Affirmative Cue Words in Task-Oriented Dialogue

Agustín Gravano*
Universidad de Buenos Aires

Julia Hirschberg**
Columbia University

Štefan Beňuš†
Constantine the Philosopher University
and Institute of Informatics,
Slovak Academy of Sciences

We present a series of studies of affirmative cue words—a family of cue words such as "okay" or "alright" that speakers use frequently in conversation. These words pose a challenge for spoken dialogue systems because of their ambiguity: They may be used for agreeing with what the interlocutor has said, indicating continued attention, or for cueing the start of a new topic, among other meanings. We describe differences in the acoustic/prosodic realization of such functions in a corpus of spontaneous, task-oriented dialogues in Standard American English. These results are important both for interpretation and for production in spoken language applications. We also assess the predictive power of computational methods for the automatic disambiguation of these words. We find that contextual information and final intonation figure as the most salient cues to automatic disambiguation.

1. Introduction

CUE PHRASES are linguistic expressions that may be used to convey explicit information about the discourse or dialogue, or to convey a more literal, semantic contribution. They aid speakers and writers in organizing the discourse, and listeners and readers in processing it. In previous literature, these constructions have also been termed discourse markers, pragmatic connectives, discourse operators, and clue words. Examples of cue phrases include *now*, *well*, *so*, *and*, *but*, *then*, *after all*, *furthermore*, *however*, *in consequence*, *as a matter of fact*, *in fact*, *actually*, *okay*, *alright*, *for example*, and *incidentally*.

The ability to correctly determine the function of cue phrases is critical for important natural language processing tasks, including anaphora resolution (Grosz and Sidner 1986), argument understanding (Cohen 1984), plan recognition (Grosz and Sidner 1986; Litman and Allen 1987), and discourse segmentation (Litman and Passonneau 1995).

* Departamento de Computación, FCEyN, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria, (C1428EGA) Buenos Aires, ARGENTINA. E-mail: gravano@dc.uba.ar.
** E-mail: julia@cs.columbia.edu.
† E-mail: sbenus@ukf.sk.

Submission received: 9 December 2009; revised submission received: 2 February 2011; accepted for publication: 13 March 2011.

Furthermore, correctly determining the function of cue phrases using features of the surrounding text can be used to improve the naturalness of synthetic speech in text-to-speech systems (Hirschberg 1990).

In this study, we focus on a subclass of cue phrases that we term **affirmative cue words** (hereafter, ACWs), and that include *alright*, *mm-hm*, *okay*, *right*, and *uh-huh*, inter alia. These words are frequent in spontaneous conversation, especially in task-oriented dialogue, and are heavily overloaded: Their possible discourse/pragmatic functions include agreeing with what the interlocutor has said, displaying interest and continued attention, and cueing the start of a new topic. Some ACWs (e.g., *alright*, *okay*) are capable of conveying as many as ten different functions, as described in Section 3. Whereas ACWs thus form a subset of more general classes of utterances which have been studied in more general studies of *cue words*, *cue phrases*, *discourse markers*, *feedback utterances*, *linguistic feedback*, *acknowledgments*, *grounding acts*, our focus is on this particular subset of lexical items which *may* convey an affirmative response—but which may also convey many different meanings. The disambiguation of these meanings we believe is critical to the success of spoken dialogue systems.

In the studies presented here, our goal is to extend our understanding of ACWs, in particular by finding descriptions of the acoustic/prosodic characteristics of their different functions, and by assessing the predictive power of computational methods for their automatic disambiguation. This knowledge should be helpful in spoken language generation and understanding tasks, including interactive spoken dialogue systems and applications doing off-line analyses of conversational data, such as meeting segmentation and summarization. For example, spoken dialogue systems lacking a model of the appropriate realization of different uses of these words are likely to have difficulty in understanding and communicating with their users, either by producing cue phrases in a way that does not convey the intended meaning or by misunderstanding users' productions.

This article is organized as follows. Section 2 reviews previous literature. In Section 3 we describe the materials used in the present study from the Columbia Games Corpus. Section 4 presents a statistical description of the acoustic, prosodic, and contextual characteristics of the functions of ACWs in this corpus. In Section 5 we describe results from a number of machine learning experiments aimed at investigating how accurately ACWs may be automatically classified into their various functions. Finally, in Section 6 we summarize and discuss our main findings.

2. Previous Work

Cue phrases have received extensive attention in the computational linguistics literature. Early work by Cohen (1984) presents a computational justification for the importance of cue phrases in discourse processing. Using a simple propositional framework for analyzing discourse, Cohen claims that, in some cases, cue phrases decrease the number of operations required by the listener to process "coherent transmissions"; in other cases, cue phrases are necessary to allow the recognition of "transmissions which would be incoherent (too complex to reconstruct) in the absence of clues" (page 251). Reichman (1985) proposes a model of discourse structure in which discourse comprises a collection of basic constituents called **context spaces**, organized hierarchically according to semantic and logical relations called **conversational moves**. In Reichman's model, cue phrases are portrayed as mechanisms that signal context space boundaries, specifying the kind of conversational move about to take place. Grosz and Sidner (1986) introduce an alternative model of discourse structure formed by three interrelated

components: a linguistic structure, an intentional structure, and an attentional state. In this model, cue phrases play a central role, allowing the speaker to provide information about all of the following to the listener:

> 1) that a change of attention is imminent; 2) whether the change returns to a previous focus space or creates a new one; 3) how the intention is related to other intentions; 4) what precedence relationships, if any, are relevant (page 196).

In a corpus study of spontaneous conversations, Schiffrin (1987) describes cue phrases as syntactically detachable from a sentence, commonly used in initial position within utterances, capable of operating at both local and global levels of discourse, and having a range of prosodic contours. As other authors, Schiffrin observes that cue phrases provide contextual coordinates for an utterance in the discourse—that is, they indicate the discourse segment to which an utterance belongs. However, she suggests that cue phrases only *display* discourse structure relations; they do not create them. In a critique of Schiffrin's work, Redeker (1991) proposes defining cue phrases as phrases "uttered with the primary function of bringing to the listener's attention a particular kind of linkage of the upcoming utterance with the immediate discourse context" (page 1169).

Prior work on the automatic classification of cue phrases includes a series of studies performed by Hirschberg and Litman (Hirschberg and Litman 1987, 1993; Litman and Hirschberg 1990), which focus on differentiating between the **discourse** and **sentential** senses of single-word cue phrases such as *now*, *well*, *okay*, *say*, and *so* in American English. When used in a discourse sense, a cue phrase explicitly conveys information about the discourse structure; when used in a sentential sense, a cue phrase instead conveys semantic information. Hirschberg and Litman present two manually developed classification models, one based on prosodic features, and one based on textual features. This line of research is further pursued by Litman (1994, 1996), who incorporates machine learning techniques to derive classification models automatically. Litman uses different combinations of prosodic and text-based features to train decision-tree and rule learners, and shows that machine learning constitutes a powerful tool for developing automatic classifiers of cue phrases into their sentential and discourse uses. Zufferey and Popescu-Belis (2004) present a similar study on the automatic classification of *like* and *well* into discourse and sentential senses, achieving a performance close to that of human annotators.

Besides the binary division of cue phrases into discourse vs. sentential meanings, the Conversational Analysis (CA) literature describes items it terms **linguistic feedback** or **acknowledgments**. These include not only the computational linguists' cue phrases but also expressions such as *I see* or *oh wow*, which CA research describes in terms of attention, understanding, and acceptance by the speaker of a proposition uttered by another conversation participant (Kendon 1967; Yngve 1970; Duncan 1972; Schegloff 1982; Jefferson 1984). Such items typically occur at the second position in common **adjacency pairs** and include **backchannels** (also referred to as **continuers**), which "exhibit on the part of [their] producer an understanding that an extended unit of talk is underway by another, and that it is not yet, or may not be (even ought not yet be) complete; [they take] the stance that the speaker of that extended unit should continue talking" (Schegloff 1982, page 81), and **agreements**, which indicate the speaker's agreement with a statement or opinion expressed by another speaker. Allwood, Nivre, and Ahlsen (1992) distinguish four basic communicative functions of linguistic feedback which enable conversational partners to exchange information: **contact**, **perception**, **understanding**, and **attitudinal reactions**. These correspond respectively

to whether the interlocutor is willing and able to continue the interaction, perceive the message, understand the message, and react and respond to the message. Allwood, Nivre, and Ahlsen posit that "simple feedback words, like *yes*, [...] involve a high degree of context dependence" (page 5), and suggest that their basic communicative function strongly depends on the type of speech act, factual polarity, and information status of the immediately preceding communicative act. Novick and Sutton (1994) propose an alternative categorization of linguistic feedback in task-oriented dialogue, which is based on the structural context of exchanges rather than on the characteristics of the preceding utterance. The three main classes in Novick and Sutton's catalogue are: (i) *other* $\rightarrow$ *ackn*, where an acknowledgment immediately follows a contribution by *other* speaker; (ii) *self* $\rightarrow$ *other* $\rightarrow$ *ackn*, where *self* initiates an exchange, *other* eventually completes it, and *self* utters an acknowledgment; and (iii) *self* + *ackn*, where *self* includes an acknowledgment in an utterance independently of *other*'s previous contribution.

Substantial attention has been paid to subsets and supersets of words we include in our class of ACWs in the psycholinguistic literature in studies of **grounding**—the process by which conversants obtain and maintain a **common ground** of mutual knowledge, mutual beliefs, and mutual assumptions over the course of a conversation (Clark and Schaefer 1989; Clark and Brennan 1991). Computational work on grounding has been pursued for a number of years by Traum and colleagues (e.g., Traum and Allen 1992; Traum 1994), who recently have described a corpus-based study of lexical and semantic evidence supporting different degrees of grounding (Roque and Traum 2009).

Our ACWs often occur in the process of establishing such common ground. Prosodic characteristics of the responses involved in grounding have been studied in the Australian English Map Task corpus by Mushin et al. (2003), who find that these utterances often consist of acknowledgment contributions such as *okay* or *yeh* produced with a "non-final" intonational contour, and followed by speech by the same speaker which appears to continue the intonational phrase. Studies by Walker of **informationally redundant utterances** (IRUs) (Walker 1992, 1996), utterances which express "a proposition already entailed, presupposed or implicated by a previous utterance in the same discourse situation" (Walker 1993a, page 12), also include some of our ACWs, such as IRU **prompts** (e.g., *uh-huh*), which, according to Walker, "add no new propositional content to the common ground" (Walker 1993a, page 32). Walker adopts the term "continuer" from the Conversational Analysis school to further describe these prompts (Walker 1993a). Walker describes some intonational contours which are used to realize IRUs in generation in Walker (1993a) and in Walker (1993b), examining 63 IRU tokens and finding five different types of contour used among them.

As part of a larger project on automatically detecting discourse structure for speech recognition and understanding tasks in American English, Jurafsky et al. (1998) present a study of four particular discourse/pragmatic functions, or **dialog acts** (Bunt 1989; Core and Allen 1997), closely related to ACWs: backchannel, agreement, **incipient speakership** (indicating an intention to take the floor), and **yes-answer** (affirmative answer to a *yes–no* question). The authors examine 1,155 conversations from the Switchboard database (Godfrey, Holliman, and McDaniel 1992), and report that the vast majority of these four dialogue acts are realized with words like *yeah*, *okay*, or *uh-huh*. They find that the lexical realization of the dialogue act is the strongest cue to its identity (e.g., backchannel is the preferred function for *uh-huh* and *mm-hm*), and report preliminary results on some prosodic differences across dialogue acts: Backchannels are shorter in duration, have lower pitch and intensity, and are more likely to end in a rising intonation than agreements. Two related studies, part of the same project, address the automatic classification of dialogue acts in conversational speech (Shriberg et al. 1998; Stolcke

et al. 2000). The results of their machine learning experiments, conducted on the same subset of Switchboard used previously, indicate a high degree of confusion between agreements and backchannels, because both classes share words such as *yeah* and *right*. They also show that prosodic features (including duration, pause, and intensity) can aid the automatic disambiguation between these two classes: A classifier trained using both lexical and prosodic features slightly yet significantly outperforms one trained using just lexical features.

There is also considerable evidence that linguistic feedback does not take place at arbitrary locations in conversation; rather, it mostly occurs at or near **transition-relevance places** for turn-taking (Sacks, Schegloff, and Jefferson 1974; Goodwin 1981). Ward and Tsukahara (2000) describe, in both Japanese and American English, a region of low pitch lasting at least 110 msec which may function as a prosodic cue inviting the realization of a backchannel response from the interlocutor. In a corpus study of Japanese dialogues, Koiso et al. (1998) find that both syntax and prosody play a central role in predicting the occurrence of backchannels. Cathcart, Carletta, and Klein (2003) propose a method for automatically predicting the placement of backchannels in Scottish English conversation, based on pause durations and part-of-speech tags, that outperforms a random baseline model. Recently, Gravano and Hirschberg (2009a, 2009b, 2011) describe six distinct prosodic, acoustic, and lexical events in American English speech that tend to precede the occurrence of a backchannel by the interlocutor.

Despite their high frequency in spontaneous conversation, the set of ACWs we examine here have seldom, if ever, been an object of study in themselves, as a separate subclass of cue phrases or dialogue acts. Some have attempted to model other types of cue phrases (e.g., *well*, *like*) or cue phrases in general; others discuss discourse/pragmatic functions that may be conveyed through ACWs, but which may also be conveyed through other types of expressions (e.g., agreements may be communicated by single words such as *yes* or longer cue phrases such as *that's correct*). Subsets of ACWs have been studied in very small corpora, with some proposals about their prosodic and functional variations. For example, Hockey (1993) examines the prosodic variation of two ACWs, *okay* and *uh-huh* (66 and 77 data points, respectively) produced as full intonational phrases in two spontaneous task-oriented dialogues. She groups the F_0 contours visually and auditorily, and shows that instances of *okay* produced with a high-rise contour are significantly more likely to be followed by speech from the other speaker than from the same speaker. The results of a perception experiment conducted by Gravano et al. (2007) suggest that, in task-oriented American English dialogue, contextual information (e.g., duration of surrounding silence, number of surrounding words) as well as word-final intonation figure as the most salient cues to disambiguation of the function of the word *okay* by human listeners. Also, in a study of the function of intonation in Scottish English task-oriented dialogue, Kowtko (1996) examines a corpus of 273 instances of single-word utterances, including affirmative cue words such as *mm-hm*, *okay*, *right*, *uh-huh*, and *yes*. Kowtko finds a significant correlation between discourse function and intonational contour: The **align** function (which checks that the listener's understanding aligns with that of the speaker) is shown to correlate with rising intonational contours; the **ready** function (which cues the speaker's intention to begin a new task) and the **reply-y** function (which "has an affirmative surface and usually indicates agreement"; Kowtko 1996, page 59) correlate with a non-rising intonation; and the **acknowledge** function (which indicates having heard and understood) presents all types of final intonation. It is important to note, however, that different dialects and different languages have distinct ways of realizing different discourse/pragmatic functions, so it is unclear how useful these results are for American English.

Although broader studies focusing on the pragmatic function of cue phrases, discourse markers, linguistic feedback, and dialogue acts do shed light on the particular subset of utterances we are studying, and although there is some information on particular lexical items we include here in our study, the class of ACWs itself has received little attention. Particularly given the frequency of ACWs in dialogue, it is important to identify reliable and automatically extractable cues to their disambiguation, so that spoken dialogue systems can recognize the pragmatic function of ACWs in user input and can produce ACWs that are less likely to be misinterpreted in system output.

3. Materials

The materials for all experiments in this study were taken from the Columbia Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from 13 native speakers (6 female, 7 male) of Standard American English (SAE). A detailed description of this corpus is given in Appendix A. In each session, two subjects were paid to play a series of computer games requiring verbal communication to achieve joint goals of identifying and moving images on the screen. Each subject used a separate laptop computer; they sat facing each other in a soundproof booth, with an opaque curtain hanging between to allow only verbal communication.

Each session contains an average of 45 minutes of dialogue, totaling roughly 9 hours of dialogue in the corpus. Trained annotators orthographically transcribed the recordings and manually aligned the words to the speech signal, yielding a total of 70,259 words and 2,037 unique words in the corpus. Additionally, self repairs and certain non-word vocalizations were marked, including laughs, coughs, and breaths. For roughly two thirds of the corpus, intonational patterns and other aspects of the prosody were identified by trained annotators using the ToBI transcription framework (Beckman and Hirschberg 1994; Pitrelli, Beckman, and Hirschberg 1994).

3.1 Affirmative Cue Words in the Games Corpus

Throughout the Games Corpus, subjects made frequent use of **affirmative cue words**: The 5,456 instances of affirmative cue words *alright*, *gotcha*, *huh*, *mm-hm*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yes*, and *yup* account for 7.8% of the total words in the corpus. Because the usage of these words seems to vary significantly in meaning, we asked three labelers to independently classify all occurrences of these 11 words in the entire corpus into the ten discourse/pragmatic functions listed in Table 1.

Among the distinctions we make in these pragmatic functions, we note particularly that our categories of **Agr** and **BC** differ primarily in that **Agr** is defined as indicating belief in or agreement with the interlocutor (e.g., a response to a *yes–no* question), whereas **BC** indicates only continued attention.[1]

1 Our definition of **BC** is similar to definitions of backchannel and continuer as discussed by a number of authors in the Conversational Analysis and spoken language processing communities (e.g., Stolcke et al.'s [2000] "a short utterance that plays discourse structuring roles, e.g., indicating that the speaker should go on talking" [page 345]; and Cathcart et al.'s [2003] "utterances, with minimal content, used to clearly signal that the speaker should continue with her current turn" [page 51]). Although some definitions of **BC** also include the notion that the speaker is indicating understanding, we did not ask annotators to make this distinction. We note further that, although it is also possible (cf. Clark and Schaefer 1989) to signal understanding without agreement, in the process of designing the labeling scheme we did not find instances of ACWs that seemed to us to have this function in our corpus; nor did our labelers find such cases. Hence we did not include this distinction among our classes.

Table 1
Labeled discourse/pragmatic functions of affirmative cue words.

Agr	**Agreement.** Indicates *I believe what you said*, and/or *I agree with what you say*.
BC	**Backchannel.** Indicates only *I hear you and please continue*, in response to another speaker's utterance.
CBeg	**Cue beginning discourse segment.** Marks a new segment of a discourse or a new topic.
CEnd	**Cue ending discourse segment.** Marks the end of a current segment of a discourse or a current topic.
PBeg	**Pivot beginning (Agr+CBeg).** Functions both to agree and to cue a beginning segment.
PEnd	**Pivot ending (Agr+CEnd).** Functions both to agree and to cue the end of the current segment.
Mod	**Literal modifier.** Examples: *I think that's okay; to the right of the lion*.
BTsk	**Back from a task.** Indicates *I've just finished what I was doing and I'm back*.
Chk	**Check.** Used with the meaning *Is that okay?*
Stl	**Stall.** Used to stall for time while keeping the floor.
?	Cannot decide.

Labelers were given examples of each category, and annotated with access to both transcript and speech source. The guidelines used by the annotators are presented in Appendix B. Appendix C includes some examples of each class of ACWs, as labeled by our annotators. Inter-labeler reliability was measured by Fleiss's κ (Fleiss 1971) as Substantial at 0.745.[2] We define the **majority label** of a token as the label chosen for that token by at least two of the three labelers; we assign the "?" label to a token either when its majority label is "?", or when it was assigned a different label by each labeler. Of the 5,456 affirmative cue words in the corpus, 5,185 (95%) have a majority label other than "?." Table 2 shows the distribution of discourse/pragmatic functions over ACWs in the whole corpus.

3.2 Data Downsampling

Some of the word/function pairs in Table 2 are skewed to contributions from a few speakers. For example, for backchannel (**BC**) *uh-huh*, as many as 65 instances (44%) are from one single speaker, and the remaining 83 are from seven other speakers. In cases like this, using the whole sample would pose the risk of drawing false conclusions on the usage of ACWs, possibly influenced by stylistic properties of individual speakers. Therefore, we downsampled the tokens of ACWs in the Games Corpus to obtain a balanced data set, with instances of each word and function coming in similar proportions from as many speakers as possible. Specifically, we downsampled our data using the following procedure: First, we discarded all word/function pairs with tokens from fewer than four different speakers; second, for each of the remaining word/function pairs, we discarded tokens (at random) from speakers who contributed more than 25% of its tokens. In other words, the resulting data set meets two conditions: For each word/

2 The κ measure of agreement above chance is interpreted as follows: 0 = None, 0–0.2 = Small, 0.2–0.4 = Fair, 0.4–0.6 = Moderate, 0.6–0.8 = Substantial, 0.8–1 = Almost perfect.

Table 2
Distribution of function over ACW. Rest = {*gotcha, huh, yep, yes, yup*}.

	alright	*mm-hm*	*okay*	*right*	*uh-huh*	*yeah*	Rest	Total
Agr	76	58	1,092	111	18	754	116	2,225
BC	6	395	120	14	148	69	5	757
CBeg	83	0	543	2	0	2	0	630
CEnd	6	0	6	0	0	0	0	12
PBeg	4	0	65	0	0	0	0	69
PEnd	11	12	218	2	0	20	15	278
Mod	5	0	18	1,069	0	0	0	1,092
BTsk	7	1	32	0	0	0	0	40
Chk	1	0	6	49	0	1	6	63
Stl	1	0	15	1	0	2	0	19
?	36	12	150	10	3	55	5	271
Total	236	478	2,265	1,258	169	903	147	5,456

function pair, (a) tokens come from at least four different speakers, and (b) no single speaker contributes more than 25% of the tokens. The two thresholds were found via a grid search, and were chosen as a trade-off between size and representativeness of the data set. With this procedure we discarded 506 tokens of ACWs, or 9.3% of such words in the corpus. Table 3 shows the resulting distribution of discourse/pragmatic functions over ACWs in the whole corpus after downsampling the data. The κ measure of inter-labeler reliability was practically identical for the downsampled data, at 0.751.

3.3 Feature Extraction

We extracted a number of lexical, discourse, timing, phonetic, acoustic, and prosodic features for each target ACW, which we use in the statistical analysis and machine learning experiments presented in the following sections. Tables 4 through 8 summarize the full feature set. For simplicity, in those tables each line may describe one or more features. Features that may be extracted by on-line applications are marked with letter $\mathcal{O}$; this is further explained later in this section.

Table 3
Distribution of function over ACW, after downsampling. Rest = {*gotcha, huh, yep, yes, yup*}.

	alright	*mm-hm*	*okay*	*right*	*uh-huh*	*yeah*	Rest	Total
Agr	76	58	1,092	74	16	754	87	2,157
BC	0	395	120	0	101	58	0	674
CBeg	61	0	543	0	0	0	0	604
CEnd	0	0	4	0	0	0	0	4
PBeg	0	0	64	0	0	0	0	64
PEnd	10	4	218	0	0	18	0	250
Mod	4	0	18	1,069	0	0	0	1,091
BTsk	5	0	28	0	0	0	0	33
Chk	0	0	5	49	0	0	4	58
Stl	0	0	15	0	0	0	0	15
Total	156	457	2,107	1,192	117	830	91	4,950

Table 4
Lexical and discourse features. Each line may describe one or more features. Features marked $\mathcal{O}$ may be available in on-line conditions.

Lexical features	
$\mathcal{O}$	Lexical identity of the target word (w).
$\mathcal{O}$	Part-of-speech tag of w, original and simplified.
$\mathcal{O}$	Word immediately preceding w, and its original and simplified POS tags. If w is preceded by silence, this feature takes value '#'.
$\mathcal{O}$	Word immediately following w, and its original and simplified POS tags. If w is followed by silence, this feature takes value '#'.
Discourse features	
$\mathcal{O}$	Number of words in w's IPU.
$\mathcal{O}$	Number and proportion of words in w's IPU before and after w.
$\mathcal{O}$	Number of words uttered by the other speaker during w's IPU.
$\mathcal{O}$	Number of words in the previous turn by the other speaker.
	Number of words in w's turn.
	Number and proportion of words and IPUs in w's turn before and after w.
	Number and proportion of turns in w's task before and after w.
	Number of words uttered by the other speaker during w's turn.
	Number of words in the following turn by the other speaker.
	Number of ACWs in w's turn other than w.

Our lexical features consist of the lexical identity and the part-of-speech (POS) tag of the target word (w), the word immediately preceding w, and the word immediately following w (see Table 4). POS tags were labeled automatically for the whole corpus using Ratnaparkhi, Brill, and Church's (1996) maxent tagger trained on a subset of the Switchboard corpus (Charniak and Johnson 2001) in lower-case with all punctuation removed, to simulate spoken language transcripts. Each word had an associated POS tag from the full Penn Treebank tag set (Marcus, Marcinkiewicz, and Santorini 1993), and one of the following simplified tags: noun, verb, adjective, adverb, contraction, or other.

For our discourse features, listed in Table 4, we define an **inter-pausal unit** (IPU) as a maximal sequence of words surrounded by silence longer than 50 msec. A **turn** is a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor.[3,4] Boundaries of IPUs and turns are computed automatically from the time-aligned transcriptions. A **task** in the Games Corpus corresponds to a simple game played by the subjects, requiring verbal communication to achieve a joint goal of identifying and moving images on the screen (see Appendix A for a description of these game tasks). Task boundaries are extracted from the logs collected automatically during the sessions, and subsequently checked by hand. Our discourse features are intended to capture discrete positional information of the target word, in relation to its containing IPU, turn, and task.

3 Here "between" refers strictly to the time after the end point of the former IPU and before the start point of the latter.

4 Note that our operational definition of "turn" here includes all speaker utterances, including backchannels, which are typically not counted as turn-taking behaviors. We use this more inclusive definition of "turn" here to avoid inventing a new term to encompass "turns and backchannels".

Our timing features (Table 5) are intended to capture positional information of a temporal nature, such as the duration (in milliseconds) of *w* and its containing IPU and turn, or the duration of any silence before and after *w*. These features also contain information about the target word relative to the other speaker's speech, including the duration of any overlapping speech, and the latencies between *w*'s conversational turn and the other speaker's preceding and subsequent turns.

Prosody was annotated following the ToBI system (Beckman and Hirschberg 1994; Pitrelli, Beckman, and Hirschberg 1994), which consists of annotations at four time-linked levels of analysis: an *orthographic tier* of time-aligned words; a *tonal tier* describing targets in the fundamental frequency (F0) contour; a *break index tier* indicating degrees of juncture between words; and a *miscellaneous tier*, in which phenomena such as disfluencies may be optionally marked. The tonal tier describes events such as **pitch accents**, which make words intonationally prominent and are realized by increased F0 height, loudness, and duration of accented syllables. A given word may be accented or not and, if accented, may bear different tones, or different degrees of prominence, with respect to other words. Five types of pitch accent are distinguished in the ToBI system for American English: two simple accents **H*** and **L***, and three complex ones, **L*+H**, **L+H***, and **H+!H***. An **L** indicates a low tone and an **H**, a high tone; the asterisk indicates which tone of the accent is aligned with the stressable syllable of the lexical item bearing the accent. Some pitch accents may be **downstepped**, such that the pitch range of the accent is compressed in comparison to a non-downstepped accent. Downsteps are indicated by the '**!**' diacritic (e.g., **!H***, **L+!H***). **Break indices** define two levels of phrasing: Level 3 corresponds to Pierrehumbert's (1980) intermediate phrase and level 4 to Pierrehumbert's intonational phrase. Level 4 phrases consist of one or more level 3 phrases, plus a high or low **boundary tone** (**H%** or **L%**) indicated in the tonal tier at the right edge of the phrase. Level 3 phrases consist of one or more pitch accents, aligned with the stressed syllable of lexical items, plus a **phrase accent**, which also may be high (**H-**) or low (**L-**). For example, a standard declarative contour consists

Table 5
Timing features. Each line may describe one or more features. Features marked $\mathcal{O}$ may be available in on-line conditions.

	Timing features
$\mathcal{O}$	Duration (in msec) of *w* (raw, normalized with respect to all occurrences of the same word by the same speaker, and normalized with respect to all words with the same number of syllables and phonemes uttered by the same speaker).
$\mathcal{O}$	Flag indicating whether there was any overlapping speech from the other speaker.
$\mathcal{O}$	Duration of *w*'s IPU.
$\mathcal{O}$	Latency (in msec) between *w*'s turn and the previous turn by the other speaker.
$\mathcal{O}$	Duration of the silence before *w* (or 0 if the *w* is not preceded by silence), its IPU, and its turn.
$\mathcal{O}$	Duration and proportion of *w*'s IPU elapsed before and after *w*.
$\mathcal{O}$	Duration of *w*'s turn before *w*.
$\mathcal{O}$	Duration of any overlapping speech from the other speaker during *w*'s IPU.
$\mathcal{O}$	Duration of the previous turn by the other speaker.
	Duration of the silence after *w* (or 0 if *w* is not followed by silence), its IPU, and its turn.
	Latency between *w*'s turn and the following turn by the other speaker.
	Duration of *w*'s turn, as a whole and after *w*.
	Duration of any overlapping speech from the other speaker during *w*'s turn.
	Duration of the following turn by the other speaker.

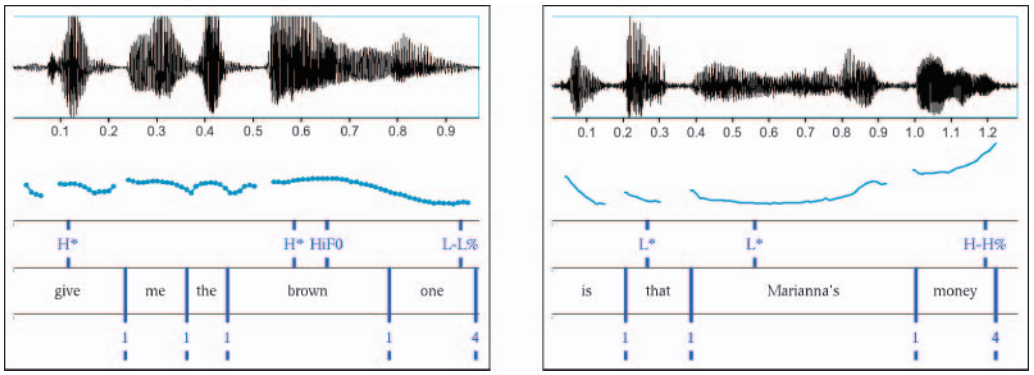

Figure 1
A standard declarative contour (left), and a standard *yes–no* question contour. The top panes show the waveform and the fundamental frequency (F0) track.

of a sequence of **H*** pitch accents ending in a low phrase accent and low boundary tone (**L-L%**); likewise, a standard *yes–no* question contour consists of a sequence of **L*** pitch accents ending in **H-H%**. These are illustrated in Figure 1. In our study, prosodic features include the ToBI labels as specified by the annotators, and also a simplified version of the labels, considering only high and low pitch targets (i.e., **H*** vs. **L*** for pitch accents, **H-** vs. **L-** for phrase accents, and **H%** vs. **L%** for boundary tones), and simplified break indices (0–4). These are listed in Table 6.

All acoustic features were extracted automatically for the whole corpus using the Praat toolkit (Boersma and Weenink 2001). These include pitch, intensity, stylized pitch, ratio of voiced frames to total frames, jitter, shimmer, and noise-to-harmonics ratio (NHR) (see Table 7). Pitch features capture how high the speaker's voice sounds or how low. Intensity is correlated with how loud the speaker sounds to a hearer. The voiced-frames ratio roughly approximates the speaking rate. Jitter and shimmer correspond to variability in the frequency and amplitude of vocal-fold vibration, respectively. NHR is the energy ratio of noise to harmonic components in the voiced speech signal. Jitter, shimmer, and NHR correlate with perceptual evaluations of voice quality, such as harsh, whispery, creaky, and nasalized, inter alia. Pitch slope features capture elements of the intonational contour, and were computed by fitting least-squares linear regression models to the F_0 data points extracted from given portions of the signal, such as a full word or its last 200 msec. This procedure is illustrated in Figure 2, which shows the pitch track of a sample utterance (blue dots) with three linear regressions, computed over the whole utterance (solid black line), and over the final 300 and 200 msec ('A' and 'B' dashed lines, respectively). We used a similar procedure to compute

Table 6
Prosodic features. In all cases, both original and simplified ToBI labels were considered. Each line may describe one or more features. Features marked $\mathcal{O}$ may be available in on-line conditions.

ToBI prosodic features
– Phrase accent, boundary tone, break index, and pitch accent on w.
– Phrase accent, boundary tone, break index, and final pitch accent on the final intonational phrase of the previous turn by the other speaker (these features are defined only when w is turn initial).

Table 7
Acoustic features. Each line may describe one or more features. Features marked $\mathcal{O}$ may be available in on-line conditions.

Acoustic features
$\mathcal{O}$ w's mean, maximum, and minimum pitch and intensity (raw and speaker normalized).
$\mathcal{O}$ Jitter and shimmer, computed over the whole word and over the first and second syllables, computed over just the voiced frames (raw and speaker normalized).
$\mathcal{O}$ Noise-to-harmonics ratio (NHR), computed over the whole word and over the first and second syllables (raw and speaker normalized).
$\mathcal{O}$ w's ratio of voiced frames to total frames (raw and speaker normalized).
$\mathcal{O}$ Pitch slope, intensity slope, and stylized pitch slope, computed over the whole word, its first and second halves, its first and second syllables, the first and second halves of each syllable, and the word's final 100, 200, and 300 msec (raw and normalized with respect to all other occurrences of the same word by the same speaker).
$\mathcal{O}$ w's mean, maximum, and minimum pitch and intensity, normalized with respect to three types of context: w's IPU, w's immediately preceding word by the same speaker, and w's immediately following word by the same speaker.
$\mathcal{O}$ Voiced-frames ratio, jitter, and shimmer, normalized with respect to the same three types of context.
$\mathcal{O}$ Mean, maximum, and minimum pitch and intensity, ratio of voiced frames, (all raw and speaker normalized), jitter, and shimmer, calculated over the final 500, 1,000, 1,500 and 2,000 msec of the previous turn by the other speaker (only defined when w is turn initial but not task initial).
$\mathcal{O}$ Pitch slope, intensity slope, and stylized pitch slope, calculated over the final 100, 200, 300, 500, 1,000, 1,500, and 2,000 msec of the previous turn by the other speaker (only defined when w is turn initial but not task initial).

intensity slopes (which capture changes in perceived loudness) and stylized pitch slopes (which capture more coarse-grained characteristics of the intonational contour). Stylized pitch curves were obtained using the algorithm provided in Praat: Look up the pitch point p that is closest to the straight line L that connects its two neighboring points; if p is further than four semitones away from L, end; otherwise, remove p and start over.

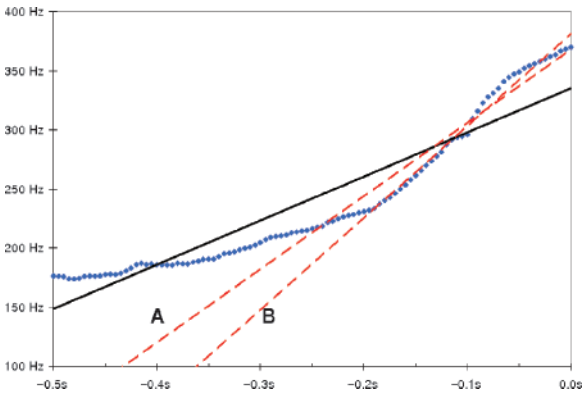

Figure 2
Sample pitch track with three linear regressions: computed over the whole IPU (bold line), and over the final 300 msec (A) and 200 msec (B).

All features related to absolute (i.e., unnormalized) pitch values, such as maximum pitch or final pitch slope, are not comparable across genders because of the different pitch ranges of female and male speakers—roughly 75–500 kHz and 50–300 kHz, respectively. Therefore, before computing those features we applied a linear transformation to the pitch track values, thus making the pitch range of speakers of both genders approximately equivalent. We refer to this process as **gender normalization**. All other normalizations were calculated using z-scores: $z = (x - \mu)/\sigma$, where x is a raw measurement to be normalized (e.g., the duration of a particular word), and μ and σ are the mean and standard deviation of a certain population (e.g., all instances of the same word by the same speaker in the whole conversation).

For our phonetic features (listed in Table 8), we trained an automatic phone recognizer based on the Hidden Markov Model Toolkit (HTK) (Young et al. 2006), using three corpora as training data: the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al. 1993), the Boston Directions Corpus (Hirschberg and Nakatani 1996), and the Columbia Games Corpus. With this recognizer, we obtained automatic time-aligned phonetic transcriptions of each instance of *alright*, *mm-hm*, *okay*, *right*, *uh-huh*, and *yeah* in the corpus. To improve accuracy, we restricted the recognizer's grammar to accept only the most frequent variations of each word, as shown in Table 9. We extracted our phonetic features, such as phone and syllable durations, from the resulting time-aligned phonetic transcriptions. The remaining five ACWs in our corpus (*gotcha*, *huh*, *yep*, *yes*, and *yup*) had too low counts to contain meaningful phonetic variation; thus, we did not compute phonetic features for those words.

Finally, our session-specific features include the session of the Games Corpus in which the target word was produced, along with the identity and gender of both

Table 8
Phonetic and session-specific features. Each line may describe one or more features. Features marked $\mathcal{O}$ may be available in on-line conditions.

Phonetic features	
$\mathcal{O}$	Identity of each of w's phones.
$\mathcal{O}$	Absolute and relative duration of each phone.
$\mathcal{O}$	Absolute and relative duration of each syllable.
Session-specific features	
–	Session number.
–	Identity and gender of both speakers.

Table 9
Restricted grammars for the automatic speech recognizer. Phones in square brackets are optional.

ACW	ARPAbet Grammar
alright	`(aa|ao|ax) r (ay|eh) [t]`
mm-hm	`m hh m`
okay	`[aa|ao|ax|m|ow] k (ax|eh|ey)`
right	`r (ay|eh) [t]`
uh-huh	`(aa|ax) hh (aa|ax)`
yeah	`y (aa|ae|ah|ax|ea|eh)`

speakers (Table 8). These features were solely intended for searching for speaker or dialogue dependencies.

Also, to simulate the conditions of on-line applications, which process speech as it is produced by the user, we distinguish a subset of features that may typically be extracted from the speech signal only up to the IPU containing the target ACW. In Tables 4 through 8 these features are marked with letter $\mathcal{O}$ (for *on-line*). All on-line features can be computed automatically in real time by state-of-the-art speech processing applications, although it should be noted that all of our lexical and discourse features strongly rely on a speech recognizer output, which typically has a high error rate for spontaneous productions. All on-line features are also available in off-line conditions; the remaining features (those not tagged $\mathcal{O}$ in Tables 4 through 8) are normally available only in offline conditions. We distinguish online features for the machine learning experiments described in Section 5, in which we assess, among other things, the usefulness of information contained in different feature sets, simulating the conditions of actual on-line and off-line applications.

In the following sections, we use the features described here in several ways. We first perform a series of statistical tests to find differences across the various functions of ACWs. Subsequently, we experiment with machine learning techniques for the automatic classification of the function of ACWs, training the models with different combinations of features.

4. Characterizing Affirmative Cue Words

In this section we present results of a series of statistical tests aimed at identifying contextual, acoustic, and prosodic differences in the production of the various discourse/pragmatic functions of affirmative cue words. This kind of characterization is important both for interpretation and for production in spoken language applications: If we can find reliable features that effectively distinguish the various uses of these words, we can hope to interpret them automatically and generate them appropriately.

4.1 Position in IPU and Turn

We begin this analysis by looking at the discourse position of the various discourse/pragmatic functions of ACWs. Because these words help shape, or at least reflect, the structure of conversations, we expect to find positional differences between their functions. Figure 3 shows the distribution of the six most frequent ACWs in the corpus (*alright*, *okay*, *yeah*, *mm-hm*, *uh-huh*, and *right*) with respect to their position in their IPU. An **IPU-initial** word is one that occurs in the first position in its corresponding IPU; that is, it is preceded by at least 50 msec of silence and followed by another word. An **IPU-final** word occurs last in its IPU. An **IPU-medial** word is both immediately preceded and followed by other words. Lastly, a **single-word IPU** is an individual word both preceded and followed by silence. Figure 3 also depicts the distribution of discourse/pragmatic functions within each of these four categories. For example, roughly 40% of all tokens of *alright* in the corpus occur as IPU initial; of those, about half are agreements (**Agr**), half are cues to beginning discourse segments (**CBeg**), and a marginal number convey other functions.

Similarly, Figure 4 shows the distribution of the same six ACWs with respect to their position in the corresponding conversational turn. **Turn-initial**, **turn-medial**, and **turn-final** words, and **single-word turns** are defined analogously to the four IPU-related categories defined previously, but considering conversational turns instead of IPUs.

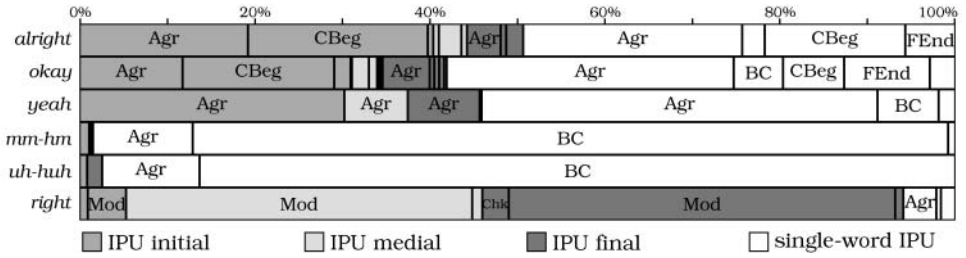

Figure 3
Position of the target word in its IPU.

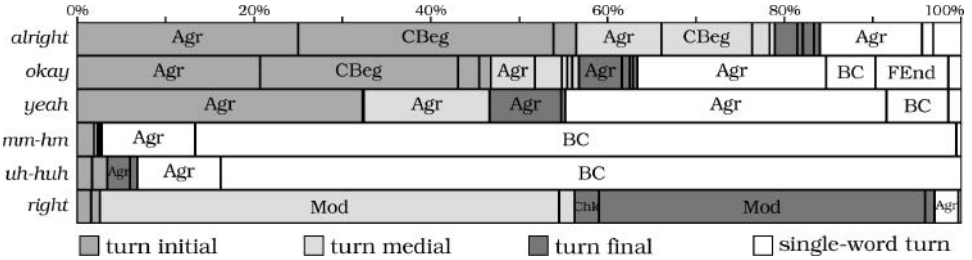

Figure 4
Position of the target word in its turn.

From these figures we observe several interesting aspects of the discourse position of ACWs in the Games Corpus. Only a minority of these words occur as IPU medial or IPU final. The only exception appears to be *right*, for which a high proportion of instances do occur in such positions—mainly tokens with the literal modifier (**Mod**) meaning, but also tokens used to check with the interlocutor (**Chk**), which take place at the end of a turn (and thus, of an IPU).

The default function of ACWs, agreement (**Agr**), occurs for *alright*, *okay*, *yeah*, and *right* in all possible positions within the IPU and the turn; for *mm-hm* and *uh-huh*, agreements occur mostly as full conversational turns. Nearly all backchannels (**BC**) occur as separate turns, with only a handful of exceptions: In four cases, the backchannel is followed by a pause in which the interlocutor chooses not to continue speaking, and the utterer of the backchannel takes the turn; in two other cases, two backchannels are uttered in fast repetition (e.g., *uh-huh uh-huh*).

From the six lexical items analyzed in Tables 3 and 4, two pairs of words seem to pattern similarly. The first such pair consists of *mm-hm* and *uh-huh*, which show very similar distributions and are realized almost always as single-word turns, as either **Agr** or **BC**. The second pair of words with comparable patterns of IPU and turn position are *alright* and *okay*. These are precisely the only two ACWs used to convey all ten discourse/pragmatic functions in the Games Corpus (recall from Table 2). This result suggests that the lexical items in these two pairs may be used interchangeably in conversation. The word *yeah* presents a pattern analogous to that of *alright* and *okay*, albeit with fewer meanings.

In all, these findings confirm the existence of large differences in the discourse position of ACWs between their functional types, as well as between their lexical types. We will revisit this topic in Section 5, where we discuss the predictive power

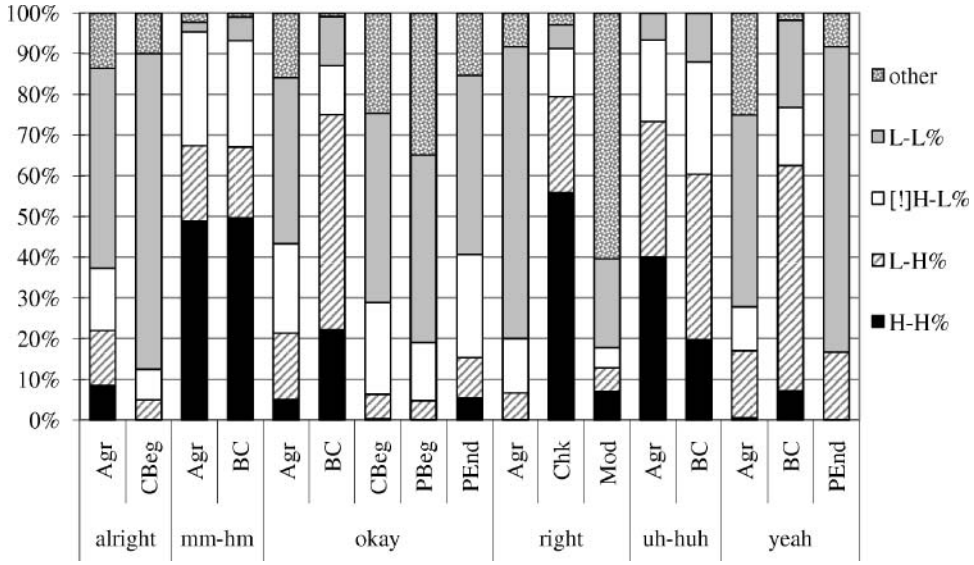

Figure 5
ToBI phrase accents and boundary tones. The 'other' category consists of cases with no phrase accent and/or boundary tone present at the target word.

of discourse features in the automatic classification of the function of ACWs. Given the observed positional differences, we expect these features to play a prominent role in such a task.

4.2 Word-Final Intonation

Shifting our attention to acoustic/prosodic characteristics of ACWs, we examine next the manner in which word-final intonation varies across ACW functions. First we look at two categorical variables in the ToBI framework which capture the final pitch incursion: phrase accent and boundary tone. Figure 5 shows the distribution of ToBI labels for each of the six most frequent ACWs and their corresponding functions (see Section 3.3 for a description of the ToBI labeling conventions). The distributions for *alright*, *okay*, *right*, and *yeah* depart significantly from random (*alright*: Fisher's Exact test, $p = 0.0483$; *okay*: Pearson's Chi-squared test, $\chi^2(24) = 261$, $p \approx 0$; *right*: Pearson, $\chi^2(8) = 220$, $p \approx 0$; *yeah*: Fisher, $p \approx 0$).[5,6] For *right*, considering just its discourse/pragmatic functions (i.e., excluding its **Mod** instances), the distribution also significantly differs from random (Fisher, $p \approx 0$). On the other hand, the distributions for *mm-hm* and *uh-huh* do not depart significantly from random.

5 Fisher's Exact test was used whenever the accuracy of Pearson's Chi-squared test was compromised by data sparsity.

6 We performed statistical tests for approximately 35 variables on the same data set. Applying the Bonferroni correction, the alpha value should be lowered from the standard 0.05 to $0.05/35 \approx 0.0014$ to maintain the familywise error rate. Thus, a result would be significant when $p < 0.0014$. According to this, most tests are still significant in the current section; however, the Tukey post hoc tests following our ANOVA tests are not: most of these have a confidence level of 95%, and significant differences begin to disappear when considering a confidence level of 99%.

The first clear pattern we find is that the backchannel function (**BC**) shows a marked preference for a high-rising (**H-H%** in the ToBI conventions) or low-rising (**L-H%**) pitch contour towards the end of the word. Those two contours account for more than 60% of the backchannel instances of *mm-hm*, *okay*, *uh-huh*, and *yeah*. For the other ACWs there are not enough instances labeled **BC** in the corpus for statistical comparison. The predominance of **H%** found for backchannels is consistent with the **openness** that such boundary tone has been hypothesized to indicate (Hobbs 1990; Pierrehumbert and Hirschberg 1990). The utterer of a backchannel understands that (i) there is more to be said, and (ii) it is the speaker holding the conversational turn who must say it.

The default function of ACWs, agreement (**Agr**) is produced most often with falling (**L-L%**) or plateau final intonation (**[!]H-L%**) in the case of *alright*, *okay*, *right*, and *yeah*. The **L%** boundary tone is believed to indicate the opposite of **H%**, a sense of **closure**, separating the current phrase from a subsequent one (Pierrehumbert and Hirschberg 1990). In our case, by agreeing with what the speaker has said, the listener indicates that enough information has been provided and that any subsequent phrases may refer to a different topic. In other words, such closure might mean that the proposition preceding the ACW has been added to the current context space (Reichman 1985), or that a new focus space is about to be created (Grosz and Sidner 1986).

Notably, **Agr** instances of *mm-hm* and *uh-huh* present a very different behavior from the other lexical items, with a distribution of final intonations that closely resembles that of backchannels. In particular, over 60% of the **Agr** tokens of *mm-hm* and *uh-huh* are produced with final rising intonation (either **L-H%** or **H-H%**). As we will see in the following sections, the realization of *mm-hm* and *uh-huh* as **Agr** or **BC** seems to be very similar along several dimensions besides intonation.

Alright and *okay* are the only two ACWs in the corpus that are used to cue the beginning of a new discourse segment, either combined with an agreement function (**PBeg**) or in its pure form (**CBeg**). These two functions typically have a falling (**L-L%**) or sustained (**[!]H-L%**) final pitch contour. Additionally, the instances of *okay* and *yeah* used to cue a discourse segment ending (**PEnd**) tend to be produced with a **L-L%** contour, and also with **[!]H-L%** in the case of *okay*. This predominance of **L%** for ACWs conveying a discourse boundary function is consistent with the previously mentioned closure that such boundary tone is believed to indicate.

Lastly, the only ACW used frequently in the corpus for checking with the interlocutor (the **Chk** function) is *right*, as illustrated in the following exchange:

A: *and the top's not either, right*?
B: *no*
A: *okay*

Such instances of *right* in the corpus normally end in a high-rising pitch contour, or **H-H%**. This fact is probably explained by the close semantic resemblance of this construction to *yes–no* questions, which typically end in the same contour type (Pierrehumbert and Hirschberg 1990).

In addition to the categorical prosodic variables described previously, word final intonation may also be studied by measuring the slope of the word-final pitch track (see Section 3.3 for a description of how pitch slopes are calculated). A high positive value of pitch slope corresponds to a rising intonation; a value close to zero, to a flat intonation; a high negative value, to a falling intonation. Final pitch slope has the advantage of being automatically computable; ToBI labels, on the other hand, still must be

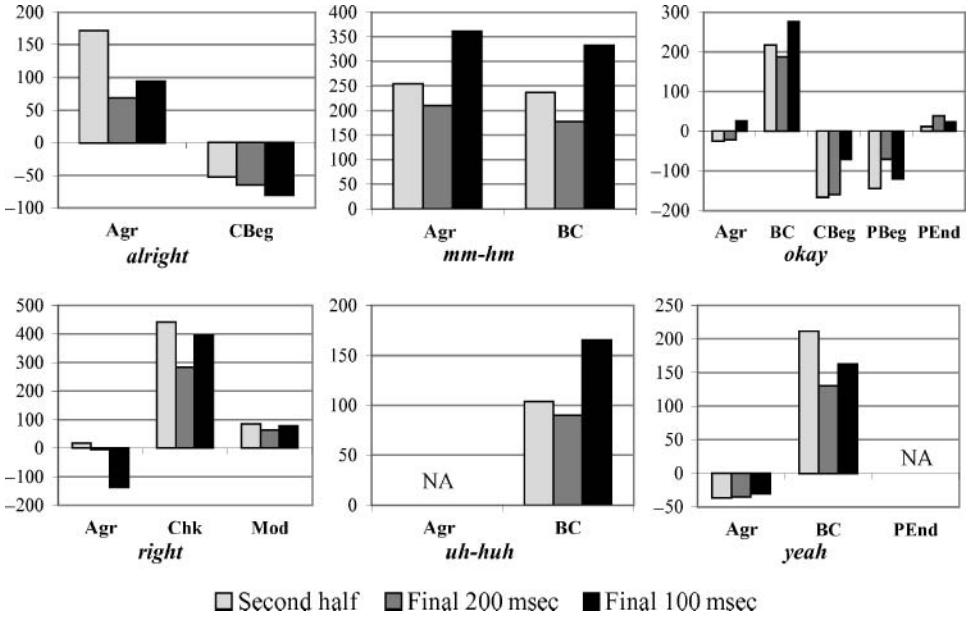

Figure 6
Final pitch slope, computed over the second half and over the final 100 and 200 msec of the target word. In all cases, the vertical axis represents the change in Hertz per second. Significant differences: For *okay*: BC>all; CBeg<Agr, BC, PEnd. For *right*: Chk>Agr. For *yeah*: BC>Agr.

manually annotated—although ongoing research may change this fact in the near future (Rosenberg and Hirschberg 2009; Rosenberg 2010a, 2010b). Therefore, it is important to verify that the results obtained using ToBI labels—if they are to be of practical use—are also observable when considering numeric measures such as pitch slope. Figure 6 shows, for the same ACWs and functions discussed earlier,[7] the mean pitch slope computed over the second half of the word and over its final 100 and 200 msec, and gender-normalized as described in Section 3.

The comparison of these numeric acoustic features across discourse/pragmatic functions confirms that the observations made previously for categorical prosodic features also hold when considering numeric features such as pitch slope, thus making the likelihood that such observations will be of practical use in actual systems. For *okay*, the three measures of word-final pitch slope are significantly higher for backchannels (**BC**) than for all other functions, and significantly lower for **CBeg** than for **Agr**, **BC**, and **PEnd** (RMANOVA for each of the three variables: between-subjects $p > 0.3$, within subjects $p \approx 0$; Tukey test confidence: 95%).[8] **BC** tokens of *yeah* are also significantly higher than **Agr**, with similar p-values. Figure 6 shows that **BC** instances of *mm-hm*

7 For **PEnd** instances of *yeah* and **Agr** instances of *uh-huh*, the number of tokens with no errors in the pitch track and pitch slope computations is too low for statistical consideration.

8 Repeated-measures analysis of variance (RMANOVA) tests estimate the existence of both within-subjects effects (i.e., differences between discourse/pragmatic functions) and between-subjects effects (i.e., differences between speakers). When the between-subjects effects are negligible, we may safely draw conclusions across multiple speakers in the corpus, with low risk of a bias from the behavior of a particular subset of speakers.

and *uh-huh* also have comparably high final pitch slopes. Again, for *mm-hm* we find no significant difference in final pitch slope between agreements and backchannels.

Although Figure 6 shows that **Chk** tokens of *right* tend to end in a steeply rising pitch, the RMANOVA tests yield between-subjects p-values of 0.01 or lower, indicating substantial speaker effects. In other words, even though the general tendency for these tokens, as indicated by both the numeric and categorical variables, seems to be to end in a high-rising intonation, there is evidence of different behavior for some individual speakers, which keeps us from drawing general conclusions about this pragmatic function of *right*.

4.3 Intensity

The next feature we find to vary significantly with the discourse/pragmatic function of ACWs is word intensity. Commonly referred to as loudness or volume, intensity generally functions to make words more salient or prominent. Figure 7 shows the maximum and mean intensity for the most frequent ACWs and functions, computed over the whole word and speaker normalized using *z*-scores.

The two types of differences we find are related to the discourse functions of ACWs. For *okay* and *yeah*, both maximum and mean intensity are significantly lower for instances cueing the end of a discourse segment (**PEnd**) than instances of all other functions (for both variables and both words, RMANOVA tests report between-subjects $p > 0.4$ and within-subjects $p \approx 0$; Tukey 95%). For ACWs cueing a beginning discourse segment, the opposite is true. Instances of *alright* and *okay* labeled **CBeg** or **PBeg** have a maximum and mean intensity significantly higher than all other functions (for *alright*,

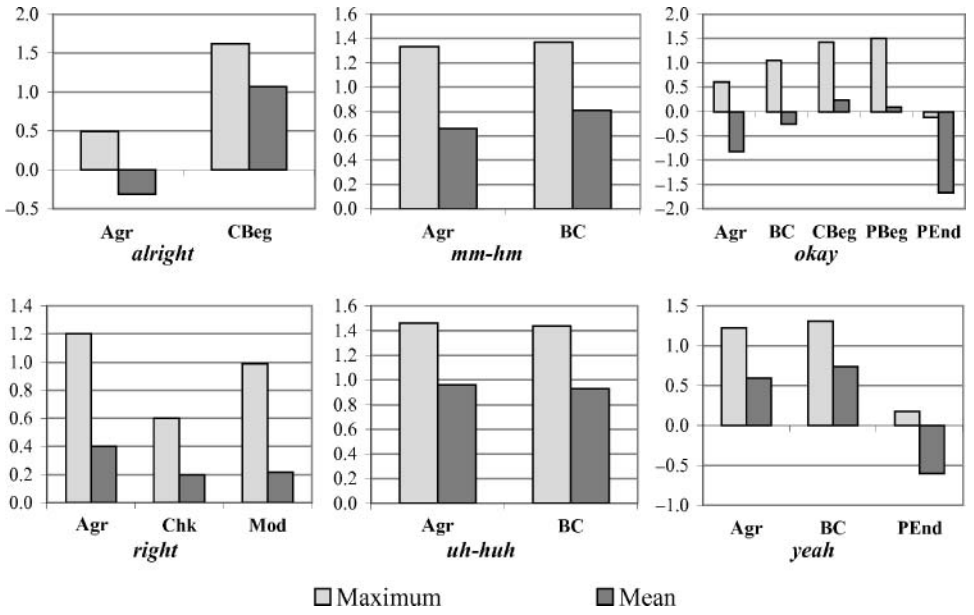

Figure 7
Word maximum and mean intensity, speaker normalized using *z*-scores. All vertical axes represent *z*-scores. Significant differences: For *alright*: Agr<CBeg. For *okay*: PEnd<all; Agr<CBeg, PBeg, BC; BC<CBeg. For *yeah*: PEnd<Agr, BC.

a RMANOVA test reports between-subjects $p > 0.12$ and within-subjects $p \approx 0$). These results are consistent with previous studies of prosodic variation relative to discourse structure, which find intensity to increase at the start of a new topic and decrease at the end (Brown, Currie, and Kenworthy 1980; Hirschberg and Nakatani 1996). Because by definition **CBeg**/**PBeg** ACWs begin a new topic and **CEnd**/**PEnd** end one, it is then not surprising to find that the former tend to be produced with higher intensity, and the latter with lower.

Finally, for *mm-hm* and *uh-huh* we find no significant differences in intensity between their unique functions, agreement (**Agr**) and backchannel (**BC**). Recall from the previous section that we find no differences in final intonation either. This further suggests that these two lexical types tend to be produced with similar acoustic/prosodic features, independently of their function.

4.4 Other Features

For the remaining acoustic/prosodic features analyzed, we find a small number of significant or approaching-significance differences between the functions of ACWs. These differences are related to duration, mean pitch, and voice quality. The first set of findings corresponds to the **duration** of ACWs, normalized with respect to all words with the same number of syllables and phonemes uttered by the same speaker. For *alright* and *okay*, instances cueing a beginning (**CBeg**) tend to be shorter than the other functions (for both words, RMANOVA: between-subjects $p > 0.5$, within-subjects $p < 0.05$, Tukey 95%). We also find tokens of *right* used to check with the interlocutor (**Chk**) to be on average shorter than the other two functions of *right* (RMANOVA, between-subjects $p > 0.7$, within-subjects $p = 0.001$; Tukey 95%). Note that these two functions are relatively simple: **CBeg** calls for the listener's attention, and is frequently conveyed with a filled pause (*uh*, *um*); **Chk** asks the interlocutor for confirmation, which may alternatively be achieved via a high-rising intonation. Thus, it is not surprising that these functions take less time to be realized than other more pragmatically loaded functions, such as agreement.

Speaker-normalized **mean pitch** over the whole word also presents significant differences for *okay* and *yeah*. Instances labeled **PEnd** (agreement and cue ending discourse segment) present a higher mean pitch than the other functions (for both words, RMANOVA: between-subjects $p > 0.6$, within-subjects $p < 0.01$; Tukey 95%). This is rather unexpected, because as noted in Section 4.2 around 70% of **PEnd** ACWs in the corpus end in a **L%** boundary tone, and thus they would plausibly be uttered with a low pitch level. What our data indicate, however, is that speakers tend to reset and raise their pitch range when producing **PEnd** instances of ACWs.

Finally, we find some evidence of differences in **voice quality**. Both *alright* and *okay* show a lower shimmer over voiced portions when starting a new segment (**CBeg**) (RMANOVA: between-subjects $p > 0.9$ for *alright*, $p = 0.09$ for *okay*; within-subjects $p < 0.001$ for both words). Also, *okay* and *yeah* present a lower noise-to-harmonics ratio (NHR) for backchannels (RMANOVA: between-subjects $p > 0.3$ for *okay*, $p = 0.04$ for *yeah*; within-subjects $p < 0.005$ for both words). A lower value of shimmer and NHR has been associated with the perception of a better voice quality (Eskenazi, Childers, and Hicks 1990; Bhuta, Patrick, and Garnett 2004). Our results suggest, then, that voice quality may constitute another dimension along which speakers vary their productions to convey the intended discourse/pragmatic meaning. Notice though that for these two variables some of the between-subjects p-values are low enough to suggest significant

speaker effects. Therefore, our results related to differences in voice quality should be considered preliminary.

5. Automatic Classification of Affirmative Cue Words

In this section we present results from machine learning (ML) experiments aimed at investigating how accurately affirmative cue words may be classified automatically into their various discourse/pragmatic functions. If spoken dialogue systems are to interpret and generate ACWs reliably, we must identify reliable cues. With this goal in mind, we explore several dimensions of the problem: We consider three classification tasks, simulating the conditions of different speech applications, and study the performance of different ML algorithms and feature sets on each task. We note that previous studies have attempted to disambiguate between the sentential and discourse uses of cue phrases such as *now*, *well*, and *like*, in corpora containing comparable numbers of instances of each class. For ACWs in the Games Corpus dialogues, sentential uses are rare, with the sole exception of *right*. Therefore, disambiguating between discourse and sentential uses appears to be less useful than distinguishing among different discourse functions.

The first ML task we consider consists in the general classification of any ACW (*alright*, *gotcha*, *huh*, *mm-hm*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yes*, *yup*) into any function (**Agr**, **BC**, **CBeg**, **PBeg**, **CEnd**, **PEnd**, **Mod**, **BTsk**, **Chk**, **Stl**; see Table 1), a critical task for spoken dialogue systems seeking to interpret user input in general. The second task involves identifying instances of these words used to signal the beginning (**CBeg**, **PBeg** in our labeling scheme) or ending (**CEnd**, **PEnd**) of a discourse segment, which is important for applications that must segment speech into coherent units, such as meeting browsing systems and turn-taking components of spoken dialogue systems. The third task consists in identifying tokens conveying some degree of acknowledgment: (**Agr**, **BC**, **PBeg**, and **PEnd**), a function especially important to spoken dialogue systems, for which it is critical to know that a user has heard the system's output.

Speech processing applications operate in disparate conditions. On-line applications such as spoken dialogue systems process information as it is generated, having access to a limited amount of context, normally up to the last IPU uttered by the user. On the other hand, off-line applications, such as meeting transcription and browsing systems, have the whole audio file available for processing. We simulate these two conditions in our experiments, assessing how the limitations of online systems affect performance. We also group the features described in Section 3.3 into five sets—lexical (`LX`), discourse (`DS`), timing (`TM`), acoustic (`AC`), and phonetic (`PH`); see Tables 4 through 8—to determine the relative importance of each feature set in the various classification tasks. For example, this approach allows us to simulate the conditions of the understanding component of a spoken dialogue system, which can use only the information up through the current IPU to detect the function of a user's ACW. Such a system may have access only to ASR transcription or it may have access to acoustic and prosodic information; we note that our analysis does not take into account the possibility that transcriptions are likely to contain some errors. Our approach also allows us to simulate a text-to-speech (TTS) system which might be used to produce a spoken version of an on-line chat room. In order to choose the appropriate acoustic/prosodic realization of each ACW, the TTS system will first need to determine its function based on features extracted solely from the input text (in our taxonomy, `LX` and `DS`).

We conduct our ML experiments using three well-known algorithms with very different characteristics: the decision tree learner **C4.5** (Quinlan 1993), the propositional

Table 10
Error rate of each classifier on the general task using different feature sets; F-measures of the SVM classifier; and error rate and F-measures of two baselines and human labelers. For the classifier error rates: † Significantly different from full model. § Significantly different from SVM. (Wilcoxon signed rank sum test, $p < 0.05$.) Significance was not tested for the classifier F-measures.

	Error Rate			SVM F-Measure					
Feature Set	C4.5	Ripper	SVM	Agr	BC	CBeg	PEnd	Mod	Chk
LX DS TM AC PH	**16.6%** §	**16.3%** §	**14.3%**	**.86**	**.81**	**.89**	**.50**	**.97**	**.39**
DS TM AC PH	21.3% †§	17.2% †	16.5% †	.84	.82	.87	.44	.94	0
LX TM AC PH	20.3% †§	20.1% §	17.0% †	.84	.80	.83	.16	.97	.21
LX DS AC PH	17.1% §	18.1% †§	14.8% †	.86	.81	.89	.38	.97	.35
LX DS TM PH	15.2% †	16.3%	16.2% †	.85	.80	.86	.16	.97	.33
LX DS TM AC	17.0% §	16.9% §	14.7%	.86	.80	.89	.48	.97	.35
LX	23.7% †§	22.7% †	22.3% †	.79	.80	.65	0	.96	.03
DS	22.8% †§	24.0% †§	25.3% †	.76	.67	.82	0	.87	0
TM	29.5% †§	27.3% †§	36.2% †	.70	0	.57	0	.83	0
AC	44.8% †§	29.8% †§	41.3% †	.67	.66	.14	0	.58	0
PH	56.4% †§	26.5% †§	45.4% †	.65	.08	.13	0	.64	0
Majority class baseline ER			56.4%	.61	0	0	0	0	0
Word-based baseline ER			27.7%	.75	.79	0	0	.94	.13
Human labelers ER (estimate 1)			9.3%	.92	.91	.94	.51	.99	.67
Human labelers ER (estimate 2)			11.0%	.90	.89	.93	–	.99	–

rule learner RIPPER (Cohen 1995), and support vector machines (SVM) (Cortes and Vapnik 1995; Vapnik 1995). We use the implementation of these algorithms provided in the WEKA machine learning toolkit (Witten and Frank 2000), known respectively as **J48**, **JRIP**, and **SMO**. We also use 10-fold cross-validation in all experiments.[9]

5.1 Classifiers and Feature Types

To assess the predictive power of the five feature types (LX, DS, TM, AC, and PH) we exclude one type at a time and compare the performance of the resulting set to that of the full model. Table 10 displays the error rate of each ML classifier on the general task, classifying any ACW into any of the most frequent discourse/pragmatic functions (**Agr**, **BC**, **CBeg**, **PEnd**, **Mod**, **Chk**). Table 11 shows the same results for the other two tasks: the detection of a discourse boundary function—cue beginning (**CBeg PBeg**), cue ending (**CEnd**, **PEnd**), or no-boundary (all other labels); and the detection of an acknowledgment function—**Agr**, **BC**, **PBeg**, or **PEnd**, vs. all other labels.[10]

9 In the case of SVM, prior to the actual tests we experimented with two kernel types: polynomial ($K(x, y) = (x + y)^d$) and Gaussian radial basis function (RBF) ($K(x, y) = exp(-\gamma||x - y||^2)$ for $\gamma > 0$). We performed a grid search for the optimal arguments for either kernel using the data portion left out after downsampling the corpus (see Section 3.2). The best results were obtained using a polynomial kernel with exponent $d = 1.0$ (i.e., a linear kernel) and model complexity $C = 1.0$.

10 We note that performance on new data may be somewhat worse than the results reported here, because we did exclude approximately 5% of tokens in our corpus due to lack of annotator agreement on labels.

Table 11
Error rate of each classifier on the detection of discourse boundary functions and acknowledgment functions, using different feature sets. † Significantly different from full model. § Significantly different from SVM. (Wilcoxon signed rank sum test, $p < 0.05$.)

	Disc. Boundary			Acknowledgment		
	{**CBeg**, **PBeg**} vs. {**CEnd**, **PEnd**} vs. Rest			{**Agr**, **BC**, **PBeg**, **PEnd**} vs. Rest		
Feature Set	C4.5	Ripper	SVM	C4.5	Ripper	SVM
LX DS TM AC PH	**6.9%**	**8.1%** §	**6.9%**	**5.8%**	**5.9%** §	**4.5%**
DS TM AC PH	7.6% †	8.0%	7.6% †	8.5% †§	5.5% §	6.4% †
LX TM AC PH	10.4% †	10.1% †	9.5% †	8.7% †§	8.7% †§	6.5% †
LX DS AC PH	8.0% †	8.7% §	7.5% †	5.3%	5.7% §	4.9%
LX DS TM PH	6.6% §	7.9%	8.9% †	5.4%	5.4%	5.1%
LX DS TM AC	7.1%	8.3% §	7.0%	5.8% §	5.6% §	4.6%
LX	14.2% †	14.5% †§	13.9% †	11.4% †	11.4% †	11.7% †
DS	7.8% §	8.6% §	10.9% †	8.4% †§	8.9% †	9.4% †
TM	12.2% †§	11.2% †§	14.7% †	12.8% †§	13.5% †	14.5% †
AC	17.3% †§	14.3% †§	18.5% †	26.7% †	16.6% †§	28.4% †
PH	18.6% †	17.6% †	18.6% †	36.5% †§	14.1% †§	25.4% †
Majority class baseline ER		18.6%			36.5%	
Word-based baseline ER		18.6%			15.3%	
Human labelers ER (est. 1)		5.3%			2.9%	
Human labelers ER (est. 2)		5.6%			3.0%	

In both tables, the first line corresponds to the full model, with all five feature types. The subsequent five lines show the performance of models with just four feature types, excluding one feature type at a time, and the following five lines show the performance of models with exactly one feature type—these are two methods for assessing the predictive power of each feature set. For the error rates of our classifiers, the † symbol indicates that the given classifier performs significantly worse when trained on a particular feature set than when trained on the full set.[11] The § symbol indicates that the difference between SVM and the given classifier, either C4.5 or Ripper, is significant. For example, the second line (DS TM AC PH) in Table 10 indicates that, for the general classification task, the three models trained on all but lexical features perform significantly worse than the respective full models; also, the performance of C4.5 is significantly worse than SVM, and the difference between Ripper and SVM is not significant.

The bottom parts of Tables 10 and 11 show the error rate of two baselines, as well as two estimates of the error rate of human labelers. We consider two types of baseline: one a majority-class baseline, and one that employs a simple rule based

11 All accuracy comparisons discussed in this section are tested for significance with the Wilcoxon signed rank sum test (a non-parametric alternative to Student's t-test) at the $p < 0.05$ level, computed over the error rates of the classifiers on the ten cross-validation folds. These tests provide evidence that the observed differences in mean accuracy over cross-validation folds across two models are not attributable to chance.

on word identity. In the general classification task, the majority class is **Agr**, and the best performing word-based rule is *huh*→**Chk**, *mm-hm*→**Mod**, *uh-huh*→**BC**, *right*→**Mod**, others→**Agr**. For the identification of a discourse boundary function, the majority class is no-boundary, and the word-based rule also assigns no-boundary to all tokens. For the detection of an acknowledgment function, the majority class is acknowledgment, and the word-based rule is *right*, *huh*→no-acknowledgment; others→acknowledgment.

The error rates of human labelers are estimated using two different approaches. Our first estimate compares the labels assigned by each labeler and the majority labels as defined in Section 3.1. Because each labeler's labels are used for calculating both the error rate and the gold standard, this estimate is likely to be over-optimistic. Our second estimate considers the subset of cases in which two annotators agree, and compares those labels with the third labeler's. Tables 10 and 11 show that these two estimates yield similar results; for **PEnd** and **Chk**, there are not enough counts for computing the F-measure of estimate 2.

The right half of Table 10 shows the F-measure of the SVM classifier for each individual ACW function, for the general task. The highest F-measures correspond to **Agr**, **BC**, **CBeg**, and **Mod**, precisely the four functions with the highest counts in the Games Corpus. For **PBeg** and **Chk** the F-measures are much lower (and equal to zero for the four remaining functions, not included in the table) due very likely to their low counts, which prevent a better generalization during the learning stage. Future research could investigate incorporating boosting and bootstrapping techniques to reduce the negative effect on classification of low counts for some of the discourse/pragmatic functions of ACWs.

For the three classification tasks, SVM outperforms, or performs at least comparably to, the other two classifiers whenever acoustic features (AC) are taken into account together with other feature types. When used alone, though, acoustic features perform poorly in all three tasks. Moreover, when acoustic features are excluded, SVM's accuracy is comparable to, or worse than, C4.5 and Ripper. This is probably due to the fact that SVM's mathematical model is better suited to exploit larger amounts of continuous numerical variables, and thus makes a difference when including acoustic features.

For the first two tasks, the SVM classifier seems to take advantage of all but one feature type, as shown by the significantly lower performance resulting from removing any of the feature types from the full model—the sole exception is the phonetic type (PH), whose removal in no case negatively affects the accuracy of any classifier. C4.5 and Ripper, on the other hand, appear to take more advantage of some feature types than others. For the third task, lexical (LX) and discourse (DS) features apparently have more predictive power for both C4.5 and SVM than the other types. Note also that for the second and third tasks, the error rates of our full-model SVM classifiers closely approximate the estimated error rates of human labelers.

For the general task of classifying any ACW into any discourse/pragmatic function, our full-model SVM classifier achieves the best overall results. To take a closer look at the performance of this model, we compute its F-measure for the discourse/pragmatic functions of each individual lexical item, as shown in Table 12. We observe that the classifier achieves better results for word–function pairs with higher counts in the Games Corpus, such as *yeah*-**Agr** or *right*-**Mod** (cf. Table 2). Again, the low counts for the remaining word–function pairs may prevent a better generalization during the learning stage, a problem that could be attenuated in future work with boosting and bootstrapping techniques.

5.2 Session-Specific and ToBI Prosodic Features

When including session-specific features in the full model, such as identity and gender of both speakers (see Table 8), the error rate of the SVM classifier is significantly reduced for the general task (13.3%) and for the discourse boundary function identification task (6.4%) (Wilcoxon, $p < 0.05$). For the detection of an acknowledgment function, the error rate is not modified when including those features (4.5%). This suggests the existence of speaker differences in the production of at least some functions of ACWs that may be exploited by ML classifiers. Finally, the inclusion of categorical prosodic features based on the ToBI framework, such as type of pitch accent and break index on the target word (see Table 5), does not improve the performance of the SVM-based full models in any of the classification tasks.

5.3 Individual Features

To estimate the importance of individual features in our classification tasks, we rank them according to an information-gain metric. We find that for the three tasks, lexical (`LX`), discourse (`DS`), and timing (`TM`) features dominate. The highest ranked features are the ones capturing the position of the target word in its IPU and in its turn. Lexical identity and POS tags of the previous, target, and following words, and duration of the target word, are also ranked high. Acoustic features appear lower in the ranking; the best performing ones are word intensity (range, mean, and standard deviation), pitch (maximum and mean), pitch slope over the final part of the word (200 msec and second half), voiced-frames ratio, and noise-to-harmonics ratio. All phonetic features are ranked very low. Note that, whereas durational features at the word level are ranked high, durational features at the phonetic level are not, because the latter only capture the duration of each phone relative to the word duration—apparently not an informative attribute for these classification tasks. These results confirm the existence of large positional differences across functions of ACWs, as seen in Section 4. Additionally, whereas several acoustic/prosodic features extracted from the target word contain useful information for the automatic disambiguation of ACWs, it is positional information that provides the most predictive power.

5.4 Online and Offline Tasks

To simulate the conditions of online applications, which process speech as it is produced by the user, we consider a subset of features that may typically be extracted from the

Table 12
F-measure achieved by our full-model SVM classifier for the different discourse/pragmatic functions of each lexical item.

	Agr	BC	CBeg	PBeg	PEnd	Mod	BTsk	Chk	Stl
alright	.88	–	.93	–	.33	–	–	–	–
mm-hm	.35	.94	–	–	–	–	–	–	–
okay	.82	.51	.88	.27	.63	.53	0	–	.18
right	.84	–	–	–	–	.98	–	.53	–
uh-huh	.35	.93	–	–	–	–	–	–	–
yeah	.96	.54	–	–	.17	–	–	–	–

speech signal only up to the IPU containing the target ACW. These features are marked in Tables 4 through 8 with letter $\mathcal{O}$. With these features, we train and evaluate an SVM classifier for the three tasks described previously. Table 13 shows the results, comparing the performance of each classifier to that of the models trained on the full feature set, which simulate the conditions of off-line applications. In all three cases the on-line model performs significantly worse than its offline correlate, but also significantly better than the baseline (Wilcoxon, $p < 0.05$).

Table 13 also shows the error rates of on-line and off-line classifiers trained using solely text-based features—that is, only features of lexical (LX) or discourse (DS) types. Text-based models simulate the conditions of spoken dialogue systems with no access to acoustic and prosodic information, or generation systems attempting to realize text-based exchanges in speech. They reflect the importance of text information alone in training such systems to recognize the function of ACWs on-line and off-line and to produce appropriate realizations from limited or full transcription.

Our on-line and off-line text-based models perform significantly worse than the corresponding models that use the whole feature set, but they still outperform the baseline models in all cases (Wilcoxon, $p < 0.05$). Finally, the off-line text-based models also outperform their on-line correlates in all three tasks (Wilcoxon, $p < 0.05$). These results indicate the important role that other classes of cues play in recognition, while indicating the level of performance we can expect from TTS systems which have only text available.

5.5 Backchannel Detection

The correct identification of backchannels is a desirable capability for speech processing systems, as it would allow us to distinguish between two quite distinct speaker intentions: the intention to take the conversational floor, and the intention to backchannel.

We first consider an off-line binary classification task—namely, classifying all ACWs in the corpus into backchannels vs. the rest, using information from the whole conversation. In such a task, an SVM classifier achieves a 4.91% error rate, slightly yet significantly outperforming a word-based baseline (*mm-hm*, *uh-huh*→**BC**; others→ no-**BC**), with 5.17% (Wilcoxon, $p < 0.05$).

On-line applications such as spoken dialogue systems need to classify every new speaker contribution immediately after (or even while) it is uttered, and certainly without access to any subsequent context. The Games Corpus contains approximately 6,700 turns following speech from the other speaker, all of which begin as potential backchannels and need to be disambiguated by the listener. Most of these candidates can be trivially discarded using a simple observation about backchannels: By definition

Table 13
Error rate of the SVM classifier on online and offline tasks.

	All Functions		Disc. Boundary		Acknowledgment	
Feature Set	Online	Offline	Online	Offline	Online	Offline
LX DS TM AC PH (Full model)	17.4%	14.3%	10.1%	6.9%	6.7%	4.5%
LX DS (Text-based)	21.4%	16.8%	13.5%	9.1%	10.0%	5.9%
Word-based baseline	27.7%		18.6%		15.3%	

they are short, isolated utterances, and consist normally in just one ACW. Of the 6,700 candidate turns in the corpus, only 2,351 (35%) begin with an isolated ACW, including 753 of the 757 backchannels in the corpus.[12] Thus, an on-line classification procedure would only need to identify backchannels in those 2,351 turns. At this point, we explore using a binary classifier for this task. The same word-based majority baseline described earlier achieves an error rate of 11.56%. An SVM classifier trained on features extracted from up to the current IPU (to simulate the on-line condition of a spoken dialogue system) fails to improve over this baseline, achieving an error rate of 11.51%, not significantly different from the baseline. A possible explanation for this might be that backchannels seem to be difficult to distinguish from agreements in many cases, leading to an increase in the error rate. Recall, from the statistical analyses in the previous section, the positional and acoustic/prosodic similarities of tokens with these two functions for *mm-hm* and *uh-huh*, for example. Shriberg et al. (1998) report the same difficulty in distinguishing these two word functions. We conclude that further research is needed to develop novel approaches to this crucial problem of spoken dialogue systems.

5.6 Comparison with Previous Work

In an effort to provide a general frame of reference for our results, we discuss here what we believe to be the most relevant results from related studies. Note, however, that comparing these results directly to the results of our classification experiments is difficult because the type of corpora, definitions used, features examined, and/or methodology employed vary greatly among the studies. The current study focuses exclusively on the discourse/pragmatic functions of ACWs whereas other studies have either a broader or narrower scope.

Among the cue words tested in Litman (1996) is *okay*, one of the ACWs we also investigate. Litman describes the automatic classification of cue words in general (including, e.g., *now*, *well*, *say*, and *so*), classifying these into discourse and sentential uses using a corpus of monologue. In this classification task, which is not performed in our study, the best results are reached by decision-tree learners trained on prosodic and text-based features, with an error rate of 13.8%.

The most relevant study to ours is that of Stolcke et al. (2000), which presents experiments on the automatic disambiguation of dialogue acts (DA) on 1,155 spontaneous telephone conversations from the Switchboard corpus, labeled using the DAMSL (Dialogue Act Markup in Several Layers) annotation scheme (Core and Allen 1997). For the subtask of identifying the **Agreement** and **Backchannels** tags collapsed together, the authors report an error rate of 27.1% when using prosodic features, 19.0% when using features extracted from the text, and 15.3% when using all features. Other DA classifications also include some of the functions of ACWs discussed in our current study. For instance, Reithinger and Klesen (1997) employ a Bayesian approach for classifying 18 classes of DAs in transcripts of 437 German dialogues from the VERBMOBIL corpus (Jekat et al. 1995). The DA tags examined include Accept, Confirm, and Feedback, all of which are related to the functions of ACWs discussed here. For the Accept DA tag, the authors report an F-measure of 0.69; for Feedback, 0.48; and for Confirm, 0.40. These

12 The four remaining backchannels correspond to a rare phenomenon in which the speaker overlaps the interlocutor's last phrase with a short agreement, followed by an optional short pause and a backchannel. Example: A: *but it doesn't overlap *them.* B: *right* yeah yeah # okay.*

experiments are repeated on transcripts of 163 English dialogues from the same corpus, yielding an F-measure of 0.78 for the Accept DA tag, and 0 for the other two tags due to data sparsity.

As part of a study aimed at assessing the effectiveness of machine learning for this type of task, Core (1998) experiments with hand-coded decision trees for classifying five high-level dialogue act classes, including AGREEMENT and UNDERSTANDING, following the DAMSL annotation scheme. On 19 dialogues from the TRAINs corpus (discussions related to solving transportation problems), Core reports an accuracy of 70% for both the Agreement and the Understanding DA classes, using only the previous utterance's DAMSL tag as a feature in the decision trees. This use of DA context in classifying ACWs would appear to be promising, assuming an accurate automatic classification of all DAs in the corpus.

Finally, Lampert, Dale, and Paris (2006) describe a statistical classifier trained on text-based features for automatically predicting eight different speech acts derived from a taxonomy called Verbal Response Modes (VRM). The experiments are conducted on transcripts of 1,368 utterances from 14 dialogues in English. For the Acknowledgment speech act (which "conveys receipt of or receptiveness to other's communication; simple acceptance, salutations; e.g., *yes*" [page 37]), the classifier yields an F-measure of 0.75.

Again, all of these studies differ significantly from our own, in their task definition, in their methodology, and in the domain they examine. However, we expect this brief summary to serve as a general frame of reference for our own classification results.

6. Discussion

In this work we have undertaken a comprehensive study of affirmative cue words, a subset of cue phrases such as *okay*, *yeah*, or *alright* that may be utilized to convey as many as ten different discourse/pragmatic functions, such as indicating continued attention to the interlocutor or cueing the beginning of a new topic. Considering the high frequency of ACWs in task-oriented dialogue, it is critical for some spoken language processing applications such as spoken dialogue systems to model the usage of these words correctly, from both an understanding and a generation perspective.

Section 4 presents statistical evidence of a number of differences in the production of the various discourse/pragmatic functions of ACWs. The most notable contrasts in acoustic/prosodic features relate to word final intonation and word intensity. Backchannels typically end in a rising intonation; agreements and cue beginnings, in a falling intonation. Cue beginnings tend to be produced with a high intensity, and cue endings with a very low one. Other acoustic/prosodic features—duration, mean pitch, and voice quality—also seem to vary with the word usage. Our findings related to final intonation are consistent with previous results obtained by Hockey (1993) and Jurafsky et al. (1998) for American English. For Scottish English, Kowtko (1996) reports a non-rising intonation for cue beginnings and for her 'reply-y' function, a subclass of our agreement function. Kowtko also reports observing all types of final intonation in her 'acknowledge' function, whose definition overlaps both our agreements and backchannels. Thus, we find no apparent contradictions between Kowtko's results for Scottish English and ours for American English.

The word *okay* is the most heavily overloaded ACW in our corpus. Our corpus includes instances conveying each of the ten identified meanings, and this item shows the highest degree of variation along the acoustic/prosodic features we have examined.

We speculate from this finding that the more ambiguous an ACW, the more a speaker needs to vary acoustic/prosodic features to differentiate its meaning.

Our statistical analysis of ACWs also shows that these words display substantial positional differences across functions, such as the position of the word in its conversational turn, or whether the word is preceded and/or followed by silence. Such large differences bring support to Novick and Sutton's (1994) claim that the discourse/pragmatic role of these expressions strongly depends on their basic structural context. For example, in Novick and Sutton's words, an ACW in turn-initial position is "clearly not serving as a prompt for the other speaker to continue" (page 97).

Previous studies on the automatic disambiguation of other types of cue words, such as *now*, *well*, or *like*, present the problem as a binary classification task: Each cue word has either a discourse or a sentential sense (e.g., Litman 1996; Zufferey and Popescu-Belis 2004). In the study of automatic classification of ACWs presented in Section 5 we show that for spoken task-oriented dialogue, the simple discourse/sentential distinction is insufficient. In consequence, we define two new classification tasks besides the general task of classifying any ACW into any function. Our first task, the detection of an acknowledgment function, has important implications for the language management component in spoken dialogue systems, which must keep track of which material has reached mutual belief in a conversation (Bunt, Morante, and Keizer 2007; Roque and Traum 2009). Our second task, the detection of a discourse segment boundary function, should help in discourse segmentation and meeting processing tasks (Litman and Passonneau 1995). Our SVM models based on lexical, discourse, timing, and acoustic features approach the error rate of trained human labelers in all tasks, while our automatically computed phonetic features offer no improvement. Previous studies indicate that the pragmatic function of ambiguous expressions may be effectively predicted by models that combine information extracted from various sources, including lexical and prosodic (e.g., Litman 1996; Stolcke et al. 2000). Our results support this, and extend the list of useful information sources to include discourse and timing features that may be easily extracted from the time-aligned transcripts.

Additionally, our machine learning study includes experiments with several combinations of feature sets, in an attempt to simulate the conditions of different applications. Models that are trained using features extracted only from the speech signal up to the IPU containing the target word simulate on-line applications such as spoken dialogue systems with access to acoustic/prosodic features. Although such models perform worse than "off-line" models, which make use of left *and* right context, they still significantly outperform our baseline classifiers. Models that simulate the conditions of current spoken dialogue systems with access only to lexical features (although perhaps errorful) and TTS systems synthesizing spoken conversations, which have access only to features extracted from the input text, also significantly outperform our baseline classifiers.

Interactions between state-of-the-art spoken dialogue systems and their users appear to contain very few instances of backchannel responses from either conversational partner. On the system's side, the absence of this important element of spoken communication may be due to the difficulty of detecting appropriate moments where a backchannel response would be welcome by the user. Recent advances on that research topic (Ward and Tsukahara 2000; Cathcart, Carletta, and Klein 2003; Gravano and Hirschberg 2009a) have encouraged research on ways to equip systems with the ability to signal to the user that the system is still listening (Maatman, Gratch, and Marsella 2005; Bevacqua, Mancini, and Pelachaud 2008; Morency, de Kok, and Gratch 2008)—for example, when the user is asked to enter large amounts of information. On the

user's side, an important reason for *not* backchanneling may lie in the unnaturalness of such systems, often described as "confusing" or even "intimidating" by users, as well as their inability to recognize backchannels as such. Nonetheless, recent Wizard-of-Oz experiments conducted by Hjalmarsson (2009, 2011) show that humans appear to react to turn-management cues produced by a synthetic voice in the same way that they react to cues produced by another human. This important finding suggests that users of spoken dialogue systems could be cued to produce backchannel responses, for example to determine if they are still paying attention. In that case, it will be crucial for systems to be able to distinguish backchannels from other pragmatic functions (Shriberg et al. 1998). In Section 5.5 we present results on the task of automatically identifying backchannel ACWs from the other possible functions. Our models improve over the baseline in an off-line condition (e.g., for meeting processing tasks), but fail to do so in an on-line setting (e.g., for spoken dialogue systems). Practically all of the confusion of this on-line model comes from misclassifying agreements (**Agr**) as backchannels (**BC**) and vice versa. The reliability of our human labelers for distinguishing these two classes was measured by Fleiss's κ at 0.570, a level considerably lower than the 0.745 achieved for the general labeling task, which indicates that the backchannel identification task is difficult for humans as well, at least when they are not engaged in the conversation itself but only listening to it after the fact. Although we asked our annotators to distinguish the agreement function of ACWs from "continued attention," there are clearly cases where people disagree about whether speakers are indicating agreement or not. In future research we will investigate this issue in more detail, given the relevance of on-line identification of backchannels in spoken dialogue systems.

In summary, in this study we have identified a number of characterizations of affirmative cue words in a large corpus of SAE task-oriented dialogue. The corpus on which our experiments were conducted, rich in ACWs conveying a wide range of discourse/pragmatic functions, has allowed us to systematically investigate many dimensions of these words, including their production and automatic disambiguation. Besides the value of our findings from a linguistic modeling perspective, we believe that incorporating these results into the production and understanding components of spoken dialogue systems should improve their performance and increase user satisfaction levels accordingly, getting us one step closer to the long-term goal of effectively emulating human behavior in dialogue systems.

Appendix A: The COLUMBIA GAMES CORPUS

The COLUMBIA GAMES CORPUS is a collection of 12 spontaneous task-oriented dyadic conversations elicited from native speakers of Standard American English. The corpus was collected and annotated jointly by the Spoken Language Group at Columbia University and the Department of Linguistics at Northwestern University. In each of the 12 sessions, two subjects were paid to play a series of computer games requiring verbal communication to achieve joint goals of identifying and moving images on the screen. Each subject used a separate laptop computer and could not see the screen of the other subject. They sat facing each other in a soundproof booth, with an opaque curtain hanging between them, so that all communication was verbal. The subjects' speech was not restricted in any way, and it was emphasized at the session beginning that the game was *not* timed. Subjects were told that their goal was to accumulate as many points as possible over the entire session, since they would be paid additional money for each point they earned.

A.1 Game Tasks

Subjects were first asked to play three instances of the CARDS game, where they were shown cards with one to four images on them. Images were of two sizes (small or large) and various colors, and were selected to contain primarily voiced consonants, which facilitates pitch track computation (e.g., *yellow lion*, *blue mermaid*). There were two parts to each Cards game, designed to vary genre from primarily monologue to dialogue.

In the *first part* of the Cards game, each player's screen displayed a stack of 9 or 10 cards (Figure A1a). Player A was asked to describe the top card on her pile, while Player B was asked to search through *his* pile to find the same card, clicking a button when he found it. This process was repeated until all cards in Player A's deck were matched. In all cases, Player B's deck contained one additional card that had no match in Player A's deck, to ensure that she would need to describe all cards.

In the *second part* of the Cards game, each player saw a board of 12 cards on the screen (Figure A1b), all initially face down. As the game began, the first card on one player's (the DESCRIBER's) board was automatically turned face up. The Describer was told to describe this card to the other player (the SEARCHER), who was to find a matching card from the cards on his board. If the Searcher could not find a card exactly matching the Describer's card, but *could* find a card depicting one or more of the objects on that card, the players could decide whether to declare a partial match and receive points proportional to the numbers of objects matched on the cards. At most three cards were visible to each player at any time, with cards seen earlier being automatically turned face down as the game progressed. Players switched roles after each card was described and the process continued until all cards had been described. The players were given additional opportunities to earn points, based on other characteristics of the matched cards, to make the game more interesting and to encourage discussion.

After completing all three instances of the Cards game, subjects were asked to play a final game, the OBJECTS game. As in the Cards game, all images were selected to have likely descriptions which were as voiced and sonorant as possible. In the Objects game, each player's laptop displayed a game board with 5 to 7 objects (Figure A1c). Both players saw the same set of objects at the same position on the screen, except for one (the TARGET). For the DESCRIBER, the target object appeared in a random location among other objects on the screen; for the FOLLOWER, the target object appeared at the bottom of the screen. The Describer was instructed to describe the position of the target object on her screen so that the Follower could move his representation to the same location on his own screen. After players negotiated what they believed to be their best

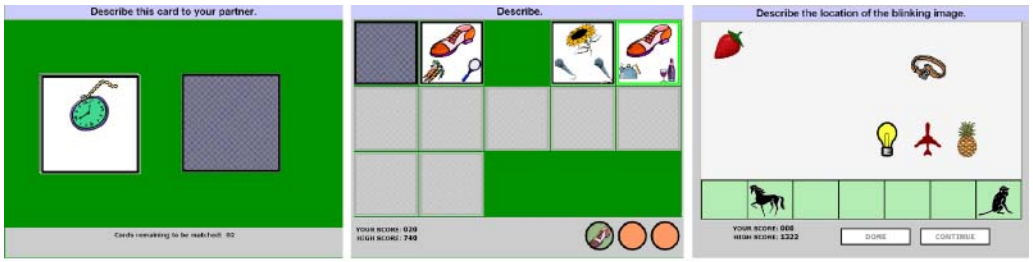

Figure A1
Sample screens from the Cards games (a, b) and Objects games (c).

location match, they were awarded 1 to 100 points based on how well the Follower's target location matched the Describer's.

The Objects game proceeded through 14 tasks. In the initial four tasks, one of the subjects always acted as the Describer, and the other one as the Follower. In the following four tasks their roles were inverted: The subject who played the Describer role in the initial four tasks was now the Follower, and vice versa. In the final six tasks, they alternated the roles with each new task.

A.2 Subjects and Sessions

Thirteen subjects (six women, seven men) participated in the study, which took place in October 2004 in the Speech Lab at Columbia University. Eleven of the subjects participated in two sessions on different days, each time with a different partner. All subjects reported being native speakers of Standard American English and having no hearing impairments. Their ages ranged from 20 to 50 years (mean, 30.0; standard deviation, 10.9), and all subjects lived in the New York City area at the time of the study. They were contacted through the classified advertisements Web site `craigslist.org`.

We recorded twelve sessions, each containing an average of 45 minutes of dialogue, totaling roughly 9 hours of dialogue in the corpus. Of those, 70 minutes correspond to the first part of the Cards game, 207 minutes to the second part of the Cards game, and 258 minutes to the Objects game. Each subject was recorded on a separate channel of a DAT recorder, at a sample rate of 48 kHz with 16-bit precision, using a Crown head-mounted close-talking microphone. Each session was later downsampled to 16 kHz, 16-bit precision, and saved as one stereo .wav file with one player per channel, and also as two separate mono .wav files, one for each player.

Trained annotators orthographically transcribed the recordings of the Games Corpus and manually aligned the words to the speech signal, yielding a total of 70,259 words and 2,037 unique words in the corpus. Additionally, self repairs and certain non-word vocalizations were marked, including laughs, coughs, and breaths. Intonational patterns and other aspects of the prosody were identified using the ToBI transcription framework (Beckman and Hirschberg 1994; Pitrelli, Beckman, and Hirschberg 1994): Trained annotators intonationally transcribed all of the Objects portion of the corpus (258 minutes of dialogue) and roughly one third of the Cards portion (90 minutes).

Appendix B: ACW Labeling Guidelines

These guidelines for labeling the discourse/pragmatic functions of affirmative cue words were developed by Julia Hirschberg, Štefan Beňuš, Agustín Gravano, and Michael Mulley at Columbia University.

Classification Scheme

Most of the labels are defined using *okay*, but the definitions hold for all of these words: *alright*, *gotcha*, *huh*, *mm-hm*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yes*, *yup*. If you really have no clue about the function of a word, label it as **?**.

[Mod] Literal Modifiers: In this case the words are used as modifiers. Examples:
*"I think that's **okay**."*
*"It's **right** between the mermaid and the car."*
*"Yeah, that's **right**."*

[Agr] Acknowledge/Agreement: The function of *okay* that indicates "I believe what you said", and/or "I agree with what you say." This label should also be used for *okay* after another *okay* or after an evaluative comment like "Great" or "Fine" in its role as an acknowledgment.[13] Examples:

A: *Do you have a blue moon?*
B: ***Yeah***.
A: *Then move it to the left of the yellow mermaid.*
B: ***Okay***, ***gotcha***. *Let's see...* (Here, both *okay* and *gotcha* are labeled **Agr**.)

[CBeg] Cue Beginning: The function of *okay* that marks a new segment of a discourse or a new topic. Test: could this use of okay be replaced by "Now"?

[PBeg] Pivot Beginning: (Agr+CBeg) When *okay* functions as both a cue word and as an Acknowledge/Agreement. Test: Can okay be replaced by "Okay now" with the same pragmatic meaning?

[CEnd] Cue Ending: The function of *okay* that marks the end of a current segment of a discourse or a current topic. Example: "So that's done. **Okay**."

[PEnd] Pivot Ending: (Agr+CEnd) When *okay* functions as both a cue word and as an Acknowledge/Agreement, but ends a discourse segment.

[BC] Backchannel: The function of *okay* in response to another speaker's utterance that indicates only "I'm still here / I hear you and please continue."

[Stl] Stall: *Okay* used to stall for time while keeping the floor. Test: Can *okay* be replaced by an elongated "Um" or "Uh" with the same pragmatic meaning? "So I yeah I think we should go together."

[Chk] Check: *Okay* used with the meaning "Is that okay?" or "Is everything okay?" For example, *"I'm stopping now,* ***okay****?"*

[BTsk] Back from a task: "I've just finished what I was doing and I'm back." Typical case: One subject spends some time thinking, and then signals s/he is ready to continue the discourse.

Special Cases

(1) *"okay so" / "okay now" / "okay then"* / and so forth, where both words are uttered together, *okay* seems to convey **Agr**, and *so / now / then* seems to convey **CBeg**. Because we do not label words like *so*, *now*, or *then*, we label *okay* as **PBeg**.

(2) If you encounter a rapid sequence of the same word several times in a row, all of them uttered in one "burst" of breath, mark only the first one the corresponding label, and label the others with "?". Example: "*okay yeah yeah yeah*" should be labeled as: "*okay*:Agr *yeah*:Agr *yeah*:? *yeah*:?".

13 Throughout this article we have used the term 'agreement' to avoid confusion with other definitions of 'acknowledgment' found in the literature.

Appendix C: ACW Labeling Examples

This appendix lists a number of examples of each type of ACWs from the Columbia Games Corpus, as labeled by our annotators. Each ACW is highlighted and annotated with its majority label. Overlapping speech segments are embraced by square brackets, and additional notes are given in parentheses.

———

A: *it's aligned to the f- to the foot of the M&M guy like to the bottom of the iron*
B: ***okay***$_{Agr}$ *lines up*
A: ***yeah***$_{Agr}$ *it's it's almost it's just barely like over*
B: ***okay***$_{Agr}$

———

A: *the tail*
B: ***mm-hm***$_{BC}$
A: *of the iron*
B: ***mm-hm***$_{BC}$
A: *is past the it's a little bit past the mermaid's body*

———

A: *when you look at the lower left corner of the iron*
B: *[**okay**$_{BC}$]*
A: *[where] the turquoise stuff is [and you]*
B: *[**mm-hm**$_{BC}$]*
A: *know the bottom point out to the farthest left for that region*

———

A: *the blinking image is a lawnmower*
B: ***okay***$_{BC}$
A: *and it's gonna go below the yellow lion and above the bl- blue lion*
B: ***mm-hm***$_{BC}$

———

A: *the bottom black part is almost aligned to the white feet of the M&M guy*
B: *[**okay**$_{Agr}$]*
A: *[**yeah**$_{PEnd}$]* (end-of-task)

———

A: ***okay***$_{CBeg}$ *um the blinking image is the iron*

———

A: ***okay***$_{CBeg}$ *it's uh the l- I guess the lime that's blinking*

———

A: *nothing lined up real well*
B: ***yeah***$_{Agr}$ *that's* ***right***$_{Mod}$
A: *that was good* ***okay***$_{CEnd}$

———

A: *that's awesome*
B: *you're still the ace* ***alright***$_{CEnd}$

———

A: *his beak's kinda orange* ***right***$_{Chk}$
B: ***uh-huh***$_{Agr}$
A: *you can't see any of that*

———

A: *that's like a smaller amount than it is on the* ***right***$_{Mod}$ *side to the ear [*$\mathbf{right}_{Chk}$*]*
B: *[*$\mathbf{right}_{Agr}$*]*
A: $\mathbf{okay}_{Agr}$

A: *the lower* ***right***$_{Mod}$ *corner*
B: $\mathbf{yeah}_{Agr}$ *the lower* ***right***$_{Mod}$ *corner*

A: *let's start over*
B: $\mathbf{okay}_{Agr}$
A: $\mathbf{okay}_{PBeg}$ *so you have your crescent moon*

A: *but not any of the yellow [part]*
B: *[*$\mathbf{okay}_{PBeg}$*] so would the top of the ear be aligned to like where*

A: *the like head of the lion to like the where the grass shoots out there's that's a significant difference*
B: $\mathbf{okay}_{PBeg}$ *so there's definitely a bigger space from the blue lion to the lawnmower than there is from the handle to the feet of the yellow*

A: ***alright****? I'll try it* (7.81 sec) $\mathbf{okay}_{BTsk}$
B: $\mathbf{okay}_{CBeg}$ *the owl is blinking*

A: *that thing is gonna be like* (0.99 sec) $\mathbf{okay}_{Stl}$ (0.61 sec) *one pixel to the* ***right***$_{Mod}$ *of the edge*

Acknowledgments

This work was supported in part by NSF IIS-0307905, NSF IIS-0803148, ANPCYT PICT-2009-0026, UBACYT 20020090300087, CONICET, and the Slovak Agency for Science and Research Support (APVV-0369-07). We thank Fadi Biadsy, Héctor Chávez, Enrique Henestroza, Jackson Liscombe, Shira Mitchell, Michael Mulley, Andrew Rosenberg, Elisa Sneed German, Ilia Vovsha, Gregory Ward, and Lauren Wilcox for valuable discussions and for their help in collecting, labeling, and processing the data.

References

Allwood, J., J. Nivre, and E. Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–30.

Beckman, Mary E. and Julia Hirschberg. 1994. The ToBI annotation conventions. Available on-line at `http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html`.

Bevacqua, E., M. Mancini, and C. Pelachaud. 2008. A listening agent exhibiting variable behavior. In B. H. Prendinger, J. Lester, and M. Ishizuka, editors, *Intelligent Virtual Agents*, pages 262–269. Springer, Berlin.

Bhuta, T., L. Patrick, and J. D. Garnett. 2004. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, 18(3):299–304.

Boersma, Paul and David Weenink. 2001. Praat: Doing phonetics by computer. Available at `http://www.praat.org`.

Brown, G., K. L. Currie, and J. Kenworthy. 1980. *Questions of Intonation*. University Park Press, Baltimore, MD.

Bunt, H. C. 1989. Information dialogues as communicative actions in relation to user modelling and information processing. In M. M. Taylor, F. Neel, and D. G. Bouwhuis, editors, *The Structure of Multimodal Dialogue*, pages 47–73. Elsevier, Amsterdam.

Bunt, H. C., R. Morante, and S. Keizer. 2007. An empirically based computational model of grounding in dialogue. In

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 283–290, Antwerp.

Cathcart, N., J. Carletta, and E. Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 51–58, Budapest.

Charniak, Eugene and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 118–126, Pittsburgh, PA.

Clark, H. H. and Susan Brennan. 1991. Grounding in communication. In L. Resnick, J. Levine, and S. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association (APA), Hyattsville, MD.

Clark, H. H. and E. F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Cohen, Robin. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 22nd Annual Meeting Association for Computational Linguistics (ACL)*, pages 251–258, Stanford, CA.

Cohen, William C. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123, Tahoe City, CA.

Core, Mark G. 1998. Analyzing and predicting patterns of DAMSL utterance tags. In *Working Notes of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 18–24, Stanford, CA.

Core, M. G. and J. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA.

Cortes, Corinna and Vladimir Vapnik. 1995. Support vector networks. *Machine Learning*, 20(3):273–297.

Duncan, Starkey. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Eskenazi, L., D. G. Childers, and D. M. Hicks. 1990. Acoustic correlates of vocal quality. *Journal of Speech, Language and Hearing Research*, 33(2):298–306.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. Ldc93s1: Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

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

Goodwin, C. 1981. *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, New York.

Gravano, Agustín, Stefan Benus, Héctor Chávez, Julia Hirschberg, and Lauren Wilcox. 2007. On the role of context and prosody in the interpretation of 'okay'. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 800–807, Prague.

Gravano, Agustín and Julia Hirschberg. 2009a. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech*, pages 1019–1022, Brighton.

Gravano, Agustín and Julia Hirschberg. 2009b. Turn-yielding cues in task-oriented dialogue. In *Proceedings of the 10th SIGdial Workshop on Discourse and Dialogue*, pages 253–261, London.

Gravano, Agustín and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(3):601–634.

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

Hirschberg, J. 1990. Accent and discourse context: Assigning pitch accent in synthetic speech. In *Proceedings of the 8th National Conference on Artificial Intelligence*, volume 2, pages 952–957, Boston, MA.

Hirschberg, Julia and Diane Litman. 1987. Now let's talk about now: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 163–171, Stanford, CA.

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

Hirschberg, Julia and Christine Nakatani. 1996. A prosodic analysis of discourse

segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting Association for Computational Linguistics (ACL)*, pages 286–293, Santa Cruz, CA.

Hjalmarsson, Anna. 2009. On cue—Additive effects of turn-regulating phenomena in dialogue. In *Proceedings of Diaholmia—13th Workshop on the Semantics and Pragmatics of Dialogue*, pages 27–34, Stockholm.

Hjalmarsson, Anna. 2011. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

Hobbs, Jerry R. 1990. The Pierrehumbert-Hirschberg theory of intonational meaning made simple: Comments on Pierrehumbert and Hirschberg. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, MA, pages 313–323.

Hockey, B. A. 1993. Prosody and the role of 'okay' and 'uh-huh' in discourse. In *Proceedings of the Eastern States Conference on Linguistics*, pages 128–136, Columbus, OH.

Jefferson, G. 1984. Notes on a systematic deployment of the acknowledgement tokens "yeah"; and "mm hm". *Research on Language & Social Interaction*, 17(2):197–216.

Jekat, S., A. Klein, E. Maier, I. Maleck, M. Mast, and J. J. Quantz. 1995. Dialogue acts in VERBMOBIL. Technical report Verbmobil-Report 65, Universitaet Erlangen, Berlin.

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

Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Koiso, H., Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech: Special Issue on Prosody and Conversation*, 41(3-4):295–321.

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

Lampert, A., R. Dale, and C. Paris. 2006. Classifying speech acts using verbal response modes. In *Proceedings of the Australasian Language Technology Workshop*, pages 34–41, Sydney.

Litman, Diane. 1994. Classifying cue phrases in text and speech using machine learning. In *Proceedings of the 12th National Conference on Artificial Intelligence - AAAI*, pages 806–813, Seattle, WA.

Litman, Diane. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence*, 5:53–94.

Litman, Diane and Julia Hirschberg. 1990. Disambiguating cue phrases in text and speech. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 251–256, Helsinki.

Litman, D. J. and J. F. Allen. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11(2):163–200.

Litman, D. J. and R. J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 108–115, Cambridge, MA.

Maatman, R. M., J. Gratch, and S. Marsella. 2005. Natural behavior of a listening agent. In *5th International Conference on Intelligent Virtual Agents*, pages 25–36, Kos.

Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Morency, L. P., I. de Kok, and J. Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents*, pages 176–190, Tokyo.

Mushin, I., L. Stirling, J. Fletcher, and R. Wales. 2003. Discourse structure, grounding, and prosody in task-oriented dialogue. *Discourse Processes*, 35(1):1–31.

Novick, D. G. and S. Sutton. 1994. An empirical model of acknowledgment for spoken-language systems. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 96–101, Morristown, NJ.

Pierrehumbert, Janet and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, MA, pages 271–311.

Pierrehumbert, J. B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Pitrelli, John F., Mary E. Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference of Spoken Language Processing (ICSLP)*, pages 123–126, Yokohama.

Quinlan, John Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Waltham, MA.

Ratnaparkhi, A., E. Brill, and K. Church. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, Philadelphia, PA.

Redeker, G. 1991. Review article: Linguistic markers of linguistic structure. *Linguistics*, 29(6):1139–1172.

Reichman, Rachel. 1985. *Getting Computers to Talk like You and Me*. MIT Press, Cambridge, MA.

Reithinger, N. and M. Klesen. 1997. Dialogue act classification using language models. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2235–2238, Rhodes.

Roque, A. and D. Traum. 2009. Improving a virtual human using a model of degrees of grounding. In *Proceedings of the 21st International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1537–1542, Pasadena, CA.

Rosenberg, A. and J. Hirschberg. 2009. Detecting pitch accent at the word, syllable and vowel level. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT) Conference*, pages 81–84, Boulder, CO.

Rosenberg, Andrew. 2010a. AuToBI – A tool for automatic ToBI annotation. In *Proceedings of Interspeech*, pages 146–149, Makuhari.

Rosenberg, Andrew. 2010b. Classification of prosodic events using quantized contour modeling. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT) Conference*, pages 721–724, Los Angeles, CA.

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

Schegloff, E. A. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Tannen D, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. APA, Hyattsville, MD.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge University Press, Cambridge, UK.

Shriberg, E., R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):443–492.

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

Traum, David. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Rochester University, Rochester, NY.

Traum, David and James Allen. 1992. A speech acts approach to grounding in conversation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 137–140, Banff.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Walker, M. A. 1992. Redundancy in collaborative dialogue. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 345–351, Morristown, NJ.

Walker, M. A. 1993a. *Informational Redundancy and Resource Bounds in Dialogue*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Walker, M. A. 1993b. When given information is accented: Repetition, paraphrase and inference in dialogue. In *LSA Annual Meeting*, pages 231–240, Los Angeles, CA.

Walker, M. A. 1996. Inferring acceptance and rejection in dialogue. *Language and Speech*, 39(2-3).

Ward, N. and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Witten, I. H. and E. Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, Waltham, MA.

Yngve, V. H. 1970. On getting a word in edgewise. In *Proceedings of the 6th Regional Meeting of the Chicago Linguistic Society*, volume 6, pages 657–677, Chicago, IL.

Young, S., G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. The HTK Book, version 3.4. Available on-line at `http://htk.eng.cam.ac.uk`.

Zufferey, S. and A. Popescu-Belis. 2004. Towards automatic identification of discourse markers in dialogs: The case of 'like'. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 63–71, Boston, MA.