other dialogue acts.

4.3 Cues

A 'cue' is a surface feature that is probabilistically associated with some speech or dialogue act. Commonly-studied features include lexical, syntactic, prosodic, and discourse factors, but cues may also involve more sophisticated and complex knowledge, such as speaker-specific or dyad-specific modeling.

4.3.1 Lexical or Syntactic Cues

Lexical and syntactic cues have been widely described, at least for the most commonly studied speech acts. In a useful typological study, Sadock and Zwicky (1985) mention the existence of such cues for 'declarative' acts as declarative particles (in Welsh or Hidatsa), or different inflectional forms used specifically in declarative acts (Greenlandic).

Cross-linguistically common lexical or syntactic cues for imperatives include sentence-initial or sentence-final particles, verbal clitics, special verb morphology in the verb stem, subject deletion, and special subject pronoun forms that are used specifically in the imperative (Sadock and Zwicky 1985).

A similar inventory of cue types applies to lexical or syntactic cues for *yes-no* QUESTIONS, including sentence-initial or sentence-final particles, special verb morphology, and word order.

In addition to these cross-linguistic universals for the major acts, more recent work has begun to examine lexical and syntactic cues for minor acts. Michaelis (2001) shows that EXCLAMATIVES, for example, are characterized cross-linguistically by anaphoric degree adverbs, as well as various surface cues associated with information questions. Michaelis and Lambrecht (1996) discuss the wide variety of surface syntactic features which can characterize EXCLAMATIVES in English, including extraposition, bare complements, and certain kinds of definite noun phrases.

I have seen these same kinds of cues in my own work and that of my colleagues. Studies of CHECKs, for example, have shown that, like the examples above, they are most often realized with declarative structure (i.e., no aux-inversion), and they often have a following question tag, usually *right*, (Quirk *et al.* 1985:810-814), as in example (4) above. They also are often realized as fragments (subsentential words or phrases) (Weber 1993).

In the Switchboard corpus, a very common type of check is the REFORMULATION. A reformulation, by repeating back some summarized or rephrased version of the interlocutor's talk, is one way to ask "is this an acceptable summary of your talk?". Our examination of 960 reformulations in Switchboard (Jurafsky and Martin 2000) show that they have a very specific microgrammar. They generally have declarative word order, often with *you* as the subject (31% of the cases), often beginning with *so* (20%) or *oh*, and sometimes ending with *then*. Some examples:

(7) Oh so you're from the Midwest too.

(8) So you can steady it.

(9) You really rough it then.

This kind of microgrammar was originally noted by Goodwin (1996), in his discussion of ASSESSMENTS. Assessments are a particular kind of evaluative act, used to ascribe positive or negative properties

(10) That's good.

(11) Oh that's nice.

(12) It's great.

Goodwin (1996) found that assessments often display the following format:

(13) *Pro Term + Copula + (Intensifier) + Assessment Adjective*

Jurafsky *et al.* (1998b) found an even more constrained, and more lexicalized, microgrammar for the 1150 assessments with overt subjects in Switchboard. They found that the vast majority (80%) of the Pro Terms were *that*, that only 2 types of intensifiers occurred (*really* and *pretty*), and that the range of assessment adjective was quite small, consisting only of the following: *great, good, nice, wonderful, cool, fun, terrible, exciting, interesting, wild, scary, hilarious, neat, funny, amazing, tough, incredible, awful*.

4.3.2 Prosodic Cues

Prosody is another important cue for dialogue act identity. The final pitch rise of *yes-no* questions in American English (Sag and Liberman 1975; Pierrehumbert 1980) as well as cross-linguistically (Sadock and Zwicky 1985) is well known. Similarly well studied is the realization of final lowering in declaratives and *wh*-questions in English (the H*L L% tune) (Pierrehumbert and Hirschberg 1990).

Prosody plays an important role in other dialogue acts. Shriberg *et al.* (1998) and Weber (1993), for example, found that CHECKs, like other questions, are also most likely to have rising intonation. Curl and Bell (2001) examined the dialogue-act coded portion of Switchboard for three dialogue acts which can all be realized by the word *yeah*: AGREEMENTS, YES-ANSWERS, and BACKCHANNELS. They found that *yeah* agreements are associated with high falling contour, and *yeah* backchannels with low falling or level contours.

Sadock and Zwicky (1985) mention various other types of prosodic cues that occur cross-linguistically, including special stress in the first word of a *yes-no* QUESTION in Hopi, and a glottal stop in the last word of a *yes-no* QUESTION in Hidatsa.

Pierrehumbert and Hirschberg (1990) offer a much more compositional kind of cue-based theory for the role of prosody in semantic interpretation in general. In their model, pitch accents convey information about such things as the status of discourse referents, phrase accents convey information about the semantic relationship between intermediate phrases, and boundary tones convey information about the directionality of interpretation. Presumably these kinds of intonational meaning cues, and others such as, e.g., the rejection contour of Sag and Liberman (1975) or the uncertainty/incredulity contour of Ward and Hirschberg (1985,1988), could be used to build a model of prosodic cues specifically for dialogue acts.

4.3.3 Discourse Cues and Summary

Finally, discourse structure is obviously an important cue for dialogue act identity. A dialogue act which functions as the second part of an adjacency-pair (for example the YES-ANSWER), obviously depends on the presence of the first part (in this case a QUESTION). This is even true for sequences that aren't clearly adjacency pairs. Allwood (1995) points out that the utterance "No it isn't" is an AGREEMENT after a negative statement like "It isn't raining" but a DISAGREEMENT after a positive statement like "It is raining".

The importance of this contextual role of discourse cues has been a main focus of the conversation analysis tradition. For example Schegloff (1988) focuses on the way that the changing discourse context and the changing understanding of the hearer affects their interpretation of the discourse function of the utterance. Schegloff gives the following example utterance:

(14) Do you know who's going to that meeting?

which occurs in the following dialogue:

Mother:	Do you know who's going to that meeting?
Russ:	Who?
Mother:	I don't kno:w
Russ:	Oh:: Prob'ly Missiz McOwen...

Mother had meant her first utterance as a REQUEST. But Russ misinterprets it as a PRE-ANNOUNCEMENT, and gives an appropriate response to such pre-announcements, by asking the question word which was included in the pre-announcement ('Who?'). Mother's response ('I don't know') makes it clear that her utterance was a REQUEST rather than a PRE-ANNOUNCEMENT. In Russ's second utterance, he uses this information to reanalyze and re-respond to Mother's utterance.

This example shows that complex discourse information, such as the fact that an interlocutor has displayed a problem with a previous dialogue act interpretation, can play a role in future dialogue act interpretation.

In summary, we have seen three kinds of cues that can be used to help determine the dialogue act type of an utterance: prosodic cues, lexical and grammatical cues, and discourse structure cues. The next section discusses how cue-based algorithms make use of these cues to recognize dialogue acts.

4.4 The cue-based algorithms

The cue-based algorithm for speech act interpretation is given as input an utterance, and produces as output the most probable dialogue act for that utterance. In a sense, the idea of the cue-based models is to treat every utterance as if it has no literal force. Determining the correct force is treated as a task of probabilistic reasoning, in which different cues at different levels supply the evidence.

In other words, I and other proponents of the cue-based model believe that the literal force hypothesis is simply wrong; that there is not a literal force for each surface sentence type. Certainly it is the case that some surface cues are more commonly associated with certain dialogue act types. But rather than model this commonality as a fact about literal meaning (the 'Literal Force Hypothesis') the cue-based models treat it as a fact about a probabilistic relationship between cue and dialogue act; the probability of a given dialogue act may simply be quite high given some particular cue.

In discussing these cue-based approaches, I will draw particularly on research in which I have participated and hence with which I am familiar, such as Stolcke *et al.* (2000) and Shriberg *et al.* (1998). As we will see, these algorithms are mostly designed to work directly from input speech waveforms. This means that they are of necessity based on heuristic approximations to the available cues. For example, a useful prosodic cue might come from a perfect ToBI phonological parse of an input utterance. But the computational problem of deriving a perfect ToBI parse from speech input is unsolved. So we will see very simplistic approximations to the syntactic, discourse, and prosodic knowledge that we will someday have better models of.

The models we will describe generally use supervised machine-learning algorithms, trained on a corpus of dialogues that is hand-labeled with dialogue acts for each utterance. That is, these algorithms are statistical classifiers. We train a "QUESTION-classifier" on many instances of QUESTIONS, and it learns to recognize the combination of features (prosodic, lexical, syntactic, and discourse) which suggest the presence of a question. We train a "REQUEST-classifier" on many instances of REQUESTs, a BACKCHANNEL-classifier on many instances of BACKCHANNELs, and so on.

Let's begin with lexical and syntactic features. The simplest way to build a probabilistic model which detects lexical and phrasal cues is simply to look at which words and phrases occur more often in one dialogue act than another. Many scholars, beginning with Nagata and Morimoto (1994), realized that simple statistical grammars based on words and short phrases could serve to detect local structures indicative of particular dialogue acts. They implemented this intuition by modeling each dialogue act as having its own separate N-gram grammar (see e.g., Suhm and Waibel 1994; Mast *et al.* 1996; Jurafsky *et al.* 1997a; Warnke *et al.* 1997; Reithinger and Klesen 1997; Taylor *et al.* 1998). An N-gram grammar is a simple Markov model which stores, for each word, what its probability of occurrence is given one or more particular previous words.

These systems create a separate mini-corpus from all the utterances which realize the same dialogue act, and then train a separate N-gram grammar on each of these mini-corpora. (In practice, more sophisticated N-gram models are generally used, such as backoff, interpolated, or class N-gram language models). Given an input utterance consisting of a sequence of words W, they then choose the dialogue act d whose N-gram grammar assigns the highest likelihood to W. Technically, the formula for this maximization problem is as follows (although the non-probabilistic reader can safely ignore the formulas):

$$(15)\quad d^* = \underset{d}{\operatorname{argmax}}\, P(d|W)$$

$$(16) \qquad = \operatorname*{argmax}_d P(d)P(W|d)$$

Equation 15 says that our estimate of the best dialogue act d^* for an utterance is the dialogue act d which has the highest probability given the string W. By Bayes rule, that can be rewritten as equation 16. This says that the dialogue act which is most probable given the input is the one which maximizes the product of two factors: the prior probability of a particular dialogue act $P(d)$ and the probability $P(W|d)$, which expresses, given that we had picked a certain dialogue act d, the probability it would be realized as the string of words W.

This N-gram approach, while only a local heuristic to more complex syntactic constraints, does indeed capture much of the microgrammar. For example *yes-no* QUESTIONS often have bigram pairs indicative of aux-inversion (*do you*, *are you*, *was he*, etc). Similarly, the most common bigrams in REFORMULATIONS are very indicative pairs like *so you*, *sounds like*, *so you're*, *oh so*, *you mean*, *so they*, and *so it's*.

While this N-gram model of microgrammar has proved successful in practical implementations of dialogue act detection, it is obviously a gross simplification of microgrammar. It is possible to keep the idea of separate, statistically-trained microgrammars for each dialogue act while extending the simple N-gram model to more sophisticated probabilistic grammars. For example Jurafsky *et al.* (1998c) show that the grammar of some dialogue acts, like APPRECIATIONS, can be captured by building probabilistic grammars of lexical category sequences. Alexandersson and Reithinger (1997) propose even more linguistically sophisticated grammars for each dialogue act, such as probabilistic context-free grammars. To reiterate this point: the idea of cue-based processing does not require that the cues be simplistic Markov models. A complex phrase-structural or configurational feature is just as good a cue. The model merely requires that these features be defined probabilistically.

Prosodic models of dialogue act microgrammar rely on phonological features like pitch or accent, or their acoustic correlates like f0, duration, and energy. We mentioned above that features like final pitch rise are commonly used for questions and fall for assertions. Indeed, computational approaches to dialogue act prosody modeling have mostly focussed on f0. Many studies have successfully shown an increase in the ability to detect *yes-no* questions by combining lexical cues with these pitch-based cues (Waibel 1988; Daly and Zue 1992; Kompe *et al.* 1993; Taylor *et al.* 1998).

One such system, Shriberg *et al.* (1998), trained CART-style decision trees on simple acoustically-based prosodic features such as the slope of f0 at the end of the utterance, the average energy at different places in the utterance, and various duration measures. They found that these features were useful, for example, in distinguishing four broad clusters of dialogue acts STATEMENT (S), *yes-no* QUESTION (QY), DECLARATIVE-QUESTIONS like CHECKS (QD) and *wh*-QUESTIONS (QW) from each other. Figure 1 shows the decision tree which gives the posterior probability $P(d|F)$ of a dialogue act d type given a sequence of acoustic features F. Each node in the tree shows four probabilities, one for each of the four dialogue acts in the order S, QY, QW, QD; the most likely of the four is shown as the label for the node. Via the Bayes rule, this probability can be used to compute the likelihood of the acoustic features given the dialogue act: $P(f|d)$.

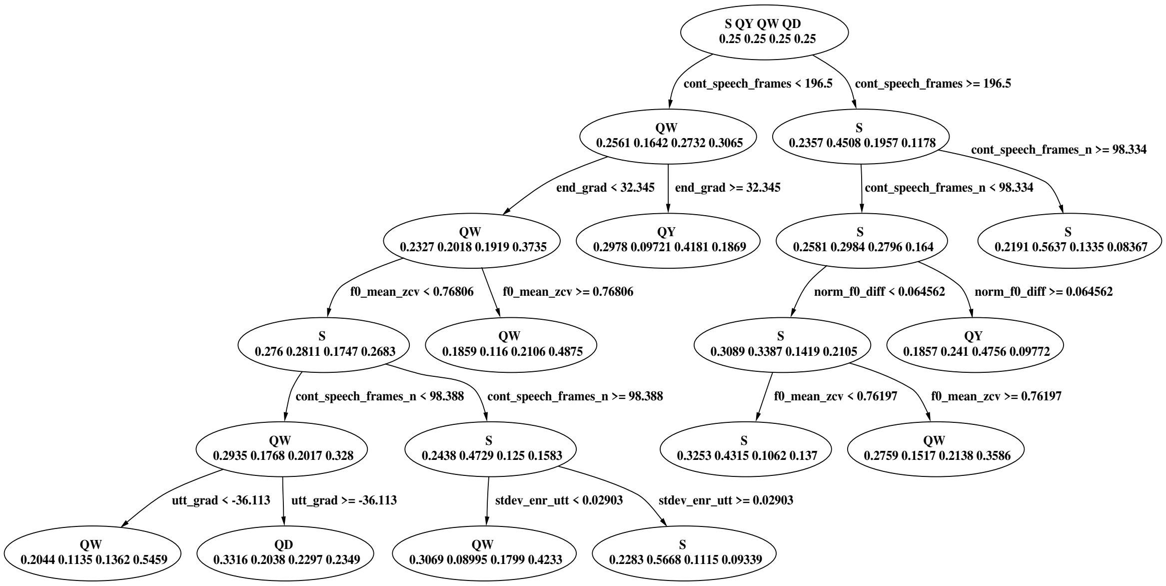

Figure 1: Decision tree for the classification of STATEMENT (S), *yes-no* QUESTIONS (QY), *wh*-QUESTIONS (QW) and DECLARATIVE QUESTIONS (QD), after Shriberg *et al.* (1998). Note that the difference between S and QY toward the right of the tree is based on the feature `norm_f0_diff` (normalized difference between mean f0 of end and penultimate regions), while the difference between QW and QD at the bottom left is based on `utt_grad`, which measures f0 slope across the whole utterance.

In general, most such systems use phonetic rather than phonological cues, modeling f0 patterns with techniques such as vector quantization and Gaussian classifiers on acoustic input (Kießling *et al.* 1993; Kompe *et al.* 1995; Yoshimura *et al.* 1996). But some more recent systems actually attempt to directly model phonological cues such as pitch accent and boundary tone sequence (Taylor *et al.* 1997).

A final important cue for dialogue act interpretation is conversational structure. One simple way to model conversational structure, drawing on the idea of adjacency pairs (Schegloff 1968; Sacks *et al.* 1974) introduced above, is as a probabilistic sequence of dialogue acts. As first proposed by Nagata (1992), and in a follow-up paper (Nagata and Morimoto 1994), the identity of the previous dialogue acts can be used to help predict upcoming dialogue acts. For example, BACKCHANNELS or AGREEMENTS might be very likely to follow STATEMENTS. ACCEPTS or REJECTS might be more likely to follow REQUESTS, and so on. Woszczyna and Waibel (1994) give the dialogue automaton shown in Figure 2, which models simple *N*-gram probabilities of dialogue act sequences for a Verbmobil-like appointment scheduling task.

Of course this idea of modeling dialogue act sequences as '*N*-grams' of dialogue acts only captures the effects of simple local discourse context. As I mentioned earlier in my discussion concerning syntactic cues, a more sophisticated model will need to take into account hierarchical discourse structure of various kinds. Indeed, the deficiencies of an *N*-gram model of dialogue structure are so great and so obvious that it might have been a bad idea for me to start this dialogue section of the chapter with them. But the fact is that the recent work on dialogue act interpretation that I describe here relies only on such simple cues. Once again, the fact that the examples we give all involve simple Markov models of dialogue structure should not be taken to imply that cue-based models of dialogue structure have to be simple. As the field progresses, presumably we will develop more complex probabilistic models of how speakers act in dialogue situations. Indeed, many others have already begun to enhance this *N*-gram approach (Nagata and Morimoto 1994; Suhm and Waibel 1994; Warnke *et al.* 1997; Stolcke *et al.* 1998; Taylor *et al.* 1998). Chu-Carroll (1998), for example, has shown how to model subdialogue structure in a cue-based model. Her model deals with hierarchical dialogue structures like insertion sequences (in which a question is followed by another question) and other kinds of complex structure. It's also important to note that a cue-based model doesn't disallow non-probabilistic knowledge sources; certainly not all dialogue structural information is probabilistic. For example, a REJECTION (a 'no' response) is a *dispreferred* response to a REQUEST. I suspect this isn't a probabilistic fact; rejections may be always dispreferred. Studying how to integrate non-probabilistic knowledge of this sort into a cue-based model is a key problem that I return to in the conclusion.

I have now talked about simple statistical implementations of detectors for three kinds of cues for dialogue acts: lexical/syntactic, prosodic, and discourse structural. How can a dialogue act interpreter combine these different cues to find the most likely correct sequence of correct dialogue acts given a conversation?

One way to combine these statistical cues into a single probabilistic cue-based model is to treat a conversation as a Hidden Markov Model (HMM), an idea that

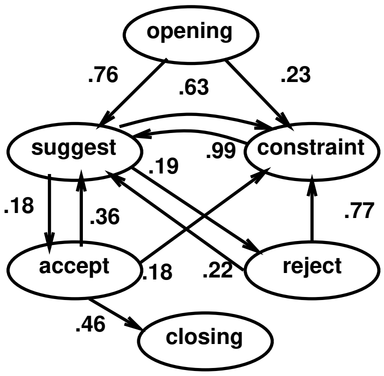

Figure 2: A dialogue act HMM for simple appointment scheduling conversations (after Woszczyna and Waibel (1994))

seems to have been first suggested by Woszczyna and Waibel (1994) and Suhm and Waibel (1994). A Hidden Markov Model is a kind of probabilistic automaton in which a series of states in an automaton probabilistically generate sequences of symbols. Since the output is probabilistic, it is not possible to be certain from a given output symbol which state generated it; hence the states are 'hidden'. The intuition behind using an HMM for a dialogue is that the dialogue acts play the role of the hidden states. The words, syntax, and prosody act as observed output symbols.

HMMs can be viewed as generative or as interpretive models. As a generative model, given that the automaton is about to generate a particular dialogue act, the probabilistic cue-models give the probabilities of different words, syntax, and prosody being produced. As an interpretive model, given a known sequence of words, syntax and prosody for an utterance, the HMM can be used to choose the single dialogue act which was most likely to have generated that sequence.

Stolcke *et al.* (2000) and Taylor *et al.* (1998) apply the HMM intuition of Woszczyna and Waibel (1994) to treat the dialogue act detection process as HMM-parsing. Given all available cues C about a conversation, the goal is to find the dialogue act sequence $D = \{d_1, d_2 \ldots, d_N\}$ that has the highest posterior probability $P(D|C)$ given those cues (here we are using capital letters to mean *sequences* of things). Applying Bayes' Rule we get

$$\begin{aligned} D^* &= \operatorname*{argmax}_D P(D|C) \\ &= \operatorname*{argmax}_D \frac{P(D)P(C|D)}{P(C)} \end{aligned}$$

$$(17) \qquad = \quad \underset{D}{\operatorname{argmax}} P(D)P(C|D)$$

Equation (17) should remind the reader of equation (16). It says that we can estimate the best series of dialogue acts for an entire conversation by choosing that dialogue act sequence which maximizes the product of two probabilities, $P(D)$ and $P(C|D)$.

The first, $P(D)$, is the probability of a sequence of dialogue acts. Sequences of dialogue acts which are more coherent will tend to occur more often than incoherent sequences of dialogue acts, and will hence be more probable. Thus $P(D)$ essentially acts as a model of conversational structure. One simple way to compute an approximation to this probability is via the dialogue act N-grams introduced by Nagata and Morimoto (1994).

The second probability which must be considered is the likelihood $P(C|D)$. This is the probability, given that we have a particular dialogue act sequence D, of observing a particular set of observed surface cues C. This likelihood $P(C|D)$ can be computed from two sets of cues. First, the microsyntax models (for example the different word-N-gram grammars for each dialogue act) can be used to estimate $P(W|D)$, the probability of the sequence of words W given a particular sequence of dialogue acts D. Next, the microprosody models (for example the decision tree for the prosodic features of each dialogue act), can be used to estimate $P(F|D)$, the probability of the sequence of prosodic features F. If we make the simplifying (but of course incorrect) assumption that the prosody and the words are independent, we can thus estimate the cue likelihood for a sequence of dialogue acts D as follows:

$$(18) \quad P(C|D) \quad = \quad P(F|D)P(W|D)$$

We can compute the most likely sequence of dialogue acts D^* by substituting equation (18) into equation (17), thus choosing the dialogue act sequence which maximizes the product of the three knowledge sources (conversational structure, prosody, and lexical/syntactic knowledge):

$$D^* \quad = \quad \underset{D}{\operatorname{argmax}} P(D)P(F|D)P(W|D)$$

Standard HMM-parsing techniques (like Viterbi) can then be used to search for this most-probable sequence of dialogue acts given the sequence of input utterances.

The HMM method is only one way of solving the problem of cue-based dialogue act identification. The link with HMM tagging suggests another approach, treating dialogue acts as tags, and applying other part-of-speech tagging methods based on various cues in the input. Samuel *et al.* (1998), for example, applied Transformation-Based Learning to dialogue act tagging.

As we conclude this section on the cue-based approach, it's worth taking a moment to distinguish the cue-based approach from what has been called the *idiom* or *conventional* approach. The idiom approach assumes that a sentence structure like *Can you give me a list?* or *Can you pass the salt?* is ambiguous between a literal meaning as a *yes-no* QUESTION and an idiomatic meaning as a REQUEST. The grammar of English would simply list REQUEST as one meaning of *Can you X*. The cue-based model does share some features of the idiom model; certain surface cues are directly linked to certain discourse functions. The difference is that the pure idiom model is by definition non-compositional and non-probabilistic. A certain surface sentence type is linked with a certain set of discourse function, one of which must be chosen. The cue-based model can capture some generalizations which the idiom approach cannot; certain cues for questions, say, may play a role also in requests. We can thus capture the link between questions and requests by saying that a certain cue plays a role in both dialogue acts.

5 Conclusion

In summary, the BDI and cue-based models of computational pragmatics are both important, and both will continue to play a role in future computational modeling. The BDI model focuses on the kind of rich, sophisticated knowledge and reasoning that is clearly necessary for building conversational agents that can interact. Agents have to know why they are asking questions, and have to be able to reason about complex pragmatic and world-knowledge issues. But the depth and richness of this model comes at the expense of breadth; current models only deal with a small number of speech acts and situations. The cue-based model focuses on statistical examination of the surface cues to the realization of dialogue acts. Agents have to be able to make use of the rich lexical, prosodic, and grammatical cues to interpretation. But the breadth and coverage of this model come at the expense of depth; current algorithms are able to model only very simplistic and local heuristics for cues.

As I mentioned earlier, I chose speech act interpretation as the topic for this chapter because I think of it as a touchstone task. Thus this same dialectic between logical models based on knowledge-based reasoning and probabilistic models based on statistical interpretation applies in other computational pragmatic areas like reference resolution and discourse structure interpretation.

This dialectic is also important for the field of linguistics as well. While linguistics has traditionally embraced the symbolic, structural, and philosophical paradigm implicit in the BDI model, it has only recently begun to flirt with the probabilistic paradigm. The cue-based model shows one way in which the probabilistic paradigm can inform our understanding of the relationship between linguistic form and linguistic function.

It is clear that both these models of computational pragmatics are in their infancy. I expect significant progress in both areas in the near future, and I look forward to a comprehensive and robust integration of the two methods.

6 Acknowledgements

Work on this chapter was partially supported by NSF CAREER-IIS-9733067. I'm very grateful to the editors and to Andy Kehler, Laura Michaelis, Barbara Fox, Brad Miller, and Jim Martin for advice on and discussion about this material.

Notes

[1]ATIS, or Air Travel Information System, is a corpus of sentences spoken to a computerized speech understanding system for planning air travel. It is available as part of the Penn Treebank Project (Marcus *et al.* 1999).

References

ALEXANDERSSON, JAN, and NORBERT REITHINGER. 1997. Learning dialogue structures from a corpus. EUROSPEECH-97, volume 4, 2231–2234.

ALLEN, JAMES. 1995. Natural Language Understanding. Menlo Park, CA: Benjamin Cummings.

——, and MARK CORE, 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

——, and C. RAYMOND PERRAULT. 1980. Analyzing intention in utterances. Artificial Intelligence 15.143–178.

ALLWOOD, JENS. 1995. An activity-based approach to pragmatics. Gothenburg Papers in Theoretical Linguistics 76.

——, JOAKIM NIVRE, and ELISABETH AHLSÉN. 1992. On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9.1–26.

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 34.351–366.

BUNT, HARRY. 1994. Context and dialogue control. Think 3.19–31.

——. 1995. Dynamic interpretation and dialogue theory. The structure of multimodal dialogue, ed. by M. M. Taylor, F. Neel, and D. G. Bouwhuis. Amsterdam: John Benjamins.

——, and BILL BLACK. 2000. The ABC of computational pragmatics. Computational pragmatics: Abduction, belief and context, ed. by Harry C. Bunt and William Black. Amsterdam: John Benjamins.

CARLETTA, J., N. DAHLBÄCK, N. REITHINGER, and M. A. WALKER. 1997a. Standards for dialogue coding in natural language processing. Technical Report Report no. 167, Dagstuhl Seminars. Report from Dagstuhl seminar number 9706.

CARLETTA, JEAN, AMY ISARD, STEPHEN ISARD, JACQUELINE C. KOWTKO, GWYNETH DOHERTY-SNEDDON, and ANNE H. ANDERSON. 1997b. The reliability of a dialogue structure coding scheme. Computational Linguistics 23.13–32.

CHU-CARROLL, JENNIFER. 1998. A statistical model for discourse act recognition in dialogue interactions. Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium. Tech. rep. SS-98-01, ed. by Jennifer Chu-Carroll and Nancy Green, 12–17. AAAI Press, Menlo Park, CA.

CLARK, HERBERT H., and EDWARD F. SCHAEFER. 1989. Contributing to discourse. Cognitive Science 13.259–294.

COHEN, PHILIP R., and C. RAYMOND PERRAULT. 1979. Elements of a plan-based theory of speech acts. Cognitive Science 3.177–212.

CORE, MARK, MASATO ISHIZAKI, JOHANNA D. MOORE, CHRISTINE NAKATANI, NORBERT REITHINGER, DAVID TRAUM, and SYUN TUTIYA.

1999. The report of the third workshop of the Discourse Resource Initiative, Chiba University and Kazusa Academia Hall. Technical Report No.3 CC-TR-99-1, Chiba Corpus Project, Chiba, Japan.

CURL, TRACI S., and ALAN BELL, 2001. Yeah, yeah, yeah: Prosodic differences of pragmatic functions. Submitted manuscript.

DALY, NANCY A., and VICTOR W. ZUE. 1992. Statistical and linguistic analyses of F_0 in read and spontaneous speech. Proceedings of the International Conference on Spoken Language Processing (ICSLP-92), volume 1, 763–766.

FIKES, RICHARD E., and NILS J. NILSSON. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. Artificial Intelligence 2.189–208.

GAZDAR, G. 1981. Speech act assignment. Elements of discourse understanding, ed. by Aravind Joshi, Bonnie Webber, and Ivan Sag. Cambridge, MA: Cambridge University Press.

GODFREY, J., E. HOLLIMAN, and J. MCDANIEL. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing (IEEE ICASSP-92), 517–520, San Francisco. IEEE.

GOODWIN, CHARLES. 1996. Transparent vision. Interaction and grammar, ed. by Elinor Ochs, Emanuel A. Schegloff, and Sandra A. Thompson. Cambridge: Cambridge University Press.

GORDON, DAVID, and GEORGE LAKOFF. 1971. Conversational postulates. CLS-71, 200–213. University of Chicago. Reprinted in Peter Cole and Jerry L. Morgan (Eds.), *Speech Acts: Syntax and Semantics Volume 3*, Academic, 1975.

GRICE, H. P. 1957. Meaning. Philosophical Review 67.377–388. Reprinted in *Semantics*, edited by D. D. Steinberg & L. A. Jakobovits (1971), Cambridge University Press, pages 53–59.

HINTIKKA, JAAKO. 1969. Semantics for propositional attitudes. Philosophical logic, ed. by J. W. Davis, D. J. Hockney, and W. K. Wilson, 21–45. Dordrecht, Holland: D. Reidel.

JEKAT, S., A. KLEIN, E. MAIER, E. MALECK, M. MAST, and J. QUANTZ, 1995. Dialogue Acts in VERBMOBIL verbmobil–report–65–95.

JURAFSKY, DANIEL, REBECCA BATES, NOAH COCCARO, RACHEL MARTIN, MARIE METEER, KLAUS RIES, ELIZABETH SHRIBERG, ANDREAS STOLCKE, PAUL TAYLOR, and CAROL VAN ESS-DYKEMA. 1997a. Automatic detection of discourse structure for speech recognition and understanding. Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding, 88–95, Santa Barbara.

JURAFSKY, DANIEL, REBECCA BATES, NOAH COCCARO, RACHEL MARTIN, MARIE METEER, KLAUS RIES, ELIZABETH SHRIBERG, ANDREAS STOLCKE, PAUL TAYLOR, and CAROL VAN ESS-DYKEMA. 1998a. Switchboard discourse language modeling project report. Research Note 30, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD.

JURAFSKY, DANIEL, and JAMES H. MARTIN. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

JURAFSKY, DANIEL, ELIZABETH SHRIBERG, and DEBRA BIASCA. 1997b. Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Technical Report 97-02, University of Colorado Institute of Cognitive Science. Also available as `http://www.colorado.edu/ling/jurafsky/manual.august1.html`.

JURAFSKY, DANIEL, ELIZABETH E. SHRIBERG, BARBARA FOX, and TRACI CURL. 1998b. Lexical, prosodic, and syntactic cues for dialog acts. In Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers, 114–120. ACL.

——, ——, ——, and ——. 1998c. Lexical, prosodic, and syntactic cues for dialog acts. In Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers, 114–120. ACL.

KEHLER, ANDY. 2000. Chapter 18: Discourse. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, ed. by Daniel Jurafsky and James H. Martin, 669–718. Prentice Hall.

KIESSLING, A., R. KOMPE, H. NIEMANN, E. NÖTH, and A. BATLINER. 1993. "Roger", "Sorry", "I'm still listening": Dialog guiding signals in informational retrieval dialogs. ESCA Workshop on Prosody, ed. by D. House and P. Touati, 140–143, Lund, Sweden.

KOMPE, R., A. KIESSLING, T. KUHN, M. MAST, H. NIEMANN, E. NÖTH, K. OTT, and A. BATLINER. 1993. Prosody takes over: A prosodically guided dialog system. EUROSPEECH-93, volume 3, 2003–2006, Berlin.

KOMPE, R., A. KIESSLING, H. NIEMANN, E. NÖTH, E. G. SCHUKAT-TALAMAZZINI, , A. ZOTTMANN, and A. BATLINER. 1995. Prosodic scoring of word hypothesis graphs. EUROSPEECH-95, 1333–1336.

LABOV, WILLIAM, and DAVID FANSHEL. 1977. Therapeutic Discourse. New York: Academic Press.

LEVINSON, STEPHEN C. 1983. Pragmatics. Cambridge: Cambridge University Press.

MACWHINNEY, B., E. BATES, and R. KLIEGL. 1984. Cue validity and sentence interpretation in English, German, and Italian. Journal of Verbal Learning and Verbal Behavior 23.127–150.

MACWHINNEY, BRIAN. 1987. The competition model. Mechanisms of language acquisition, ed. by Brian MacWhinney, 249–308. Hillsdale, NJ: Lawrence Erlbaum.

MARCUS, MITCHELL P., BEATRICE SANTORINI, MARY ANN MARCINKIEWICZ, and ANN TAYLOR. 1999. Treebank-3. Linguistic Data Consortium (LDC). Catalog #LDC99T42.

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. Proceedings of the International Conference on Spoken Language Processing (ICSLP-96), volume 3, 1732–1735, Philadelphia, PA.

MICHAELIS, LAURA A. 2001. Exclamative constructions. Language typology and universals: An international handbook, ed. by Martin Haspelmath *et al.* Berlin: Walter de Gruyter.

——, and KNUD LAMBRECHT. 1996. Toward a construction-based theory of language function: The case of nominal extraposition. Language 72.215–247.

NAGATA, MASAAKI. 1992. Using pragmatics to rule out recognition errors in cooperative task-oriented dialogues. Proceedings of the International Conference on Spoken Language Processing (ICSLP-92), 647–650, Banff, Canada.

——, 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.

PERRAULT, C. RAYMOND, and JAMES ALLEN. 1980. A plan-based analysis of indirect speech acts. American Journal of Computational Linguistics 6.167–182.

PIERREHUMBERT, J., and J. HIRSCHBERG. 1990. The meaning of intonational contours in the interpretation of discourse. Intentions in communication, ed. by P. R. Cohen, J. Morgan, and M. Pollack, 271–311. Cambridge, MA: MIT Press.

PIERREHUMBERT, JANET, 1980. The Phonology and Phonetics of English Intonation. MIT dissertation.

POWER, R. 1979. The organization of purposeful dialogs. Linguistics 17.105–152.

QUIRK, RANDOLPH, SIDNEY GREENBAUM, GEOFFREY LEECH, and JAN SVARTVIK. 1985. A Comprehensive Grammar of the English Language. London: Longman.

REITHINGER, NORBERT, and MARTIN KLESEN. 1997. Dialogue act classification using language models. EUROSPEECH-97, volume 4, 2235–2238.

SACKS, HARVEY, EMANUEL A. SCHEGLOFF, and GAIL JEFFERSON. 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50.696–735.

SADOCK, JERROLD M., and ARNOLD M. ZWICKY. 1985. Speech act distinctions in syntax. Language typology and syntactic description, volume 1, ed. by Timothy Shopen, 155–196. Cambridge University Press.

SAG, IVAN A., and MARK LIBERMAN. 1975. The intonational disambiguation of indirect speech acts. Cls-75, 487–498. University of Chicago.

SAMUEL, KEN, SANDRA CARBERRY, and K. VIJAY-SHANKER. 1998. Dialogue act tagging with transformation-based learning. COLING/ACL-98, volume 2, 1150–1156, Montreal. ACL.

SCHEGLOFF, EMANUEL A. 1968. Sequencing in conversational openings. American Anthropologist 70.1075–1095.

—— 1988. Presequences and indirection: Applying speech act theory to ordinary conversation. Journal of Pragmatics 12.55–62.

——, GAIL JEFFERSON, and HARVEY SACKS. 1977. The preference for self-correction in the organization of repair in conversation. Language 53.361–382.

SEARLE, JOHN R. 1975. Indirect speech acts. Speech acts: Syntax and semantics volume 3, ed. by Peter Cole and Jerry L. Morgan, 59–82. New York: Academic Press.

SHRIBERG, ELIZABETH, REBECCA BATES, PAUL TAYLOR, ANDREAS STOLCKE, 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 (Special Issue on Prosody and Conversation) 41.439–487.

STOLCKE, ANDREAS, KLAUS RIES, NOAH COCCARO, ELIZABETH SHRIBERG, REBECCA BATES, DANIEL JURAFSKY PAUL TAYLOR, RACHEL MARTIN, MARIE METEER, and CAROL VAN ESS-DYKEMA. 2000. Dialog act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26.339–371.

STOLCKE, ANDREAS, E. SHRIBERG, R. BATES, N. COCCARO, D. JURAFSKY, R. MARTIN, M. METEER, K. RIES, P. TAYLOR, and C. VAN ESS-DYKEMA. 1998. Dialog act modeling for conversational speech. Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium. Tech. rep. SS-98-01, ed. by Jennifer Chu-Carroll and Nancy Green, 98–105, Stanford, CA. AAAI Press.

SUHM, B., and A. WAIBEL. 1994. Toward better language models for spontaneous speech. Proceedings of the International Conference on Spoken Language Processing (ICSLP-94), volume 2, 831–834.

SWINNEY, DAVID A., and ANNE CUTLER. 1979. The access and processing of idiomatic expressions. Journal of Verbal Learning and Verbal Behavior 18.523–534.

TAYLOR, PAUL, SIMON KING, STEPHEN ISARD, and HELEN WRIGHT. 1998. Intonation and dialog context as constraints for speech recognition. Language and Speech 41.489–508.

——, ——, ——, ——, and JACQUELINE KOWTKO. 1997. Using intonation to constrain language models in speech recognition. EUROSPEECH-97, 2763–2766, Rhodes, Greece.

WAIBEL, A. 1988. Prosody and Speech Recognition. San Mateo, CA: Morgan Kaufmann.

WALKER, MARILYN A., ELISABETH MAIER, JAMES ALLEN, JEAN CARLETTA, SHERRI CONDON, GIOVANNI FLAMMIA, JULIA HIRSCHBERG, STEVE ISARD, MASATO ISHIZAKI, LORI LEVIN, SUSANN LUPERFOY, DAVID TRAUM, and STEVE WHITTAKER, 1996. Penn multiparty standard coding scheme: Draft annotation manual. `www.cis.upenn.edu/~ircs/discourse-tagging/newcoding.html`.

WARD, GREGORY, and JULIA HIRSCHBERG. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. Language 61.747–776.

——, and ——. 1988. Intonation and propositional attitude: The pragmatics of l*+h l h Proceedings of the Fifth Eastern States Conference on Linguistics, 512–522, Berkeley, CA.

WARNKE, V., R. KOMPE, H. NIEMANN, and E. NÖTH. 1997. Integrated dialog act segmentation and classification using prosodic features and language models. EUROSPEECH-97, volume 1, 207–210.

WEBER, ELIZABETH G. 1993. Varieties of Questions in English Conversation. Amsterdam: John Benjamins.

WOSZCZYNA, M., and A. WAIBEL. 1994. Inferring linguistic structure in spoken language. Proceedings of the International Conference on Spoken Language Processing (ICSLP-94), 847–850, Yokohama, Japan.

YOSHIMURA, TAKASHI, SATORU HAYAMIZU, HIROSHI OHMURA, and KAZUYO TANAKA. 1996. Pitch pattern clustering of user utterances in human-machine dialogue. Proceedings of the International Conference on Spoken Language Processing (ICSLP-96), volume 2, 837–840, Philadelphia, PA.