Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue

Agustín Gravano

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2009

ABSTRACT

Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue

Agustín Gravano

As interactive voice response systems spread at a rapid pace, providing an increasingly more complex functionality, it is becoming clear that the challenges of such systems are not solely associated to their synthesis and recognition capabilities. Rather, issues such as the coordination of turn exchanges between system and user, or the correct generation and understanding of words that may convey multiple meanings, appear to play an important role in system usability. This thesis explores those two issues in the Columbia Games Corpus, a collection of spontaneous task-oriented dialogues in Standard American English.

We provide evidence of the existence of seven turn-yielding cues — prosodic, acoustic and syntactic events strongly associated with conversational turn endings — and show that the likelihood of a turn-taking attempt from the interlocutor increases linearly with the number of cues conjointly displayed by the speaker. We present similar results related to six backchannel-inviting cues — events that invite the interlocutor to produce a short utterance conveying continued attention.

Additionally, we describe a series of studies of affirmative cue words — a family of cue words such as *okay* or *alright* that speakers use frequently in conversation for several purposes: for acknowledging what the interlocutor has said, or for cueing the start of a new topic, among others. We find differences in the acoustic/prosodic realization of such functions, but observe that contextual information figures prominently in human disambiguation of these words. We also conduct machine learning experiments to explore the automatic classification of affirmative cue words. Finally, we examine a novel measure of speaker entrainment related to the usage of these words, showing its association with task success and dialogue coordination.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my profound gratitude to my advisor, Julia Hirschberg, whose expertise, patience and understanding have given me the best guidance I could possibly imagine through all these years.

I also want to thank my dissertation committee — Kathleen McKeown, Rebecca Passonneau, Maxine Eskenazi, and Amanda Stent — for generously devoting their time and effort to the review of this document, and for their valuable comments and suggestions.

I am very thankful to my colleagues at the Speech Lab, with whom it has been an immense pleasure to work: Stefan Benus, Fadi Biadsy, Sasha Caskey, Bob Coyne, Frank Enos, Martin Jansche, Jackson Liscombe, Sameer Maskey, and Andrew Rosenberg.

Also, I would like to thank the following people, who have collaborated in different aspects of this thesis: Gregory Ward and Elisa Sneed German (Northwestern University); Ani Nenkova (University of Pennsylvania); Héctor Chávez, David Elson, Michel Galley, Enrique Henestroza, Hanae Koiso, Shira Mitchell, Michael Mulley, Kristen Parton, Ilia Vovsha, and Lauren Wilcox.

Finally, I would never have gotten very far without the unconditional love and support of my future wife, Mariana Obertello, and of my family and friends. To all of you, thank you from the bottom of my heart.

Para Mari

Chapter 1

Introduction

The last few decades have witnessed considerable advances in text-to-speech (TTS), automatic speech recognition (ASR) and other speech technologies. Consequently, applications based on interactive voice response (IVR) systems have spread at a rapid pace, and their functionality has become increasingly more complex. However, interactions with state-of-the-art IVR systems are often described by users as “confusing” and even “intimidating”. As synthesis and recognition capabilities continue to improve, it is becoming clear that such negative judgments may be found in other aspects of the systems as well.

A possible explanation for part of the unsatisfactory user experience is coordination problems in the exchange of speaking turns between system and user. For example, currently the most common method for determining when the user has yielded the speaking turn consists in waiting for a long pause. However, this strategy is rarely used by humans, who rely instead on other types of cues, including syntactic, prosodic and acoustic ones, to anticipate turn transitions. If such cues could be modeled and incorporated into IVR systems, it would be possible to make faster and more accurate turn-taking decisions, thus making interaction more fluent.

Another dimension of spoken language that is important for IVR systems to model are expressions such as *by the way*, *however* or *after all* that humans use frequently for structuring discourse and shaping conversation, rather than for making a semantic contribution. A particular subclass of such expressions is especially frequent in task-oriented dialogue: individual words such as *okay*, *yeah* and *alright*, which we term AFFIRMATIVE CUE WORDS.

These words may be used in conversation for several purposes: for acknowledging what the interlocutor has said, for displaying interest and continued attention, or for cueing the start of a new topic, among others. IVR systems lacking a model of the usage of these words are likely to run into communication problems, either by producing them improperly and thus confusing the user, or by misunderstanding the users' productions.

As we progressively understand and incorporate these and other factors into our models, the quality of IVR systems should tend to improve, approaching human behavior. Bearing this long-term goal in mind, the present work represents a comprehensive attempt to (i) model contextual, acoustic and prosodic cues for anticipating the end of speaking turns, which may aid conversational partners in engaging in synchronized conversation; and (ii) characterize the contextual, acoustic and prosodic differences in the realization of affirmative cue words, which may aid listeners in disambiguating their meaning. Our hope is that it will later be possible to incorporate the resulting models into IVR systems, thus improving their performance.

This work makes no strong cognitive claims about the degree of awareness of speakers when producing any of the mentioned cues, or about the degree of awareness of listeners when perceiving and/or using such cues. We do not propose a mental model of the interactions between conversational partners. Rather, our goal consists merely in finding and describing associations between observed phenomena (such as turn-taking decisions) and objective, measurable events (such as variations in features such as pitch or intensity).

Additionally, this study briefly explores a promising research topic in Computational Linguistics that investigates how speakers tend to adapt their speech to match their conversational partners'. We examine a novel dimension of this phenomenon related to the usage of high-frequency words, including affirmative cue words, and show its association with task success and dialogue coordination, results that could have a substantial impact on the quality of IVR systems.

All experiments described in this thesis are performed on the Columbia Games Corpus, a collection of spontaneous task-oriented dialogues in Standard American English (SAE). Thus, our conclusions may not necessarily generalize to other populations; e.g. to other conversation genres, or to other English variants. Future research should verify the validity of our findings in different settings. However, note that, since most IVR applications are task-oriented, our results should apply at least to such systems.

This thesis is organized as follows. Part I introduces the Columbia Games Corpus, describing how the data were collected and subsequently annotated. Part II presents statistical studies aimed at identifying individual and complex cues for anticipating conversational turn endings. Part III describes a series of experiments on the production and perception of affirmative cue words.

Part I

The Columbia Games Corpus

Chapter 2

Corpus Description

The materials for all experiments in this thesis were taken from the Columbia Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from native speakers of Standard American English (SAE). The corpus was collected and annotated jointly by the Spoken Language Group at Columbia University and the Department of Linguistics at Northwestern University, as part of an ongoing project of prosodic variation in SAE (NSF IIS-0307905). The following sections describe the collection and annotation processes. Appendix B provides additional information, including the complete instructions screens shown to the subjects and the full sets of images used in each game. The Games Corpus was originally designed to test a set of hypotheses regarding how accentuation patterns are affected by grammatical function and information status of discourse entities. Appendix B also describes such hypotheses in detail.

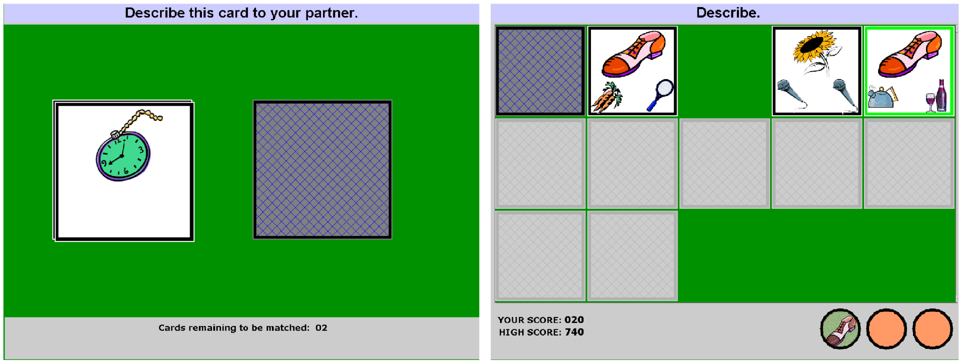
2.1 Corpus collection

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 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. The complete script read to the subjects at the session beginning is available in Appendix B.1.

2.1.1 Cards Game

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 have descriptions as voiced and sonorant as possible (e.g., *yellow lion*, *blue mermaid*), to improve pitch track computations. Appendix B.3 shows all the images used in the Cards games, arranged as they were presented to subjects on their screens. There were two parts to each Cards game, each with different rules, but both designed to test the same hypotheses.



Describe this card to your partner.

Cards remaining to be matched: 02

Describe.

YOUR SCORE: 020
HIGH SCORE: 740

(a) (b)

Figure 2.1: Sample screens from the Cards Games.

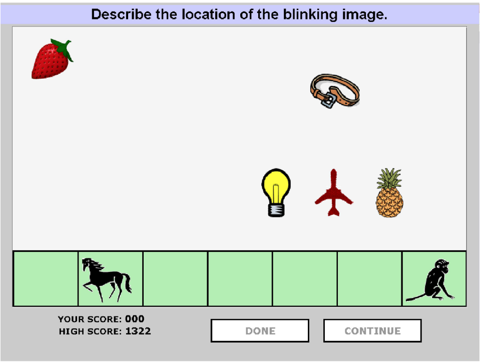
In the **first part** of the Cards game, each player's screen displayed a pile of 9 or 10 cards (Figure 2.1.a). Player A was asked to describe the top card on her pile, while Player B was asked to search through **his** pile to find the same card, clicking a button to indicate accomplishment. This process 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 prevent subjects from not describing the final card.

In the **second part** of the Cards game, each player saw a board of 12 cards on the screen (Figure 2.1.b), 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 similar card from the cards on his board. If the Searcher could find a card depicting one or more of the objects described by the Describer, the players could decide whether to declare a 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 earlier cards 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. The complete instructions are given in Appendix B.1

2.1.2 Objects Game

After completing all three instances of the Cards game, subjects were asked to play the Objects game, which we describe in this section. As in the Cards game, all images were selected to have descriptions as voiced and sonorant as possible. Appendix B.4 shows all the images used in the Objects game, arranged as they were presented to subjects on their screens.



Describe the location of the blinking image.

YOUR SCORE: 000
HIGH SCORE: 1322

Figure 2.2: Sample screen from the Objects Games.

In the Objects game, each player’s laptop displayed a game board with 5 to 7 objects (Figure 2.2). 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 their best 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 they inverted their roles: the subject that 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.

2.1.3 Subjects and sessions

Thirteen subjects (six female, seven male) 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 website `craigslist.org`. Table 2.1 shows detailed information of the sessions participants.

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. On average, the first part of each Cards game took 1.9 minutes; the second part, 5.8 minutes; and the Objects game, 21.5 minutes.

Additionally, before the actual games, subjects played one short version of each game to become familiar with the environment. The curtain was removed during these PRELIMINARY GAMES, so there could be visual communication between the players, and they were allowed to ask questions to the experimenter. The total duration of the preliminary games was 110

Session no.	Speaker A			Speaker B		
01	101	Male	25	102	Male	25
02	103	Female	25	104	Male	25
03	105	Female	25	106	Male	30
04	107	Male	30	108	Male	45
05	109	Female	50	101	Male	25
06	108	Male	45	109	Female	50
07	110	Female	50	111	Female	20
08	102	Male	25	105	Female	25
09	113	Male	20	112	Female	20
10	111	Female	20	103	Female	25
11	112	Female	20	110	Female	50
12	106	Male	30	107	Male	30

Table 2.1: Number, gender and approximate age of the participants of the twelve sessions.

minutes. These data were not used in any of the experiments presented in this thesis.

Each subject was recorded on a separate channel of a DAT recorder, at a sample rate of 48kHz with 16-bit precision, using a Crown head-mounted close-talking microphone. Each session was later downsampled to 16k, 16 bits, and saved as one stereo wav file with one player per channel, and also as two separate mono wav files, one for each player.

2.2 Corpus annotation

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 2037 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 (Pitrelli et al., 1994; Beckman and Hirschberg, 1994; see Appendix A for a brief description). All of the Objects portion of the corpus (260 minutes of dialogue) and roughly one third of the Cards portion

(60 minutes) were intonationally transcribed by trained annotators.

Part-of-speech tags were labeled automatically for the whole corpus using Ratnaparkhi et al.'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 et al., 1993), and one of the following simplified tags: noun, verb, adjective, adverb, contraction or other.

We define an INTER-PAUSAL UNIT (IPU) as a maximal sequence of words surrounded by silence longer than 50 milliseconds. 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. Boundaries of IPUs and turns are computed automatically from the time-aligned transcriptions. We classified the beginning of each turn in the Games Corpus into one of several turn-taking categories, including smooth switch, overlap, interruption, butting-in, backchannel, and others. These categories are defined in Chapter 5, along with a detailed description of the corpus annotation.

Throughout the Games Corpus, we noted that subjects made frequent use of AFFIRMATIVE CUE WORDS: the 5456 instances of such words account for 7.8% of the total words in the corpus. The most frequent affirmative cue word in the corpus is *okay*, with 2265 instances, followed by *right* (1258), *yeah* (903), *mm-hm* (478), *alright* (236), *uh-huh* (169), *yes* (53), *yep* (47), *gotcha* (26), *yup* (11), and *huh* (10). Since the usage of these words apparently varies significantly in meaning, we asked three labelers to independently classify all occurrences of the 11 words listed above in the entire corpus into several discourse/ pragmatic functions, including acknowledgment/agreement, backchannel, and literal modifier, among others. Definitions of these functions, as well as a detailed description of the labeling task, are provided in Chapter 12.

Finally, trained annotators identified all questions in the Objects portion of the Games Corpus, subsequently categorizing them according to their form (e.g., *yes-no* question, *wh*-question) and function (e.g., information request, rhetorical question). This labeling task is described in more detail in Appendix B.5.

2.2.1 Acoustic features

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.

Pitch slopes 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 milliseconds. This procedure is illustrated in Figure 2.3, 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 200ms ('A' and 'B' dashed lines, respectively). We used a similar procedure to compute the slope of intensity and stylized pitch measurements.

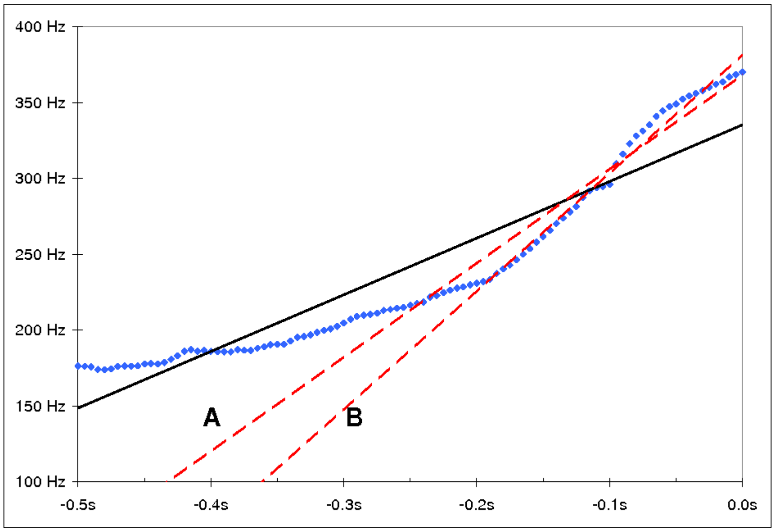


Figure 2.3: Sample pitch track with three linear regressions: computed over the whole IPU (bold line), and over the final 300ms (A) and 200ms (B).

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 4 semitones away from L, end; otherwise, remove p and start over.

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 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).

Part II

Turn-Taking

Chapter 3

Motivation and Research Goals

Interactions with state-of-the-art interactive voice response (IVR) systems are often described by users as "confusing" and even "intimidating". As speech technology continues to improve, it is becoming clear that such negative judgments are not due solely to errors in the speech recognition and synthesis components. Rather, coordination problems in the exchange of speaking turns between system and user are a plausible explanation for part of the deficient user experience (Bohus and Rudnicky, 2003; Raux et al., 2006).

For example, currently the most common method for determining when the user is willing to yield the conversational floor consists in waiting for a silence longer than a prespecified threshold, typically ranging from 0.5 to 1 second (Ferrer et al., 2002). However, this strategy is rarely used by humans, who rely instead on cues from sources such as syntax, acoustics and prosody to anticipate turn transitions (Yngve, 1970). If such TURN-YIELDING CUES can be modeled and incorporated in IVR systems, it should be possible to make faster, more accurate turn-taking decisions, thus leading to a more fluent interaction. Additionally, a better understanding of the mechanics of turn-taking could be used to vary the speech output of IVR systems to (i) produce turn-yielding cues when the system is finished speaking and the user is expected to speak next, and (ii) avoid producing such cues when the system has more things to say.

Another source of problems for state-of-the-art IVR systems are backchannel responses uttered by the user. BACKCHANNELS are short expressions, such as *uh-huh* or *mm-hm*, uttered by listeners to convey that they are paying attention, and to encourage the speaker to continue (Duncan, 1972; Ward and Tsukahara, 2000). When the user utters a backchannel while the system is talking, that input is typically interpreted as a turn-taking attempt, or BARGE-IN, thus leading the system to stop and listen — the opposite of the user's intention. Therefore, knowing the characteristics of backchannels should be a valuable tool for distinguishing them from utterances that initiate longer contributions.

A related issue is backchannel responses uttered by the system. In situations in which users are expected to enter large amounts of information, such as lists or long descriptions, the ability for the system to output backchannel responses should improve the coordination between the two parties. To achieve this, the system needs first to be capable of detecting acceptable points to produce backchannels, possibly following the speaker's production of hypothetical BACKCHANNEL-INVITING CUES conveying that a subsequent backchannel response would be welcome. The system should also know the appropriate acoustic/prosodic properties needed for backchannels to be interpreted correctly as backchannels rather than as attempts to take the turn.

These and other issues of current IVR systems can be summarized in the following empirical questions:

Q1. The system wants to keep the floor; how should it formulate its output to avoid an interruption from the user?

Q2. The system wants to keep the floor, ensuring that the user is paying attention; how should it formulate its output to give the user an opportunity to utter a backchannel?

Q3. The system wants to yield the floor to the user; how should it formulate its output to invite the user to take the turn?

Q4. The user has produced a short segment of speech; how can the system tell whether that was a backchannel or an attempt to take the turn?

Q5. The user is speaking; how can the system know when it is an appropriate moment to take the turn?

Q6. The user is speaking; how can the system know whether and when it should produce a backchannel as positive feedback to the user?

Q7. The user is speaking and the system wants to produce a backchannel response; how should it formulate its output for the backchannel to be interpreted correctly?

These questions guide us throughout the research on turn-taking phenomena presented in this thesis. Our hope is that our findings will help improve the naturalness and usability of IVR systems in the short term, as well as open new research directions for further advances in the field.

It is important to note that we make no strong cognitive claims about the awareness of speakers when producing turn-taking cues, or of listeners when perceiving and/or using such cues. Rather than proposing a mental model of the interactions between conversational partners, we aim at finding and describing associations between turn-taking phenomena (e.g., turn changes or backchannels) and objective, measurable events (e.g., variations in features such as pitch or intensity), hoping that such associations will eventually be useful in speech processing applications.

Chapter 4

Previous Research on Turn-Taking

In influential work, Sacks et al. (1974) present a characterization of turn-taking in conversations between two or more persons. After providing a detailed description of fourteen "grossly apparent facts" about human conversation, such as "speaker change recurs" or "one party talks at a time", they enunciate a basic set of rules governing turn construction: At every TRANSITION-RELEVANCE PLACE (TRP),

(a) if the current speaker (CS) selects a conversational partner as the next speaker, then such partner must speak next;

(b) if CS does not select the next speaker, then anyone may take the next turn;

(c) if no one else takes the next turn, then CS may take the next turn.

The authors do not provide a formal definition of TRPs, but conjecture that these tend to occur at syntactic "possible completion points", with intonation playing a decisive role.

The question of what types of cues humans exploit for engaging in synchronized conversation has been addressed repeatedly over the past decades. Yngve (1970) shows that pausing in itself is **not** a turn-yielding signal, in clear opposition to the strategy used in most of today's IVR systems.

In a series of analyses of face-to-face conversations in Standard American English (SAE), Duncan (1972; 1973; 1974; 1975; Duncan and Fiske, 1977) conjectures that speakers display complex signals at turn endings, composed of at least one of six discrete behavioral cues:

(1) any phrase-final intonation other than a sustained, intermediate pitch level; (2) a drawl on the final syllable of a terminal clause; (3) the termination of any hand gesticulation; (4) a stereotyped expression like *you know*; (5) a drop in pitch and/or loudness in conjunction with such a stereotyped expression; (6) the completion of a grammatical clause. The central finding of these studies is that the likelihood of a turn-taking attempt by a listener increases linearly with the number of turn-yielding cues conjointly displayed. Duncan's work has been criticized for two reasons (Beattie, 1981; Cutler and Pearson, 1986). First, it lacks a formal description of the cues under observation. No metric, specific procedure or inter-labeler reliability measure is provided, suggesting that the author merely recorded his subjective impressions. Second, the robustness of its statistical analysis is at least questionable. Duncan reports a correlation of 0.96 ($p < 0.01$) between number of turn-yielding cues displayed and percentage of auditor turn-taking attempts, but this computation is based on a reduced sample size. For example, as little as nine instances of the simultaneous display of five cues are reported, and therefore a small fluctuation in the data may change the results substantially. Nonetheless, Duncan is the first to posit the existence of complex turn-yielding signals formed by individual cues such that, the more complex the signal, the higher the likelihood of a speaker change. This crucial finding has laid the groundwork for a number of subsequent studies of turn-taking that confirm many of Duncan's claims.

In one such study, Ford and Thompson (1996) seek to formalize two of Duncan's individual cues, grammatical completion and intonation, and study their correlation with speaker changes in two naturally occurring conversations in SAE. For grammatical completion, Ford and Thompson define SYNTACTIC COMPLETION POINTS as those points at which an utterance could possibly be interpreted as syntactically complete "so far" in the discourse context, independent of intonation or pause (see Figure 4.1 for a few examples). For intonation, they consider a binary distinction between *final* (either rising or falling) or *non-final* (all other). They find that syntactic completion points operate together with a rising or falling final intonation as an important turn-yielding cue. Also, they show that while almost all (98.8%) intonationally complete utterances are also syntactically complete,[1] only half

[1] Ford and Thompson (1996) use a perceptual definition of intonational unit by Du Bois et al. (1993): "a stretch of speech uttered under a single coherent intonation contour"; and rely on acoustic, prosodic and

V: *and his knee was being worn/ okay/ wait/ it was bent/ that way/*

D: *I mean it's it's not like wine/ it doesn't taste like wine/ but it's*
W: *fermented/*
D: *white/ and milky/ but it's fermented/*

Figure 4.1: Examples of syntactic completion points, indicated by slashes. Taken from Ford and Thompson (1996) [p. 144].

(53.6%) of syntactically complete utterances are intonationally complete, thus highlighting the prominent role played by intonation in marking discourse and dialogue structure.

Wennerstrom and Siegel (2003) enrich Ford and Thompson's technique with a more precise definition of final intonation based on the system developed by Pierrehumbert (1980), a predecessor of ToBI. They use six phrase-final intonational categories: *high rise* (H-H% in the ToBI system), *low* (L-L%), *plateau* (H-L%), *low rise* (L-H%), *partial fall* (also L-L%),[2] and *no boundary.* They find *high rise* intonation to be a strong cue of turn finality, with 67% of its occurrences coinciding with turn shifts, followed by *low*, with 40%. The remaining four intonational categories strongly correlate with turn holds. Additionally, Wennerstrom and Siegel analyze the interaction between intonation and Ford and Thompson's syntactic completion, and report similar findings in line with the hypothesized existence of complex turn-yielding signals.

A potential problem of observational studies such as the ones presented above is that they only collect indirect evidence of turn-yielding cues, arising from the fact that conversational decisions are **optional**. A listener who intends to let the speaker continue to hold the floor may choose not to act on turn-yielding cues displayed by the speaker. Furthermore, when using corpora of spontaneous conversations, it is extremely difficult to obtain

timing cues to manually identify unit boundaries, independently of syntax.

[2] The *partial fall* category is described as a "downward sloping pitch contour that subsided before reaching the bottom of the speaker's range" [p. 84], and corresponds to a special type of L-L% in the ToBI system called 'suspended fall' (Pierrehumbert, 1980).

a balanced set of utterances controlling for the diverse features under study; e.g., utterance pairs from the same speaker, with the same syntactic and semantic meaning, but one half in turn-medial position and the other half in turn-final position. To address these issues, there have been several production and perception experiments aimed at replicating in the laboratory the turn-taking decisions made by speakers. In a typical production study, participants read or enact fabricated dialogues with controlled target utterances; in a typical perception study, subjects classify a set of utterances into turn-medial or turn-final according to the believed speaker's intentions. These settings give the experimenter a great amount of control over the experimental conditions.

For instance, Schaffer (1983) presents a perception study to compare non-visual turn-taking cues in face-to-face and non-face-to-face conversations in SAE. She finds no significant differences, but reports that syntactic and lexical information appears to be more useful to listeners in judging turn boundaries than prosodic information in both conditions. Also, listeners show a great amount of variability in their perception of intonation as a turn-yielding cue. In a production and perception study of turn-taking in British English, Cutler and Pearson (1986) obtain the same results: a wide listener variability in perception of intonation as a turn-yielding cue. They also find a slight tendency to characterize a "downstep in pitch" towards the phrase end as a turn-yielding cue, and an "upstep in pitch" as a TURN-HOLDING CUE (that is, a cue that typically prevents turn-taking attempts from the listener), seemingly conflicting with Duncan's hypothesis. The subsequent findings by Wennerstrom and Siegel (2003) described above, relating *high rises* to turn shifts and *low rises* to turn holds, seem to provide a plausible explanation for this apparent contradiction.

In two perception experiments designed to study intonation and syntactic completion in British English turn-taking, Wichmann and Caspers (2001) find only mild support for Duncan's claim that both syntactic completion and anything but a high level tone work as turn-yielding cues. It is important to note, however, that it is reasonable to expect different dialects and cultures to have different turn-taking behaviors. Therefore, findings even for languages within the same group, like British vs. American English, could differ substantially.

As mentioned in the previous chapter, a related topic is backchannel-inviting cues — that is, events in the current speaker's speech that invite the listener to produce a backchannel response. This research topic has received less attention than turn-yielding cues. Ward and Tsukahara (2000) describe a region of low pitch lasting at least 110 milliseconds as a backchannel-inviting cue. They show that, in a corpus of spontaneous non-face-to-face dyadic conversations in SAE, 48% of backchannels follow a low-pitch region, while only 18% of such regions precede a backchannel response.

Shifting our attention to implementation issues, several more recent studies investigate ways of improving the turn-taking decisions made by IVR systems, by incorporating some of the features shown in previous studies to correlate with turn or utterance endings. Ferrer et al. (2002; 2003) present an approach for online detection of utterance boundaries (defined similarly to transition-relevance places), combining decision trees trained with prosodic features (related mainly to pitch level, pitch slope and phone durations) and n-gram language models. Edlund et al. (2005) experiment with a hand-crafted rule for detecting utterance boundaries: If a long-enough pause follows a long-enough speech segment that does not end in a level pitch slope, then mark the pause as an utterance end. Schlangen (2006), Atterer et al. (2008) and Baumann (2008) conduct a series of experiments using machine learning classifiers trained on prosodic and acoustic features to detect utterance boundaries. Raux and Eskenazi (2008) present an algorithm to dynamically set the threshold used for determining that a silence follows a turn boundary, based on a number of features extracted from the immediately preceding user turn. The models presented in these studies share that all of them improve over the silence-based techniques for predicting points where the speaker has finished the current utterance, a knowledge that should also improve the performance and naturalness of IVR systems (Ward et al., 2005; Raux et al., 2006). The positive results obtained by these studies encourage further research on the field.

Chapter 5

Turn-Taking in the Games Corpus

The Games Corpus (see Part I) offers an excellent opportunity to study the turn-taking management mechanisms occurring in spontaneous conversation, and to provide answers to the research questions posited in Chapter 3. A superficial analysis of the corpus reveals it to be rich in all kinds of turn-taking phenomena, as all subjects became engaged in active conversation to achieve the highest possible performance in the various game tasks, all designed to be interesting and challenging.

All conversations in the corpus are between two people collaborating to perform a common task, and take place with no visual contact between the participants. These conditions roughly replicate the typical settings of current telephone IVR systems, in which a person is assisted by a remote computer using natural speech over the telephone to perform relatively simple tasks, such as making travel reservations or requesting banking information.

Conversations involving not just two, but three or more participants are very frequent in every day life, and a better understanding of their turn-taking mechanisms will be useful in speech processing tasks such as, for example, automatic meeting summarization. Even though previous studies of turn-taking (e.g. Sacks et al., 1974) do not restrict the number of conversation participants, the question of whether the rules governing turn-taking in dialogue also apply to multi-party exchanges is yet to be addressed. Therefore, there is currently no reason to assume that the results presented in this thesis generalize (or, do not generalize) beyond dyadic conversations. Further research will indeed be needed to answer this empirical question.

When visual contact is permitted between the conversation participants, a whole new dimension of complexity is introduced to the analysis of turn-taking phenomena. For instance, eye gaze and hand gesticulation are known to be strong turn-taking cues (Kendon, 1972; Duncan, 1972; McNeill, 1992). When collecting the Games Corpus, visual contact was impeded by hanging a curtain between the two participants, thus forcing all communicational to be verbal. The lack of visual contact allows us to effectively isolate audio-only cues, the central object of study in our experiments.

Finally, we take several steps to achieve results as general as possible — i.e., not true only for a specific set of speakers, but generalizable to a larger population. First, the corpus contains twelve conversations recorded from thirteen different people, as opposed to smaller numbers used in previous studies, typically limited to two or three conversations. Second, the participants of each conversation had never met each other before the recording session. This allows us to avoid any potential communicational codes or behaviors arising from preexisting acquaintances between the subjects, and that are also beyond the scope of our study. Third, in the statistical studies presented in the following chapters, we pay great attention to speaker variation. Specifically, for each result holding for all thirteen speakers together, we check and report whether the same results holds for each individual speaker.

5.1 Labeling scheme

As discussed in Chapter 3, our main research goal is to investigate the existence of acoustic, prosodic, lexical and syntactic turn-yielding and backchannel-inviting cues. That is, we search for events in the speech produced by the person holding the conversational floor that may cue the listener about an imminent turn boundary, or that may invite the listener to utter a backchannel response. With this goal in mind, we need first to define and identify various types of turn-taking phenomena in the corpus, which we later analyze separately. For example, in our search for turn-yielding cues, we need to define and identify turn boundaries, to later compare turn-final utterances against turn-medial ones. In this section we consider a number of labeling systems adopted by previous works, and describe in detail the one we choose for our experiments.

In an approach adopted by a number of studies, all exchanges are collapsed into a single CHANGE category, defined as a transition from a turn by the participant currently holding the floor to a new turn by the other participant (see Figure 5.1).[1] Some studies further subdivide this category into CHANGE WITH OVERLAP and CHANGE WITHOUT OVERLAP, depending on whether the two contributions have a non-empty temporal intersection. The second main class in this approach is the HOLD category, defined as a transition between two adjacent IPUs within a turn by the same speaker. The change and hold categories are typically contrasted to look for turn-yielding cues, with the assumption that instances of the former are more likely to contain such cues than instances of the latter. The main advantage of these simple binary and ternary distinctions is that they can be computed automatically from the speech signal: turn boundaries can be estimated using an energy-based silence detector, provided that each speaker has been recorded on a separate channel. In our case, this labeling system oversimplifies the problem, since we need to be able to differentiate phenomena such as backchannels and interruptions from regular turn changes. In other words, we need a finer grained categorization of speaker changes.

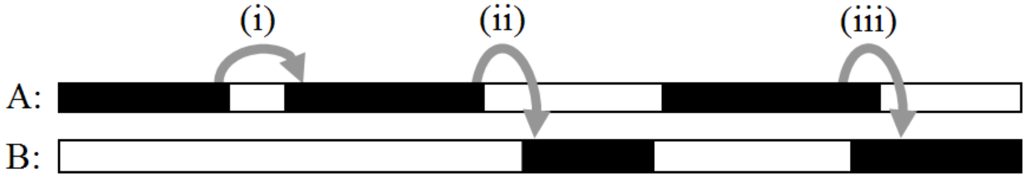
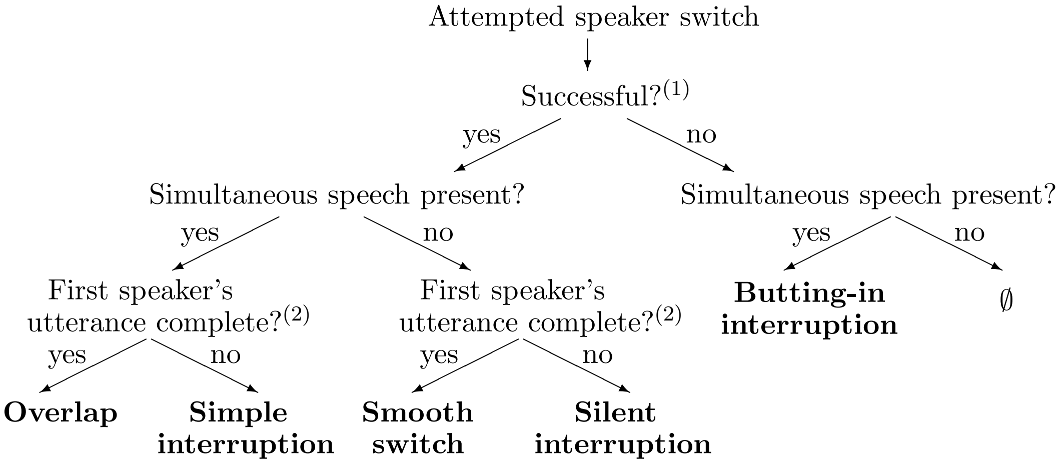


Figure 5.1: Simple 3-way definition of turn exchanges. Black segments represent speech; white segments, silence. (i) Hold, (ii) Change without overlap, (iii) Change with overlap.

One such categorization is introduced by Ferguson (1977) for a study of behavioral psychology that investigates simultaneous speech and interruptions as measures of dominance in family interaction. Beattie (1982) adopts the same system in a study of two political interviews comparing the turn-taking styles of former British Prime Ministers Jim Callaghan and Margaret Thatcher, and proposes the decision tree shown in Figure 5.2 as a systematic

[1] Recall from Chapter 2 that we define a TURN as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. An INTER-PAUSAL UNIT (IPU) is defined as a maximal sequence of words surrounded by silence longer than 50 ms.



(1) By successful it is meant that "the initiator of the attempted speaker switch gains the floor".

(2) Completeness is "judged intuitively, taking into account the intonation, syntax and meaning of the utterance".

Figure 5.2: Turn-taking labeling scheme proposed by Beattie (1981).

procedure for the manual annotation of turn exchange types. Beattie reports an almost perfect inter-labeler agreement using this labeling scheme, with a Cohen's κ score (Cohen, 1960) of 0.89. This system is better suited for our experiments on turn-yielding cues than the ones using binary and ternary distinctions. It distinguishes two exchange types (smooth switches and overlaps) in which turn-yielding cues are likely to be present, given that a turn exchange occurs and the first speaker (i.e., the one originally holding the floor) manages to finish the utterance. The remaining three types (simple, silent and butting-in interruptions) are less likely to contain turn-yielding cues, given that the first speaker is interrupted and does not manage to finish the utterance. Additionally, the difference between smooth switches and overlaps is that some simultaneous speech is present in the latter. In such cases, the listener effectively anticipates the end of a turn and starts speaking right before the interlocutor finishes, but without actually causing an interruption in the conversational flow. These cases are useful for looking for turn-yielding cues that may occur **before** the final part of the turn, and that may aid the listener in projecting the turn boundary.

We adopt a slightly modified version of Beattie's labeling scheme, depicted in Figure 5.3. The left half of the decision tree is equivalent to Beattie's scheme, but rearranged in a different order. The decision "Simultaneous speech present?" is placed higher up in the tree, as it is pre-computed automatically based on the manual orthographic transcripts of the conversations. Backchannels play an important role in our research goals, but Beattie explicitly excludes them from his study. Therefore, we incorporate backchannels in the labeling scheme by adding the decision marked (1) at the root of the decision tree. Since backchannels were identified by annotators of the function of affirmative cue words (as described in detail in Chapter 12, on page 96), we use these labels, and annotators of turn-taking are not asked to make this decision. For the decision marked (2) in Figure 5.3, we use Beattie's informal definition of utterance completeness: "Completeness [is] judged intuitively, taking into account the intonation, syntax, and meaning of the utterance" [p. 100]. Additionally, we identify three cases that do not correspond to actual turn exchanges, and thus receive special labels:

For each turn by speaker S2, where S1 is the other speaker, label S2's turn as follows:

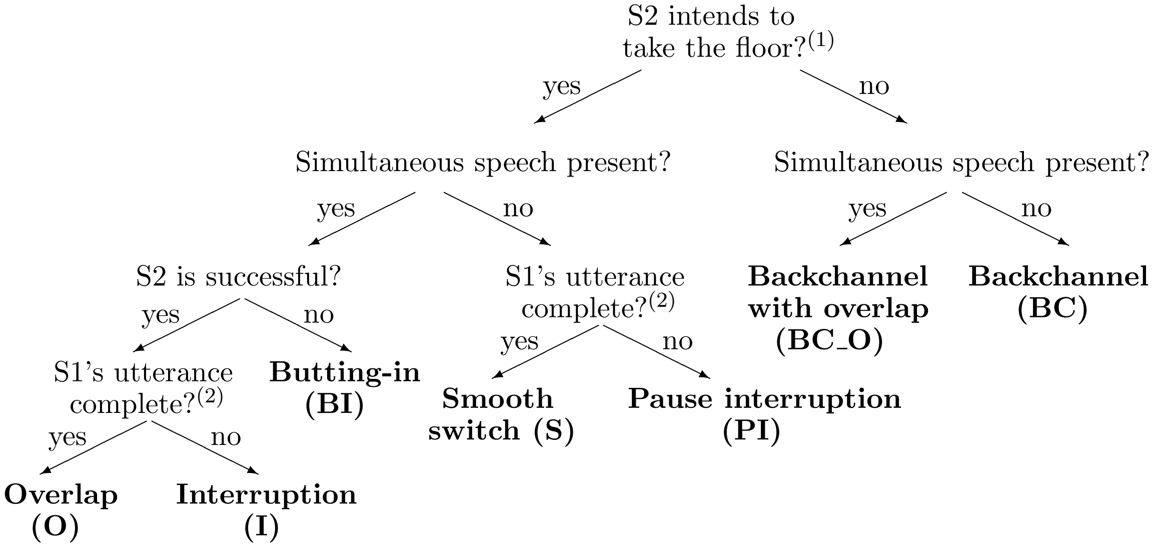
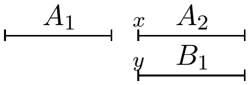

S2 intends to take the floor?(1)
yes
no
Simultaneous speech present?
yes
no
Simultaneous speech present?
yes
no
S2 is successful?
yes
no
S1's utterance complete?(2)
yes
no
Backchannel with overlap (BC_O)
Backchannel (BC)
S1's utterance complete?(2)
yes
no
Butting-in (BI)
Smooth switch (S)
Pause interruption (PI)
Overlap (O)
Interruption (I)

Figure 5.3: Turn-taking labeling scheme.

- **Task beginnings:** Turns beginning a new game task are labeled **X1**.

- **Continuation after a backchannel:** If a turn t is a continuation after a **BC** or **BC_O** from the other speaker, it is labeled **X2_O** if t overlaps the backchannel, or **X2** if not.

- **Simultaneous start:** Fry (1975) reports that humans require at least 210 milliseconds to react verbally to a verbal stimulus. Thus, if two turns begin within 210 ms of each other, they are most probably connected to preceding events than to one another. In Figure 5.4, A_1, A_2 and B_1 represent turns from speakers A and B. Most likely, A_2 is simply a continuation from A_1, and B_1 occurs in response to A_1. Thus, B_1 is labeled with respect to A_1 (not A_2), and A_2 is labeled **X3**.



$$0 < |y - x| < 210\text{ms}$$

Figure 5.4: Simultaneous start.

Finally, all continuations from one IPU to the next within the same turn are labeled automatically with the special label **H**, for ‘hold’.

Needless to say, the categories defined in this taxonomy are too broad to accommodate the wide spectrum of variation in human conversation. However, they are well suited for our turn-taking experiments, as they allow us to look for turn-yielding cues by contrasting the places where such cues are likely to occur (e.g. before smooth switches) against the places where they are not likely to occur (e.g. before holds or interruptions). Furthermore, more fine-grained distinctions, albeit closer to representing the full diversity of turn-taking events present in spontaneous dialogue, would have the cost of data sparsity, thus compromising the statistical significance of the results.

Two trained annotators labeled the whole Objects portion of the corpus separately,[2] with a Cohen’s κ score (Cohen, 1960) of 0.913 corresponding to ‘almost perfect’ agreement.[3] [4]

[2] The complete guidelines used by the annotators are presented in Appendix D.

[3] 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.

[4] Note that this κ score does not include the identification of backchannels, performed by different annotators as described in Chapter 12.

Subsequently, we performed the following steps to correct potential labeling errors. The cases with dissimilar judgments were marked for revision and given back to one of the annotators (*ANN1*), without specifying the labels assigned by the other annotator (*ANN2*). *ANN1* corrected what he considered were errors in his labels, and the process was repeated for *ANN2*, who revised the remaining differences, again blind to *ANN1*'s choices. At the end of this process, the κ score improved to 0.9895. Given the high inter-labeler agreement obtained in the Objects portion of the corpus, the Cards portion was labeled by just one trained annotator. Table 5.1 shows the distribution of turn-taking labels in the entire corpus. Additionally, there are 8123 instances of 'hold' transitions (**H**) in the Games Corpus, as defined above.

Label	Count	Percentage
BC	553	6.8%
BC_O	202	2.5%
BI	104	1.3%
I	158	1.9%
O	1067	13.1%
PI	275	3.4%
S	3247	39.9%
X1	1393	17.1%
X2	449	5.5%
X2_O	59	0.7%
X3	590	7.3%
?	37	0.5%
Total	8134	100.0%

Table 5.1: Distribution of turn-taking labels in the Games Corpus.

Chapter 6

Turn-Yielding Cues

We begin our study of turn-taking in the Columbia Games Corpus by investigating turn-yielding cues — events from acoustic, prosodic or syntactic sources, inter alia, produced by the speaker when approaching the potential end of a conversational turn, that may be used by the listener to detect, or even anticipate, an opportunity to take the floor. We adopt the assumption proposed by Duncan (1972) that individually identifiable cues may be combined together to form a complex turn-yielding signal. As discussed in the previous sections, a number of non-visual turn-yielding cues have been hypothesized in the literature: any final intonation other than a sustained pitch level; a drawl on the final syllable of a terminal clause; a drop in intensity and pitch levels; stereotyped expressions such as *you know* or *I think*; and the completion of a grammatical clause. In this chapter we examine each of these individual cues in the Games Corpus. We also present results introducing two turn-yielding cues rarely mentioned in the literature, related to voice quality (Ogden, 2002) and IPU duration (Cutler and Pearson, 1986). After considering individual cues, we describe how they are combined together to form a complex signal, and show the manner in which the likelihood of a turn switch increases with the number of cues present in such a signal.

Our general approach consists in contrasting IPUs immediately preceding smooth switches (**S**) with those immediately preceding holds (**H**). We hypothesize that turn-yielding cues are more likely to occur before **S** than before **H**. It is important to emphasize the optionality of all turn-taking phenomena and decisions: For **H**, turn-yielding cues — whatever their nature — may still be present; and for **S**, they may be sometimes absent. However, we hypothesize that their likelihood of occurrence should be much higher before **S**.

Finally, as mentioned above, we make no claims regarding whether speakers intend to produce turn-yielding cues, or whether listeners consciously perceive and/or use them to aid their turn-taking decisions. Instead, we find and describe associations between turn exchanges and a number of objective, measurable events — such as variations in pitch or intensity, or lexical and syntactic patterns, which may eventually be useful in modeling human-like behavior in IVR systems and other speech processing applications.

6.1 Individual turn-yielding cues

6.1.1 Intonation

IPU-final intonation is the turn-yielding cue most frequently mentioned in the literature (Duncan, 1972; Cutler and Pearson, 1986; Ford and Thompson, 1996; Wennerstrom and Siegel, 2003; inter alia). Anything other than a plateau (i.e., a sustained pitch level, neither rising nor falling; a H- phrase accent followed by a L% boundary tone according to the ToBI system: H-L%) has been characterized as a turn-yielding cue. In this section, we investigate the existence of this cue in the Games Corpus using manual prosodic annotations, as well as automatic computations of the IPU-final pitch slope.

First, we analyze the categorical prosodic labels in the portion of the corpus annotated using the ToBI conventions. We tabulate the phrase accent and boundary tone labels assigned to the end of each IPU, and compare their distribution for the **S** and **H** turn exchange types, as shown in Table 6.1. A chi-square test reports a significant departure from a random distribution ($\chi^2 = 1102.5$, $d.f. = 5$, $p \approx 0$). Only 13.2% of all IPUs immediately preceding a smooth switch (**S**) — where turn-yielding cues are most likely present — end in a plateau (H-L%); the majority of the remaining ones end in either a falling pitch (L-L%) or a high rise (H-H%). For IPUs preceding a Hold (**H**) the counts approximate a uniform distribution, with the plateau contours ([!]H-L%) being the most common. In other words, a smooth switch rarely follows a plateau contour, while in seven out of ten cases it follows either a high-rising or a falling contour. On the other hand, the

	S		**H**	
H-H%	**484**	**(22.1%)**	513	(9.1%)
[!]H-L%	289	(13.2%)	**1680**	**(29.9%)**
L-H%	309	(14.1%)	646	(11.5%)
L-L%	**1032**	**(47.2%)**	1387	(24.7%)
no boundary tone	16	(0.7%)	1261	(22.4%)
other	56	(2.6%)	136	(2.4%)
total	2186	(100%)	5623	(100%)

Table 6.1: ToBI phrase accent and boundary tone for IPUs preceding **S** and **H**.

high counts for the falling contour preceding a hold (24.7%) may be explained by the fact that, as discussed above, taking the turn is optional for the listener, who may choose not to act upon hearing some turn-yielding cues. Still, plateau is the contour with the highest count before holds, supporting Duncan's (1972) hypothesis that it works as a turn-holding cue. It is not entirely clear, though, what the role of the low-rising contour (L-H%) is, as it occurs in similar proportions in both cases. Finally, we note that the absence of a boundary tone works as a strong indication that the speaker has not finished speaking, since nearly all (98%) IPUs without a boundary tone precede a hold.

As an objective acoustic approximation of this perceptual feature, we use the slope of linear regression models fitted to the pitch track, both raw and stylized, computed over the final 200 and 300 milliseconds of each IPU (see Section 2.2 on page 9 for a detailed explanation). This gives us four acoustic approximations of the IPU-final intonation. The case of a plateau contour, or a sustained pitch, would correspond to a value of F_0 slope in the vicinity of zero; the second case, either a rising or a falling pitch, would correspond to a high positive or a high negative value of F_0 slope. Therefore, we use the **absolute value** of the F_0 slope calculations to differentiate these two cases.

Figure 6.1 shows the absolute value of the speaker-normalized F_0 slope,[1] both raw and

[1] All normalizations by speaker were calculated using z-scores: $z = (X - mean)/stdev$, where mean and standard deviation were computed for a given speaker over the full conversation.

stylized, computed over the final 200 and 300 milliseconds of IPUs immediately preceding smooth switches (**S**) or holds (**H**). Anova tests reveal the final slope for **S** to be significantly higher (at $p < 0.01$) than for **H** in all cases. This indicates that IPUs preceding a hold tend to be produced with a flatter final intonation, while IPUs preceding a smooth switch tend to be produced with either a rising or falling intonation. These findings provide additional support to the hypothesis that falling and high-rising final intonations tend to be associated with turn endings.

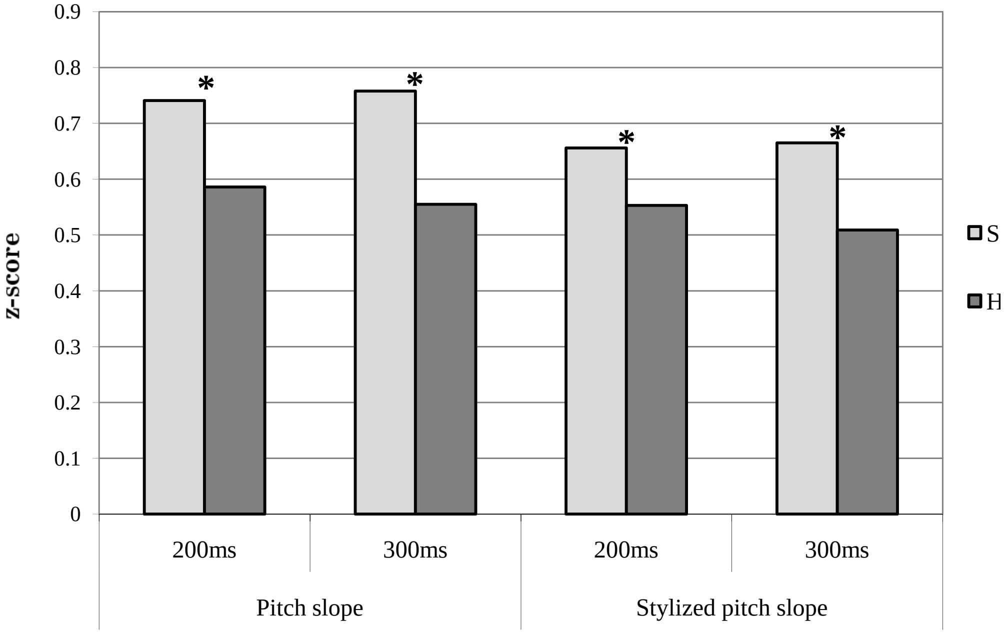


Figure 6.1: Absolute value of speaker-normalized F_0 slope, both raw and stylized, computed over the IPU's final 200 and 300 ms. Significant differences at the $p < 0.01$ level are marked with an asterisk ('$*$').

Speaker variation: For each individual speaker, we compare the absolute value of the F_0 slope over the final 300 ms of IPUs preceding **S** and **H** turn exchange types. For 12 out of the 13 speakers, this variable is significantly higher for **S** than for **H** ($p < 0.05$); for the remaining speaker (id 110), the same relation approaches significance at $p = 0.056$. This suggests that the findings reported above are valid across individual speakers. The results for each individual speaker are detailed in Appendix E.1.

Summary of findings: The results presented in this section support the hypothesis that plateau final intonation is most likely to be produced when the speaker plans to continue talking. On the other hand, smooth switches are more likely to occur following IPUs with falling or high-rising intonation. The meaning of low-rising intonation is not clear, though, as it appears to be related to switches and holds in similar proportions. Additionally, we find the lack of a boundary tone to be strongly related to turn holds.

6.1.2 Speaking rate

Duncan (1972) hypothesizes a "drawl on the final syllable or on the stressed syllable of a terminal clause" [p. 287] as a turn-yielding cue. Such a drawl would probably lead to a noticeable decrease in the speaking rate. However, preliminary exploratory analyses we have run on our corpus suggest an **increase** in speaking rate just before turn changes. In the following paragraphs we try to shed some light on this apparent contradiction.

We begin our analysis using two common definitions of speaking rate: syllables per second, and phonemes per second. Both syllable and phoneme counts were estimated using dictionaries, and word durations were extracted from the manual orthographic alignments. Figure 6.2 shows the speaker-normalized speaking rate, computed over the whole IPU and over its final word, for IPUs preceding smooth switches (**S**) or holds (**H**). The first thing that becomes clear from these results is that both measures of speaking rate, computed either over the whole IPU or over its final word, are significantly faster before **S** than before **H** (ANOVA tests, $p < 0.01$), thus indicating an increased speaking rate before turn boundaries.

Furthermore, the speaking rate is in both cases (before **S** and before **H**) significantly slower on the final word than over the final IPU. This finding is in line with phonological theories that predict a segmental lengthening near prosodic phrase boundaries (Beckman and Edwards, 1990; Wightman et al., 1992; inter alia), and may account for the drawl or lengthening described by Duncan before turn boundaries. However, it seems to be the case — at least for our corpus — that the final lengthening tends to occur at all phrase final positions, not just at turn endings. In fact, our results indicate that the final lengthening is

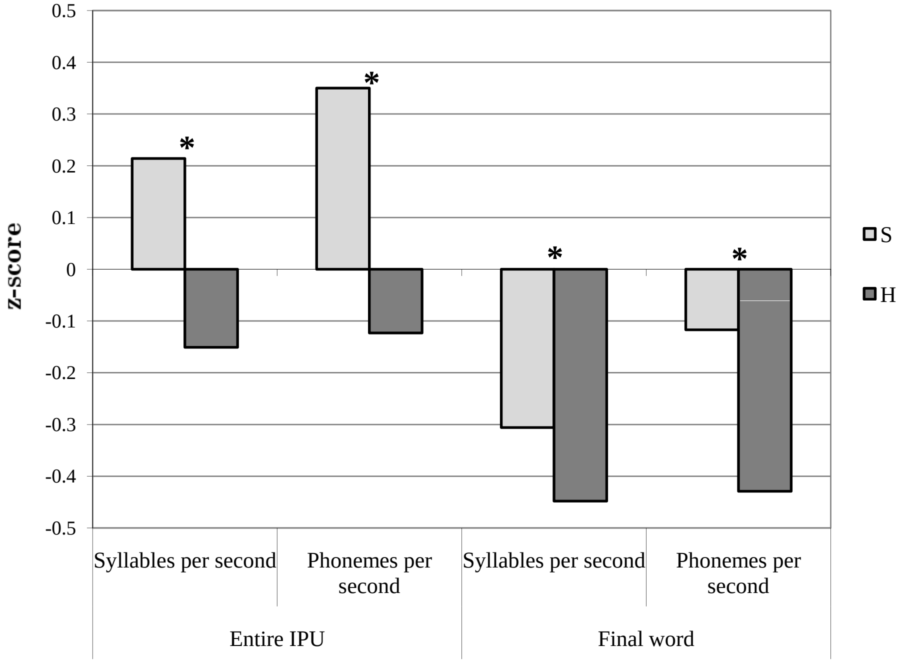


Figure 6.2: Speaker-normalized number of syllables and phonemes per second, computed over the whole IPU and over its final word.

more prominent in turn-medial IPUs than in turn-final ones, in contradiction to Duncan's hypothesis.

To investigate this issue in more detail, we look next at the most frequent IPU-final bigrams and trigrams preceding either a smooth switch (**S**) or a hold (**H**) — that is, instances of IPUs that share the final two or three lexical items. For example, 29 IPUs preceding **S**, and 52 preceding **H**, end in the final trigram *the bottom left*. For each bigram and trigram with high enough counts to perform a statistical comparison, we compare the duration of each word across turn-taking types using ANOVA tests. This way, we can compare the speaking rate while controlling for lexical variation. The results are summarized in Table 6.2. The table on the left shows the most frequent IPU-final bigrams (e.g., *hand corner*, *the iron*); the table on the right, the trigrams (e.g., *the bottom left*, *the bottom right*). For each bigram and trigram, these tables show the speaker-normalized duration of each word preceding **S** and **H**, along with the relation holding between the mean of the two groups ('less than' or 'greater than') and the *p*-value of the corresponding ANOVA test. Significant results

word	**S**		**H**	p
hand	-0.055	>	-0.252	0.379
corner	-0.246	<	0.358	**0.001**
the	-0.110	<	-0.075	0.773
iron	-0.021	<	0.382	**0.069**
the	-0.124	<	0.122	**0.080**
onion	-0.399	<	0.728	**0.000**
the	-0.372	<	-0.358	0.922
ruler	0.069	<	0.357	0.194
crescent	-0.283	<	-0.275	0.977
moon	-0.064	<	0.129	0.556

word	**S**		**H**	p
the	-0.563	<	-0.497	0.376
bottom	-0.443	<	0.001	**0.004**
left	0.244	<	0.494	**0.099**
the	-0.591	<	-0.453	**0.035**
bottom	-0.411	<	-0.208	0.165
right	-0.135	<	0.528	**0.013**
the	-0.482	<	-0.308	0.132
lower	0.014	<	0.810	**0.005**
right	0.386	<	0.464	0.768
the	-0.405	>	-0.611	**0.007**
lower	-0.467	<	-0.330	0.183
left	0.420	<	0.841	**0.004**
on	-0.382	>	-0.582	0.328
the	-0.495	>	-0.523	0.785
right	0.252	<	0.515	0.435

Table 6.2: Speaker-normalized word duration for IPU-final bigrams (e.g., *hand corner*, *the iron*) and trigrams (e.g., *the bottom left*, *the bottom right*). Significant p-values are highlighted.

are highlighted in bold font.[2] In all cases with a significant difference between the two groups, the duration of the word preceding **S** is shorter than that of the word preceding **H**. Additionally, we observe that, before holds, almost all content words have longer duration than the speaker mean (i.e., with a z-score greater than zero), probably due to the final lengthening mentioned above. However, this effect is attenuated before smooth switches, and even disappears in some cases. These findings provide further support to the hypotheses enunciated above, that (a) IPU-final words tend to be lengthened, but (b) such lengthening decreases when the IPU is in turn-final position, followed by a smooth switch.

[2] In this case we use $p < 0.1$, given the low counts in the groups being compared.

Speaker variation: All 13 speakers in the corpus show a significantly faster speaking rate — measured both as syllables per second and as phonemes per second — before smooth switches (**S**) than before holds (**H**), mirroring the results obtained when considering all subjects together. Furthermore, 10 speakers also tend to produce IPU-final words significantly faster before holds than before smooth switches, whereas the remaining three subjects show no significant difference. This indicates that our general results for speaking rate also seem to hold for individual subjects. Detailed results for each individual speaker are shown in Appendix E.1.

Summary of findings: We find that speakers tend to decrease their speaking rate toward the end of IPUs, in correspondence with a final lengthening predicted by theories of phonology. In our data, such lengthening appears to be more pronounced before holds than before smooth switches. Therefore, when comparing the speaking rate before each of these two turn-taking categories, we find that speakers tend to speak faster before turn switches. In other words, our results suggest that a reduced lengthening of IPU-final words may function as a turn-yielding cue.

One plausible explanation for this contradiction of Duncan's hypothesis is the differences in genre and in experimental setup. In Duncan's materials, both conversations are face-to-face, the first between a therapist and a psychotherapy applicant, the second between two therapists discussing another intake interview; the Games Corpus contains non-face-to-face task-oriented collaborative conversations. Duncan's dialogues do not involve performing tasks like the computer games in the Games Corpus, and they do not necessarily require collaboration between the participants. Additionally, although prior studies do not report substantial differences between face-to-face and non-face-to-face conversations, it is certainly not inconceivable that participants could modify their usage pattern of particular turn-yielding cues depending on the availability of visual contact. In any case, further research is needed to address these questions.

6.1.3 Intensity and pitch levels

A third hypothesized turn-yielding cue consists in a drop in intensity and pitch levels towards the end of the turn, in conjunction with a stereotyped expression such as *you know.* In this section, we study such a drop as a more general turn-yielding cue, independently of the lexical items at the end of the target IPU.

We analyze intensity and pitch, measured over all of each IPU, and over its final 500 and 1000 milliseconds. This way, we can study how these two acoustic features vary both across and within IPUs. Subsequently, we compare the mean value of each variable across smooth switches (**S**) and holds (**H**) as summarized in Figure 6.3.

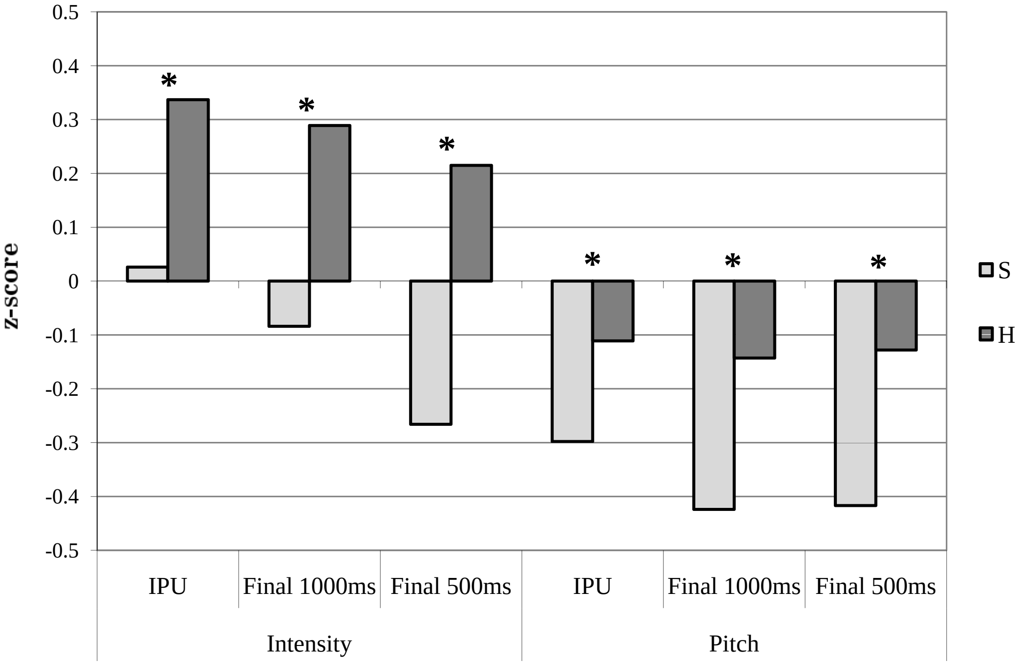


Figure 6.3: Speaker-normalized mean intensity and pitch, computed over the whole IPU and over its final 500 and 1000 ms.

For intensity, IPUs followed by **S** have a mean intensity significantly lower than those followed by **H** (ANOVA, $p < 0.01$). Also, the differences increase when moving towards the end of the IPU. This suggests that speakers tend to lower their voices towards potential turn boundaries, whereas they reach turn-internal pauses with a higher intensity. Thus, intensity level may aid listeners in detecting, or even anticipating, turn endings.

Phonological theories conjecture a declination in the pitch level, which tends to decrease gradually within utterances, and across utterances within the same discourse segment, as a consequence of a gradual compression of the pitch range (Pierrehumbert and Hirschberg, 1990). For conversational turns, then, we would expect to find that speakers tend to lower their pitch level as they reach potential turn boundaries. This hypothesis is verified by the dialogues in the Games Corpus, where we find that for pitch, IPUs preceding **S** have a significantly lower mean pitch than those preceding **H** (ANOVA, $p < 0.01$). In consequence, pitch level may also work as a turn-yielding cue.

Speaker variation: We look for individual speaker differences in mean intensity and mean pitch, computed over the final 500 milliseconds of IPUs preceding **S** and **H**. All but one speaker (id 101) show the same marked difference in intensity as reported above. For pitch, such difference exists only for seven speakers; for the other six we find no significant differences. Therefore, while a drop in intensity before turn boundaries is consistent across speakers, the evidence of a drop in the pitch level is less strong, although we find no evidence against such cue. Detailed results for each individual speaker are shown in Appendix E.1.

Summary of findings: In the Games Corpus dialogues, participants tend to produce turn endings with lower intensity and pitch levels than those showed before turn-internal pauses. While previous studies present a drop in intensity and pitch levels as a turn-yielding cue when displayed in conjunction with a stereotyped expression, we show that such a drop can actually function as a more general turn-yielding cue, independently of the lexical items.

6.1.4 Lexical cues

Stereotyped expressions such as *or something*, *you know* or *I think* — sometimes referred to as SOCIOCENTRIC SEQUENCES — have been portrayed in the literature as lexical turn-yielding cues. We look next for uses of such expressions in the Games Corpus, along with their relation to turn-taking phenomena.

Table 6.3 lists the 25 most frequent IPU-final bigrams preceding smooth switches (**S**) and holds (**H**). Note that some of the entries in this table are actually unigrams, since they do not have any preceding words in the turn — i.e., they correspond to turn-initial single-word

	S	Count	Perc.	**H**	Count	Perc.
1	*okay*	241	7.4%	*okay*	402	4.9%
2	*yeah*	167	5.1%	*on top*	172	2.1%
3	*lower right*	85	2.6%	*um*	136	1.7%
4	*bottom right*	74	2.3%	*the top*	117	1.4%
5	*the right*	59	1.8%	*of the*	67	0.8%
6	*hand corner*	52	1.6%	*blue lion*	57	0.7%
7	*lower left*	43	1.3%	*bottom left*	56	0.7%
8	*the iron*	37	1.1%	*with the*	54	0.7%
9	*the onion*	33	1.0%	*the um*	54	0.7%
10	*bottom left*	31	1.0%	*yeah*	53	0.7%
11	*the ruler*	30	0.9%	*the left*	48	0.6%
12	*mm-hm*	30	0.9%	*and*	48	0.6%
13	*right*	28	0.9%	*lower left*	46	0.6%
14	*right corner*	27	0.8%	*uh*	45	0.6%
15	*the bottom*	26	0.8%	*oh*	45	0.6%
16	*the left*	24	0.7%	*and a*	45	0.6%
17	*crescent moon*	23	0.7%	*alright*	44	0.5%
18	*the lemon*	22	0.7%	*okay um*	43	0.5%
19	*the moon*	20	0.6%	*the uh*	42	0.5%
20	*tennis racket*	20	0.6%	*the right*	41	0.5%
21	*blue lion*	19	0.6%	*the bottom*	39	0.5%
22	*the whale*	18	0.6%	*I have*	39	0.5%
23	*the crescent*	18	0.6%	*yellow lion*	37	0.5%
24	*the middle*	17	0.5%	*the middle*	37	0.5%
25	*of it*	17	0.5%	*I've got*	34	0.4%

Table 6.3: 25 most frequent final bigrams preceding each turn-taking type.

IPUs. Such unigrams comprise mostly affirmative cue words such as *okay*, *yeah*, or *alright*. These words are strongly overloaded, in the sense that they may perform very different functions. For example, they may start a new discourse segment (thus holding the floor), or finish the current discourse segment (thus potentially releasing the floor). Therefore, the occurrence of these words does not constitute a turn-yielding or turn-holding cue *per se*; rather, additional contextual, acoustic and prosodic information is needed to disambiguate their meaning. Affirmative cue words are studied in detail in Part III of this thesis.

Most of the top IPU-final bigrams preceding smooth switches and holds are specific to the computer games in which the subjects participated. The cards used in the Cards game tend to be spontaneously described by subjects from top to bottom and from left to right; for example,

> A: *I have a blue lion on top # with a lemon in the bottom left # and a yellow crescent moon in- # i- # in the bottom right*
>
> B: *oh okay* [...]

In consequence, bigrams such as *lower right* and *bottom right* are common before **S**, while *on top* or *bottom left* are common before **H**. These are all task-specific lexical constructions and do not constitute stereotyped expressions in the traditional sense.

Affirmative cue words and game-specific expressions cover the totality of the 25 most frequent IPU-final bigrams listed in Table 6.3. Further down in the list, we find some rare uses of stereotyped expressions preceding smooth switches, all with only marginal counts: *I guess* (6 instances, or 0.18% of the total), *I think* (4), and *you know* (2). Notably, there were more instances of each of these expressions before holds: 6, 5 and 21, respectively, challenging the idea that the mere occurrence of these expressions works as a strong turn-yielding cue. As with affirmative cue words, more information from other sources seems to be necessary to disambiguate the meaning of these expressions.

While we do not find clear examples of lexical turn-yielding cues in our task-oriented corpus, we do find two lexical turn-holding cues: word fragments and filled pauses. As depicted in Table 6.4, both are much rarer before smooth switches (**S**) than before holds (**H**). This suggests that, after a word fragment or a filled pause, the speaker is much more likely to intend to continue holding the floor. This notion of disfluencies serving as a turn-

	S		H	
Word fragments	10	(0.3%)	549	(6.7%)
Filled pauses	31	(1.0%)	764	(9.4%)
Total IPUs	3246	(100%)	8123	(100%)

Table 6.4: Distribution of IPU-final word fragments and filled pauses preceding each turn-taking type.

taking device has been studied by Goodwin (1981), who shows that they may be used to secure the listener's attention at turn beginnings.

Summary of findings: We find no evidence in the Games Corpus that stereotyped expressions, such as *you know* or *I think*, represent lexical turn-yielding cues. In fact, affirmative cue words, such as *okay* or *yeah*, and game-specific expressions, such as *lower right* or *on top*, cover all of the most frequent IPU-final unigrams and bigrams, preceding both smooth switches and holds. Affirmative cue words are overloaded, used both to initiate and to end discourse segments, among other functions; thus, they do not represent lexical turn-yielding cues in themselves. While game-specific expressions are likely to aid listeners in detecting or anticipating turn endings, they are particular to the computer games played by the subjects in the Games Corpus, and thus not generalizable to other task-oriented dialogues. However, our findings suggest that participants in task-oriented dialogues tend to structure their utterances in a way that facilitates the processing by the listener, a way that may even be negotiated and agreed upon — either implicitly or explicitly — by both participants at the beginning of the conversation. For example, in the Games Corpus, subjects tend to describe the cards in a top to bottom, left to right fashion. When such a structure is available, listeners may effectively use it as a turn-yielding cue to detect or anticipate turn boundaries.

6.1.5 Textual completion

Several authors (Duncan, 1972; Sacks et al., 1974; Ford and Thompson, 1996; Wennerstrom and Siegel, 2003, inter alia) claim that some sort of completion independent of intonation and interactional import functions as a turn-yielding cue. Although some call this *syntactic completion*, all authors acknowledge the need for semantic and discourse information in judging utterance completion: "we judged an utterance to be syntactically complete if, in its discourse context, it could be interpreted as a complete clause" (Ford and Thompson, 1996, p. 143); "context could also influence coding decisions" (Wennerstrom and Siegel, 2003, p. 85). Therefore, we choose the more neutral term TEXTUAL COMPLETION for this phenomenon.

In this section we describe how we manually annotated a portion of the corpus using a simple definition of textual completion. These data were subsequently used to train a machine learning (ML) classifier, with which we automatically labeled the whole Games Corpus. Finally, we present results relating both manual and automatic textual completion labels to turn-taking phenomena.

6.1.5.1 Manual labeling

In conversation, listeners judge textual completion incrementally and without access to future phrases. To simulate the same conditions in the labeling task, annotators were asked to judge the textual completion of a turn up to a target pause, and did not have access to the transcripts after the target pause. Annotators had access only to the written transcript of the current turn up to the target pause, and also the full previous turn by the other speaker (if any). These are a few sample tokens:

A: *the lion's left paw our front*

B: *yeah and it's th- right so the*

——

A: *and then a tea kettle and then the wine*

B: *okay well I have the big shoe and the wine*

——

A: —

B: *okay there is a belt in the lower right a microphone in the lower left*

——

A: *so when you say directly above you really mean directly above the right arrow the the arrow the owl*

B: *the owl yeah*

We selected 400 tokens at random from the Games Corpus. The target pauses were also chosen at random. To obtain a good coverage of the variation present in the corpus, tokens were selected in such a way that 100 of them were followed by speech from the same speaker (i.e., preceding a hold, or **H**), 100 by a backchannel from the other speaker (**BC**), 100 by a smooth switch to the other speaker (**S**), and 100 by a pause interruption by the other speaker (**PI**). Three annotators labeled each token independently as either complete or incomplete according to these guidelines:

> Determine whether you believe what speaker B has said up to this point could constitute a complete response to what speaker A has said in the previous turn/segment.
> Note: If there are no words by A, then B is beginning a new task, such as describing a card or the location of an object.

To avoid biasing the results, annotators were not given the turn-taking labels of the tokens.

Inter-annotator reliability is measured by Fleiss' κ at 0.8144, which corresponds to the 'almost perfect' agreement category. The mean pairwise agreement between the three subjects is 90.8%. For the cases in which there is disagreement between the three annotators, we adopt the MAJORITY LABEL as our gold standard; that is, the label chosen by two annotators.

6.1.5.2 Automatic classification

Next, we train a machine learning model using the 400 manually annotated tokens as training data, to automatically classify all IPUs in the corpus as either complete or incomplete. For each IPU we extract a number of lexical and syntactic features from the current turn up to the IPU itself:

- lexical identity of the IPU-final word (w);
- POS tag of w;
- simplified POS tag of w (Noun, Verb, Adjective, Adverb, Contraction, Other);

- POS tags of the IPU-final bigram;
- simplified POS tags of the IPU-final bigram;
- number of words in the IPU;
- a binary flag indicating if *w* is a word fragment;
- size and type of the biggest (*bp*) and smallest (*sp*) phrase that end in *w*;
- binary flags indicating if each of *bp* and *sp* is a major phrase (NP, VP, PP, ADJP, ADVP);
- binary flags indicating if *w* is the head of each of *bp* and *sp*.

We choose these features in order to capture as much lexical and syntactic information as possible from the transcripts. The motivation for lexical identity and part-of-speech features is that complete utterances are unlikely to end in expressions such as *the* or *but there*, and more likely to finish in nouns, for example. Since fragments indicate almost by definition that the utterance is incomplete, we also include a flag indicating if the final word is a fragment. As for the syntactic features, our intuition is that the boundaries of textually complete utterances tend to occur between large syntactic phrases — a similar approach is used by Koehn et al. (2000) for predicting intonational phrase boundaries in raw text. The syntactic features are computed using two different parsers: Collins (Collins, 2003), a high-performance statistical parser; and CASS (Abney, 1996), a partial parser especially designed for use with noisy text.

We experiment with several learners, including the propositional rule learner Ripper (Cohen, 1995), the decision tree learner C4.5 (Quinlan, 1993), Bayesian networks (Heckerman et al., 1995; Jensen, 1996) and support vector machines (SVM) (Vapnik, 1995; Cortes and Vapnik, 1995). We use the implementation of these algorithms provided in the Weka machine learning toolkit (Witten and Frank, 2000). Table 6.5 shows the accuracy of the majority-class baseline and of each classifier, using 10-fold cross validation on the 400 training data points, and the mean pairwise agreement by the three human labelers. The linear-kernel SVM classifier achieves the highest accuracy, significantly outperforming the

Classifier	Accuracy
Majority-class (‘complete’)	55.2%
C4.5	55.2%
Ripper	68.2%
Bayesian networks	75.7%
SVM, RBF kernel	78.2%
SVM, linear kernel	80.0%
Human labelers (mean agreement)	90.8%

Table 6.5: Mean accuracy of each classifier for the textual completion labeling task, using 10-fold cross validation on the training data.

majority-class baseline, and approaching the mean agreement of human labelers. However, there is still margin for further improvement. New approaches could include features capturing information from the previous turn by the other speaker, which was available to the human labelers but not to the ML classifiers. Also, the sequential nature of this classification task might be better exploited by more advanced graphical learning algorithms, such as Hidden Markov Models (HMM; Rabiner, 1989) and Conditional Random Fields (CRF; Lafferty et al., 2001).

6.1.5.3 Results

First we examine the 400 tokens that were manually labeled by three human annotators, considering the majority label as the gold standard. Of the 100 tokens followed by a smooth switch, 91 were labeled textually complete, an overwhelming proportion compared to those followed by a hold (42%). A chi-square test reports that this distribution departs significantly from random ($\chi^2 = 51.7, d.f. = 1, p \approx 0$), suggesting that textual completion as defined earlier in this section constitutes a necessary, but not sufficient, turn-yielding cue.

The analysis of tokens automatically annotated for textual completion provides additional support for this hypothesis. We used the highest performing classifier, the linear-kernel SVM, to label all IPUs in the corpus. Of the 3246 IPUs preceding a smooth switch, 2649 (81.6%) were labeled textually complete; while just about half of all IPUs preceding a hold (4272/8123, or 52.6%) were labeled complete. These numbers depart significantly from a random distribution ($\chi^2 = 818.7, d.f. = 1, p \approx 0$), confirming the predominance of textual completion before smooth switches.

Speaker variation: To investigate speaker variation for the textual completion cue, we compute the proportion of complete IPUs preceding smooth switches (**S**) and holds (**H**) for each speaker. In all cases, the proportion before **S** ranges from 71.4% to 88.5%, and before **H**, from 46.5% to 60.9%, indicating that our general findings are valid across speakers. Detailed results for each speaker are provided in Appendix E.1.

Summary of findings: We provide a definition of textual completion, as well as a procedure for manual annotation that achieves a high inter-labeler agreement rate. Subsequently, we show how a relatively small manually labeled data set may be utilized to train a ML classifier that approaches human performance. When examining both manually and automatically labeled data, we find that textual completion seems to work almost as a necessary condition before smooth switches, but not before holds. A possible interpretation is that textual completion functions as a turn-yielding cue, with listeners more likely to take the speaking turn after completion points.

6.1.6 Voice quality

Voice quality has received some attention in the literature in connection to turn-taking. For instance, Ogden (2002; 2004) annotates voice quality impressionistically, and finds creaky voice to be a turn-yielding cue in Finnish, independent of syntactic, lexical and intonational cues. In this section we examine the relation between turn-taking phenomena and three objective measures of voice quality: jitter, shimmer and noise-to-harmonics ratio (NHR). 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. Measurements of these features have been shown to correlate with perceptual evaluations of voice quality (Eskenazi et al., 1990; Kitch et al., 1996; Bhuta et al., 2004; inter alia).

Using the Praat toolkit, we compute the three features for each IPU over the entire segment and over the final 500 and 1000 ms, and subsequently speaker-normalize them using z-scores. We compute jitter and shimmer over just the voiced portions of the signal for improved robustness. Figure 6.4 summarizes the comparison of these features for IPUs immediately preceding smooth switches (**S**) and holds (**H**). For all three features, the mean value for IPUs preceding **S** is significantly higher than for IPUs preceding **H** ($p < 0.01$), with the difference increasing towards the end of the IPU. In other words, the likelihood of a turn-taking attempt from the interlocutor increases with higher values of jitter, shimmer and NHR towards the end of an IPU, suggesting that voice quality plays a role as a turn-yielding cue.

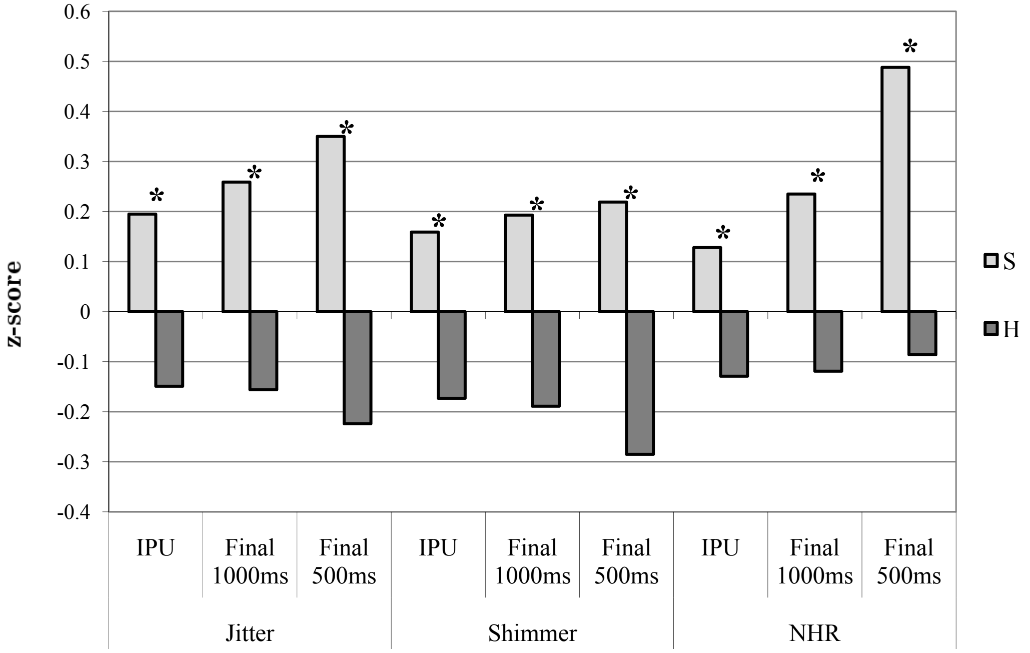


Figure 6.4: Speaker-normalized jitter, shimmer and noise-to-harmonics ratio, over the whole IPU and over its final 500 and 1000 ms.

Speaker variation: When comparing the mean jitter, shimmer and NHR over the final 500 milliseconds of IPUs preceding **S** and **H** for each individual speaker, we find that 12 of the 13 speakers show the same significant differences for jitter, all 13 speakers for shimmer, and all 13 speakers for NHR. For jitter, the remaining speaker (id 113) shows the same relation between the group means, but does not reach significance. This supports that our findings for voice quality are also true across speakers. Detailed results for each individual speaker are shown in Appendix E.1.

Summary of findings: The examination of three acoustic features associated with the perception of voice quality — jitter, shimmer and NHR — reveals that all three of them show significantly higher values before turn boundaries than before turn-internal pauses. Therefore, voice quality seems to function as a turn-yielding cue, potentially aiding listeners in detecting and/or anticipating turn endings. To the best of our knowledge, this is the first work to propose voice quality cues in SAE and to test them empirically. Future work should explore additional features, such as relative average perturbation (RAP), soft phonation index (SPI), and amplitude perturbation quotient (APQ), all of which have been shown to capture different aspects of voice quality.

6.1.7 IPU duration

A final feature that we investigate as a turn-yielding cue is the duration of the IPU, measured in seconds or in number of words. Cutler and Pearson (1986) find mild evidence of longer utterances being judged as turn-final by listeners. Our results are summarized in Figure 6.5. The number of words in IPUs preceding smooth switches (**S**) is significantly smaller than in IPUs preceding holds (**H**) (ANOVA, $p < 0.01$). For duration in seconds, such difference in means between **S** and **H** is also significant at $p = 0.016$.

Speaker variation: All 13 speakers show a significantly larger number of words in IPUs preceding smooth switches than in those preceding holds. Likewise, for nine speakers such difference is also significant when considering the IPU duration in seconds; for the other four speakers, the differences are not significant. Appendix E.1 provides detailed results for each individual speaker.

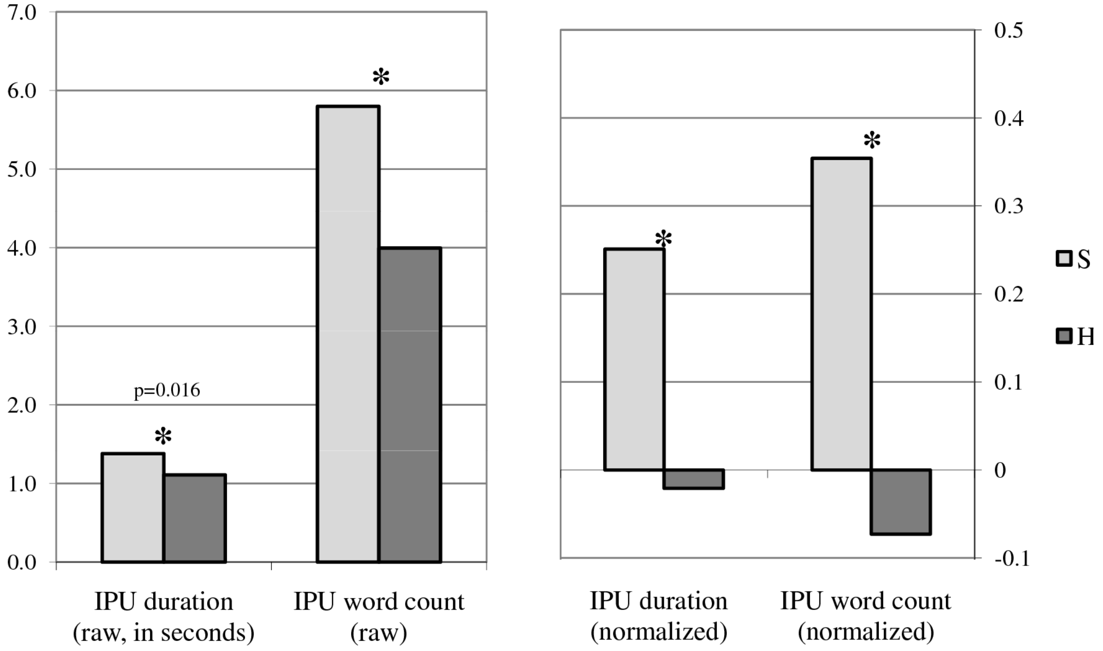


Figure 6.5: IPU duration in seconds and in number of words, both raw and speaker-normalized.

Summary of findings: Turn-medial IPUs tend to be shorter than turn-final ones, suggesting that IPU duration could function as a turn-yielding cue, and supporting similar findings by Cutler and Pearson (1986). We obtain similar results when measuring duration in seconds or in number of words.

6.1.8 Speaker variation

Table 6.6 summarizes the evidence found of the existence of the seven turn-yielding cues described above, for each of the thirteen speakers in the Games Corpus. Six speakers show evidence of all seven cues, while the remaining seven speakers show at least six cues. Pitch level is the least reliable cue, present only for seven subjects. Notably, the cues related to intonation, speaking rate, textual completion, voice quality, and IPU duration are present for all thirteen speakers.

Speaker	101	102	103	104	105	106	107	108	109	110	111	112	113
Intonation	√	√	√	√	√	√	√	√	√	√	√	√	√
Speaking rate	√	√	√	√	√	√	√	√	√	√	√	√	√
Intensity level		√	√	√	√	√	√	√	√	√	√	√	√
Pitch level	√	√			√		√		√	√		√	
Textual completion	√	√	√	√	√	√	√	√	√	√	√	√	√
Voice quality	√	√	√	√	√	√	√	√	√	√	√	√	√
IPU duration	√	√	√	√	√	√	√	√	√	√	√	√	√

Table 6.6: Presence of turn-yielding cues for each speaker.

6.2 Combining turn-yielding cues

So far, we have shown strong evidence supporting the existence of individual acoustic, prosodic and textual turn-yielding cues. Now we shift our attention to the manner in which they combine together to form more complex turn-yielding signals. We consider two approaches: in the DISCRETE APPROACH, individual cues may be either present or absent; in the CONTINUOUS APPROACH, individual cues range from 0 (absent) to 1 (present). The discrete approach is similar to the one proposed by Duncan (1972); the continuous approach represents a natural generalization. Below we describe both approaches and the results obtained with each.

6.2.1 Discrete approach

For each individual cue type, we choose two or three features shown to correlate strongly with smooth switches, as seen earlier in this chapter. These features are summarized in Table 6.7. For example, the individual turn-yielding cue related to IPU-final intonation is represented by two objective measures of F_0 slope, computed over the final 200 and 300 milliseconds of the IPU.

Next, we estimate the presence or absence on a given IPU of each of the individual cues in the left column of Table 6.7 using the procedure depicted in Figure 6.6. This procedure first defines the default case (or null hypothesis), that the cue is absent. The cue is present

Individual cues	**Acoustic features**
Intonation	Absolute value of the F_0 slope over the IPU-final 200 ms
	Absolute value of the F_0 slope over the IPU-final 300 ms
Speaking rate	Syllables per second over the whole IPU
	Phonemes per second over the whole IPU
Intensity level	Mean intensity over the IPU-final 500 ms
	Mean intensity over the IPU-final 1000 ms
Pitch level	Mean pitch over the IPU-final 500 ms
	Mean pitch over the IPU-final 1000 ms
IPU duration	IPU duration in ms
	Number of words in the IPU
Voice quality	Jitter over the IPU-final 500 ms
	Shimmer over the IPU-final 500 ms
	Noise to harmonics ratio over the IPU-final 500 ms

Table 6.7: Features used to estimate the presence of individual turn-yielding cues. All features were speaker normalized using z-scores.

```
$present \leftarrow false$
for each feature $f$ modeling $c$:
    $f_S \leftarrow$ mean $f$ across all IPUs preceding a smooth switch (S)
    $f_H \leftarrow$ mean $f$ across all IPUs preceding a hold (H)
    $f_u \leftarrow u$'s value for $f$
    if $|f_u - f_S| < |f_u - f_H|$ then $present \leftarrow true$
end for
return $present$
```

Figure 6.6: Procedure to estimate the presence or absence of cue c on IPU u (discrete approach).

if, for any of its corresponding features, the value for the given IPU is closer to the mean value of all IPUs preceding a smooth switch (**S**) than that of all IPUs preceding a hold (**H**). In other words, if any feature related to a particular cue shows a value close to that of a turn boundary, then the null hypothesis is discarded and the cue is considered to be present.

Additionally, we automatically annotate all IPUs in the corpus for textual completion using the linear-kernel SVM classifier described in Section 6.1.5. IPUs classified as complete are considered to bear the textual completion turn-yielding cue. Since this feature is essentially binary, no further processing is necessary.

We first analyze the frequency of occurrence of conjoined individual turn-yielding cues. Table 6.8 shows the top ten frequencies for IPUs immediately before smooth switches (**S**), holds (**H**), pause-interruptions (**PI**) and backchannels (**BC**). For IPUs preceding a smooth switch (**S**), the most frequent cases correspond to all, or almost all, cues present at once. For IPUs preceding a hold (**H**), the opposite is true: those with no cues, or with just one or two, represent the most frequent cases. Two different things seem to happen before pause interruptions (**PI**): some of the IPUs present four or even five conjoined cues; others present practically none, as before **H**. This is consistent with two plausible explanations for a **PI** to occur in the first place: (1) that the speaker displays — possibly involuntarily — one or more turn-yielding cues, thus leading the listener to believe that a turn boundary has been reached; or (2) that the listener chooses to break in, regardless of any turn-yielding cues. Finally, the distribution of cues before **BC** does not show a clear pattern, suggesting that backchannel-inviting cues differ from turn-yielding cues. Backchannel-inviting cues are discussed in detail in Chapter 7.

Table 6.9 shows the same results, now grouping together all IPUs with the same **number** of cues, independently of the cue types. Again, we observe that larger proportions of IPUs preceding **S** present more conjoined cues than IPUs preceding **H**, **PI** and **BC**.

Next we look at how the likelihood of turn-taking attempts varies with respect to the number of individual cues displayed by the speaker, a relation hypothesized to be linear by Duncan (1972). Figure 6.7 shows the proportion of IPUs with 0-7 cues present that are followed by a turn-taking attempt from the interlocutor — namely, the number of **S**

S		**H**		**PI**		**BC**	
Cues	Count	Cues	Count	Cues	Count	Cues	Count
1234567	267	...4...	392	.23456.	17	.2..5.7	53
.234567	226	7	247	...4...	13	.2....7	29
1234.67	138	....	223	...45..	12	12..5.7	23
.234.67	109	...4..7	218	....	9	.2.45.7	23
.23..67	98	...45..	178	123..6.	7	12..567	21
..34567	94	.2....7	166	.234.6.	7	.2..5..	21
123..67	93	1234.67	163	.2.4.6.	7	12.4567	18
.2.4567	73	.2..5.7	157	..3456.	7	.2.4567	17
.2.45.7	73	123..67	133	..34.6.	7	1234567	16
12.4.67	70	1234567	130	...4..7	7	12....7	16
	...		...		...		...
Total	3246	Total	8123	Total	274	Total	553

Table 6.8: Top 10 frequencies of complex turn-yielding cues for IPUs preceding **S**, **H**, **PI** and **BC**. For each of the seven cues, a digit indicates presence, and a dot, absence. 1: Intonation; 2: Speaking rate; 3: Intensity level; 4: Pitch level; 5: IPU duration; 6: Voice quality; 7: Textual completion.

and **PI** divided by the number of **S**, **PI**, **H** and **BC**, for each cue count.[3] The dashed line corresponds to a linear model fitted to the data (Pearson's correlation test: $r^2 = 0.969$), and the continuous line, to a quadratic model ($r^2 = 0.995$). The high correlation coefficient of the linear model supports Duncan's hypothesis, that the likelihood of a turn-taking attempt by the interlocutor increases linearly with the number of individual cues displayed by the speaker. However, an ANOVA test reveals that the quadratic model fits the data significantly better than the linear model ($F(1, 5) = 23.014$; $p = 0.005$), even though the curvature of the quadratic model is only moderate, as can be observed in the figure.

[3] In this analysis we only consider non-overlapping exchanges, thus leaving out **O**, **I**, **BI** and **BC_O**; overlapping exchanges are addressed in Chapter 8. Also, note that backchannels are not considered turn-taking attempts.

# Cues	**S**		**H**		**PI**		**BC**	
0	4	(0.1%)	223	(2.7%)	9	(3.3%)	1	(0.2%)
1	52	(1.6%)	970	(11.9%)	33	(12.0%)	15	(2.7%)
2	241	(7.4%)	1552	(19.1%)	59	(21.5%)	82	(14.8%)
3	518	(16.0%)	1829	(22.5%)	59	(21.5%)	140	(25.3%)
4	740	(22.8%)	1666	(20.5%)	53	(19.3%)	137	(24.8%)
5	830	(25.6%)	1142	(14.1%)	46	(16.8%)	113	(20.4%)
6	594	(18.3%)	611	(7.5%)	12	(4.4%)	49	(8.9%)
7	267	(8.2%)	130	(1.6%)	3	(1.1%)	16	(2.9%)
Total	3246	(100%)	8123	(100%)	274	(100.0%)	553	(100.0%)

Table 6.9: Distribution of number of turn-yielding cues displayed in IPUs preceding **S**, **H**, **PI** and **BC**.

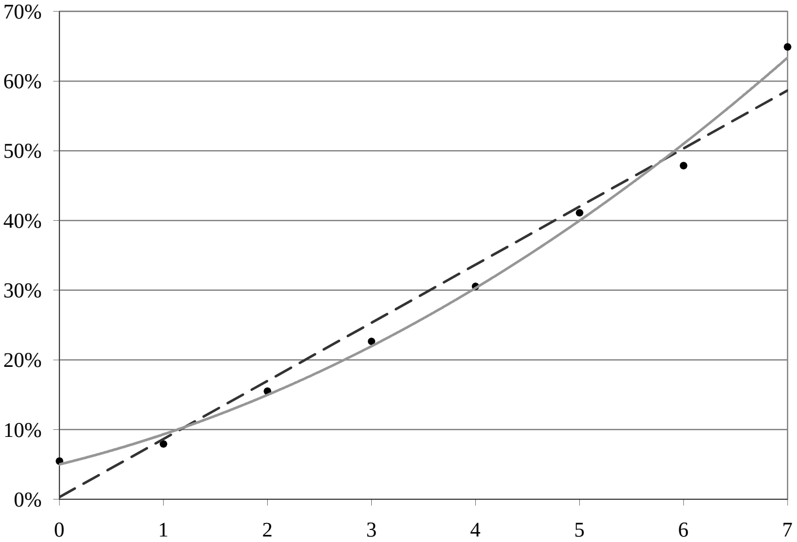


Figure 6.7: Percentage of turn-taking attempts (either **S** or **PI**) following IPUs with 0-7 turn-yielding cues.

We repeat the same analysis for each speaker separately. Figure 6.8 plots, for each of the 13 speakers in the corpus, the probability of a turn-taking attempt per number of displayed cues. Table 6.10 shows the correlation coefficient r^2 of linear and quadratic

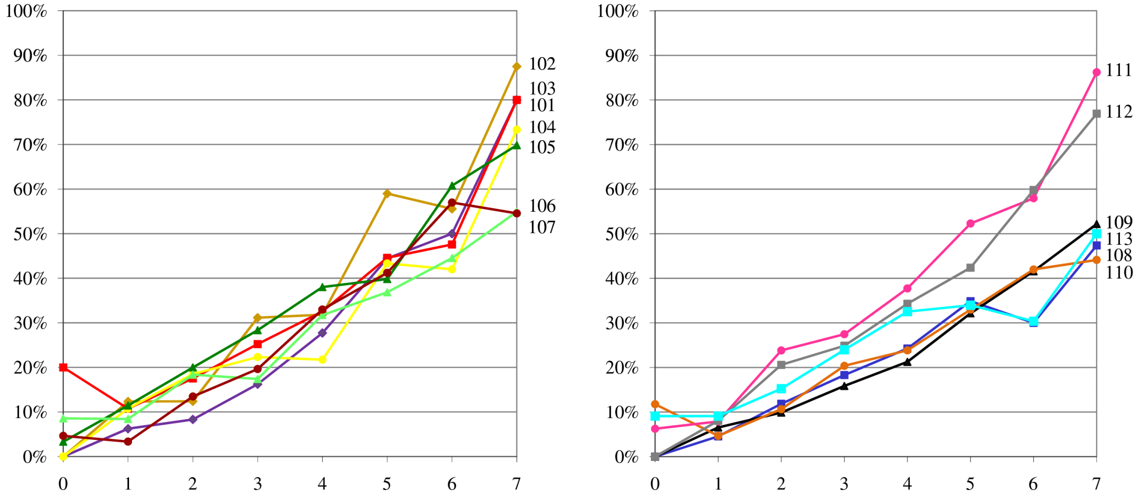


Figure 6.8: Percentage of turn-taking attempts (either **S** or **PI**) following IPUs with 0-7 turn-yielding cues, per speaker.

regressions performed separately on the data from each speaker. In all cases, the coefficients are very high, indicating that the models explain most of the variation present in the data. Additionally, the rightmost column in the table shows the *p*-values of ANOVA tests conducted to compare the goodness of fit of both regressions. The fit of the quadratic model is significantly better than that of the linear model for four speakers (101, 103, 109 and 112), and such difference approaches significance for two other speakers (106 and 111). For the remaining seven speakers, the linear and quadratic models provide statistically indistinguishable explanations of the data.

The slight curvature of the quadratic model, together with the failure of the quadratic models to improve over the linear models for all speakers, indicates that both linear and quadratic models represent good options for explaining the variation in the data. We may conclude then that, in the Games Corpus, we observe that the likelihood of a turn-taking attempt by the interlocutor increases in a nearly linear fashion with respect to the number of cues displayed by the speaker.

Speaker	LM r^2	QM r^2	LM vs. QM p-value
101	0.919	0.983	**0.007**
102	0.929	0.952	0.186
103	0.817	0.954	**0.012**
104	0.884	0.925	0.159
105	0.975	0.983	0.173
106	0.957	0.978	0.076
107	0.955	0.959	0.502
108	0.953	0.953	0.811
109	0.970	0.997	**0.002**
110	0.913	0.942	0.175
111	0.948	0.977	0.053
112	0.970	0.989	**0.035**
113	0.895	0.898	0.753
All	0.969	0.995	**0.005**

Table 6.10: Per-speaker linear and quadratic models showing the relation between number of displayed cues and likelihood of a turn-taking attempt.

6.2.2 Continuous approach

In the previous section we described the results of a discrete approach for combining individual turn-yielding cues, which assumes that each cue may be either present or absent. Now we introduce a generalization of that concept, allowing the presence of a cue to range from 0 (absent) to 1 (present), in what we call the continuous approach.

As in the discrete case, we choose for each individual cue two or three features shown to correlate strongly with smooth switches, as summarized in Table 6.7 (page 51). For each feature f in the right column of the table, its PRESENCE (p) on a given IPU (u) is a real number ranging from 0 to 1, and is defined as follows:

```
$p \leftarrow \frac{f_u - f_H}{f_S - f_H}$
**if** $p < 0$ **then** $p \leftarrow 0$
**if** $p > 1$ **then** $p \leftarrow 1$
```

where f_S is the mean value of f across all IPUs preceding a smooth switch; f_H is the mean value of f across all IPUs preceding a hold; and f_u is the mean value of f on the target IPU. Figure 6.9 illustrates how p varies as a function of f_u; note that p approaches 1 as f_u gets closer to f_S, and it approaches 0 as f_u gets closer to f_H. Finally, the PRESENCE OF A TURN-YIELDING CUE is defined simply as the maximum presence of the features modeling the cue. For example, if the presence of the two features modeling the speaking rate cue — syllables per second and phonemes per second — are 0.8 and 0.6, then the presence of such cue is 0.8.

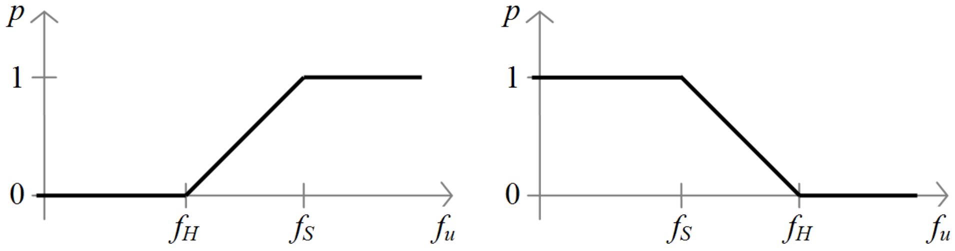


Figure 6.9: Presence (p) of a given feature, as a function of the feature's value over a given IPU (f_u).

The textual completion cue is a special case, as it is essentially binary. Therefore, we leave it as is, without transforming it into a continuous cue. Again, we use the automatic annotations of textual completion performed with the SVM-based classifier (as described in Section 6.1.5), and assign 1 to IPUs classified as 'complete', and 0 to those classified as 'incomplete'.

In the previous section, we studied how the likelihood of turn-taking attempts varies with respect to the number of individual cues displayed by the speaker. Under the continuous approach we cannot talk about the number of cues; instead, we use the **sum** of continuous cues. The resulting sum for a given IPU is a real number ranging from 0 to 7.

The results of all tests using continuous cues are nearly identical to those using discrete cues, both for all speakers together and for each speaker individually. For example, Figure 6.10 shows the proportion of IPUs with different sums of continuous cues that are followed by a turn taking attempt from the interlocutor.[4] The dashed line corresponds to a linear model fitted to the data; the continuous line, to a quadratic model. Again, both models are highly correlated with the data (Pearson's correlation tests; linear model: $r^2 = 0.963$; quadratic model: $r^2 = 0.984$), and the quadratic model has a significantly better fit when considering all speakers together ($p = 0.0016$), but not for each speaker independently (only for 6 of the 13 speakers). For simplicity, we omit all other results for continuous cues, as they would add nothing novel to our analysis.

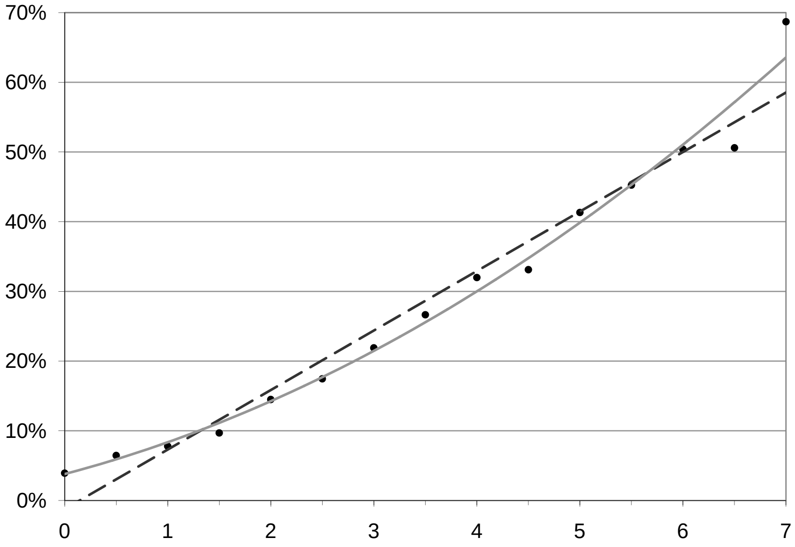


Figure 6.10: Percentage of turn-taking attempts following IPUs with a given sum of continuous turn-yielding cues.

6.3 Discussion

In this chapter we have presented evidence of the existence of seven turn-yielding cues. In other words, we have described seven measurable events that take place with a significantly higher frequency on IPUs preceding smooth switches (when the current speaker completes an utterance and the interlocutor takes the turn after a short pause) than on IPUs preceding holds (when the current speaker continues speaking after a short pause). These events may

[4] To replicate the analyses of the previous section, we binned the sums in intervals of width 0.5.

be summarized as follows:

- a falling or high-rising intonation at the end of the IPU;
- a reduced lengthening of IPU-final words;
- a lower intensity level;
- a lower pitch level;
- a point of textual completion;
- a higher value of three voice quality features: jitter, shimmer, and NHR; and
- a longer IPU duration.

Additionally, we have shown that, when several turn-yielding cues occur simultaneously, the likelihood of a subsequent turn-taking attempt by the interlocutor increases in almost a linear fashion. In the Games Corpus, the percentage of IPUs followed by a turn-taking attempt ranges from 5% when no turn-yielding cues are present, to 65% when all seven cues are present.

These findings could be used to improve the turn-taking decisions of state-of-the-art IVR systems. In particular, our model of turn-taking provides answers to three of the questions posed in Chapter 3:

Q1. The system wants to keep the floor; how should it formulate its output to avoid an interruption from the user?

According to our model, including as few as possible of the described turn-yielding cues in the system's output will decrease the likelihood that the user will take the turn. Therefore, when the system intends to continue holding the floor, it should end its IPUs in plateau intonation, with high intensity and pitch levels, leaving utterances textually incomplete (e.g., preceding pauses with expressions such as *and* or *also*), and so on.

Q3. The system wants to yield the floor to the user; how should it formulate its output to invite the user to take the turn?

This situation corresponds to the opposite of the previous question. If the system includes in its output as many of the described turn-yielding cues as possible, a turn-taking attempt by the user will be more likely to take place. Thus, if the system intends to cede the floor to the user, it should end the final IPU in either falling or high-rising intonation (e.g., depending on whether the system's message is a statement or a direct question), with low intensity and pitch levels, and so on.

Q5. The user is speaking; how can the system know when it is an appropriate moment to take the turn?

Most current systems simply wait for a long-enough pause from the user before attempting to take the turn, a technique that might be possible to improve using the findings in our study. Although the difficulty of estimating each turn-yielding cue will vary according to many implementation details, we may draft a high-level description of the turn-taking decision procedure. At every pause longer than 50 milliseconds, the system estimates the presence of as many cues as possible over the user's final IPU. Depending on the number of detected cues, the system may then make an informed turn-taking decision: If the number of detected cues is high, it may choose to conduct a turn-taking attempt immediately; otherwise, it may continue waiting, thus defaulting to its original behavior.

The addition to current IVR systems of the capabilities described in the answers to Q1, Q3 and Q5 could effectively improve their naturalness and usability, by offering users a turn-taking experience that resembles more closely the normal interaction in human-human conversation.

An implicit assumption of our study is that all turn-yielding cues are equally important, and contribute with either 0 or 1 to the total count. While this is a convenient assumption to simplify a first approach to the problem, it is also not necessarily true. For example, we have mentioned that the textual completion cue seems to work almost as a necessary condition for smooth switches, which does not appear to be the case for other cues. A possible topic for future research, then, is to explore the assignment of numeric weights to the different cues, in order to account for their relative importance.

Another future research topic is to further investigate turn-yielding cues related to voice quality. Additional features should be incorporated into the analysis, such as relative average perturbation (RAP), soft phonation index (SPI), and amplitude perturbation quotient (APQ), all of which have been shown to capture different aspects of voice quality. Furthermore, we have chosen to collapse jitter, shimmer and NHR into one simple voice quality cue, but these features could instead be used as finer grained turn-yielding cues, perhaps in combination with the numeric weights mentioned in the previous paragraph.

Finally, we do not find evidence in the Games Corpus of lexical cues related to stereotyped expressions such as *you know* or *I think*. Larger corpora should be examined for the existence of such cues. However, we do find frequent use of expressions such as *lower right* or *on top*, which appear to function as task-specific turn-taking cues. Future research should investigate this issue in more detail, as speech processing applications could benefit from it. Additionally, we find that affirmative cue words, such as *okay* or *alright*, seem to play a central role in the organization of turn-taking in conversations. These words are heavily overloaded, used to convey acknowledgment, to backchannel, and to begin or end discourse segments, among other functions. We devote Part III of this thesis to the study of affirmative cue words in the Games Corpus.

Chapter 7

Backchannel-Inviting Cues

We continue our study of turn-taking phenomena by focusing on a second set of cues produced by the speaker that may induce a particular behavior from the listener, which we term BACKCHANNEL-INVITING CUES. Backchannels are short expressions, such as *uh-huh* or *mm-hm*, uttered by the listener to convey that they are paying attention, and to encourage the speaker to continue. Normally, they are neither disruptive nor acknowledged by the speaker holding the conversational floor. Hypothetically, speakers produce a set of cues marking specific moments within speaking turns at which listeners are welcome to produce backchannel responses.

Finding out whether such cues exist and being able to model them could help answer two of the empirical questions discussed in the introduction of Part II:

Q2. The system wants to keep the floor, ensuring that the user is paying attention; how should it formulate its output to give the user an opportunity to utter a backchannel?

Q6. The user is speaking; how can the system know whether and when it should produce a backchannel as positive feedback to the user?

In this chapter we investigate the existence of lexical, acoustic and prosodic backchannel-inviting cues. Using the turn-taking categories available in our corpus, we compare IPUs preceding a backchannel (**BC**) to IPUs preceding a hold (**H**), making the strong assumption that such cues, if any exist, are more likely to occur in the former group. Additionally, we contrast IPUs before **BC** with those before a smooth switch (**S**), to study how backchannel-inviting cues differ from turn-yielding cues. The way backchannels are realized by speakers is studied in further detail in Part III of this thesis.

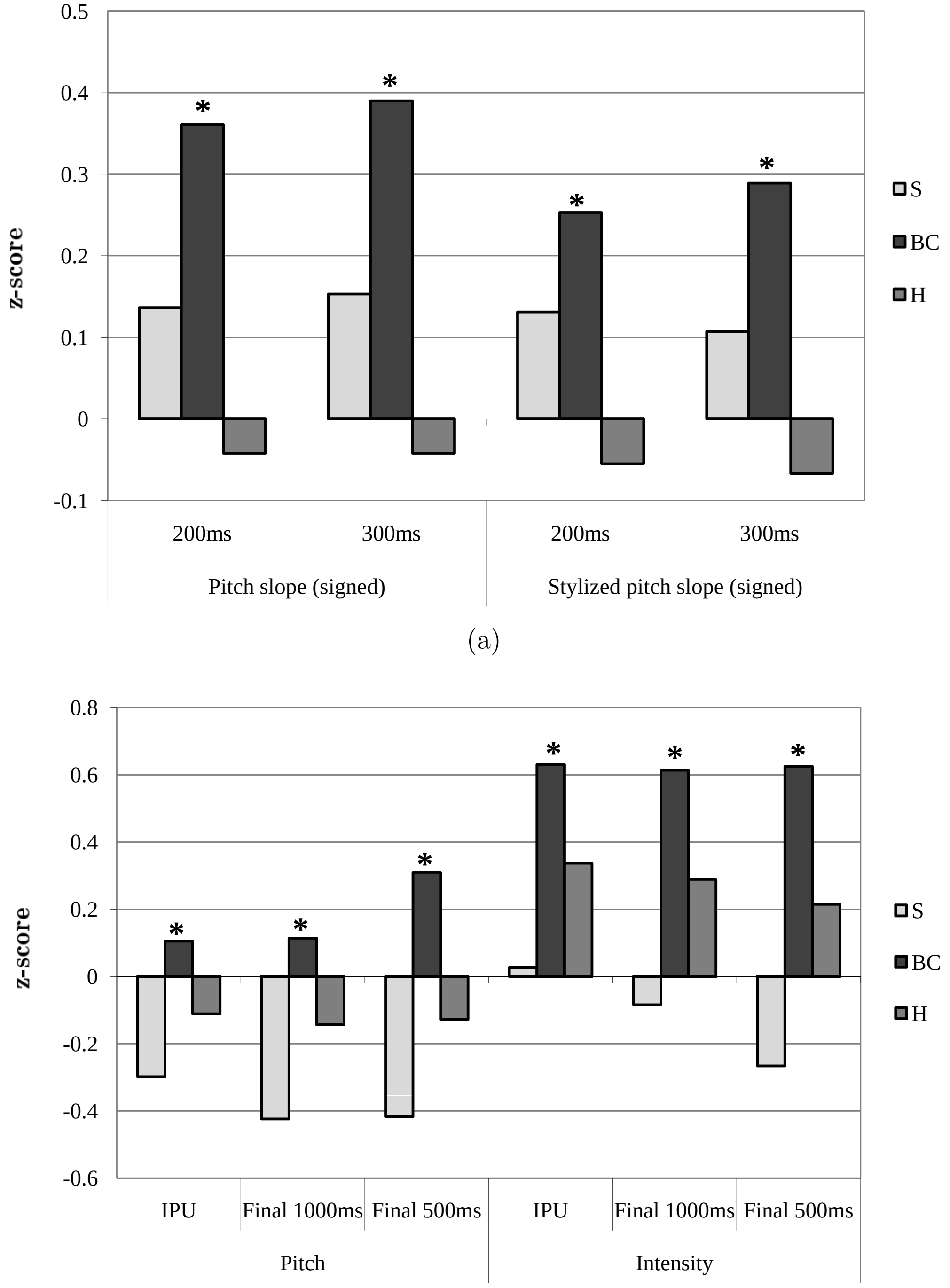
7.1 Individual cues

We repeat the procedures described in Chapter 6, now looking for individual backchannel-inviting cues instead of turn-yielding cues. We find significant differences between IPUs preceding **BC** and **H** for final intonation, pitch and intensity levels, IPU duration, and voice quality. These results are summarized in Figures 7.1 and 7.2.

IPUs immediately preceding backchannels show a clear tendency towards a final rising intonation, as hypothesized by a preliminary study on the Games Corpus by Benus et al. (2007). All pitch slope measures (raw and stylized, over the IPU-final 200 and 300 milliseconds) are significantly higher before **BC** than before **S** or **H**. As seen in Table 7.1, categorical ToBI labels support this finding. More than half of the IPUs preceding a backchannel end in a high-rise contour (H-H%), and about a quarter with a low-rise contour (L-H%). Together, these two contours account for more than 81% of all IPUs before **BC**, but only 36.2% and 20.6% of those before **S** and **H**, respectively. Thus, final intonation presents very different patterns in IPUs preceding these three turn-taking categories: either high-rising or low-rising before backchannels, either falling or high-rising before smooth

	BC		**S**		**H**	
H-H%	**257**	**55.7%**	484	(22.1%)	513	(9.1%)
[!]H-L%	27	5.9%	289	(13.2%)	1680	(29.9%)
L-H%	**119**	**25.8%**	309	(14.1%)	646	(11.5%)
L-L%	52	11.3%	1032	(47.2%)	1387	(24.7%)
No boundary tone	4	0.9%	16	(0.7%)	1261	(22.4%)
Other	2	0.4%	56	(2.6%)	136	(2.4%)
Total	461	100.0%	2186	(100.0%)	5623	(100.0%)

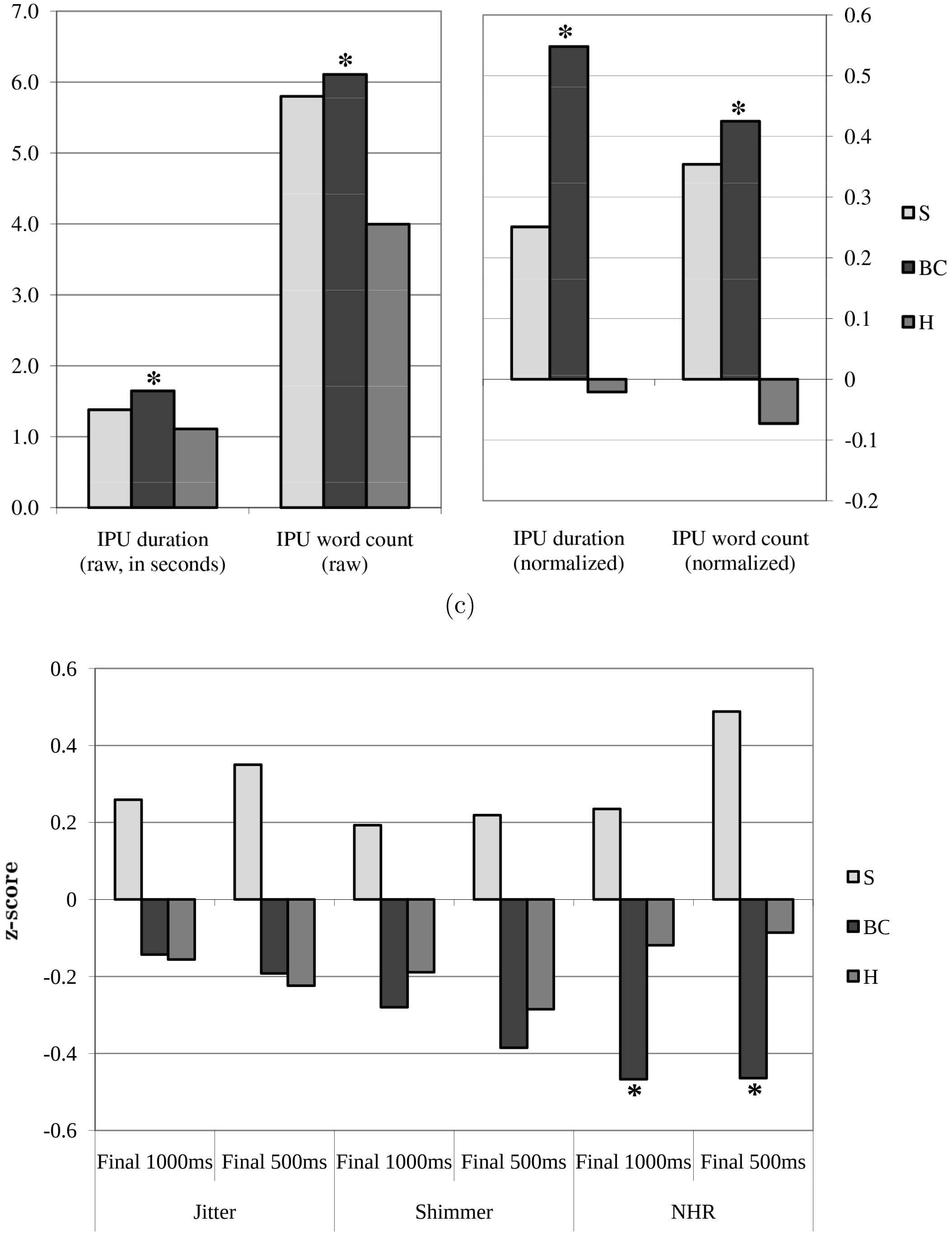
Table 7.1: ToBI phrase accent and boundary tone for IPUs preceding **BC**, **S** and **H**.



(a)

(b)

Figure 7.1: Individual backchannel-inviting cues: (a) pitch slope and stylized pitch slope; (b) pitch and intensity. *Continued in Figure 7.2.*



(c)

(d)

Figure 7.2: Individual backchannel-inviting cues: (c) IPU duration; (d) voice quality.

Continued from Figure 7.1.

switches, and plateau before holds.

Mean pitch and intensity levels tend to be significantly higher for IPUs before **BC** than before the other two categories. This suggests that backchannel-inviting cues related to these two features function in a manner opposite to turn-yielding cues.

We also find that IPUs followed by backchannels tend to be significantly longer than IPUs followed by either smooth switches or holds, both when measured in seconds and in number of words. Thus, IPU duration works not only as a potential turn-yielding cue (as we say in the previous chapter) but also as backchannel-inviting cues.

Finally, we find differences for just one of the three voice quality features under consideration. Noise-to-harmonics ratio (NHR) tends to be significantly lower in IPUs preceding **BC** than in those preceding **H**. Again, this backchannel-inviting cue is the opposite of the related turn-yielding cue, which corresponds to a high level of NHR. For the other two voice quality features, jitter and shimmer, the two groups are indistinguishable.

Next we look at lexical backchannel-inviting cues. We examine the distribution of part-of-speech tags in IPU-final phrases, and find that as many as 72.5% of all IPUs preceding backchannels end in either 'DT NN', 'JJ NN', or 'NN NN' (Table 7.2) — that is, 'determiner noun' (e.g., *the lion*), 'adjective noun', (*blue mermaid*), or 'noun noun' (*top point*). In comparison, the same three final POS bigrams account for only 31.1% and 21.3% of IPUs preceding **S** and **H**, respectively. Furthermore, the three most frequent final POS bigrams before **S** and **H** add up to just 43.7% and 29.0%, showing more spread distributions, and suggesting that the part-of-speech variability for IPUs before **BC** is relatively very low. These results strongly suggest the existence of a backchannel-inviting cue related to the part-of-speech tags of the IPU-final words.

Speaker variation: We investigate the existence of the hypothesized backchannel-inviting cues for each individual speaker. Four subjects (ids 101, 104, 107 and 109) have fewer than 20 instances of IPUs preceding **BC**, a count too low for statistical tests, and are thus excluded from the analysis. Table 7.3 summarizes the evidence found of the existence of the six backchannel-inviting cues described above, for each of the nine speakers with high

BC			**S**			**H**		
POS	#	%	POS	#	%	POS	#	%
DT NN	**234**	**42.3%**	DT NN	600	18.5%	DT NN	1093	13.5%
JJ NN	**100**	**60.4%**	UH	578	36.3%	UH	832	23.7%
NN NN	**67**	**72.5%**	JJ NN	242	43.7%	JJ NN	430	29.0%
IN NN	12	74.7%	NN NN	168	48.9%	IN DT	374	33.6%
DT JJ	12	76.9%	DT JJ	111	52.3%	UH UH	243	36.6%
IN PRP	9	78.5%	NN UH	96	55.3%	DT JJ	225	39.4%
NN RB	7	79.7%	IN PRP	90	58.1%	IN NN	214	42.0%
DT NNP	7	81.0%	UH UH	83	60.6%	NN NN	211	44.6%
VBZ VBG	6	82.1%	JJR NN	83	63.2%	DT UH	154	46.5%
NNS NN	5	83.0%	IN DT	67	65.2%	NN IN	112	47.9%
	...			...			...	
Total	553	100%	Total	3246	100%	Total	8123	100%

Table 7.2: Count and cumulative percentage of the 10 most frequent IPU-final POS bigrams preceding **BC**, **S** and **H**.

enough counts.[1] Differences in intonation, duration and voice quality are significant for the great majority of speakers, and a smaller proportion of speakers display differences for pitch and intensity. Also, all nine speakers show a marked predominance of at least two of the three final POS bigrams mentioned above ('DT NN', 'JJ NN' and 'NN NN') before backchannels. Notably, no single acoustic/prosodic cue is used by all speakers; rather, each seem to use their own combination of cues. For example, speaker 102 varies only intonation, while speaker 108 varies only intensity level and IPU duration. We conclude then that, unlike the case of turn-yielding cues, the speaker variation present in the production of backchannel-inviting cues is not insignificant, with different speakers apparently displaying different combinations of cues.

[1] Detailed results for each individual speaker are shown in Appendix E.2.

Speaker	102	103	105	106	108	110	111	112	113
Intonation	√	√	√	√		√		√	√
Pitch level							√	√	√
Intensity level		√		√	√		√	√	√
IPU duration		√	√	√	√	√	√	√	√
Voice quality		√	√	√		√	√	√	√
POS bigram	√	√	√	√	√	√	√	√	√

Table 7.3: Presence of backchannel-inviting cues for each speaker.

7.2 Combining cues

After finding evidence of the existence of individual acoustic, prosodic and textual backchannel-inviting cues, we replicate the procedures described in the previous chapter to investigate how such cues combine together to form complex signals. The results are almost identical when using the approach with discrete individual cues (either present or absent) and its generalization to continuous values. For simplicity, we present only the results of the discrete approach in this section.

For each individual cue, we choose two features shown to strongly correlate with IPUs preceding backchannels, as seen earlier in this chapter. These features are shown in Table 7.4. For example, the individual cue related to IPU-final intonation is represented by two objective measures of the F_0 slope, computed over the final 200 and 300 milliseconds of the IPU.

Next, we estimate the presence or absence in a given IPU of each of the individual cues in the left column of Table 7.4 using the same procedure described in the previous chapter (Figure 6.6, page 51). Additionally, we annotate automatically all IPUs in the corpus according to whether they end in one of the three POS bigrams found to strongly correlate with IPUs preceding a backchannel: 'DT NN', 'JJ NN' and 'NN NN'. IPUs ending in any such POS bigram are considered to bear the 'POS bigram' backchannel-inviting cue. Since this feature is essentially binary, no further processing is necessary.

We first analyze the frequency of occurrence of conjoined individual cues before each

Individual cues	**Acoustic features**
Intonation	F_0 slope over the IPU-final 200 ms F_0 slope over the IPU-final 300 ms
Intensity level	Mean intensity over the IPU-final 500 ms Mean intensity over the IPU-final 1000 ms
Pitch level	Mean pitch over the IPU-final 500 ms Mean pitch over the IPU-final 1000 ms
IPU duration	IPU duration in ms Number of words in the IPU
Voice quality	Noise to harmonics ratio over the IPU-final 500 ms Noise to harmonics ratio over the IPU-final 1000 ms

Table 7.4: Acoustic features used to estimate the presence of individual backchannel-inviting cues. All features were speaker normalized using z-scores.

turn-taking category. Table 7.5 shows the top ten frequencies for IPUs immediately before a backchannel (**BC**), a smooth switch (**S**), and a hold (**H**). For IPUs preceding **BC**, the most frequent cases correspond to all, or almost all, cues present at once. Very different is the picture for IPUs preceding **H**, which show primarily few to no cues. For IPUs preceding **S**, those with no cues, or just one or two, represent the most frequent cases. This suggests that complex signals produced by speakers to yield the turn differ considerably from signals that invite the interlocutor to utter a backchannel response.

Table 7.6 shows the same results, now grouping together all IPUs with the same **number** of cues, independently of the cue types. Again, we observe that larger proportions of IPUs preceding **BC** show more conjoined cues than IPUs preceding **S** and **H**.

Next we look at how the likelihood of the occurrence of backchannels varies with respect to the number of individual cues conjointly displayed by the speaker. Figure 7.3 shows the proportion of IPUs with 0-6 cues present that are followed by a backchannel from the interlocutor — namely, the number of **BC** divided by the number of **S**, **PI**, **H** and **BC**,

BC		**S**		**H**	
Cues	Count	Cues	Count	Cues	Count
123456	83	......	243	.2..5.	865
12.456	49	...4..	195	.23.5.	533
123.56	47	..3...	172	......	513
.23456	27	1.....	153	..3...	414
12345.	24	1..4..	123	....5.	368
123.5.	19	1.3...	113	.2.45.	344
12.45.	16	...4.6	111	.2....	330
12..56	16	1..4.6	108	1.....	256
1.3456	14	...45.	107	...45.	237
.2.456	14	.2....	94	...4..	218
	...		...		...
Total	553	Total	3246	Total	8123

Table 7.5: Top 10 frequencies of complex backchannel-inviting cues for IPUs preceding **BC**, **S** and **H**. For each of the six cues, a digit indicates presence, and a dot, absence. 1: Intonation; 2: Intensity level; 3: Pitch level; 4: IPU duration; 5: Voice quality; 6: Final POS bigram.

for each cue count.[2] The dashed line in the plot corresponds to a linear model fitted to the data ($r^2 = 0.812$); the continuous line, to a quadratic model ($r^2 = 0.993$). The fit of the quadratic model is significantly better than that of the linear model, as reported by an ANOVA test ($F(1,4) = 110.0$; $p < 0.001$). In this case, the fit of the linear model is not as good as in the case of turn-yielding cues. The quadratic model, on the other hand, achieves an almost perfect fit and shows a marked curvature, confirming that a quadratic model provides a good explanation for the relation between number of backchannel-inviting cues and occurrence of a backchannel.

We repeat the same analysis for each speaker separately. Figure 6.8 plots the probability

[2] Again, we only consider non-overlapping exchanges, thus leaving out **O**, **I**, **BI** and **BC_O**.

# Cues	**BC**		**S**		**H**	
0	4	(0.7%)	243	(7.5%)	513	(6.3%)
1	17	(3.1%)	746	(23.0%)	1634	(20.1%)
2	57	(10.3%)	912	(28.1%)	2364	(29.1%)
3	90	(16.3%)	723	(22.3%)	1960	(24.1%)
4	139	(25.1%)	379	(11.7%)	1010	(12.4%)
5	163	(29.5%)	192	(5.9%)	501	(6.2%)
6	83	(15.0%)	51	(1.6%)	141	(1.7%)
Total	553	(100%)	3246	(100%)	8123	(100%)

Table 7.6: Distribution of number of backchannel-inviting cues displayed in IPUs preceding **BC**, **S** and **H**.

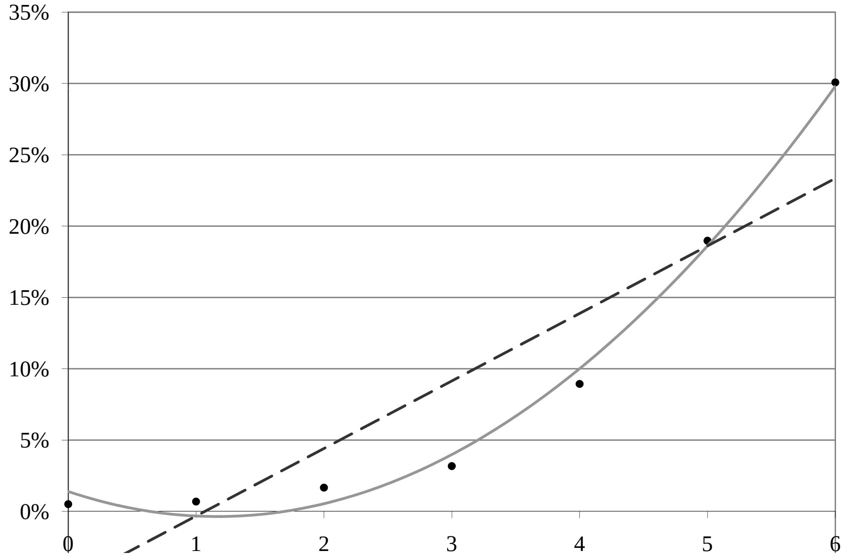


Figure 7.3: Percentage of backchannels following IPUs with 0-6 backchannel-inviting cues.

of occurrence of a backchannel per number of conjoined cues, for each of the 9 speakers with high enough counts to conduct statistical tests. Table 6.10 shows the correlation coefficient (r^2) of the linear and quadratic regressions performed separately on the data from each speaker. The fit of the linear models ranges from moderate at 0.625 to high at 0.884. In seven out of nine cases, the fit of the quadratic models is significantly better, ranging from

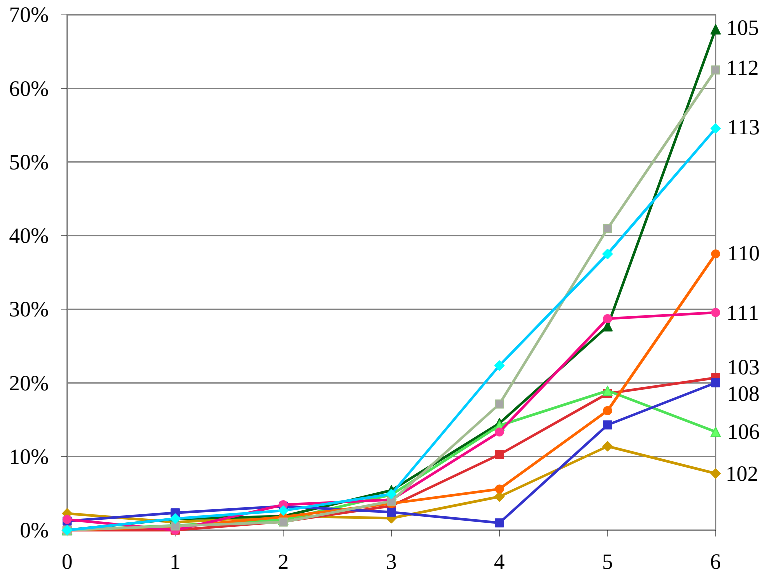


Figure 7.4: Percentage of backchannels following IPUs with 0-6 backchannel-inviting cues, for nine speakers with high enough counts.

Speaker	LM r^2	QM r^2	LM vs. QM p-value
102	0.625	0.702	0.369
103	0.884	0.962	**0.044**
105	0.715	0.954	**0.010**
106	0.799	0.799	0.990
108	0.628	0.869	**0.053**
110	0.703	0.947	**0.013**
111	0.840	0.934	**0.075**
112	0.798	0.990	**0.001**
113	0.850	0.989	**0.002**
All	0.812	0.993	< 0.001

Table 7.7: Per-speaker linear and quadratic regressions on the relation between number of displayed conjoined cues and probability of a backchannel occurrence.

0.702 to 0.990.

The fact that, for most speakers, the quadratic model fits the data better than the linear model, together with the marked curvature of the general quadratic model (as seen in Figure 7.3), suggests that the quadratic model is well suited for explaining the relation between the number of backchannel-inviting cues conjointly displayed by the speaker, and the likelihood of occurrence of a backchannel from the interlocutor.

7.3 Discussion

In this chapter we have presented evidence of the existence of six backchannel-inviting cues. That is, we have described six measurable events that take place with a significantly higher frequency on IPUs preceding backchannels than on IPUs preceding holds or smooth switches. These events may be summarized as follows:

- a rising intonation at the end of the IPU;
- a higher intensity level;
- a higher pitch level;
- a final POS bigram equal to 'DT NN', 'JJ NN' or 'NN NN';
- a lower value of noise-to-harmonics ratio (NHR); and
- a longer IPU duration.

We have also shown that, when several backchannel-inviting cues occur simultaneously, the likelihood of occurrence of a backchannel from the interlocutor increases in a quadratic fashion, ranging from only 0% of IPUs followed by a backchannel when no cues are present, to more than 30% when all six cues are present.

There are two important things worth emphasizing regarding our results. First, we noted in the previous chapter that speaker variation is very low for turn-yielding cues, with almost all speakers producing all cues. In the case of backchannel-inviting cues, however, there is considerably more speaker variation. In fact, each speaker seems to use their own combination of cues. Still, some of the findings are true across all speakers: all tend to display at least two cues, and all share the POS bigram cue. Future research should pursue this issue further, trying to shed some light on when, how and why speakers choose to use a particular set of cues.

The second comment is related to the optionality of backchannels. We have shown that a backchannel is produced by the other speaker after around 30% of IPUs containing all six backchannel-inviting cues. This number looks quite small when compared to the 65% of turn-taking attempts following IPUs with all seven turn-yielding cues. The reason for this disparity may be explained by a higher optionality of backchannels in SAE. It is perfectly conceivable that two speakers may have a successful conversation without producing any backchannels — even if doing so requires not acting upon clear backchannel-inviting cues. On the other hand, it is harder to imagine a conversation in which both speakers systematically ignore turn-yielding cues, taking the turn exclusively at places other than transition-relevance places. In our corpus, this optionality seems to be reflected in the relatively low percentage of backchannels following rich backchannel-inviting signals.

The findings presented in this chapter could be used to further improve the turn-taking decisions of state-of-the-art IVR systems. In particular, our model of backchannels provides answers to two of the questions posed in Chapter 3:

Q2. The system wants to keep the floor, ensuring that the user is paying attention; how should it formulate its output to give the user an opportunity to utter a backchannel?

According to our model, if the system includes in its output as many of the described cues as possible, the likelihood of occurrence of a backchannel from the user will increase. Thus, if the system intends to elicit a backchannel response from the user, it should end the final IPU in one of the listed part-of-speech bigrams, with rising intonation (preferably high-rising), high pitch and intensity levels, and so on.

Q6. The user is speaking; how can the system know whether and when it should produce a backchannel as positive feedback to the user?

The ability to detect points where the user invites the system to backchannel — or, at least, where backchannels would be acceptable — could be coupled with the procedure described in the previous chapter for detecting turn endings based on turn-yielding cues. Every time the system estimates the presence of turn-yielding cues over the user's final IPU, it could also estimate the presence of backchannel-inviting cues. (Note that some features may be reused, as they belong to both cue sets.) If the number of detected backchannel-inviting cues is high enough, then the system may utter a backchannel; otherwise, it may keep silent. Since at least three backchannel-inviting cues are opposite to the corresponding turn-yielding cues (intensity, pitch and NHR) there is little risk of detecting both a turn ending and a point for backchanneling at the same time.

Two of the final considerations made in the previous chapter regarding future research topics apply here as well. The assignment of numeric weighs to the different cues, according to their relative importance, might improve the model's description of the data. Also, additional features shown to capture different aspects of voice quality features should be examined as potential backchannel-inviting cues.

In this chapter we have studied the context in which backchannels are likely too occur. Part III of this thesis deals, among other things, with the acoustic, prosodic and phonetic characteristics of backchannels in the Games Corpus. Those results are intended to aid IVR systems in generating backchannels with the correct parameters, and in correctly interpreting backchannel utterances from the user.

Chapter 8

Overlapping Speech

Often in conversation speakers take the turn just before the end of their interlocutors' contribution, without interrupting the conversational flow (Sacks et al., 1974). There is evidence of the occurrence of these events in multiple languages, including Arabic, English, German, Japanese, Mandarin and Spanish (Yuan et al., 2007), and previous studies also report situational and genre differences. For example, non-face-to-face dialogues have significantly fewer speech overlaps than face-to-face ones (Bosch et al., 2005); people make fewer overlaps when talking with strangers (Yuan et al., 2007); and speakers tend to make fewer overlaps and longer pauses when performing difficult tasks (Bull and Aylett, 1998).

The existence of this phenomenon suggests that listeners are capable of anticipating possible turn endings, and poses the question of how they manage to do this. One possible explanation could be the early detection on the part of the listener of turn-yielding and backchannel-inviting cues, such as the ones discussed in previous chapters. That is, listeners may be able to perceive such signals some amount of time prior to the end of the speaker's turn. Another explanation could be the occurrence of additional cues **earlier** in the speaker's turn. Note, though, that these two hypothesis are not mutually exclusive.

This chapter describes the results of preliminary studies aimed at providing evidence for these two hypothesis. First, we review the types of overlapping speech existing in the Games Corpus. Second, we investigate the existence of the cues discussed in previous chapters on turn-final IPUs preceding transitions with overlapping speech. Third, we study the durational distribution of overlapping speech segments. Finally, we look for evidence of turn-yielding and backchannel-inviting cues occurring earlier in the speaker's turn.

8.1 Types of overlapping speech in the Games Corpus

The turn-taking labeling scheme presented in Chapter 5 includes four categories of turn exchanges with simultaneous speech present: overlap (**O**), backchannel with overlap (**BC_O**), interruption (**I**) and butting-in (**BI**). In this study we consider only the first two classes (**O** and **BC_O**), and ignore the last two, since they correspond to disruptions of the conversational flow at arbitrary points during the speaker's turn, rather than slight, unobtrusive overlapping speech segments. Note that the existence of overlapping speech is the only difference between **O** and smooth switches (**S**), and between **BC_O** and backchannels (**BC**).

Instances of **O** can be divided in two cases: FULL OVERLAPS, which take place completely within the interlocutor's turn (as depicted in the left part of Figure 8.1); and PARTIAL OVERLAPS, which begin during the interlocutor's turn but extend further after its end (right part of Figure 8.1). Fully and partially overlapping backchannels are defined analogously. In this study we consider only instances of partial **O** and **BC_O**, which are clear cases of turn endings overlapped by new turns from the interlocutor. For fully overlapping instances, we have no indication of the location of the speech portion that triggers the overlapping turn, which complicates the search for turn-taking cues. Furthermore, full overlaps correspond to complex events in which the current speaker talks — without pausing — before, during and after a complete utterance from the interlocutor. In such occasions, it seems to be the case that the two speakers briefly *share* the conversational floor, an interesting phenomenon that should be addressed specifically in future research.

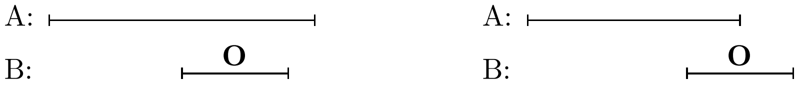


Figure 8.1: Full and partial overlap types.

In the Games Corpus, 767 of the 1067 instances of **O**, as well as 104 of the 202 tokens of **BC_O**, are partially overlapping. We use only these data in the present study. For clarity, we refer to partially overlapping **O** and **BC_O** simply as **O** and **BC_O**.

8.2 Existence of cues before O and BC_O

8.2.1 Turn-yielding cues preceding O

In Chapter 6 we presented a procedure to estimate the existence of seven turn-yielding cues before smooth switches (**S**). We begin our study of overlapping speech by searching for evidence of the same cues in IPUs preceding overlaps (**O**), and obtain the results summarized in Table 8.1. The table on the left lists the top ten frequencies of complex cues (1: Intonation; 2: Speaking rate; 3: Intensity level; 4: Pitch level; 5: IPU duration; 6: Voice quality; 7: Textual completion). Similarly to what we observe for IPUs followed by **S** (see Table 6.8 on page 53), the most frequent cases correspond to all, or almost all, cues present at once.

Cues	Count
1234567	61
.234567	50
.234.67	26
.23456.	24
..34567	24
1234.67	22
..3..67	22
123456.	21
.2.4567	20
..34.67	20
	...

# Cues	**O**	
0	1	(0.1%)
1	15	(2.0%)
2	55	(7.2%)
3	111	(14.5%)
4	163	(21.3%)
5	213	(27.8%)
6	148	(19.3%)
7	61	(8.0%)
Total	767	(100%)

Table 8.1: Left: Top 10 frequencies of complex turn-yielding cues for IPUs preceding **O** (*cf* Table 6.8 on page 53). Right: Distribution of number of turn-yielding cues in IPUs preceding **O** (*cf* Table 6.9 on page 54).

The right part of Table 8.1 shows the same results, now grouping together all IPUs with the same number of cues, independently of the cue types (see Table 6.9 on page 54). Again, we observe a marked tendency of IPUs preceding **O** to present a high number of conjoined turn-yielding cues.

These results indicate that IPUs immediately preceding smooth switches (**S**) and overlaps (**O**) show a similar behavior in terms of the occurrence of our posited turn-yielding cues. This finding is consistent with the hypothesis of an early detection of such cues by listeners, allowing them to effectively anticipate turn endings. Further research is needed to determine whether, and to what extent, listeners perceive and/or use these cues.

8.2.2 Backchannel-inviting cues preceding BC_O

We repeat the same analysis to study the presence of backchannel-inviting cues — as defined in Chapter 7 — in IPUs preceding backchannels with overlap (**BC_O**). The results are summarized in Table 8.2, and are comparable to the results obtained for backchannels without overlap (**BC**), shown in Tables 7.5 and 7.6 (pages 70 and 71). In both cases, we observe that IPUs preceding **BC** or **BC_O** tend to have a high number of conjointly displayed cues.

Cues	Count
`123456`	14
`12.456`	9
`.23456`	8
`12345.`	6
`123.56`	6
`1..456`	5
`123.5.`	4
`12.45.`	4
`.2.456`	3
`.2.45.`	3
	...

# Cues	**BC_O**	
0	1	(1.0%)
1	3	(2.9%)
2	8	(7.7%)
3	20	(19.2%)
4	28	(26.9%)
5	30	(28.8%)
6	14	(13.5%)
Total	104	(100%)

Table 8.2: Left: Top 10 frequencies of complex backchannel-inviting cues for IPUs preceding **BC_O** (*cf* Table 7.5 on page 70). Right: Distribution of number of backchannel-inviting cues in IPUs preceding **BC_O** (*cf* Table 7.6 on page 71).

These results indicate that IPUs preceding backchannels (**BC**) and backchannels with overlap (**BC_O**) present a similar behavior in terms of the occurrence of the discussed backchannel-inviting cues. Again, this finding is consistent with the hypothesis of an early detection of these cues by listeners, allowing them to anticipate the places where backchannel responses would be welcome by their interlocutors. Future research should investigate the perception and usage of these cues by listeners.

8.3 Early turn-yielding cues

In this section we investigate the second hypothesized explanation for overlapping turns: the occurrence of turn-yielding cues earlier in the current speaker's turn. First, we examine the durational distribution of overlapping segments, and find that the current turn's second-to-last intermediate phrase is a reasonable place to search for such cues. Subsequently, we identify a number of early turn-yielding cues. Given the low count of backchannels with overlap (**BC_O**) in the corpus, we restrict this preliminary study to overlaps (**O**).

8.3.1 Onset of overlaps

The annotation of turn-taking phenomena in the Games Corpus specifies only the presence or absence of overlapping speech (e.g., **O** vs. **S**). However, it does not provide information about the duration of the overlapping segments, a knowledge useful for inferring the location of cues potentially perceived and used by listeners early enough to anticipate turn endings. We investigate, then, how long overlapping turns begin before the end of the previous turns.

Figure 8.2 shows the cumulative distribution function of the duration of overlapping speech segments in overlaps (**O**). Around 60% of the instances have 200 ms or less of simultaneous speech, and 10% have 500 ms or more, although only a marginal number have more than one second. If we look at lexical rather than temporal units, we find that 613 (80%) of all instances begin during the last word in the previous turn; 100 (13%), during the second-to-last word; and the remaining 54 (7%), before that. The mean duration of the final word before overlaps is 384 ms (stdev = 180 ms); and of the second-to-last word, 376 ms (stdev = 170 ms).

Finally, looking at prosodic units, we find that over 95% of overlaps begin during the turn-final intermediate phrase (*ip*), according to the ToBI conventions.[1] The mean duration

[1] This computation, as well as the subsequent analysis of early turn-yielding cues, considers only the

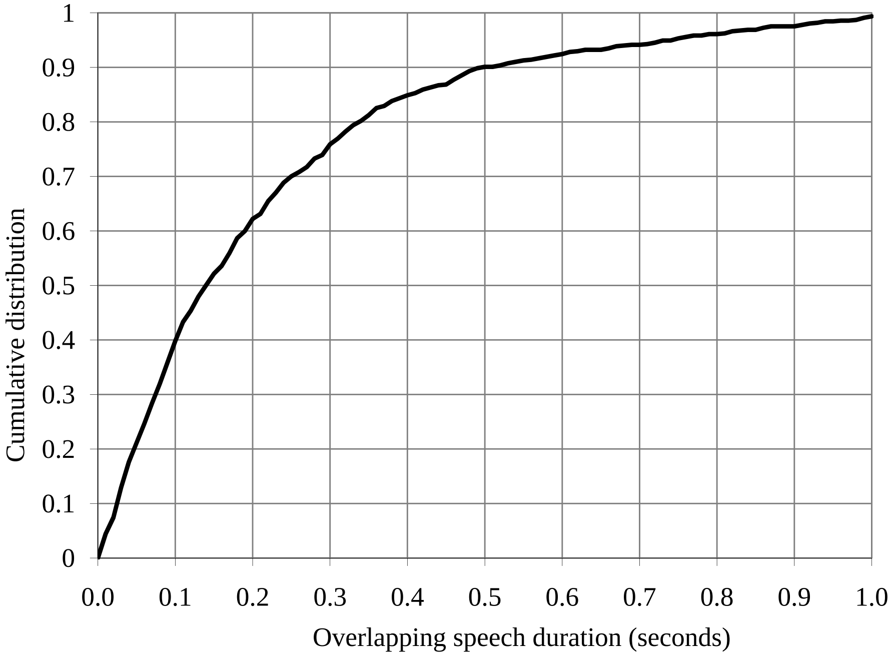


Figure 8.2: Cumulative distribution function of the duration of overlapping speech segments in overlaps (**O**).

of the final *ip* before overlaps is 747 ms (stdev = 418 ms).

These results indicate that, while in most cases the overlapping turn begins just before the end of the previous turn, in some cases the overlapping speech spans up to several words. Nonetheless, since nearly the totality of overlaps occur during the turn-final *ip*, the second-to-last *ip* appears to be a plausible place to search for early turn-yielding cues.

8.3.2 Cues in second-to-last intermediate phrases

To complete this preliminary study, we search for early turn-yielding cues in the second-to-last *ip*s preceding overlaps (**O**), using a slightly modified version of the procedure described in the previous chapters: Our current approach consists in contrasting the second-to-last *ip*s before **O** with prior turn-internal *ip*s (which we call **H**, analogously to the IPUs preceding 'hold' transitions). Any significant differences found would suggest the existence of potential

portion of the Games Corpus that is annotated using the ToBI framework, which includes 538 instances of partially-overlapping **O**.

turn-yielding cues. Additionally, we examine second-to-last *ip*s before **S**, to determine whether any such cues tend to occur in all turn endings, or whether they constitute a device that triggers or invites overlaps.

We find significant differences (ANOVA, $p < 0.05$; Tukey 95%) in speaking rate, measured in number of syllables and phonemes per second, over the whole *ip* and over its final word, as shown in Figure 8.3. The speaking rate of second-to-last *ip*s before **O** is significantly faster than that of *ip*s preceding **H**. We also find that second-to-last *ip*s preceding **O** tend to be produced with significantly lower intensity, and with higher values of three voice quality features — jitter, shimmer and NHR (Figure 8.4). Additionally, second-to-last *ip*s preceding **O** and second-to-last *ip*s preceding **S** show no significant differences with respect to all these features.

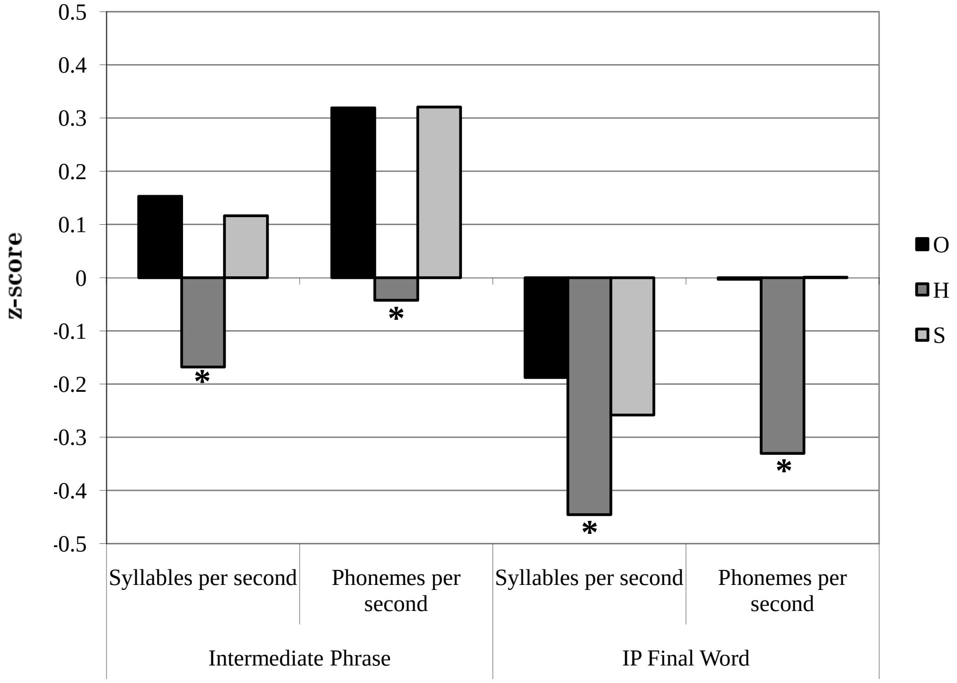


Figure 8.3: Speaker-normalized number of syllables and phonemes per second, computed over the whole intermediate phrase and over its final word.

These differences might suggest, at first sight, the existence of early turn-yielding cues related to these features. However, a closer inspection reveals that these results are equiv-

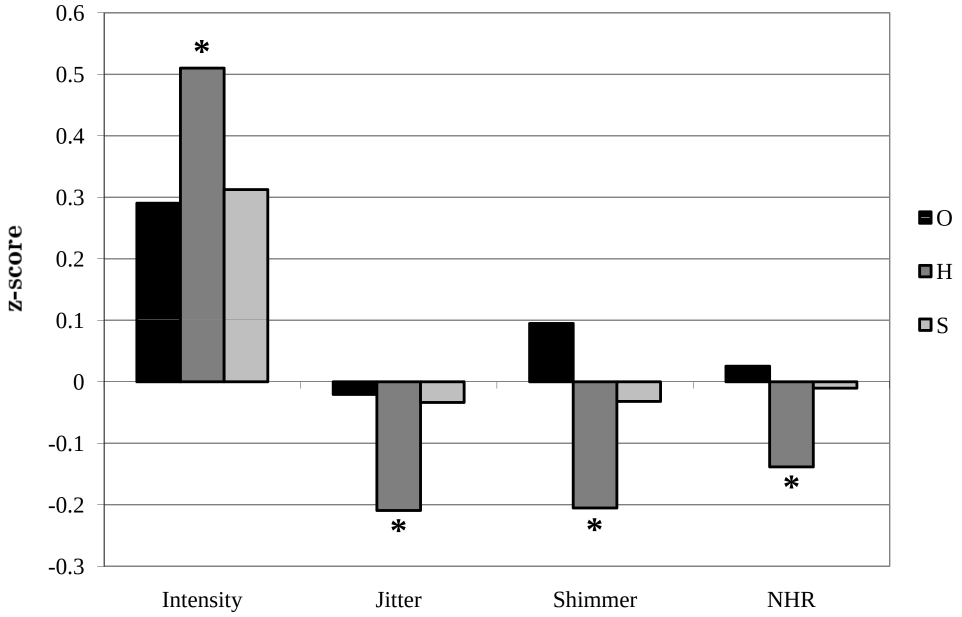


Figure 8.4: Speaker-normalized mean intensity, jitter, shimmer and NHR, computed over the whole intermediate phrase.

alent to the ones discussed in Chapter 6, according to which IPUs preceding **S** tend to be produced with faster speaking rate, lower intensity, and higher jitter, shimmer and NHR than IPUs preceding **H**. Often, turn-final IPUs contain more than one *ip*, which would explain the results presented in this section as a mere consequence of the ones presented in Chapter 6 — if something is true for an entire IPU, it will likely be true for the *ip*s that form it. However, 58% of IPUs preceding **S** and 48% of IPUs preceding **O** contain exactly one *ip*; in those cases, second-to-last *ip*s occur earlier than turn-final IPUs. In consequence, rather than the existence of distinct early turn-yielding cues, these results suggest the **prolongation** of turn-final cues further back in the turn. In other words, these turn-yielding cues apparently start to be displayed before the final IPU, probably growing in prominence as the turn gradually approaches its end (as indicated by the increasing differences observed for intensity, jitter, shimmer and NHR towards the end of the turn; see Figures 6.3 and 6.4 on pages 37 and 47).

8.4 Discussion

In this chapter we have presented the results of a preliminary study of overlapping speech in conversation. We find that IPUs preceding overlaps and smooth switches show comparable patterns of turn-yielding cues. Similarly, IPUs preceding backchannels with and without overlap show comparable patterns of backchannel-inviting cues. In other words, we find no indication of cues inviting the listener to make a contribution — either take the turn or produce a backchannel response — slightly overlapping the previous turn. If such cues existed and we were able to characterize them, IVR systems could then try to avoid producing them in their output, as a measure to prevent simultaneous speech, which poses serious difficulties for ASR systems (Shriberg et al., 2001).

Additionally, we observe that some of the turn-yielding cues described in Chapter 6 seem to originate further back in the turn, gradually increasing its prominence toward the end of the turn. This finding opens a new direction for future research, which could investigate turn-yielding and backchannel-inviting cues not as discrete events occurring at turn endings, but as phenomena that extend over entire conversational turns, starting low at turn beginnings and gradually increasing toward transition-relevance places. Graphical models such as HMM and CRF might be appropriate for this task.

Chapter 9

Conclusions and Future Work

The studies of turn-taking presented in this thesis strongly suggest the existence of seven measurable events that take place with a significantly higher frequency on IPUs preceding smooth switches (when the current speaker completes an utterance and the interlocutor takes the turn after a short pause) than on IPUs preceding holds (when the current speaker continues speaking after a short pause). These seven events may act as turn-yielding cues, such that when several cues occur simultaneously, the likelihood of a subsequent turn-taking attempt by the interlocutor increases in a close to linear manner. Additionally, we have presented similar evidence of the existence of six backchannel-inviting cues such that, when they take place simultaneously, the likelihood of occurrence of a backchannel from the interlocutor increases in a quadratic fashion.

These findings could be used to improve several turn-taking decisions of state-of-the-art IVR systems, such as how to keep the floor, either preventing interruptions from the user or inviting the user to produce backchannel responses; how to yield the floor to the user; when to take the floor from the user; and when to produce backchannel responses to encourage the user to continue speaking. An improvement in the turn-taking capabilities of IVR systems should lead to a more natural and efficient human-computer interaction.

There are several possible directions for future research. The first is to experiment with novel turn-yielding and backchannel-inviting cues. For example, voice quality seems to be a promising source to look for new cues, given the good results obtained with jitter, shimmer and noise-to-harmonics ratio. Furthermore, these three features could be used as finer grained turn-yielding cues, rather than a single voice-quality cue as in our approach.

A second direction consists in modifying the model of complex cues adopted in this study, which implicitly assumes that all cues are equally important, contributing with either 0 or 1 to the total count. Future research should explore the assignment of numeric weights to the different cues, in order to account for their relative importance.

Third, there seems to be some margin for improvement in the task of automatic classification of textual completion. Our best performing classifier, based on support vector machines, achieves an accuracy of 80%, while the agreement for humans is 90.8%. New approaches could incorporate features capturing information from the previous turn by the other speaker, which was available to the human labelers but not to the machine learning classifiers. Also, the sequential nature of this classification task might be better exploited by more advanced graphical learning algorithms, such as Hidden Markov Models and Conditional Random Fields.

Future research could also investigate turn-yielding and backchannel-inviting cues, not as discrete events occurring in the final portion of conversational turns, but as phenomena that extend over entire turns, gradually increasing as turns approach potential transition-relevance places.

Another research direction consists in running a perception study to learn more about the detection of cues by human listeners. For example, in a Wizard-of-Oz setting subjects could be asked to respond as soon as possible to the interviewer's prompts, but without breaking the conversational flow. Through controlled manipulation of output parameters, it should be possible to assess the relative perceptual importance of individual and combined cues, as well as the subjects' ability to perceive them prior to the turn boundary.

Users of IVR systems sometimes engage in an uninterrupted flow of speech which the system might want to interrupt, either because it has already collected the information needed for the task at hand, or simply because it has lost track of what the user is saying and needs to start over. In such occasions, it is crucial for the system to interrupt in an acceptable manner. Modeling the way in which people interrupt in spontaneous, collaborative conversations should aid IVR systems in this aspect of turn-taking. Since our labeling scheme distinguishes three types of interruptions (simple, pause, and barge-in interruptions) another direction for future research would be characterizing interruptions, both identifying places where interruptions are more likely to occur, and also describing the acoustic and prosodic properties of the interrupter's speech.

Lastly, we find strong indications in the Games Corpus that affirmative cue words, such as *okay* or *alright*, play a central role in the organization of turn-taking in task-oriented dialogue. These words are heavily overloaded, used to convey acknowledgment, to backchannel, and to begin or end discourse segments, among other functions. Therefore, we devote Part III of this thesis to study the realization of affirmative cue words in the Games Corpus.

Part III

Affirmative Cue Words

Chapter 10

Motivation and Research Goals

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. These constructions have received several names in the literature, such as 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, incidentally*, and countless others.

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 (Litman and Allen, 1987; Grosz and Sidner, 1986), and discourse segmentation (Litman and Passonneau, 1995). 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 the studies presented in Part III of this thesis, 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 very frequent in spontaneous conversation, especially in task-oriented dialogue. As we have seen in the description of the Games Corpus, ACWs account for almost eight percent of all words in the corpus. Also, these words appear to be heavily overloaded. Some of them (e.g., *alright*, *okay*) are capable of conveying as many as ten different discourse/pragmatic functions.

ACWs are strongly connected to turn-taking in conversation along various dimensions. First, they are the most natural choice for backchannel responses — in the Games Corpus, all backchannels are instances of ACWs. Second, they may function as explicit turn-yielding cues, as in *"right?"* or *"okay?"* at the end of a sentence. And third, ACWs may also be used to initiate a new conversational turn. Therefore, it is crucial for IVR systems to distinguish correctly between the several discourse/pragmatic functions of ACWs, both for speech generation and speech understanding tasks. In particular, a better understanding of the characteristics of backchannels would help us answer the following two questions posed in the introduction of Part II that we have not addressed yet:

Q4. The user has produced a short segment of speech; how can the system tell whether that was a backchannel or an attempt to take the turn?

Q7. The user is speaking and the system wants to produce a backchannel response; how should it formulate its output for the backchannel to be interpreted correctly?

Part III of this thesis describes a series of studies aimed at advancing our understanding of ACWs. We seek descriptions of the acoustic/prosodic characteristics of their functions, a knowledge helpful in spoken language generation tasks. Additionally, we assess the predictive power of computational methods for their automatic disambiguation, a capability useful for various spoken language understanding tasks. Lastly, we investigate speaker entrainment — or, how conversational partners tend to adapt their speech to each other's behavior — related to the usage of high-frequency words, including ACWs, and explore its connection to task success and dialogue coordination.

Chapter 11

Previous Work on Cue Phrases

Cue phrases have received extensive attention in the Computational Linguistics literature. Early work by Cohen (1984) presents a computational justification for the usefulness and the necessity of cue phrases in discourse processing. Using a simple propositional framework for analyzing discourse, the author 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" [p. 251]. Additionally, Cohen introduces a taxonomy of cue phrases consisting of six categories of connectives: parallel (e.g., *in addition*), inference (*therefore*), detail (*in particular*), summary (*in sum*), and reformulation (*in other words*).

Reichman (1985) proposes a model of discourse structure in which discourse comprises a collection of basic constituents called CONTEXT SPACES, organized hierarchically according to various kinds of semantic and logical relations called CONVERSATIONAL MOVES. In such a model, cue phrases are portrayed as mechanisms that signal context space boundaries, specifying the kind of conversational move about to take place. Reichman identifies eleven types of conversational moves, and provides a list of example cue phrases for each. For instance, expressions such as *because* and *like* function as SUPPORT conversational moves, which introduce new elements supporting previous arguments; and expressions such as *incidentally* and *by the way* function as INTERRUPTION moves, which introduce a sudden topic change.

Grosz and Sidner (1986) introduce an alternative model of discourse structure formed by three interrelated components: a LINGUISTIC STRUCTURE, which defines a hierarchy of discourse segments, an INTENTIONAL STRUCTURE, which comprises the discourse intentions that organize the discourse segments, and an ATTENTIONAL STATE, which models the attention as a stack of focus spaces. In such a 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" [p. 196]. For example, expressions such as *for example* and *moreover* push a new focus space onto the attentional stack, and create a new discourse segment subordinated to the current one; expressions such as *anyway* and *in any case* pop the existing space from the stack, and return to a previous discourse segment.

Subsequent studies propose a formal definition of cue phrases. For example, a corpus study of spontaneous conversations by 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. Schiffrin observes, like previous studies, that cue phrases provide contextual coordinates for an utterance in the discourse, but suggests nonetheless that cue phrases only **display** the discourse structure relations, rather than create them. Later on, 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" [p. 1169]. A detailed review of these and other related works can be found in Fraser (1999).

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*. When used in a discourse sense, a cue phrase explicitly conveys structural information; when used in a sentential sense, a cue phrase instead conveys semantic rather than structural information. Hirschberg and Litman present two manually developed classification models, one based on prosodic features, and one based on textual features. In the prosodic model, when a cue phrase is uttered as a single intermediate phrase, or in a larger intermediate phrase with an initial position and a L* accent or deaccented, it is classified as 'discourse'; otherwise, as 'sentential'. In the textual model, when a cue phrase is preceded by any punctuation or by a paragraph boundary (as specified in manual transcriptions of the recordings), it is classified as 'discourse'; otherwise, as 'sentential'. An evaluation of both models on a single-speaker keynote address in SAE reports an error rate of 24.6% for the prosodic model, or 14.7% when excluding all instances of conjuncts *and*, *or*, and *but* — for which classification into discourse and sentential senses by human annotators is reported to be highly unreliable. The error rate of the textual model is 19.9% in general, and 16.1% after removing conjuncts. These results significantly improve over the majority-class ('sentential') baselines, whose error rates are 38.8% and 40.8%, respectively.

This line of research is further pursued by Litman (1994; 1996), who incorporates machine learning techniques to derive classification models automatically. Litman extracts a number of prosodic features (e.g., accent type, length of intonational phrase) and textual features (e.g., part-of-speech tags, preceding punctuation symbol or paragraph boundary), and uses them to train decision-tree and rule learners on the same data from the previous studies, experimenting with different combinations of features. Litman then compares the performance of automatically and manually learned models using all prosodic features, all textual features, and all features combined, as summarized in Table 11.1. The automatic models outperform the manual models for all single-word cue phrases; when conjuncts are excluded, however, all models reach comparable error rates. In all, these studies show

Model	All cue phrases	Non-conjuncts
Manual prosodic	24.6%	14.7%
Manual textual	19.9%	16.1%
Automatic prosodic	15.5%	17.2%
Automatic textual	18.8%	19.0%
Automatic prosodic+textual	15.9%	14.6%

Table 11.1: Error rates of manual and automatic classifiers (Litman, 1996).

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 their discourse and sentential senses, achieving a performance close to that of human annotators. More recently, Lai (2008) discusses a characterization of prosodic cues for distinguishing two possible uses of the word *really*, as a question or as a backchannel.

Despite their high frequency in spontaneous conversation, affirmative cue words have been little studied as a separate subclass of cue phrases. An exception is a study by Hockey (1991; 1992) on the prosodic variation of tokens of *okay* and *uh-huh* produced as full intonational phrases in two spontaneous task-oriented dialogues. Hockey groups the F_0 contours visually and auditorily, “using characteristics such as relative F_0 height of the first and second syllables and general shapes of the two syllables (e.g. rise, fall, level, degree of rise or fall)” [p. 129]. This clustering procedure divides the intonational contours into three groups, described impressionistically by the author, which roughly match the ToBI contours H* H-L% (plateau), H+!H* L-L% (downstep), and H* H-H% (high-rise). The only result described by the author showing statistical significance is that tokens of *okay* produced with a high-rise contour are more likely to be followed by speech from the other speaker than from the same speaker, which could be the case of either a backchannel or a turn change.

In a study of the function of intonation in British English task-oriented dialogue, Kowtko (1997) examines single-word utterances, including affirmative cue words such as *mm-hm*, *okay*, *right*, *uh-huh* and *yes*. She finds a significant correlation between discourse function and intonational contour. For example, 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, correlates with non-rising intonation; and the ACKNOWLEDGE function, which indicates having heard and understood, presents overall a non-falling intonation.

As part of a larger project on automatically detecting discourse structure for speech recognition and understanding tasks, Jurafsky et al. (1998) present a study of four particular discourse/pragmatic functions, or DIALOG ACTS (Stolcke et al., 2000), closely related to ACWs: CONTINUER (short utterance indicating that the other speaker should go on talking), INCIPIENT SPEAKERSHIP (indicating an intention to take the floor), AGREEMENT (indicating the speaker's agreement with a statement or opinion expressed by another speaker), and YES-ANSWER (affirmative answer to a yes-no question).[1] The authors examine 1155 conversations from the Switchboard database (Godfrey et al., 1992), and report that the vast majority of these four dialog acts are realized with words like *yeah*, *okay*, or *uh-huh*. They find that the lexical realization of the dialog act is the strongest cue to its identity. For example, *uh-huh* is used as a continuer twice as often as *yeah*, while *yeah* is used to take the floor (incipient speakership) three times as often as *uh-huh*. They also report preliminary results on a few prosodic differences across dialog acts. Continuers tend to be shorter in duration, with a flatter contour, and lower in F_0 and intensity than agreements. When continuers end in rising intonation, however, they can be longer, and higher in F_0 and intensity. Also, falling intonation tends to be associated with agreements more often than with continuers. Interestingly, they report that some speakers tend to use a characteristic prosody on a particular lexical item to distinguish its continuer and agreement uses, while others seem to use one lexical item exclusively for continuers and another for agreements.

[1] In this thesis we refer to continuers as backchannels, a term that Jurafsky et al. (1998) use in a broader sense, to include the continuer, incipient-speakership and agreement dialog acts, among others. In the coding scheme presented in Chapter 5, incipient-speakership corresponds roughly to the cue beginning functions, **CBeg** and **PBeg**; and agreement and yes-answer are collapsed into a single class, **Ack**.

Chapter 12

ACWs in the Games Corpus

The materials for the studies of ACWs presented in this thesis were again taken from the Games Corpus. In total, this corpus has 5456 instances of affirmative cue words *alright*, *gotcha*, *huh*, *mm-hm*, *okay*, *right*, *uh-huh*, *yeah*, *yep*, *yes* and *yup*, which were labeled by three annotators into the ten different discourse/pragmatic functions listed in Table 12.1. Labelers were given examples of each category, and labeled using both transcripts and speech together. The complete guidelines used by the annotators are presented in Appendix C. Inter-labeler reliability was measured by Fleiss' κ (Fleiss, 1971) as 'substantial' at 0.69. 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 5456 affirmative cue words in the corpus, 5185 (95%) have a majority label. Table 12.2 shows the distribution of discourse/pragmatic functions over ACWs in the whole corpus.

Throughout the Games Corpus, there are 8139 conversational turns.[1] Of the 2480 turns containing just one word, 2015 (81.2%) consist of an ACW. Of the 5659 turns containing more than one word, 1520 (26.9%) begin with an ACW, and 780 (13.8%) end with one. These numbers show clearly the central role that ACWs play in turn-taking in task-oriented conversations. The wide range of discourse/pragmatic meanings associated with ACWs

[1] Recall from Chapter 2 that we define a TURN as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. An INTER-PAUSAL UNIT (IPU) is defined as a maximal sequence of words surrounded by silence longer than 50 ms.

Ack	**Acknowledgment/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 (Ack+CBeg).** Functions both to acknowledge/agree and to cue a beginning segment.
PEnd	**Pivot ending (Ack+CEnd).** Functions both to acknowledge/agree and to cue the end of the current segment.
Mod	**Literal modifier.** Example: *"I think that's okay"*.
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.

Table 12.1: Labeled discourse/pragmatic functions of affirmative cue words.

make this class of cue phrases a powerful tool for speakers to coordinate the development of tasks requiring a high degree of collaboration.

12.1 Data downsampling

Table 12.2 shows the complete distribution of ACWs and discourse/pragmatic functions in the corpus. Some of the word/function pairs in that table 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.

	alright	*mm-hm*	*okay*	*right*	*uh-huh*	*yeah*	Rest	Total
Ack	76	58	1092	111	18	754	116	2225
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	1069	0	0	0	1092
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	2265	1258	169	903	147	5456

Table 12.2: Distribution of function over ACW. Rest = {*gotcha, huh, yep, yes, yup*}

Therefore, we downsample 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. We discard tokens of ACWs until two conditions are met: for each word/function pair, (a) tokens come from at least four different speakers, and (b) no single subject 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.

This procedure leads to discarding 506 tokens of ACWs, or 9.3% of such words in the corpus. Table 12.3 shows the resulting distribution of discourse/pragmatic functions over ACWs in the whole corpus after downsampling the data.

12.2 Feature extraction

We extract a number of lexical, discourse, timing, phonetic, acoustic and prosodic features for each target ACW, which we use in the statistical analysis, machine learning experiments

	alright	*mm-hm*	*okay*	*right*	*uh-huh*	*yeah*	Rest	Total
Ack	76	58	1092	74	16	754	87	2157
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	1069	0	0	0	1091
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	2107	1192	117	830	91	4950

Table 12.3: Distribution of function over ACW, after downsampling.

Rest = {*gotcha, huh, yep, yes, yup*}

and perception studies presented in the following chapters. Tables 12.5 and 12.6 summarize the full feature set. Some considerations regarding the process of feature extraction, such as the part-of-speech tagger or the method for calculating pitch slopes, are given in the corpus description in Part I of this thesis.

Boundaries of IPUs and turns are computed automatically from the time-aligned transcriptions. A TASK in the Cards Games corresponds to matching a card, and in the Objects Games to placing an object in its correct position. Task boundaries are extracted from the logs collected automatically during the sessions, and later checked by hand.

For the phonetic features, we train 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, we obtain automatic time-aligned phonetic transcriptions of each instance of ACWs in the Columbia Games Corpus. For improved accuracy, we restrict the recognizer's grammar to accept only the most frequent variations of each word, as shown in Table 12.4.

We extract our phonetic features, such as phone and syllable durations, from the resulting time-aligned phonetic transcriptions.

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)`

Table 12.4: Restricted grammars for the automatic speech recognizer. Phones in square brackets are optional.

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) without diacritics such as ‘p’ or ‘-’.

Additionally, we categorize the features according to the portion of signal from which they were extracted: word-only (marked $\mathcal{W}$ in Tables 12.5 and 12.6), from just the target word itself; backward-looking ($\mathcal{B}$), from up to the IPU containing the target word; and all ($\mathcal{A}$), from the entire conversation. We create this taxonomy for the machine learning experiments described in Chapter 14, in which we assess, among other things, the usefulness of information extracted from each of the three sources, simulating the conditions of actual online and offline applications.

In the following chapters, we use the features described here in several ways. First, we perform a series of statistical tests to find differences in the production of the function of ACWs. Second, we experiment with machine learning techniques for the automatic classification of the function of ACWs, training the models with different combinations of features. Finally, we investigate the relative importance of contextual features in human disambiguation of ACWs.

Lexical features	
$\mathcal{WBA}$	Lexical identity of the target word (w).
$\mathcal{WBA}$	Part-of-speech tag of w, original and simplified.
$\mathcal{BA}$	Words immediately preceding and following w, and their original and simplified POS tags.
Discourse features	
$\mathcal{BA}$	Number of words in w's IPU.
$\mathcal{BA}$	Number and proportion of words in w's IPU before and after w.
$\mathcal{BA}$	Number of words uttered by the other speaker during w's IPU.
$\mathcal{BA}$	Number of words in the previous turn by the other speaker.
$\mathcal{A}$	Number of words in w's turn.
$\mathcal{A}$	Number and proportion of words and IPUs in w's turn before and after w.
$\mathcal{A}$	Number and proportion of turns in w's task before and after w.
$\mathcal{A}$	Number of words uttered by the other speaker during w's turn.
$\mathcal{A}$	Number of words in the following turn by the other speaker.
$\mathcal{A}$	Number of ACWs in w's turn other than w.
Timing features	
$\mathcal{WBA}$	Duration (in ms) 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).
$\mathcal{BA}$	Flag indicating whether there was any overlapping speech from the other speaker.
$\mathcal{BA}$	Duration of w's IPU.
$\mathcal{BA}$	Latency (in ms) between w's turn and the previous turn by the other speaker.
$\mathcal{BA}$	Duration of the silence before w (or 0 if the w is not preceded by silence), its IPU, and its turn.
$\mathcal{BA}$	Duration and proportion of w's IPU elapsed before and after w.
$\mathcal{BA}$	Duration of w's turn before w.
$\mathcal{BA}$	Duration of any overlapping speech from the other speaker during w's IPU.
$\mathcal{BA}$	Duration of the previous turn by the other speaker.
$\mathcal{A}$	Duration of the silence after w (or 0 if w is not followed by silence), its IPU, and its turn.
$\mathcal{A}$	Latency between w's turn and the following turn by the other speaker.
$\mathcal{A}$	Duration of w's turn, as a whole and after w.
$\mathcal{A}$	Duration of any overlapping speech from the other speaker during w's turn.
$\mathcal{A}$	Duration of the following turn by the other speaker.

Table 12.5: Feature set. *Continued in Table 12.6.*

Acoustic features	
$\mathcal{WBA}$	w's mean, maximum, minimum pitch and intensity (raw and speaker normalized).
$\mathcal{WBA}$	w's ratio of voiced frames to total frames (raw and speaker normalized).
$\mathcal{WBA}$	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).
$\mathcal{WBA}$	Noise-to-harmonics ratio (NHR), computed over the whole word and over the first and second syllables (raw and speaker normalized).
$\mathcal{WBA}$	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 ms (raw and normalized with respect to all other occurrences of the same word by the same speaker).
$\mathcal{BA}$	w's mean, maximum, 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.
$\mathcal{BA}$	Voiced-frames ratio, jitter and shimmer, normalized with respect to the same three types of context.
$\mathcal{BA}$	Mean, maximum, minimum pitch and intensity, ratio of voiced frames, (all raw and speaker normalized), jitter and shimmer, calculated over the final 500, 1000, 1500 and 2000 ms of the previous turn by the other speaker (only defined when w is turn initial but not task initial).
$\mathcal{BA}$	Pitch slope, intensity slope, and stylized pitch slope, calculated over the final 100, 200, 300, 500, 1000, 1500 and 2000 ms of the previous turn by the other speaker (only defined when w is turn initial but not task initial).
Phonetic features	
$\mathcal{WBA}$	Identity of each of w's phones.
$\mathcal{WBA}$	Absolute and relative duration of each phone.
$\mathcal{WBA}$	Absolute and relative duration of each syllable.
Session-specific features	
–	Session number.
–	Identity and gender of both speakers.
ToBI prosodic features	
–	Pitch accent, phrase accent, boundary tone and break index on w (original and simplified ToBI labels).
–	Pitch accent, phrase accent, boundary tone and break index on the final intonational phrase of the previous turn by the other speaker (original and simplified ToBI labels; only defined when w is turn initial).

Table 12.6: Feature set. *Continued from Table 12.5.*

Chapter 13

Descriptive Statistics

In this chapter 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. To look for such differences, for each numeric feature we conduct a repeated-measures analysis of variance (RMANOVA) test, considering the data from all speakers together. In most cases, the low count of word/function pairs for individual speakers impedes assessing those differences for each speaker separately. Therefore, instead of regular ANOVA, we use RMANOVA tests, which 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.

13.1 Context

We begin this analysis by looking at the discourse context of the various discourse/pragmatic functions of ACWs. Since these words help shape, or at least reflect, the structure of conversations, we expect to find contextual differences between their functions. Figure 13.1 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 the corresponding IPU.[1] An

[1] See Table F.2 in Appendix F for the actual numbers corresponding to this figure.

ipu-initial word is one that occurs in the first position in its corresponding IPU; i.e., it is preceded by at least 50 milliseconds 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 13.1 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 acknowledgments (**Ack**), half are cues to beginning discourse segments (**CBeg**), and a marginal number convey other functions.

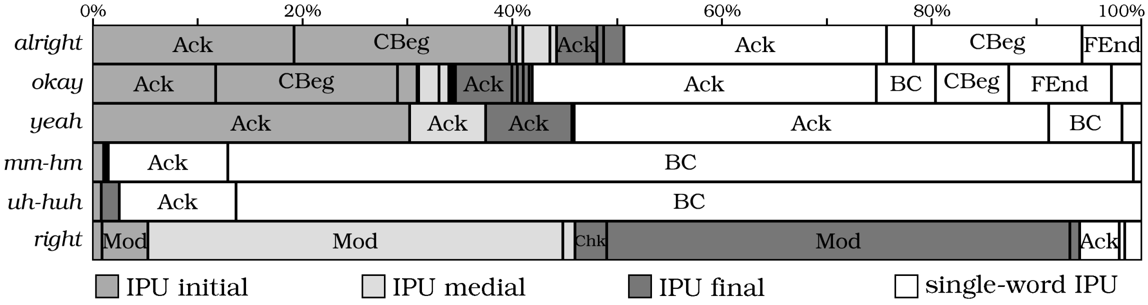


Figure 13.1: Position of the target word in its IPU.

Similarly, Figure 13.2 shows the distribution of the same six ACWs with respect to their position in the corresponding conversational turn.[2] Turn-initial, turn-medial and turn-final words, and single-word turns are defined analogously to the four IPU-related categories defined above, but considering conversational turns instead of word IPUs.

From these figures we observe several interesting aspects of the discourse context 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, acknowledgment/agreement (**Ack**), occurs for *alright*, *okay*, *yeah* and *right* in all possible positions within the IPU and the turn; for *mm-hm* and

[2] See Table F.3 in Appendix F.

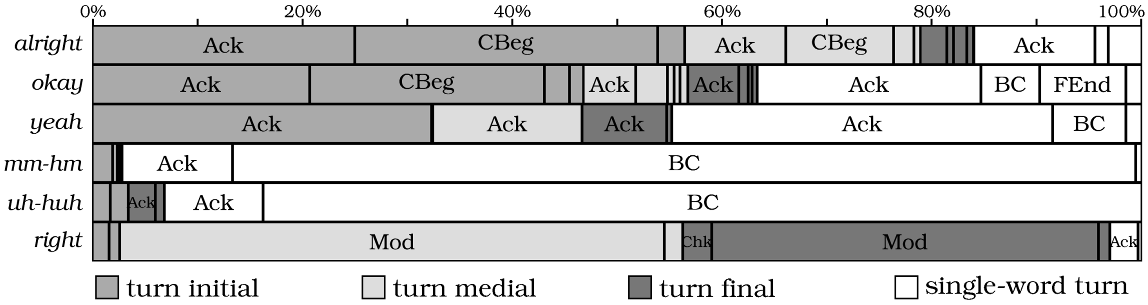


Figure 13.2: Position of the target word in its turn

uh-huh, acknowledgments 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 other two cases, two backchannels are uttered in fast repetition (e.g., *"uh-huh uh-huh"*).

In all, these preliminary results confirm the existence of large contextual differences between the discourse/pragmatic functions of ACWs, and also between their lexical types. We will revisit this topic twice in this thesis. In Chapter 14 we discuss the predictive power of contextual features in the automatic classification of the function of ACWs. Given the observed contextual differences, we expect these features to play a prominent role in such a task. Subsequently, in Chapter 15 we investigate the importance of contextual information in human perception of the function of ACWs. In particular, we study the extent to which the disambiguation process is affected by the complete lack of contextual information.

13.2 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 13.3 shows the distribution of ToBI labels for each of the six most frequent ACWs and their corresponding functions.[3] The distributions

[3] See Table F.1 in Appendix F for the actual numbers corresponding to this figure.

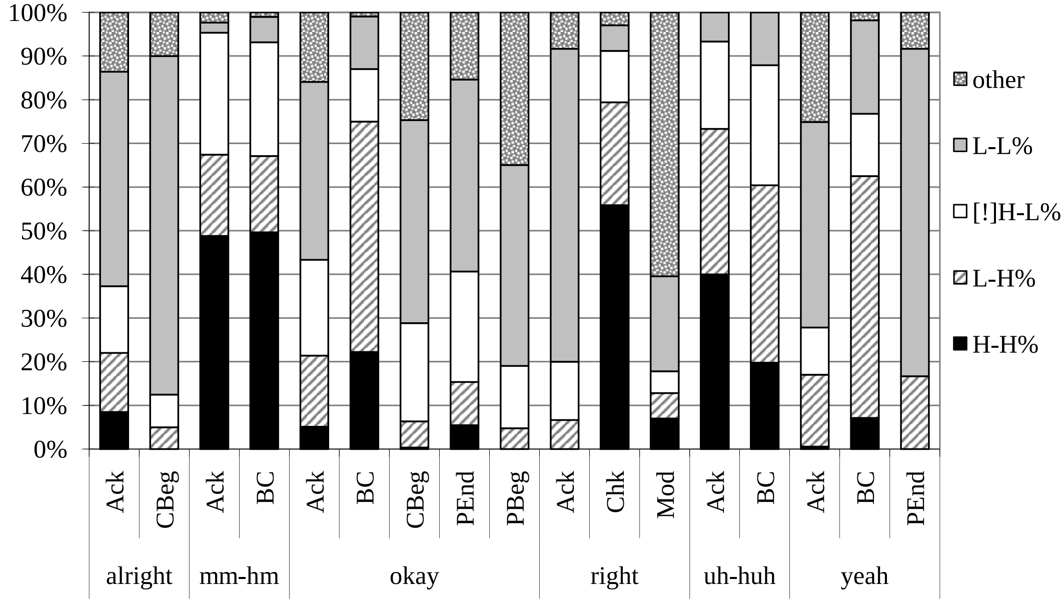


Figure 13.3: ToBI phrase accents and boundary tones. The 'other' category consists of cases with no phrase accent and/or boundary tone present at the target word.

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, $p \approx 0$). 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.

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 default function of ACWs, acknowledgment/agreement (**Ack**) is produced most often with falling (L-L%) or plateau final intonation ([!]H-L%) in the case of *alright*, *okay*, *right* and *yeah*. Notably, **Ack** instances of *mm-hm* and *uh-huh* present a very different behavior, with a distribution of final intonations that closely resembles that of backchannels.

In particular, over 60% of the tokens of *mm-hm* and *uh-huh* are produced with a final rising intonation (either L-H% or H-H%).

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 acknowledgment 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*.

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

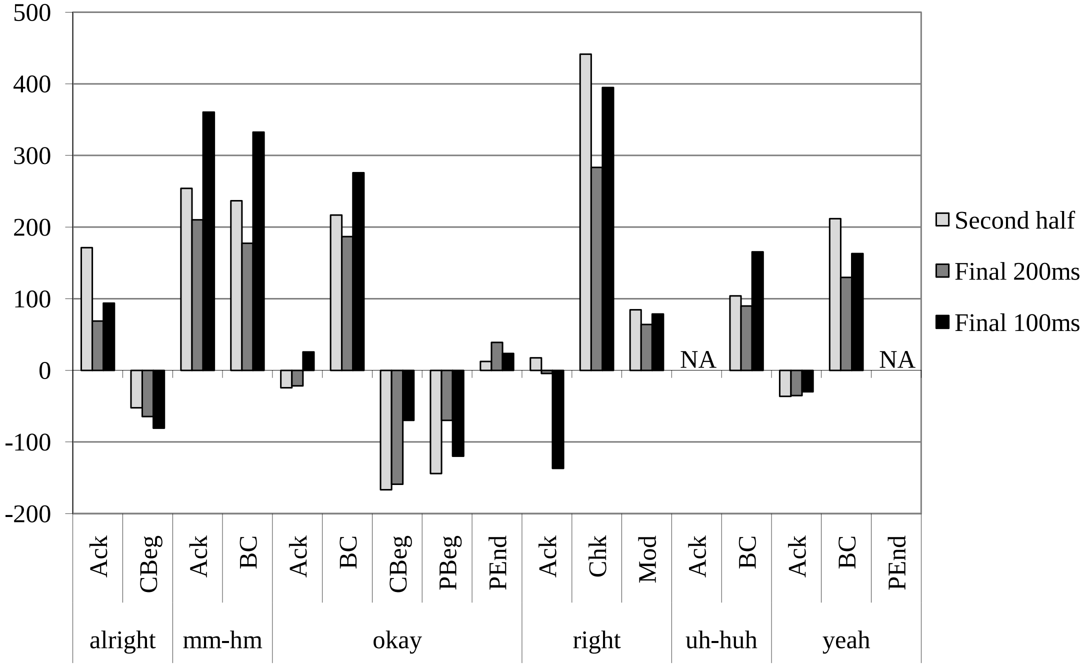
A: *and the top's not either, right?*
B: *no*
A: *okay*

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 above, word final intonation may also be studied by exploring the slope of the word-final pitch track. Figure 13.4 shows, for the same ACWs and functions discussed above,[4] the mean pitch slope computed over the second half of the word and over its final 100 and 200 milliseconds, and gender-normalized as described in Section 2.2.

The comparison of these numeric acoustic features across discourse/pragmatic functions provides additional support for the observations made above. 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 **Ack**, **BC** and **PEnd** (RMANOVA for each of the three variables: between-subjects $p > 0.3$, within subjects $p \approx 0$; Tukey test

[4] For **PEnd** instances of *yeah* and **Ack** 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.



Significant differences: For *okay*: BC>all; CBeg<Ack, BC, PEnd.

For *right*: Chk>Ack. For *yeah*: BC>Ack.

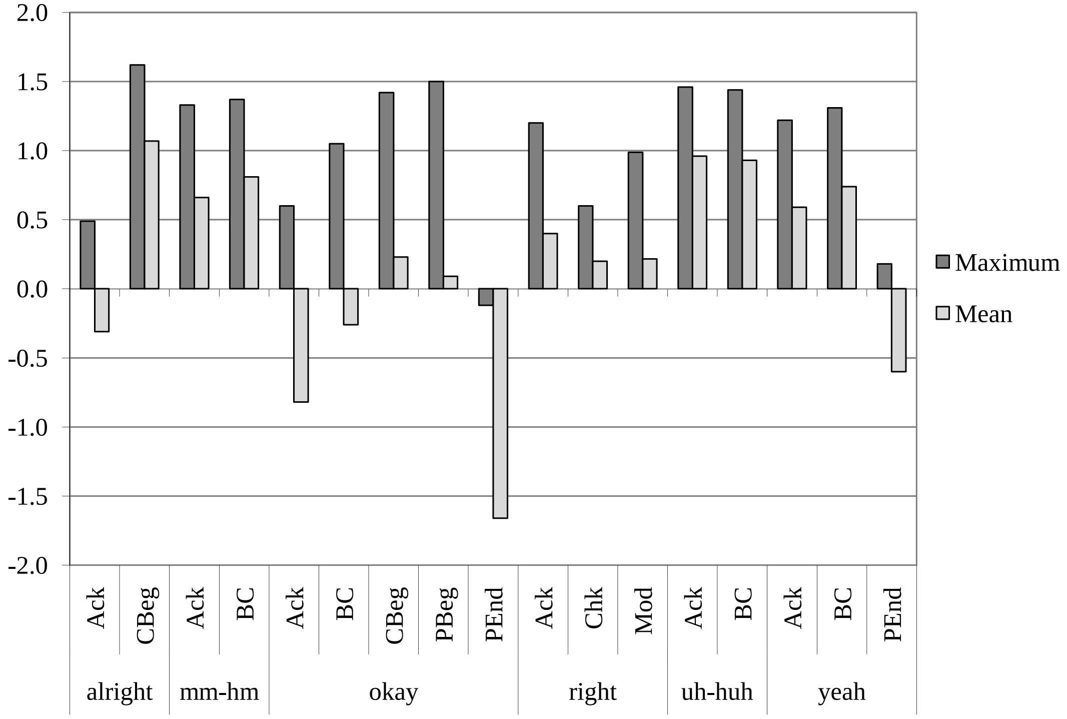
Figure 13.4: Final pitch slope, computed over the second half and the final 100 and 200 milliseconds of the target word.

confidence: 95%). **BC** tokens of *yeah* are also significantly higher than **Ack**, with similar p-values. Figure 13.4 shows that **BC** instances of *mm-hm* and *uh-huh* also have comparably high final pitch slopes. Again, for *mm-hm* we find no significant difference in final pitch slope between acknowledgments and backchannels.

Although Figure 13.4 shows that **Chk** tokens of *right* tend to end in a very high pitch slope, 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*.

13.3 Intensity

The next feature we find to vary significantly with the discourse/pragmatic function of ACWs is word intensity. Figure 13.5 shows the maximum and mean intensity for the most frequent ACWs and functions, computed over the whole word and speaker normalized using z-scores.



Significant differences: For *alright*: Ack<CBeg. For *yeah*: PEnd<Ack, BC.
For *okay*: PEnd<all; Ack<CBeg, PBeg, BC; BC<CBeg.

Figure 13.5: Word maximum and mean intensity.

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*, 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 et al., 1980; Hirschberg and Nakatani, 1996). Since by definition **CBeg/PBeg** ACWs begin a new topic and **CEnd/PEnd** end one, it is then expectable 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 two only functions, acknowledgment (**Ack**) and backchannel (**BC**). Recall from the previous section that we find no differences in final intonation either. This contributes to the hypothesis that these two lexical types tend to be produced with indistinguishable acoustic/prosodic features, independently of their function.

13.4 Other features

For the remaining acoustic/prosodic features described in Chapter 12 we find only a small number of significant differences between the functions of ACWs, 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** and **PBeg**) 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%).

Speaker-normalized **mean pitch** over the whole word also presents significant differences for *okay* and *yeah*. Instances labeled **PEnd** (acknowledgment 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%).

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, both *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). 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.

13.5 Discussion

In this chapter we have presented statistical evidence of a number of differences in the production of the various discourse/pragmatic functions of ACWs. The most marked contrasts in acoustic/prosodic features relate to word final intonation and word intensity. Backchannels typically end in a rising pitch, acknowledgments and cue beginnings in a falling pitch; cue beginnings are produced with a high intensity, cue endings with a very low one. Other acoustic/prosodic features — duration, mean pitch, shimmer and NHR — also seem to vary with the word usage.

Interestingly, every significant difference that we find for individual ACWs is also present when considering only the word *okay*. For example, if a word like *yeah* shows cue endings to have a lower intensity, then such difference is also true for the word *okay*. This suggests a plausible explanation for this finding is that (1) the mechanisms of acoustic/prosodic variation relative to word function are the same across all ACWs (*alright*, *okay*, *yeah*, etc.), and (2) the higher the ambiguity of the ACW (i.e., the more functions it may convey), the more marked such variation becomes.

This possibility gains additional support from the fact that for *mm-hm* and *uh-huh* we observe no clear differences in the production of their two main functions, backchannel and acknowledgment. These two words are used very rarely in the Games Corpus for conveying functions other than **BC** or **Ack**. Thus, listeners normally need to distinguish between two relatively similar meanings, and the production similarities between the two suggest that such distinction relies strongly on contextual cues. It is reasonable to assume that if *mm-hm* or *uh-huh* were frequently used to convey other functions, the acoustic/prosodic variation found in their productions might be more noticeable.

In this chapter, we have looked only for variation along individual features, such as word intensity and final intonation. However, there is no reason to assume that such features may not be coupled together to form more complex cues to disambiguation. In the next chapter, we employ three machine learning algorithms to explore, among other things, the effectiveness of different combinations of features in the automatic prediction of the discourse/pragmatic functions of ACWs.

As shown earlier in this chapter, ACWs also display substantial contextual 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 pose the question of whether context alone is enough for disambiguation purposes, with listeners not actually using any of the observed acoustic/prosodic variation. This question is addressed in the perception study presented in Chapter 15.

Chapter 14

Automatic Classification of ACWs

In this chapter we present results from a number of machine learning (ML) experiments aimed at investigating how accurately affirmative cue words may be classified automatically into their various discourse/pragmatic functions, a procedure from which multiple spoken language processing applications could potentially benefit. With that general goal in mind, we explore several dimensions of the problem: we consider three classification tasks, simulating the conditions in which actual applications may perform them, and study the performance of different ML algorithms and feature sets on each task.

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 (**Ack**, **BC**, **CBeg**, **PBeg**, **CEnd**, **PEnd**, **Mod**, **BTsk**, **Chk**, **Stl**; see Table 12.1). 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 could aid applications that need to segment speech into coherent units, such as meeting processing applications, or turn-taking components of IVR systems. The third task consists in identifying tokens conveying some degree of acknowledgment (**Ack**, **BC**, **PBeg**, **PEnd**), a function especially important in IVR systems for knowing that the user has understood the system's output. Previous studies disambiguate between the sentential and discourse uses of cue phrases such as *now*, *well* and *like*, for which there typically exist comparable amounts of instances conveying each use. For ACWs in the Games Corpus, sentential uses are rare, with the sole exception of *right*. Therefore, disambiguating between discourse and sentential uses appears to be less important than distinguishing among different discourse functions.

Speech processing applications operate in disparate conditions. ONLINE applications, such as IVR systems, process information as it is generated, having access to a very limited scope, normally up to the last IPU uttered by the user. On the other hand, OFFLINE applications, such as meeting transcription 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 12.2 into five sets — lexical (LX), discourse (DS), timing (TM), acoustic (AC) and phonetic (PH) — to determine the relative importance of each feature set in the various classification tasks. Among other things, this approach permits evaluating how accurately the function of ACWs may be determined based solely on textual features. TTS systems could later use such information to produce the target word with appropriate acoustic/prosodic features for its predicted function.

For our ML experiments we use three well-known algorithms with very different characteristics: the decision tree learner C4.5 (Quinlan, 1993), the propositional rule learner RIPPER (Cohen, 1995), and support vector machines (SVM; Vapnik, 1995; Cortes and 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.[1]

14.1 Classifiers and feature types

To assess the predictive power of the five feature types — lexical (LX), discourse (DS), timing (TM), acoustic (AC) and phonetic (PH) — we exclude one type at a time and compare the performance of the resulting set to that of the full model. Table 14.1 displays the error rate

[1] 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 12.1). The best results were obtained using a polynomial kernel with exponent $d = 1.0$ (i.e., a linear kernel) and model complexity $C = 1.0$.

of each ML classifier on the general task, classifying any ACW into any of the most frequent discourse/pragmatic functions (**Ack**, **BC**, **CBeg**, **PEnd**, **Mod**, **Chk**). Table 14.2 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 — **Ack**, **BC**, **PBeg** or **PEnd**, vs. all other labels).

	Error Rate			SVM F-Measure					
Feature Set	C4.5	Ripper	SVM	Ack	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	.00
`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
		Majority class baseline ER	56.4%						
		Word-based baseline ER	27.7%						
		Mean human labelers ER	9.8%						

Table 14.1: Error rate of each classifier on the general task using different feature sets; F-measures of the SVM classifier; and error rate of two baselines and human labelers. † Significantly different from full model. § Significantly different from SVM. (Wilcoxon signed rank sum test, $p < 0.05$.)

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. The '†' symbol indicates that the given classifier performs significantly worse when trained on a particular feature set than when trained on the full set.[2] The '§' symbol indicates that the difference between SVM and the given

[2] All accuracy comparisons discussed in this chapter 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.

classifier, either C4.5 or Ripper, is significant. For example, the second line (`DS TM AC PH`) in Table 14.1 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 14.1 and 14.2 show the error rate of two baselines, as well as an estimate 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 on word identity. In the general classification task, the majority class is **Ack**, and the best performing word-based rule is *huh*→**Chk**, *mm-hm*→**Mod**, *uh-huh*→**BC**, *right*→**Mod**, others→**Ack**. 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 by comparing the labels assigned by each labeler and the majority labels

	Disc. Boundary			Acknowledgment		
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%
Majority class baseline ER	18.6%			36.5%		
Word-based baseline ER	18.6%			15.3%		
Mean human labelers ER	5.7%			3.3%		

Table 14.2: 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, $p < 0.05$.)

as defined in Chapter 12.

The right half of Table 14.1 shows the F-measure of the SVM classifier for each individual ACW function, for the general task. The highest F-measures correspond to **Ack**, **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. 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 than the other two.

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.

14.1.1 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 12.6), 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 12.6), does not improve the performance of the SVM-based full models in any of the classification tasks.

14.1.2 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 ms and second half), voiced-frames ratio, and noise-to-harmonics ratio. All phonetic features are ranked very low. These results again confirm the existence of large contextual differences across functions of ACWs. Additionally, while several acoustic/prosodic features extracted from the target word contain useful information for the automatic disambiguation of ACWs, it is contextual information that provides the most predictive power.

14.2 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 extracted from the speech signal only up to the IPU containing the target ACW. These features are marked in Tables 12.5 and 12.6 (pages 101 and 102) with the letter $\mathcal{B}$ (backward looking). With these features, we train and evaluate an SVM classifier for the three tasks described above. Table 14.3 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 offline applications. In all three cases the online model performs significantly worse than its offline correlate, but also significantly better than the baseline (Wilcoxon, $p < 0.05$).

	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%	

Table 14.3: Error rate of the SVM classifier on online and offline tasks.

Table 14.3 also shows the error rates of online and offline classifiers trained using solely text-based features — i.e., only features of lexical (`LX`) or discourse (`DS`) types. Text-based models simulate the conditions of TTS systems: After determining the discourse/pragmatic function of ACWs, TTS systems may produce such words with appropriate acoustic/prosodic parameters, such as those explored in Chapter 13. For the ACW classification task, some TTS systems may only have information up to the current utterance available (online setting), while others may have the complete text available (offline setting). Our online and offline text-based models perform significantly worse than the corresponding models that use the whole feature set, but still outperform the baseline models in all cases (Wilcoxon, $p < 0.05$). Finally, the offline text-based models also outperform their online correlates in all three tasks (Wilcoxon, $p < 0.05$).

14.3 Features extracted solely from the target word

In the descriptive statistics discussed in Chapter 13, we reported evidence of strong contextual differences across the various functions of ACWs, such as the position of the word in its conversational turn, or whether the word is preceded and/or followed by silence. Based on that finding, we posed the question of whether such differences would be sufficient for the listener to disambiguate the word meaning, thus occluding the described variation along several acoustic/prosodic features of ACWs such as word final intonation and word mean intensity. We address this empirical question fully in the perception study discussed in Chapter 15. In this section, we report on an experiment aimed at answering the same questions, but for ML classifiers rather than humans: Are features extracted solely from the target ACW enough for predicting the function of ACWs, or do contextual features improve the classification performance? While the answer to this question may not directly indicate which cues humans actually perceive and/or use to disambiguate, it will tell us more about the existence, location and usefulness of automatically computable features for ML classification of ACWs.

For each of the three tasks — classification of all words into all functions, detection of a discourse boundary, and detection of an acknowledgment function, we train an SVM classifier considering only features extracted from the target word. These features are marked in Tables 12.5 and 12.6 with the letter $\mathcal{W}$, and comprise the word's lexical identity, part-of-speech tag, duration, and a number of acoustic and phonetic features. Table 14.4 contrasts the error rate of this classifier (which we call the WORD-ONLY model) to that of the full model and the word-based baseline. As in the previous experiments, the full model employs the complete feature set, extracted from the whole conversation.

Feature Set	All Functions	Disc. Boundary	Acknowledgment
Full model (`LX DS TM AC PH`)	14.3%	6.9%	4.5%
Word-only model	23.6%	14.4%	15.0%
Word-based baseline	27.7%	18.6%	15.3%

Table 14.4: Error rate of the SVM classifier trained on features extracted only from the target word.

On the one hand, the word-only model significantly outperforms the baseline in the general and discourse boundary tasks (Wilcoxon, $p < 0.05$), indicating that the target ACW itself contains a substantial amount of information useful to those two tasks, and that such information is at least partially captured by the word-only features and exploited by the SVM classifier. On the other hand, the word-only model performs significantly worse than the full model on the three tasks (Wilcoxon, $p < 0.05$). This means that the word-only features are insufficient for the SVM classifier to reach the accuracy level of the full model, and that our contextual features significantly reduce the classification error rate.

14.4 Backchannel detection

The correct identification of backchannels is a desirable capability for speech processing systems, as it would allow to distinguish between two opposite intentions of speakers' contributions: that of taking the conversational floor, and that of encouraging the interlocutor to continue talking.

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

Online applications such as IVR systems need to classify every new speaker contribution immediately after it has been uttered, and without access to any subsequent context. The Games Corpus contains approximately 6700 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 they are short, isolated utterances, and consist normally in just one ACW. Of the 6700 candidate turns in the corpus, only 2351 (35%) begin with an isolated ACW, including 753 of the 757 backchannels in the corpus.[3] At this point, we explore using a ML classifier to distinguish the backchannels from the other functions. The same word-based majority baseline described above achieves an error rate of 11.56%. An SVM classifier trained on features extracted from up to the current IPU (to simulate the online condition of an IVR 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 acknowledgments in many cases, leading to an increase in the error rate. (Recall, from the statistical analyses in the previous chapter, the acoustic/prosodic similarities of these two functions for *mm-hm* and *uh-huh*, for example.) We conclude that further research is needed to develop novel approaches to this crucial problem of IVR systems.

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

14.5 Discussion

In this study of automatic classification of ACWs we have shown that, for spoken task-oriented dialogue, the simple discourse/sentential distinction is insufficient. In consequence, we have defined two new classification tasks (the detection of an acknowledgment function, and the detection of a discourse segment boundary function), besides the general task of classifying any ACW into any function. We have shown that 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. Additionally, we have experimented with several combinations of feature sets, in an attempt to simulate the settings of real applications. All these results are intended to aid future researchers and developers in building effective classifiers of the discourse/pragmatic function of ACWs.

Finally, we have shown results suggesting that the predictive power of contextual information is much stronger than that of the acoustic, prosodic and phonetic characteristics of the target word itself. Again, this finding raises the question of whether context alone is sufficient for disambiguation purposes. The following chapter describes a perception study aimed at shedding light on this issue, investigating how humans' interpretations of ACWs varies when some or no context is available.

Chapter 15

A Perception Study of *Okay*

In this chapter, we address the question of how hearers disambiguate the discourse pragmatic function of ACWs. Our main goal is to determine the role of discourse context in this process: Can listeners classify ACW tokens reliably from listening to the word alone, or do they require contextual information? Additionally, we look for acoustic, prosodic and phonetic features potentially used by listeners in the disambiguation process.

Below we describe a perception experiment in which listeners are presented with a number of spoken productions of *okay*, both in isolation and in context, and asked to select the function of each token. Subsequently, we examine how the listeners' classifications vary across conditions, and look for acoustic, prosodic and phonetic correlates of these classifications.

15.1 Experiment design and implementation

For our perception study we choose the most frequent affirmative cue word in the Games Corpus, *okay*, for two reasons. First, as shown in Chapter 13, *okay* is the ACW that presents the highest degree of variation along the studied prosodic/acoustic features, as well as the most heavily overloaded ACW, with instances conveying each of the ten identified discourse/pragmatic functions. Second, the over 2200 instances of *okay* in the corpus allow for a balanced experimental design, with tokens uttered by several different speakers.

We choose the three most frequent simple functions of *okay*:[1] Acknowledgment/agreement (**Ack**), Backchannel (**BC**), and Cue beginning discourse segment (**CBeg**). Additionally, we choose tokens with three different degrees of potential ambiguity, based on the agreement achieved by the labelers that annotated all ACWs in the corpus. Unanimous tokens are those that were assigned the same function by the three labelers; majority tokens were assigned the same function by exactly two of the three labelers; no-agreement tokens were assigned a particular function by exactly one labeler, and two other functions by the remaining two labelers.

To obtain a good coverage of the three functions and the three degrees of ambiguity, we identify 9 categories of *okay* tokens to include in the experiment: 3 functions (**Ack**, **BC**, **CBeg**) × 3 levels of labeler agreement (unanimous, majority, no-agreement). To control for speaker variation in the stimuli, we select tokens from 6 speakers (3 female, 3 male) who produced at least one token for each of the 9 conditions, leaving a total of 54 tokens.

We prepare two versions of each token to investigate whether subjects' classifications of *okay* are dependent upon contextual information or not. The isolated versions consist of only the word *okay* extracted from the waveform. For the contextualized versions, we extract two full speaker turns for each *okay*,[2] including the full turn containing the target *okay* plus the full turn from the previous speaker. In the following three sample contexts, pauses are indicated with '#', and the target *okay*s are underlined:

A: *yeah # um there's like there's some space there's*
B: *okay # I think I got it*

A: *but it's gonna be below the onion*
B: *okay*

[1] Even though Pivot ending (**PEnd**) *okay*s were more frequent than **BC** *okay*s, we choose to avoid compound functions like the former (a combination of **Ack** and **CBeg**), using only simple functions instead.

[2] Recall from Chapter 2 that we define a turn as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor. An inter-pausal unit (IPU) is defined as a maximal sequence of words surrounded by silence longer than 50 ms.

A: *okay # alright # I'll try it # okay*

B: *okay the owl is blinking*

We present the isolated *okay* tokens in single-channel audio files; the contextualized *okay* tokens are formatted so that each speaker is presented to subjects on a different channel, with the speaker uttering the target *okay* consistently on the same channel.

The perception study is divided in two parts. In the first part (hereafter, the ISOLATED CONDITION), subjects are presented with the 54 isolated *okay* tokens, in a different random order for each subject. They are given a forced choice task to classify tokens as **Ack**, **BC**, or **CBeg**, with the corresponding function labels also presented in a random order for each token. In the second part (the CONTEXTUALIZED CONDITION), the same subjects are given 54 contextualized tokens, presented in a different random order, and asked to make the same choice.

We recruited 20 (paid) subjects for the study, 10 female and 10 male, all between the ages of 20 and 60. All subjects reported no hearing problems and were native speakers of Standard American English, except for one subject who reported being a native speaker of Jamaican English. Subjects performed the study in a quiet lab using headphones to listen to the tokens and indicating their classification decisions in a GUI interface on a lab workstation. They were given instructions on how to use the interface before each of the two parts of the study. The full instructions, as well as sample screens of the interface of the study, are given in Appendix G.

During the study, subjects could listen to the sound files as many times as they wished but were instructed not to be concerned with answering the questions "correctly", but to answer with their immediate response if possible. They were allowed though to change their selection as many times as they liked before moving to the next screen. In the contextualized condition, they were also shown an orthographic transcription of a small part of the contextualized token, aimed only at helping subjects identify the target *okay*. The mean duration of the first part of the study was 25 minutes, and of the second part, 27 minutes.

15.2 Subject ratings

The distribution of class labels in each experimental condition is shown in Table 15.1. While this distribution roughly mirrors our selection of equal numbers of tokens from each previously-labeled class, in both parts of the study more tokens were labeled as **Ack** (acknowledgment/agreement) than as **BC** (backchannel) or **CBeg** (cue to topic beginning). This supports the hypothesis that acknowledgment/agreement acts as the default interpretation of *okay*.

	Isolated		Contextualized	
Ack	426	(39%)	452	(42%)
BC	324	(30%)	306	(28%)
CBeg	330	(31%)	322	(30%)
Total	1080	(100%)	1080	(100%)

Table 15.1: Distribution of label classes in each study condition.

Next we examine inter-subject agreement using Fleiss' κ measure for multiple raters.[3] Table 15.2 shows Fleiss' κ calculated for each individual function label vs. the other two labels, and for all three labels together, in both study conditions. While there is very little overall agreement among subjects on how to classify tokens in the isolated condition, agreement is higher in the contextualized condition, reaching a moderate agreement for class

	Isolated	Contextualized
Ack vs. Rest	0.089	0.227
BC vs. Rest	0.118	0.164
CBeg vs. Rest	0.157	0.497
All	0.120	0.293

Table 15.2: Fleiss' κ for each label class in each study condition.

[3] 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.

CBeg (κ score of 0.497). This suggests that context helps distinguish the cue beginning function of *okay* more than the other two functions.

Recall from Section 15.1 that the *okay* tokens were chosen in equal numbers from three classes (unanimous, majority, and no-agreement) according to the level of agreement of our three original labelers, who had the full dialogue context available for making their decisions. Table 15.3 shows Fleiss' κ measure now grouped by level of agreement, again presented for each context condition. We see here that the inter-subject agreement also mirrors the agreement of the three original labelers. In both study conditions, tokens on which the original labelers agreed also had the highest κ scores, followed by tokens in the majority and no-agreement classes, in that order.

	Isolated	Contextualized	Original labelers
No-agreement	0.085	0.104	–
Majority	0.092	0.299	–
Unanimous	0.158	0.452	–
All	0.120	0.293	0.312

Table 15.3: Fleiss' κ in the two study conditions, grouped by level of agreement of the three original labelers.

The overall κ is small at 0.120 for the isolated condition, and fair at 0.293 for the contextualized condition. The three original labelers also achieved fair agreement at 0.312.[4] The similarity between the latter two κ scores suggests that the full context available to the original labelers and the limited context presented to the participants of the perception study offered comparable amounts of information to disambiguate between the three functions. On the other hand, the unavailability of any context clearly affected subjects' decisions. We conclude that context is of considerable importance in the interpretation of the word *okay*, although even a relatively limited context appears to suffice.

[4] For the calculation of this κ, we consider four label classes: **Ack**, **BC**, **CBeg**, and a fourth class 'other' that comprises the remaining seven discourse/pragmatic functions of ACWs. Since the existence of a fourth category may have an effect on the measurement of inter-subject agreement, these κ scores should be compared with caution.

15.3 Cues to interpretation

In this section we perform a series of statistical tests aimed at finding correlations between the discourse/pragmatic function perceived by subjects in either study condition, and a number of acoustic, prosodic, phonetic and contextual features.

For each target *okay*, we examine its duration and its maximum, mean and minimum pitch and intensity (all raw and speaker-normalized), and the slope of the pitch, intensity and stylized pitch tracks, calculated over the whole word and over its last final portion. We also consider nominal features extracted from the ToBI transcriptions of each token, such as pitch accent, phrase accent and boundary tone. All of these features are described in detail in Section 12.2 (pages 98 and following).

Additionally, two expert annotators transcribed together the phonetic realization of each token of *okay* using the International Phonetic Alphabet (IPA) conventions. In the tokens used in this experiment we find the following variations for the three phonemes (/oʊ/, /k/, /eɪ/) of *okay*:

- /oʊ/: [], [ɑ], [ɐ], [ɔ], [ɔʊ], [m], [ŋ], [ə], [əʊ].
- /k/: [ɣ], [k], [kx], [q], [x].
- /eɪ/: [e], [eɪ], [ɛ], [eə].

From the phonetic transcriptions we calculate the duration of each phone and of the velar closure, whether the target *okay* is at least partially whispered or not, and whether there is glottalization in the target *okay*.

First, for each numerical feature we compute Pearson's correlation coefficient to look for an association between the feature and the proportion of subjects that chose each label. (For example, if a particular *okay* was labeled as **Ack** by 5 subjects, as **BC** by 3, and as **CBeg** by 12, then its corresponding proportions are 5/20, 3/20 and 12/20, or 0.25, 0.15 and 0.6.) Subsequently, we compute two-sided t-tests to assess the significance of the correlations. Table 15.4 shows the significant results (two-sided t-tests, $p < 0.05$) for the isolated and contextualized conditions, respectively.

In the isolated condition, we observe that subjects tended to classify as **Ack** tokens of *okay* which had a longer realization of the /k/ phoneme; as **BC**, those with a lower intensity,

Acknowledgment/agreement	*r*
duration of realization of /k/	−0.299

Backchannel	*r*
stylized pitch slope, 2nd half 2nd syl.	0.752
pitch slope over 2nd half 2nd syl.	0.409
speaker-norm. maximum intensity	−0.372
pitch slope over last 80 ms	0.349
speaker-norm. mean intensity	−0.327
duration of realization of /eɪ/	0.278
word duration	0.277

Cue to disc. segment beginning	*r*
stylized pitch slope over whole word	−0.380
pitch slope over whole word	−0.342
pitch slope over 2nd half 2nd syllable	−0.319

Acknowledgment/agreement	*r*
latency of *Spkr A* before *Spkr B*'s turn	−0.528
duration of silence by *Spkr B* before *okay*	−0.404
number of words by *Spkr B* after *okay*	−0.277

Backchannel	*r*
pitch slope, 2nd half of 2nd syllable	0.520
pitch slope, last 80 ms	0.455
number of words by *Spkr A* before *okay*	0.451
number of words by *Spkr B* after *okay*	−0.433
duration of speech by *Spkr B* after *okay*	−0.413
latency between the two turns	−0.385
intensity slope over 2nd syllable	−0.279

Cue to disc. segment beginning	*r*
latency of *Spkr A* before *Spkr B*'s turn	0.645
number of words by *Spkr B* after *okay*	0.481
number of words by *Spkr A* before *okay*	−0.426
pitch slope over 2nd half of 2nd syllable	−0.385
pitch slope over last 80 ms	−0.377
duration of speech by *Spkr B* after *okay*	0.338

Table 15.4: Features significantly correlated to the proportion of votes for each label. Isolated (left) and contextualized conditions.

a longer duration, a longer realization of the /eɪ/ phoneme, and a final rising pitch; and as **CBeg**, those ending in a falling pitch. In the contextualized condition, we find very different correlations, nearly all of them involving contextual features, such as the latency before *Speaker B*'s turn, or the number of words by each speaker before and after the target *okay*. Notably, only one of the features showing strong correlations in the isolated condition presents the same strong correlation in the contextualized condition: word final pitch slope. In both conditions subjects tended to label tokens with a final rising pitch contour as **BC**, and tokens with a final falling pitch contour as **CBeg**.

We conduct next a series of two-sided Fisher's exact tests to find correlations between subjects' classification of *okay* and nominal features related to the phonetic and prosodic transcriptions of the tokens. We first divide the 54 tokens of each condition into three groups, according to the label assigned by a plurality of subjects,[5] and explore whether these three groups correlate with our nominal features. We find a significant association between the realization of the /oʊ/ phoneme and the perceived discourse/pragmatic function of *okay* in the isolated condition ($p < 0.005$). Table 15.5 shows that, in particular, [m] seems to be the preferred realization for **BC** *okay*s, [ə] for **Ack**, and [ɔʊ] and [ɔ] for **Ack** and **CBeg**. Notably, we do not find such significant associations in the contextualized condition. However, we do find significant correlations in both conditions between *okay* classifications and the type of phrase accent and boundary tone on the target word (Fisher's Exact Test, $p < 0.05$ for the isolated condition, $p < 0.005$ for the contextualized condition). Table 15.6 shows that L-L% tends to be associated with **Ack** and **CBeg**, H-H% with **BC**, and L-H% with **Ack** and **BC**. In this case, such correlations are present in the isolated condition, and enhanced in the contextualized condition.

	?	[ɑ]	[ɐ]	[ɔʊ]	[ɔ]	[ŋ]	[əʊ]	[ə]	[]	[m]
Ack	0	0	5	6	4	0	0	8	0	0
BC	2	0	4	1	0	1	0	1	1	5
CBeg	1	1	2	3	4	0	1	3	0	0

Table 15.5: Realization of the /oʊ/ phoneme, grouped by subject plurality label. Isolated condition only.

Summing up, for tokens of *okay* listened in isolation, with only acoustic, prosodic and phonetic properties available to the subjects, a few features seem to strongly correlate with the perception of word function. For example, maximum intensity, word duration, and realizing the /oʊ/ phoneme as [m] tend to be associated with the backchannel function, while the duration of the realization of the /k/ phoneme, and realizing the /oʊ/ phoneme as [ə] tend to be associated with the acknowledgment/agreement function.

[5] A *plurality* is also known as a *simple majority*: the candidate who gets more votes than any other candidate is the winner.

		H-H%	[!]H-L%	L-H%	L-L%	other
Isolated	**Ack**	0	2	4	8	9
	BC	3	3	1	5	3
	CBeg	1	1	0	8	5
Contextualized	**Ack**	0	2	3	10	10
	BC	4	3	2	1	2
	CBeg	0	1	0	10	5

Table 15.6: Phrase accent and boundary tone, grouped by subject plurality label.

In the second part of the study, for the contextualized version of the same tokens of *okay*, most of the strong correlations of perceived word function with acoustic, prosodic and phonetic features are replaced with correlations with contextual features, such as latency and turn duration. In other words, these results suggest that contextual features might override the effect of most other features of *okay*. There is nonetheless one notable exception: word final intonation. Captured by the pitch slope and the ToBI labels for phrase accent and boundary tone, this feature seems to play a central role in the interpretation of both isolated and contextualized *okay*s.

15.4 Discussion

In this perception study, we have presented evidence of differences in the interpretation of the discourse/pragmatic function of isolated and contextualized instances of *okay* by human listeners. We have shown that word final intonation strongly correlates with the subjects' classification of *okay*s in both conditions. Additionally, the higher degree of inter-subject agreement in the contextualized condition, along with the strong correlations found for contextualized features, suggests that context, when available, plays a central role in the disambiguation of *okay*. (Note, however, that further research is needed in order to assess whether these features are, indeed, perceptually important, both individually and combined.)

We have also presented results suggesting that acknowledgment/agreement acts as a default function for both isolated and contextualized *okay*s. Furthermore, while this function remains confusable with the backchannel function in both conditions, the availability of some context helps in distinguishing those two from the **CBeg** function.

Chapter 16

Entrainment of ACW Usage

This final chapter describes a preliminary study conducted in collaboration with Prof. Ani Nenkova (Dept. of Computer and Information Science, University of Pennsylvania), that investigates how speakers tend to adapt their usage of ACWs and other high-frequency words to match their interlocutors', and the relation of this phenomenon to task success and dialogue coordination (Nenkova et al., 2008). This study incorporates a new dimension to the analysis of ACWs and turn-taking in dialogue, by portraying each speaker not as static and behaving always in the same manner, but rather as constantly changing his/her speech according to the environment. Modeling whether and how this phenomenon takes place and identifying its potential implications will help improve our understanding of variation in human speech, and should aid IVR systems in providing a more natural user experience.

16.1 Previous research on speaker entrainment

When people engage in conversation, they adapt the way they speak to their conversational partner. For example, they often adopt a certain way of describing something based upon the way their conversational partner describes it, negotiating a common description, particularly for items that may be unfamiliar to them (Brennan, 1996). They also alter their amplitude, if the person they are speaking with speaks louder than they do (Coulston et al., 2002; Ward and Litman, 2007), or reuse syntactic constructions employed earlier in the conversation (Reitter et al., 2006). This phenomenon is known in the literature as

ENTRAINMENT.

There is a considerable body of literature which posits that entrainment may be crucial to human perception of dialogue success and overall quality, as well as to participants' evaluation of their conversational partners. Pickering and Garrod (2004) propose that the automatic alignment at many levels of linguistic representation (lexical, syntactic and semantic) is key for both production and comprehension in dialogue, and facilitates interaction. Goleman (2006) also claims that a key to successful communication is human ability to synchronize their communicative behavior with that of their conversational partner. For example, in laboratory studies of non-verbal entrainment (mimicry of mannerisms and facial expressions between subjects and a confederate), Chartrand and Bargh (1999) find not only that subjects display a strong unintentional entrainment, but also that greater entrainment/mimicry leads subjects to feel that they like the confederate more and that the overall interaction is progressing more smoothly. People who have a high inclination for empathy (understanding the point of view of the other) entrains to a greater extent than others. Reitter and Moore (2007) also find that degree of entrainment in lexical and syntactic repetitions that take place in only the first five minutes of each dialogue in the HCRC Map Task Corpus significantly predicts task success.

In the following sections, we examine a novel dimension of entrainment between conversational partners: the use of high-frequency words, such as affirmative cue words, or the most frequent words in a dialogue. We discuss experiments on the association of entrainment in the usage of such words with task success and turn-taking behavior.

16.2 Measures of entrainment

We define two measures of entrainment of the usage of a word class c. Both measures capture in different ways the differences in usage frequency of a word class c by the two speakers S_1 and S_2. The first one is the negated sum, for each word $w \in c$, of the absolute difference between the fraction of times w is used by S_1 and S_2. More formally,

$$ENTR_1(c) = -\sum_{w \in c} \left| \frac{count_{S_1}(w)}{ALL_{S_1}} - \frac{count_{S_2}(w)}{ALL_{S_2}} \right|$$

Here, ALL_{S_i} is the number of all words uttered by speaker S_i in the given conversation, and $count_{S_i}(w)$ is the number of times S_i used word w. $ENTR_1$ ranges from 0 to $-\infty$, with 0 meaning perfect match on usage of lexical items in class c. Our second measure of entrainment is defined as

$$ENTR_2(c) = -\frac{\sum\limits_{w \in c} |count_{S_1}(w) - count_{S_2}(w)|}{\sum\limits_{w \in c} (count_{S_1}(w) + count_{S_2}(w))}$$

The entrainment score defined in this way ranges from 0 to -1, with 0 meaning perfect match on lexical usage and -1 meaning perfect mismatch.

16.3 Entrainment and task success

In the Games Corpus, we hypothesize that the game score achieved by the participants is a good measure of the effectiveness of the dialogue. To determine the extent to which task success is related to the degree of entrainment in high-frequency word usage, we examine the dialogues in the Games Corpus. We compute the correlation coefficient between the game score (normalized by the highest achieved score for the game type) and our two measures of entrainment between the speakers (S_1 and S_2) in four high-frequency word classes:

- **ACW:** Affirmative cue words.
- **FP:** Filled pauses: *uh*, *um*, *mm*. The corpus contains 1845 instances of filled pauses (2.5% of all tokens).
- **25MF-G:** The 25 most frequent words in the current game.
- **25MF-C:** The 25 most frequent words over the entire corpus: *the*, *a*, *okay*, *and*, *of*, *I*, *on*, *right*, *is*, *it*, *that*, *have*, *yeah*, *like*, *in*, *left*, *it's*, *uh*, *so*, *top*, *um*, *bottom*, *with*, *you*, *to*.

The correlations between the normalized game score and these measures of entrainment are shown in Table 16.1: r is Pearson's correlation coefficient; p is the significance of the correlation estimated with two-sided t-tests. $ENTR_1$ for the 25 most frequent words, both corpus-wide and game-specific, is highly and significantly correlated with task success, with stronger results for game-specific words. For the filled pauses class, there is essentially no correlation between entrainment and task success, while for affirmative cue words there is association only under the $ENTR_2$ definition of entrainment. The difference in results between $ENTR_1$ and $ENTR_2$ suggests that the two measures of entrainment capture different aspects of dialogue coordination. Exploring novel formulations of entrainment deserves future attention.

	$ENTR_1$		$ENTR_2$	
Word class	r	p	r	p
ACW	0.230	0.116	**0.372**	**0.009**
FP	−0.080	0.591	−0.007	0.964
25MF-G	**0.376**	**0.008**	0.260	0.074
25MF-C	**0.341**	**0.018**	0.187	0.202

Table 16.1: Correlations of entrainment and game score.

16.4 Entrainment and dialogue coordination

The coordination of turn-taking in dialogue is especially important for successful interaction. Speech overlaps (**O**), might indicate a lively, highly coordinated conversation, with participants anticipating the end of their interlocutor's speaking turn. Smooth switches (**S**) with no overlapping speech are also characteristic of good coordination, in cases where these are not accompanied by long pauses between turns. On the other hand, interruptions (**I**) and long inter-turn latency — long simultaneous pauses by the speakers — are generally perceived as a sign of poorly coordinated dialogues.

To determine the relationship between entrainment and dialogue coordination, we examine the correlation between entrainment types and the proportion of interruptions, smooth switches and overlaps, in the Objects portion of the Games Corpus. We also look at the correlation of entrainment with mean latency in each dialogue. Table 16.2 summarizes the major findings.

		r	p
$ENTR_1$(25MF-C)	I	$\mathbf{-0.612}$	**0.035**
$ENTR_1$(25MF-G)	I	-0.514	0.087
$ENTR_1$(ACW)	O	**0.636**	**0.026**
$ENTR_2$(ACW)	O	**0.606**	**0.037**
$ENTR_1$(FP)	O	**0.750**	**0.005**
$ENTR_2$(25MF-G)	O	**0.605**	**0.037**
$ENTR_2$(25MF-G)	S	$\mathbf{-0.663}$	**0.019**
$ENTR_2$(ACW)	*lat*	$\mathbf{-0.757}$	**0.004**
$ENTR_2$(25MF-G)	*lat*	-0.523	0.081

Table 16.2: Correlations of entrainment with proportion of smooth switches, overlaps, interruptions, and mean latency (*lat*).

Two measures that significantly correlate with task success — $ENTR_1$(25MF-C) and $ENTR_1$(25MF-G) — also correlate negatively with the proportion of interruptions in the dialogue. Additionally, overlaps are strongly associated with entrainment in usage of ACWs, filled pauses and game-specific most frequent words. Long latency is negatively associated with entrainment in affirmative cue words and game-specific most frequent words.

Unexpectedly, smooth switches correlate negatively with entrainment in game-specific most frequent words. This result might be confounded by the presence of long latencies in some switches. While smooth switches are desirable, especially in IVR systems, long latencies between turns can indicate lack of coordination.

Overall, the higher the presence of speaker entrainment, the more engaged the participants and the better coordination there is between them, with shorter latencies, more overlaps and fewer interruptions.

16.5 Discussion

In this section we have presented a preliminary corpus study relating dialogue success and coordination with speaker entrainment on common words: affirmative cue words, filled pauses, and most frequent words in the corpus and in a dialogue. Our results suggest that entrainment over classes of frequent words strongly correlates with task success, and with engaged and coordinated turn-taking behavior.

These findings open new topics for future research, such as experimenting with novel ways of quantifying the degree of entrainment between speakers, and also with other word classes. Most importantly, future research should assess the causal relations holding between the associations described in this study. If speaker entrainment is found to **cause** task success and/or dialogue coordination, then IVR system designers could try to adapt the system's usage of high-frequency words to match the user's, aiming at improving the performance and usability of such systems. On the other hand, if entrainment is a **consequence** of task success and/or dialogue coordination, then it would constitute a valuable evaluation metric for IVR systems: measuring the degree to which the user entrains with the system could be used to estimate the performance and usability of such systems.

Chapter 17

Conclusions and Future Work

The studies of ACWs presented in this thesis provide evidence of several differences in the production of the various discourse/pragmatic functions of ACWs. We find marked contrasts in acoustic/prosodic features, such as word final intonation and word intensity, and also in contextual features, such as the position of the word in its conversational turn, or whether the word is preceded and/or followed by silence. Furthermore, in a perception study of the uses of the word *okay*, we find that such contextual differences play a central role in the disambiguation of its function by human listeners.

Our study of automatic classification of ACWs shows that the simple discourse/sentential distinction commonly used for other cue phrases is insufficient in this case. In consequence, we propose two new classification tasks (the detection of an acknowledgment function, and the detection of a discourse segment boundary), besides the general task of classifying any ACW into any function. SVM models based on lexical, discourse, timing and acoustic features approach the performance of trained human labelers in all tasks. Additionally, we have experimented with several combinations of feature sets to simulate the settings of real applications, in an attempt to aid future researchers and developers in building effective classifiers of the discourse/pragmatic function of ACWs.

Finally, we have presented a preliminary study of speaker entrainment on the usage of ACWs, filled pauses, and other classes of frequent words. Our results suggest that such entrainment strongly correlates with task success, and with engaged and coordinated turn-taking behavior.

We propose two possible directions for future research. First, the results obtained by our machine learning classifiers in the task of automatic detection of backchannels failed to significantly outperform the majority-class baseline. This is a crucial task for IVR systems, which need the capability to distinguish users' backchannels from turn-taking attempts. Therefore, future research should look into novel approaches to this problem.

A second direction is related to speaker entrainment. Our promising preliminary results encourage future research to look into new ways of capturing the degree to which speakers adapt their speech to resemble their interlocutors'. Additionally, establishing the causal relations of speaker entrainment with task success and/or dialogue coordination could provide powerful tools to IVR system designers, for either improving or evaluating the performance and usability of such systems.

Part IV

Conclusions

Chapter 18

Conclusions

In this thesis we described the results of a series of studies aimed at advancing our understanding of various aspects of spoken dialogue. We collected and annotated a large corpus of spontaneous task-oriented dyadic conversation in Standard American English, on which we studied turn-taking behavior and the usage of heavily overloaded cue words such as *okay* or *alright*. Our hope is that these findings will help improve the quality and usability of IVR systems and other spoken language processing applications.

18.1 The Columbia Games Corpus

The first main contribution of this work is the Columbia Games Corpus, which comprises twelve spontaneous task-oriented dyadic conversations in Standard American English between thirteen people, totaling nine hours of dialogue. The collection and annotation of this corpus was described in Part I of this thesis. In addition to time-aligned orthographic transcriptions, it contains manual annotations of diverse phenomena, including (1) the discourse/pragmatic function of affirmative cue words, (2) the category of turn-taking exchanges between the conversation participants, (3) intonational patterns and other aspects of the prosody (using the ToBI framework), (4) non-word vocalizations such as laughs, coughs and breaths, and (5) the form and function of questions. This corpus represents a valuable data set for future research in spoken dialogue.

18.2 Turn-taking

The second main contribution of this work is a large-corpus-based systematic study of turn-taking behavior in Standard American English dialogue. The motivation for this study consisted in developing a framework that would help improve the turn-taking decisions made by state-of-the-art IVR systems. The results were presented in Part II of this thesis.

18.2.1 Summary of findings and novel contributions

We identified and described seven turn-yielding cues — distinct events that strongly correlate with the imminent occurrence of a conversational turn boundary: (1) a falling or high-rising final intonation; (2) a reduced final lengthening; (3) a low intensity level; (4) a low pitch level; (5) a point of textual completion; (6) a high value of three voice quality features: jitter, shimmer, and noise-to-harmonics ratio; and (7) a long duration of the final inter-pausal unit. We showed that these cues combine together to form complex signals, such that the likelihood of a turn-taking attempt by the interlocutor increases almost linearly with respect to the number of cues conjointly displayed by the speaker.

To our knowledge, this is the first study to systematically examine all of these turn-yielding cues, both individually and combined together to form complex signals. An important characteristic of our results is that they were drawn from a large corpus of conversations between thirteen different people. Most previous studies of turn-yielding cues, by contrast, examine a smaller number of conversations — typically only two or three. Thus, our findings offer statistically robust evidence of the existence of these cues and support their generalizability to larger speaker populations.

Additionally, we provided a computational definition of the presence or absence of each individual cue, in contrast with the perceptual or impressionistic definitions used in most previous studies of turn-yielding cues. Using automatically computed cues eliminates a source of subjectivity from human annotators and makes the results more straightforward to incorporate into speech processing systems. In particular, we introduced a novel procedure for predicting the textual completion of speech utterances. Our SVM-based classifier, trained on lexical and syntactic features extracted from a small manually labeled data set, significantly outperformed the majority-class baseline and approached human agreement. Given the unambiguous evidence presented in this and previous studies signaling textual completion as one of the most prominent turn-yielding cues, our procedure represents an important contribution in itself to the advancement of turn-taking technologies.

Our results for the final intonation and textual completion cues, the ones most frequently examined in previous studies, are consistent with the literature: Turn switches tend to follow textually complete speech segments with falling or high-rising final intonation. For the cues related to a drop in intensity, a drop in pitch, and a longer IPU duration, our results are also consistent with the hypotheses presented in the literature, although those cues received much less attention in previous studies. In addition to providing solid evidence validating the existence of those five turn-yielding cues, we described two new cues which have **not** been previously examined for English dialogues: a high level of jitter, shimmer and noise-to-harmonics ratio — acoustic features associated with the perception of voice-quality; and a reduction or attenuation of the final lengthening that typically precedes prosodic boundaries.

We also described six backchannel-inviting cues — events in the current speaker's speech that may invite the listener to produce a short utterance conveying continued attention: (1) a rising final intonation; (2) a high intensity level; (3) a high pitch level; (4) a final POS bigram equal to 'DT NN', 'JJ NN' or 'NN NN'; (5) a low value of noise-to-harmonics ratio; and (6) a long duration of the final inter-pausal unit. We showed that the likelihood of occurrence of a backchannel from the interlocutor increases in a quadratic fashion with the number of cues conjointly displayed by the speaker. The whole of our study of backchannel-inviting cues represents a novel contribution to the field.

18.2.2 Impact

The purpose of the study of turn-taking behavior presented in this thesis was to provide a framework that would help improve several decisions of IVR systems, which should, in turn, enhance the usability and naturalness of such systems. If the system intends to keep the conversational floor, it should formulate its output in a way that includes as few as possible of the turn-yielding cues we have found to be important, a behavior that will decrease the likelihood that the user will take the turn. For example, the output of the IVR system's speech synthesis component should end its IPUs in plateau intonation, with high intensity and pitch levels, and leaving utterances textually incomplete (e.g., ending in expressions such as *and* or *also*). If the system wants to yield the floor to the user, it should formulate its output to include as many as possible of the turn-yielding cues we have found to be significant, which will more likely lead to a turn-taking attempt by the user. For example, the system's final IPU should be textually complete, have low intensity and pitch levels, and end in either falling or high-rising intonation (depending on whether the system's message is a statement or a direct question).

From the results presented in this thesis, it should also be possible to improve the detection of turn boundaries in the user's speech. Even though the difficulty of estimating each turn-yielding cue will depend on the individual system implementation, a high-level description of the turn-taking decision procedure could be as follows: At every silence longer than a threshold (e.g., 50 milliseconds), the system estimates the presence of as many cues as possible over the user's final IPU. If the number of detected cues is higher than some predefined threshold, the system may attempt to take the turn immediately; otherwise, it may continue waiting. Note that some of the mentioned cues, such as voice quality features or pitch and intensity levels, may be precomputed at regular intervals while the user is still speaking, thus reducing the processing time required at each silence.

Finally, IVR systems could benefit from our results on backchannel-inviting cues to refine additional turn-taking decisions. For example, our results suggest how the system should formulate its output to give the user an opportunity to utter a backchannel (as a way of ensuring that the user is paying attention), or how to determine when the system should produce a backchannel as positive feedback to the user. The implementation of these decisions should be analogous to the turn-yielding decisions described above.

18.2.3 Future work

Our study of turn-taking behavior opens numerous directions for future research:

- Future studies should seek novel turn-yielding and backchannel-inviting cues, aiming at enriching our current models and providing IVR systems with further information to make more informed decisions. In particular, given our clear findings for jitter, shimmer and noise-to-harmonics ratio, additional voice quality features appear to be a promising option to explore, including relative average perturbation (RAP), soft phonation index (SPI), and amplitude perturbation quotient (APQ).

- The novel procedure presented in this thesis for the automatic prediction of textual completion presents some margin for improvement. Our SVM-based classifier achieved an accuracy of 80%, while human agreement was 90.8%. New approaches could incorporate features capturing information from the previous turn by the other speaker, which was available to the human labelers but not to the machine learning classifier. Also, the sequential nature of this classification task might be better exploited by more advanced graphical learning algorithms, such as Hidden Markov Models and Conditional Random Fields.

- We presented two simple procedures (one discrete, the other continuous) for determining the presence or absence of numeric turn-yielding cues. These procedures are based on whether the values of two or three features are closer to the mean before holds (**H**) or the mean before smooth switches (**S**). This procedure could be refined, for example, by fitting a Gaussian curve to the two groups (**H** and **S**) and subsequently determining which model explains the observed values better: If the model for **S** is better suited, the cue is present; otherwise, it is absent. (The same consideration applies to the procedure for determining backchannel-inviting cues.)

- Our study implicitly assumed that all cues are equally important, contributing with either 0 or 1 to the total cue count. Future research should explore the assignment of numeric weights to the different cues, depending on their relative importance: e.g., the textual completion cue should be assigned a high weight, since, as we showed, this cue seems to work almost as a necessary condition for smooth switches. These weights could also reflect the reliability of the procedures for automatically computing the cues: e.g., the pitch slope features used for estimating the final intonation are often strongly affected by pitch tracking errors, a good reason for decreasing the relative weight of the final intonation cue.

- An examination of instances of overlapping speech in the corpus yielded preliminary results suggesting that both types of cues — turn-yielding and backchannel-inviting — may be present **before** the final part of conversational turns. Rather, they seem to extend further back in the turn. This suggests that future work might examine, for example, whether these cues extend over a longer portion of the turn, starting low at the turn onset, to gradually increase as turns approach potential transition-relevance places.

- The turn-taking labeling scheme proposed in this thesis distinguishes three types of interruptions. Future work could study these interruptions in detail, trying to understand when and how they are likely to occur, as well as both speakers' behavior before, during and after interruptions. This knowledge would be valuable for situations in which an IVR system needs to interrupt the user, either because it has already collected the necessary information, or simply because it has lost track of what the user is saying.

- While all speakers in the corpus presented seemingly homogeneous strategies for displaying turn-yielding cues, each speaker seemed to use their own combination of backchannel-inviting cues. Future research should thus seek an explanation for this large degree of speaker variability, in an attempt to understand when, how and why speakers choose a particular set of cues.

18.3 Affirmative cue words

In Part III of this thesis, we undertook 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 acknowledging the interlocutor or cueing the beginning of a new topic. Considering the high frequency of ACWs in task-oriented dialogue, it is critical for IVR systems — most of which have a task-oriented domain — to model the usage of these words correctly, from both an understanding and a generation perspective.

18.3.1 Summary of findings and novel contributions

A series of statistical experiments revealed a number of significant differences in the production of the various discourse/pragmatic functions of ACWs. Final intonation and intensity were the acoustic/prosodic features that showed the most marked differences. For example, backchannels tended to end in rising intonation; acknowledgments and cue beginnings, in falling intonation; cue beginnings tended to be produced with a high intensity; cue endings, with a low intensity. We also found strong contextual differences across functions, such as the position of the word in its conversational turn, or whether the word was preceded or followed by silence. Subsequently, a perception study of the uses of the word *okay* signaled such contextual information as the most salient cue for human disambiguation of ACWs. Final intonation was the only acoustic/prosodic feature that correlated significantly with human perception of the meaning of *okay*.

We also explored the automatic classification of ACWs, for which we conducted several machine learning experiments with varying conditions to simulate the settings of real applications. We showed that the traditional distinction between sentential and discourse uses of cue phrases is insufficient for ACWs, and presented two novel alternative classification tasks: the detection of an acknowledgment function, and the detection of a discourse boundary function. Additionally, we found that the predictive power of contextual information was stronger than that of acoustic, prosodic and phonetic features extracted from the target word itself. Still, the best performing models employed information from all of these sources.

Lastly, we investigated a new dimension of speaker entrainment — or, how conversational partners tend to adapt their speech to each other's behavior. We introduced two novel measures of entrainment related to the usage of high-frequency words, including ACWs, and showed how they strongly and positively correlated with objective measures of task success and dialogue coordination.

This is, to our knowledge, the most comprehensive study of affirmative cue words in spoken dialogue. The large corpus on which it was conducted, rich in ACWs conveying a wide range of discourse/pragmatic functions, allowed us to systematically investigate various dimensions of these words, including their production, perception, and automatic disambiguation, all of which represent novel contributions to the field.

18.3.2 Impact

The findings of our statistical experiments should aid designers of IVR systems in assigning the appropriate acoustic and prosodic features to affirmative cue words, in order to unambiguously convey the intended meaning. Moreover, the results of our perception study suggest that special attention should be paid to the context in which these words occur, given that contextual information may override the effect of acoustic/prosodic properties of the words themselves.

In the experiments on the automatic disambiguation of ACWs, we explored several variations to simulate the settings of real applications — e.g. online vs. offline settings. These tests were intended to aid future researchers and developers in building effective classifiers of the discourse/pragmatic functions of ACWs, a task important not only for IVR systems, but also for other speech processing applications, such as the automatic processing of multi-party meetings.

18.3.3 Future work

When an IVR system is speaking and the user produces a short utterance, it is critical for the system to correctly determine whether the short utterance is a backchannel — in which case the system is encouraged to continue holding the turn, or a turn-taking attempt — in which case the system should yield the turn to the user. The machine learning classifiers we trained for this task failed to significantly outperform the majority-class baseline. Among the plausible reasons for this, are the ambiguity in some conditions between the acknowledgment/agreement and backchannel functions, and the similarities in the production of those two functions for some high-frequency words such as *mm-hm* and *uh-huh*. A possible direction for future research, then, consists in seeking novel approaches to this crucial classification task.

Future research should also pursue the interesting results on speaker entrainment of high-frequency words. In particular, it should try to identify any causal relations between entrainment on one side, and task success and/or dialogue coordination on the other. Such findings could have a strong impact on the development of IVR systems, providing either guidelines to enhance their quality, or novel evaluation metrics.

18.4 Epilogue

Altogether, in this thesis we proposed a number of models of variation of human speech in task-oriented dialogue, along with several plausible directions in which to enrich them in future research. If these models can be successfully incorporated into IVR systems and other speech processing applications, it might be possible to improve their performance and user satisfaction levels, thus getting us one step closer to the long-term goal of effectively emulating human behavior.

Part V

Appendices

Appendix A

The ToBI Labeling Conventions

The ToBI system (Beckman and Hirschberg, 1994; Pitrelli et al., 1994) consists of annotations at four time-linked levels of analysis: an ORTHOGRAPHIC TIER of time-aligned words; a BREAK INDEX TIER indicating degrees of juncture between words, from 0 'no word boundary' to 4 'full intonational phrase boundary', which derives from Price et al. (1991); a TONAL TIER, where pitch accents, phrase accents and boundary tones describing targets in the F0 contour define intonational phrases, following Pierrehumbert's (1980) scheme for describing SAE; and a MISCELLANEOUS TIER, in which phenomena such as disfluencies may be optionally marked.

Break indices define two levels of phrasing: level 3 corresponds to Pierrehumbert's INTERMEDIATE PHRASE and level 4, Pierrehumbert's INTONATIONAL PHRASE, with an associated tonal tier that describes the phrase accents and boundary tones for each level. Level 4 phrases consist of one or more level 3 phrases, plus a high or low boundary tone (**H%** or **L%**) 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-**). A standard declarative contour, e.g., ends in a low phrase accent and low boundary tone, and is represented by **L-L%**; a standard *yes-no* question contour ends in **H-H%**. These are illustrated in Figure A.1.

Differences among ToBI break indices can be associated with variation in F0, PHRASE-FINAL LENGTHENING (a lengthening of the syllable preceding the juncture point), glottalization ('creaky voice') over the last syllable or syllables preceding the break, and some

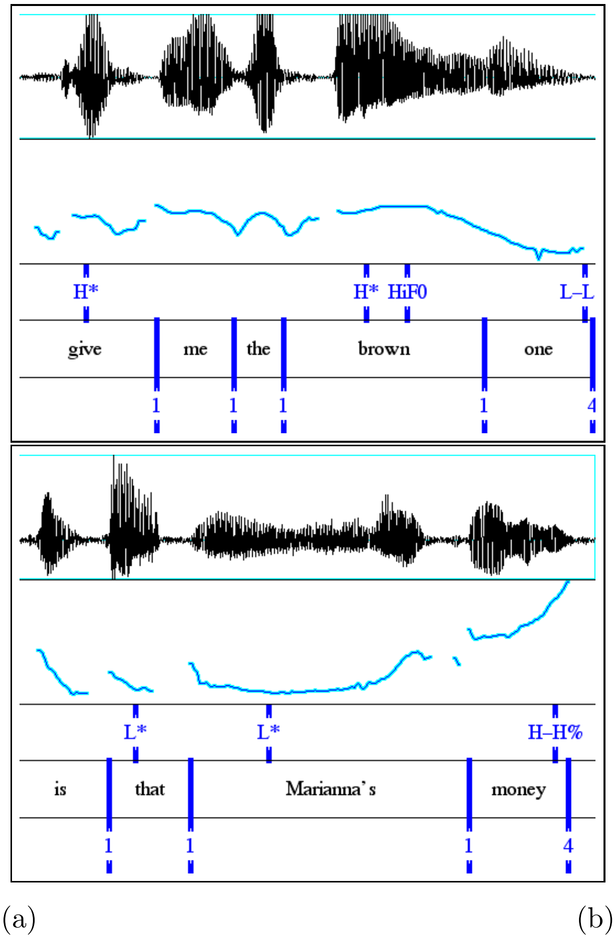


Figure A.1: (a) A **H* L-L%** contour; (b) A **L* H-H%** contour.

amount of pause. Higher number indices tend to correspond to greater evidence of these phenomena.

Pitch accents make words intonationally prominent and are realized by increased F0 height, loudness, and duration of accented syllables. A given word may be accented or deaccented and, if accented, may bear different tones, or different degrees of prominence, with respect to other words. The most prominent accent in an intermediate phrase is called the phrase's nuclear accent or nuclear stress. 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***. 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. Figure A.2 shows an example of a downstepped contour bearing two downstepped accents.

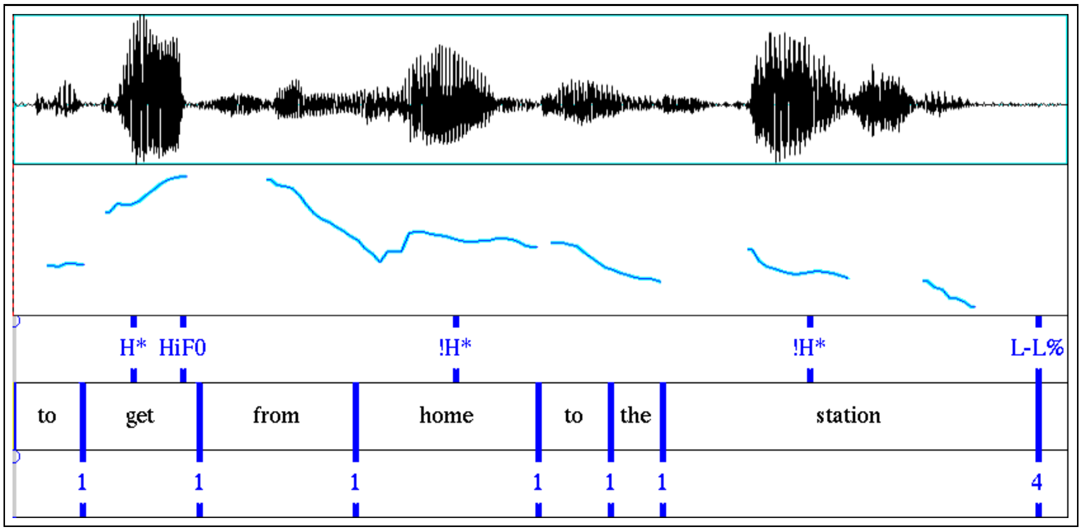


Figure A.2: A **H* !H* !H* L-L%** contour.

Appendix B

The Columbia Games Corpus

In Part I of this thesis, we described the general rules and characteristics of the computer games prepared for the collection of the Columbia Games Corpus. In this Appendix we present the detailed instructions given to the subjects, the hypotheses each game was designed to test, and the full sets of images in the same order they were presented to the subjects.

B.1 Session script and instructions screens

Subjects were read the following script by the experimenter at the beginning of the session. Actions performed by the experimenter are shown in bold typeface.

> *Today we would like you to participate in a communications experiment, which will involve playing an electronic game with a partner. We will be recording your comments to one another while you play the game. You will receive online and oral instructions on how to play the game and then will be given a chance to practice before the actual experiment begins. Feel free to ask us questions at any time.*
>
> *First, we would like to ask you to sign this consent form.*
>
> **[Give consent forms to subjects.]**
>
> *Now, we would like to fit you with recording equipment and to test some levels.*

[Set up recording equipment.]

To set our recording levels and to get you accustomed to the recording environment, we would like you to take turns asking some biographical questions of your partner. Here is the list of questions. Please alternate, so that each of you asks your partner the question and gets an answer before moving on to the next question.

[Show list of questions.]

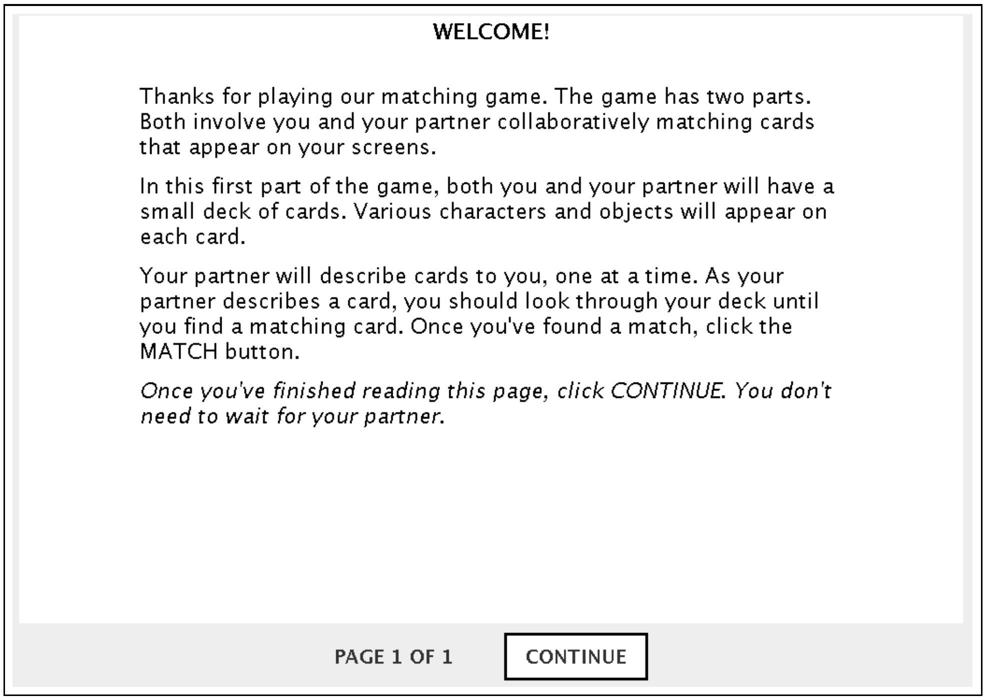
1. What is your name and why were you given your first name? Middle name?
2. Where did you grow up and did you like the place?
3. Who is your favorite relative and why?
4. What is the best movie you have seen recently, and can you give a brief summary of the plot?
5. Of all the things you do at least once a week, which do you like doing the least?
6. If you could have any occupation in the world, what would you choose and why?

Now, we'll start the games. Speak calmly and take your time. ***There is no rush.*** *This are* ***not*** *timed games.*

[Start games.]

——

The complete instructions screens given to the subjects for the first part of the Cards Game are shown in Figure B.1; for the second part of the Cards Game, in Figures B.2 and B.3; and for the Objects Game, in Figure B.5. Additionally, for the second part of the Cards Game, subjects were given a quick reference sheet, shown in Figure B.4, containing a summary of the game instructions, which they could check at any time during the game.



WELCOME!

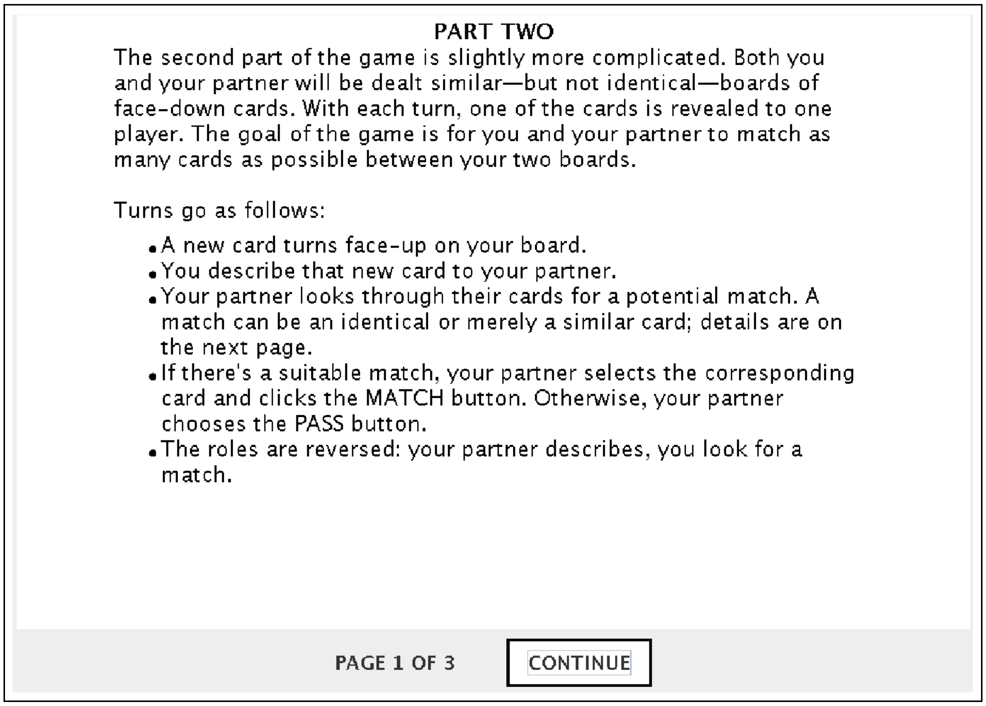
Thanks for playing our matching game. The game has two parts. Both involve you and your partner collaboratively matching cards that appear on your screens.

In this first part of the game, both you and your partner will have a small deck of cards. Various characters and objects will appear on each card.

Your partner will describe cards to you, one at a time. As your partner describes a card, you should look through your deck until you find a matching card. Once you've found a match, click the MATCH button.

Once you've finished reading this page, click CONTINUE. You don't need to wait for your partner.

Figure B.1: Instructions of the first part of the Cards Game.



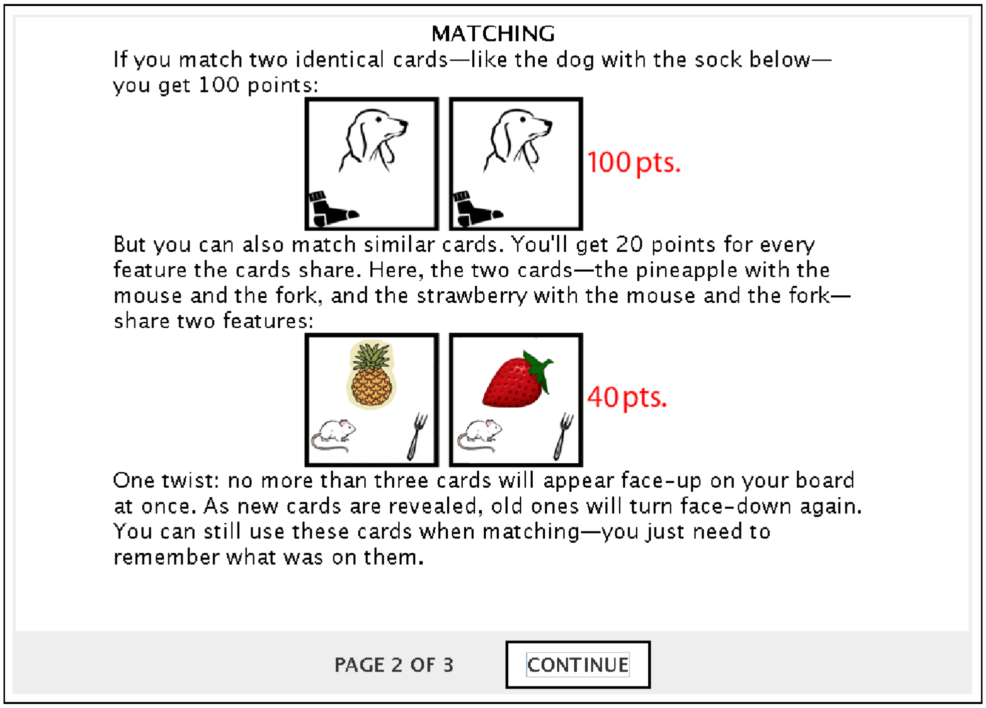
PART TWO

The second part of the game is slightly more complicated. Both you and your partner will be dealt similar—but not identical—boards of face-down cards. With each turn, one of the cards is revealed to one player. The goal of the game is for you and your partner to match as many cards as possible between your two boards.

Turns go as follows:

- A new card turns face-up on your board.
- You describe that new card to your partner.
- Your partner looks through their cards for a potential match. A match can be an identical or merely a similar card; details are on the next page.
- If there's a suitable match, your partner selects the corresponding card and clicks the MATCH button. Otherwise, your partner chooses the PASS button.
- The roles are reversed: your partner describes, you look for a match.

Figure B.2: Instructions of the second part of the Cards Game. *Continued in Figure B.3.*

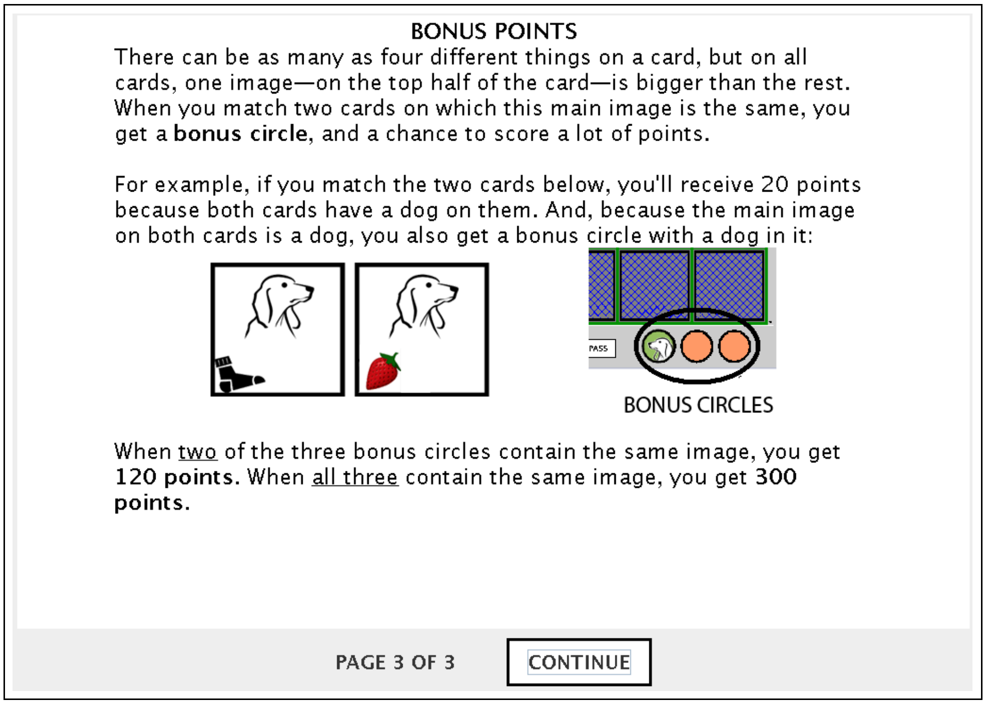


MATCHING

If you match two identical cards—like the dog with the sock below—you get 100 points:

But you can also match similar cards. You'll get 20 points for every feature the cards share. Here, the two cards—the pineapple with the mouse and the fork, and the strawberry with the mouse and the fork—share two features:

One twist: no more than three cards will appear face-up on your board at once. As new cards are revealed, old ones will turn face-down again. You can still use these cards when matching—you just need to remember what was on them.



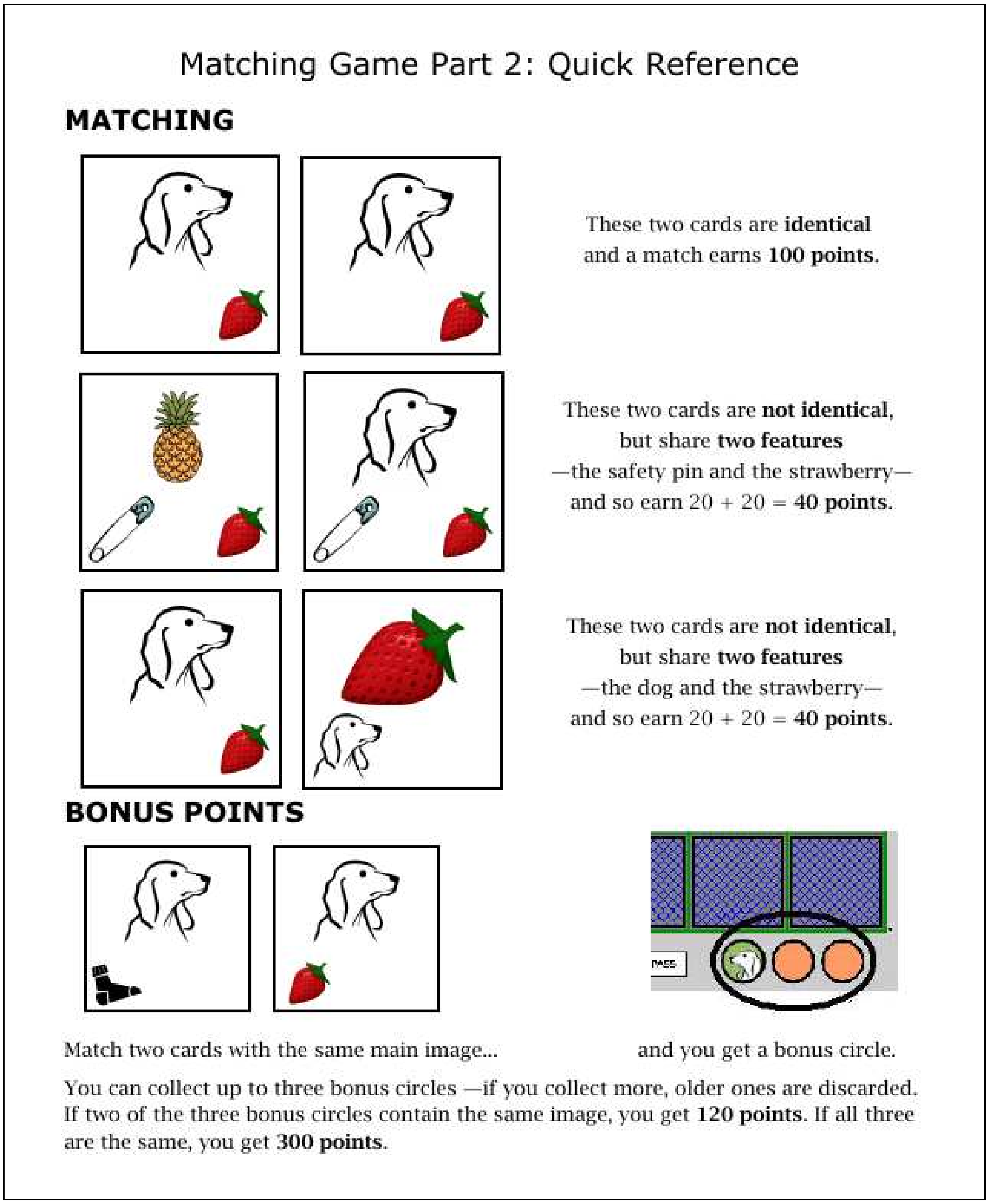
BONUS POINTS

There can be as many as four different things on a card, but on all cards, one image—on the top half of the card—is bigger than the rest. When you match two cards on which this main image is the same, you get a **bonus circle**, and a chance to score a lot of points.

For example, if you match the two cards below, you'll receive 20 points because both cards have a dog on them. And, because the main image on both cards is a dog, you also get a bonus circle with a dog in it:

When two of the three bonus circles contain the same image, you get **120 points**. When all three contain the same image, you get **300 points**.

Figure B.3: Instructions of the second part of the Cards Game. *Continued from Fig. B.2.*



Matching Game Part 2: Quick Reference

MATCHING

These two cards are **identical** and a match earns **100 points**.

These two cards are **not identical**, but share **two features** —the safety pin and the strawberry— and so earn 20 + 20 = **40 points**.

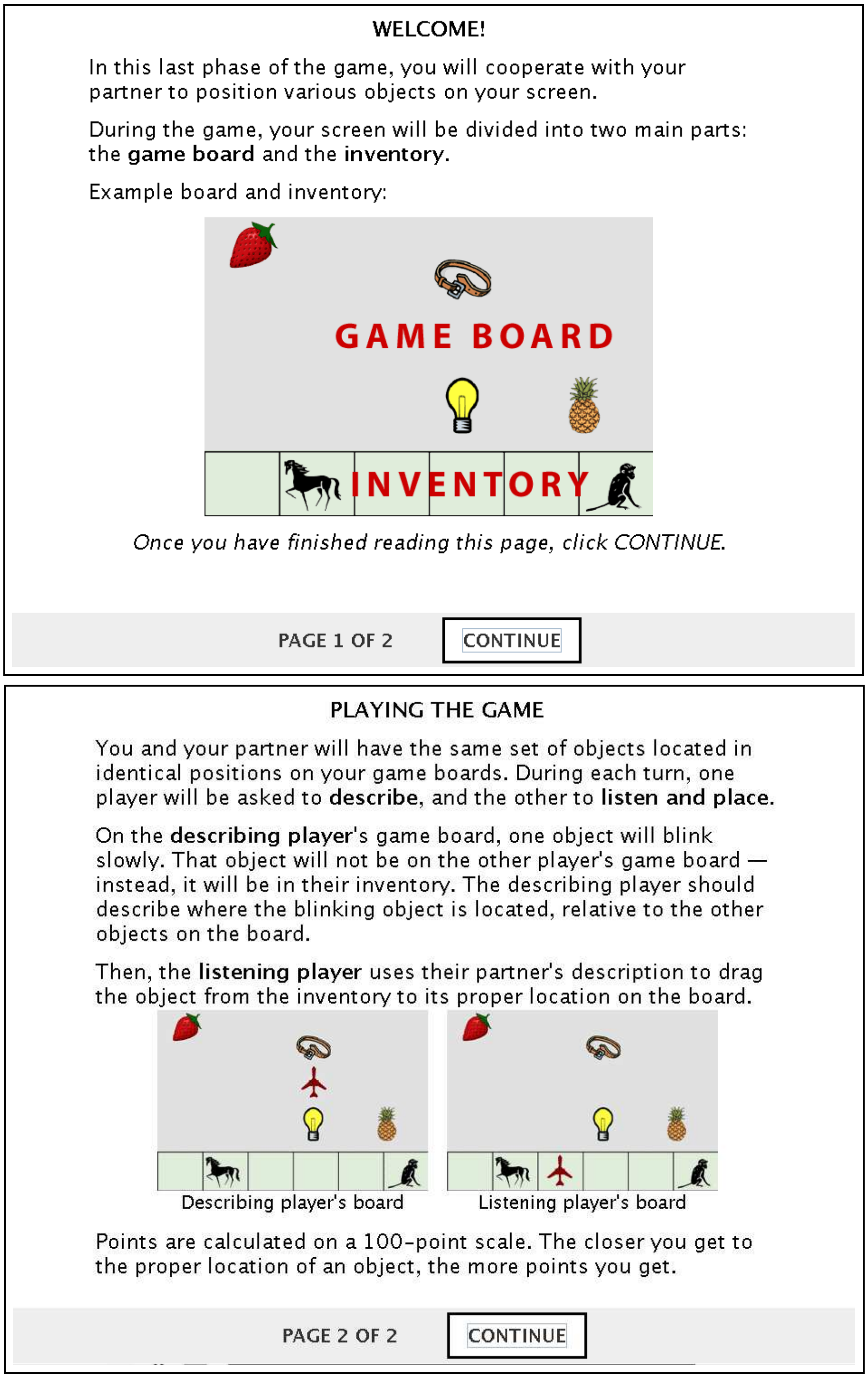
These two cards are **not identical**, but share **two features** —the dog and the strawberry— and so earn 20 + 20 = **40 points**.

BONUS POINTS

Match two cards with the same main image... and you get a bonus circle.

You can collect up to three bonus circles —if you collect more, older ones are discarded. If two of the three bonus circles contain the same image, you get **120 points**. If all three are the same, you get **300 points**.

Figure B.4: Reference sheet for the second part of the Cards Game.



WELCOME!

In this last phase of the game, you will cooperate with your partner to position various objects on your screen.

During the game, your screen will be divided into two main parts: the **game board** and the **inventory**.

Example board and inventory:

Once you have finished reading this page, click CONTINUE.

PLAYING THE GAME

You and your partner will have the same set of objects located in identical positions on your game boards. During each turn, one player will be asked to **describe**, and the other to **listen and place.**

On the **describing player**'s game board, one object will blink slowly. That object will not be on the other player's game board — instead, it will be in their inventory. The describing player should describe where the blinking object is located, relative to the other objects on the board.

Then, the **listening player** uses their partner's description to drag the object from the inventory to its proper location on the board.

Points are calculated on a 100-point scale. The closer you get to the proper location of an object, the more points you get.

Figure B.5: Instructions of the Objects Game.

B.2 Hypotheses tested

In the **first instance** of the Cards game that subjects were asked to play, we systematically varied the number of cards between occurrences of the target images: from 0 to 7 cards. This design was intended to test the hypothesis that the production of *given* information changes depending on the recency of preceding mentions. In particular, when referring to a *given* entity, is the choice between deaccentuation and a downstepped pitch accent guided by the distance to the entity's previous reference?

The **second instance** of the Cards game was designed to test the hypothesis that, the more complex (or heavier) a noun phrase, and the higher its number of *given* items, then the more likely it is to be produced with a downstepped contour. For this, subjects were shown cards containing varying numbers of images, with the *given/new* status of the images also varied systematically. For example, based on preliminary tests of the Cards game, we expected subjects to describe the first card in Figure B.6 as *"the rhinoceros with the owl and the ruler"*, and the second card as *"the rhinoceros with the owl, the ear and the ruler"*. Then the question we want to answer is, given that the second NP is heavier and has more *given* items, is it more likely to be produced with a downstepped contour than the first NP?

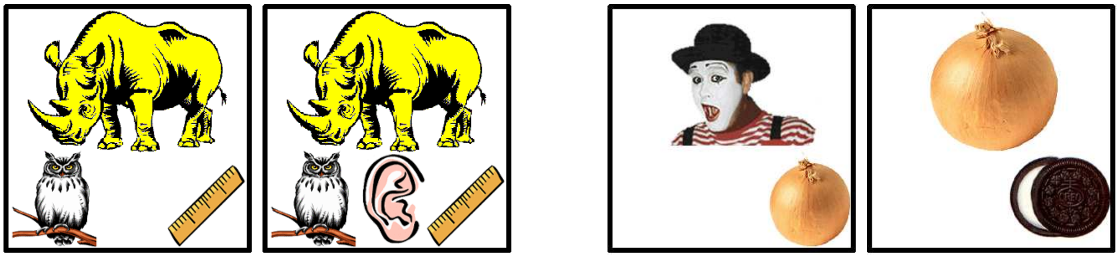
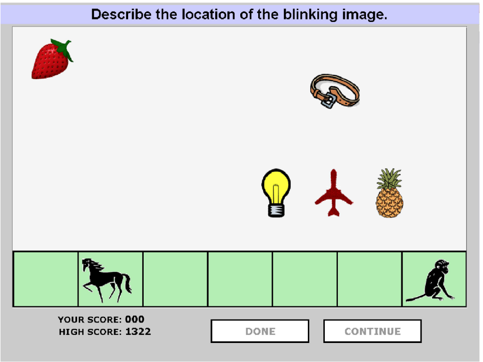


Figure B.6: Sample cards from the second and third Cards game.

We designed the **third instance** of the Cards game to study the effect of grammatical function and surface position on the production of *given* information. For example, we expected subjects to describe the third card in Figure B.6 as *"the mime with the onion"*, and the fourth card as *"the onion with the Oreo cookie"*. Then our question is, since the grammatical function of *the onion* shifts from object in the first mention to subject in the second, and its surface position from phrase-final to phrase-initial position, how will its second mention be produced? Will it bear a high pitch accent, a downstepped accent, or will it be deaccented?



Describe the location of the blinking image.

YOUR SCORE: 000
HIGH SCORE: 1322

Figure B.7: Sample screen from the Objects Game.

Finally, the Objects game too was designed to study the effect of grammatical function on the production of *given* information, although in a different way. In this case, we expected target images to be produced in subject position, and surrounding images in object position. For example, the location of the airplane in Figure B.7 could be described as *The airplane is between the lightbulb and the pineapple*, where *the airplane* appears in subject position, while *the lightbulb* and *the pineapple* are in object position.

B.3 Images of the Cards Games

B.3.1 Cards Game Number 1

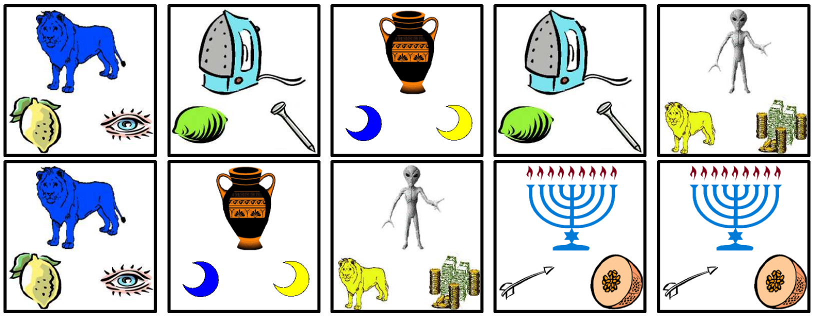


Figure B.8: Cards Game 1, first part, Describer's Deck

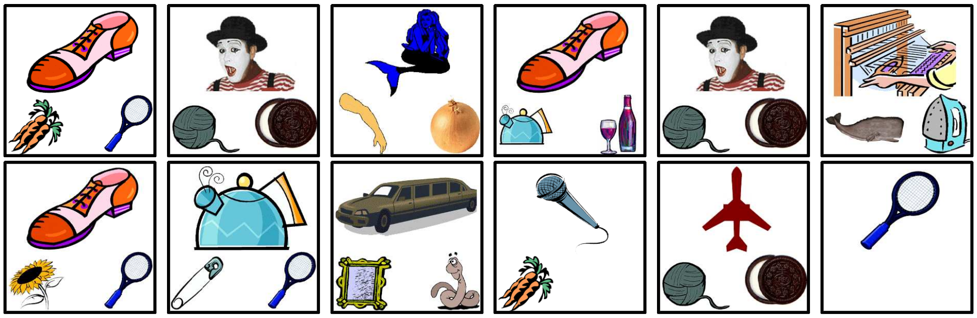


Figure B.9: Cards Game 1, second part, Player A's Board

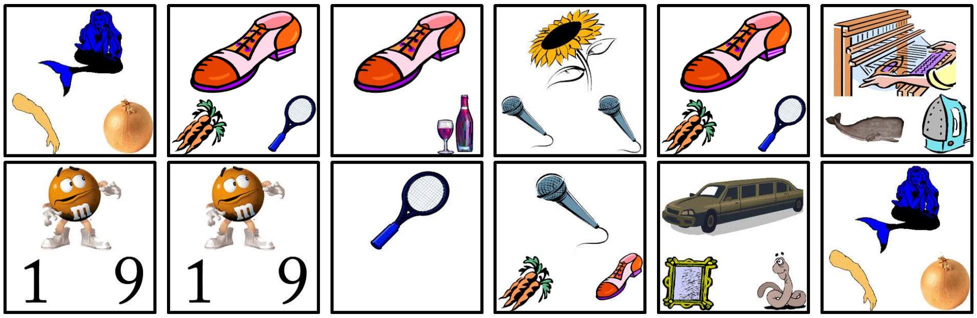


Figure B.10: Cards Game 1, second part, Player B's Board

B.3.2 Cards Game Number 2

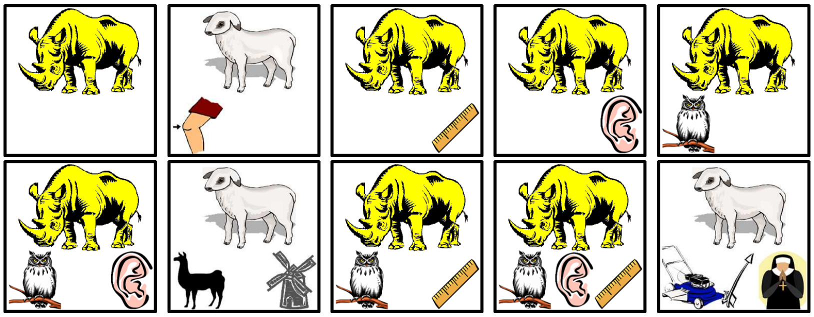


Figure B.11: Cards Game 2, first part, Describer's Deck

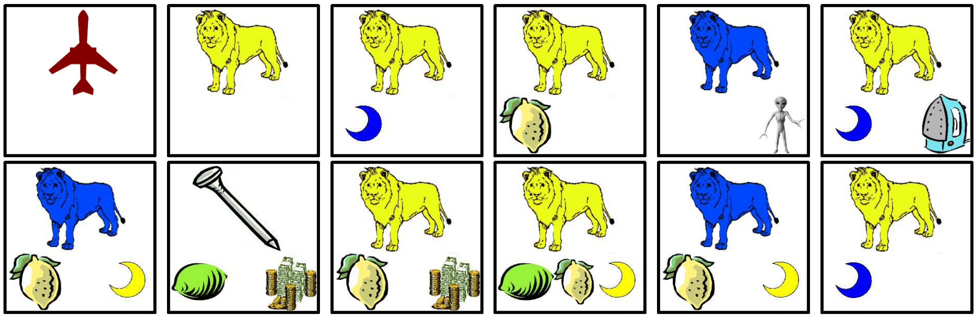


Figure B.12: Cards Game 2, second part, Player A's Board

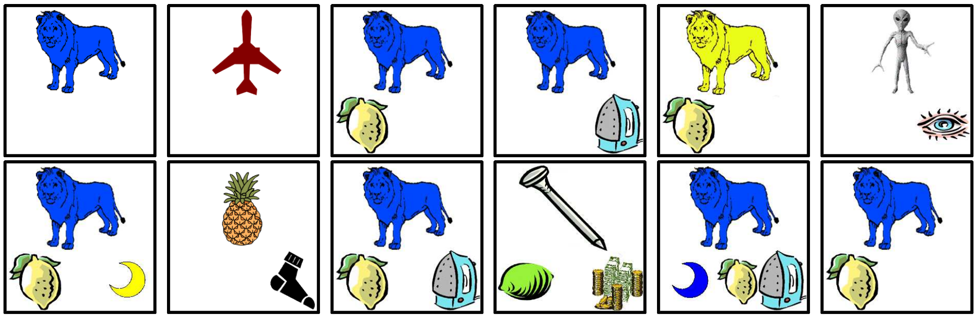


Figure B.13: Cards Game 2, second part, Player B's Board

B.3.3 Cards Game Number 3

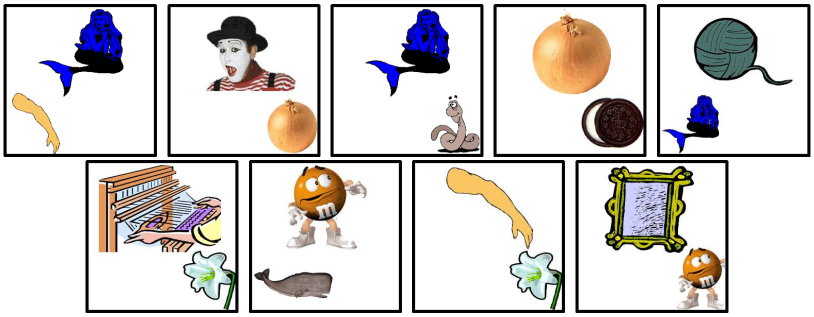


Figure B.14: Cards Game 3, first part, Describer's Deck

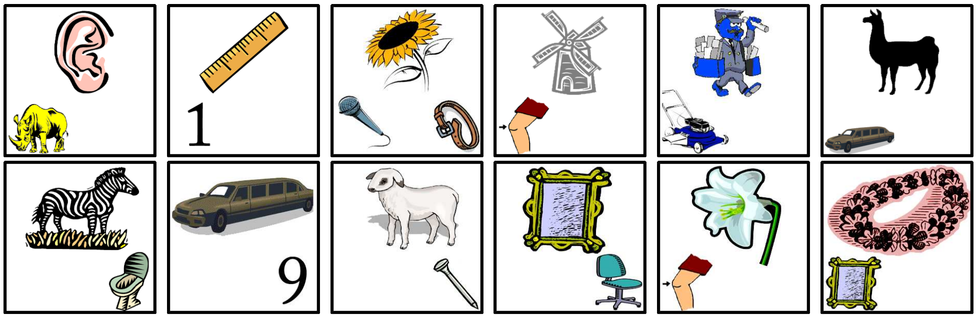


Figure B.15: Cards Game 3, second part, Player A's Board

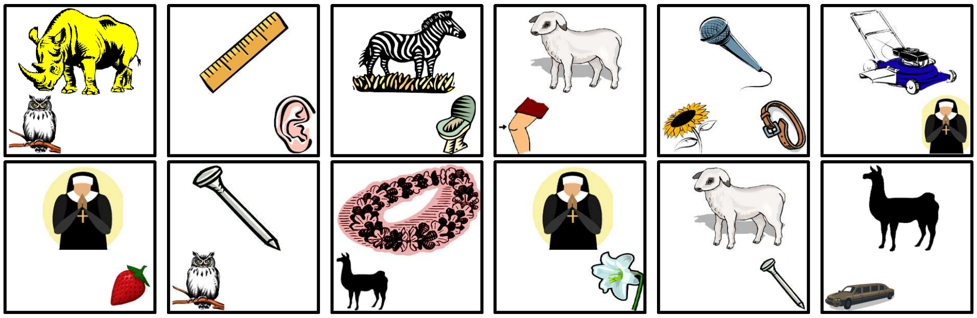
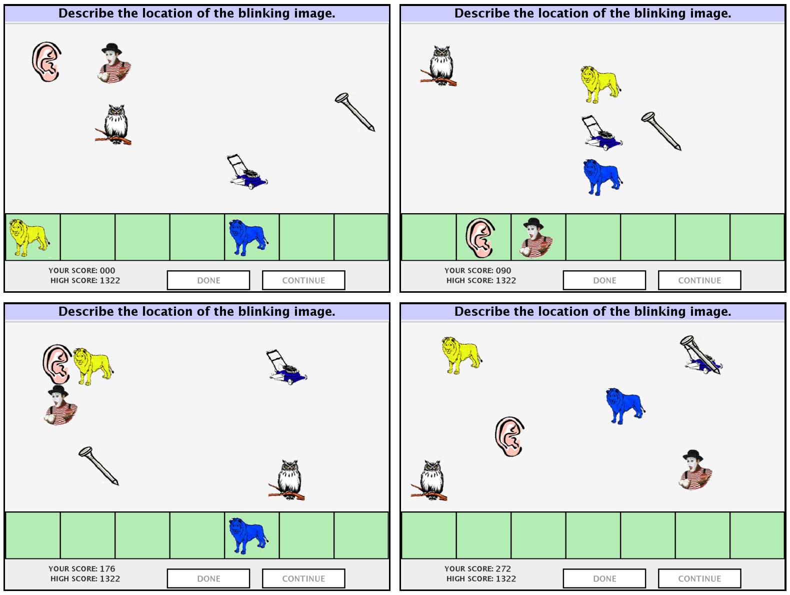


Figure B.16: Cards Game 3, second part, Player B's Board

B.4 Images of the Objects Game



Describe the location of the blinking image.

YOUR SCORE: 000
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 090
HIGH SCORE: 1322

Describe the location of the blinking image.

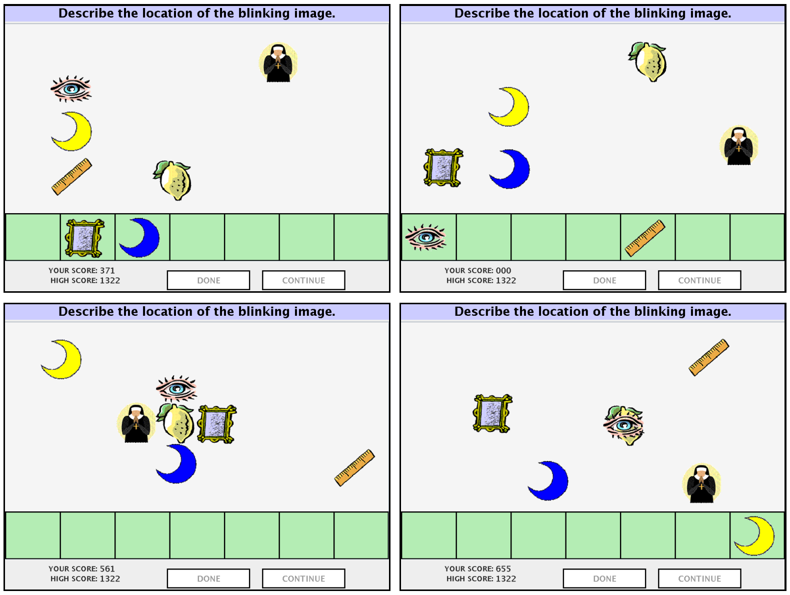
YOUR SCORE: 176
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 272
HIGH SCORE: 1322

Figure B.17: Objects Game 1, Describer's Board.

Target objects (from left to right and top to bottom): mime, lawnmower, ear, nail.



Describe the location of the blinking image.

YOUR SCORE: 371
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 000
HIGH SCORE: 1322

Describe the location of the blinking image.

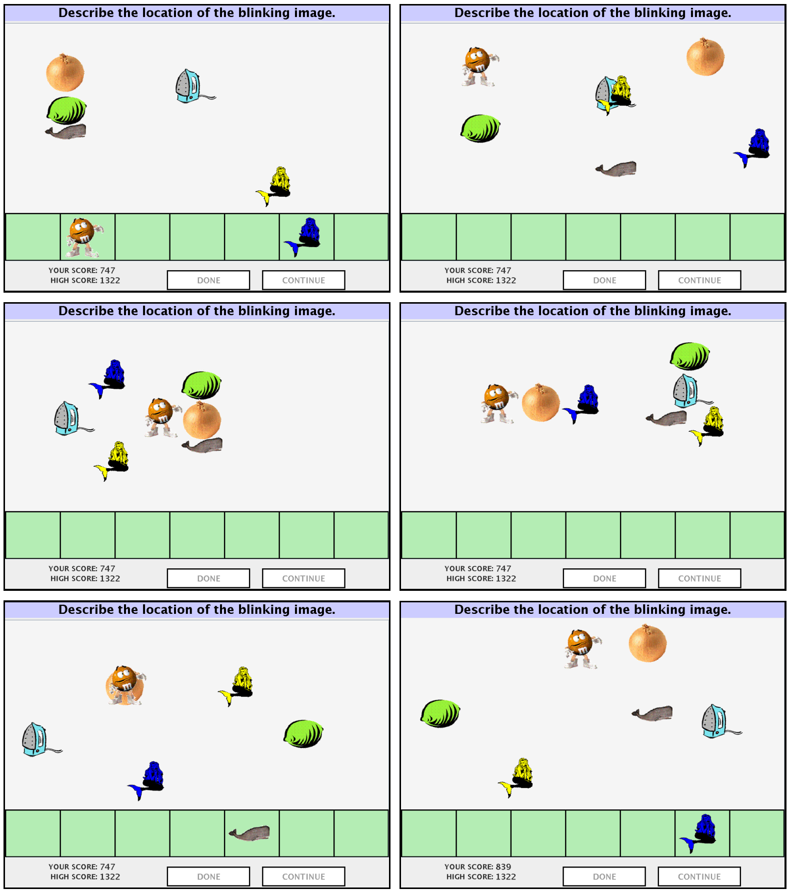
YOUR SCORE: 561
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 655
HIGH SCORE: 1322

Figure B.18: Objects Game 2, Describer's Board

Target objects: yellow moon, blue moon, lemon, eye.



Describe the location of the blinking image.

YOUR SCORE: 747
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 747
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 747
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 747
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 747
HIGH SCORE: 1322

Describe the location of the blinking image.

YOUR SCORE: 839
HIGH SCORE: 1322

Figure B.19: Objects Game 3, Describer's Board

Target objects: lime, yellow mermaid, onion, iron, M&M, whale.

B.5 Questions

Two trained annotators identified all questions in the Objects portion of the Games Corpus using a simple definition: A QUESTION is an *'utterance that requests an answer'*. Additionally, the same annotators identified all QUESTION-LIKE (QL) utterances, defined as utterances that do not fit our definition of questions, but that satisfy the following two conditions: a) there is something in the utterance that is plausibly questionable from the context, and b) the utterance allows, rather than requests, an answer.

Subsequently, two different trained annotators classified each question (not including QL utterances) according to their form and function, as shown in Tables B.1 and B.2, respectively. There are 5 types and 10 subtypes of question forms, and 4 types and 13 subtypes of question functions. The inter-labeler agreement for the question form labeling task is substantial: $\kappa = 0.719$ when considering all 10 subtypes, and $\kappa = 0.815$ when using only the 5 main types. For the question function labeling task, the inter-labeler agreement is low: $\kappa = 0.190$ when considering all 13 subcategories, and $\kappa = 0.231$ when using only the 4 main categories.

Type	**Subtype**	**Example**
Yes-no question	Declarative	*The card has a blue moon on it?*
	Canonical/full	*Is the card blinking?*
	Reduced	*You see that?*
Wh-questions	Declarative	*You're putting the lemon where?*
	Canonical	*How many cards are there?*
	Reduced	*A what?*
Alternative question	–	*Or is it more blue than green?*
Tag question	Canonical/full	*You like Mac computers, don't you?*
	Reduced	*I'm going to look at that top card, okay?*
Fragment	–	*A Lion?*

Table B.1: Question form types

Type	**Subtype**	**Example**
Action request	Indirect	*Why don't you go ahead?*
	Direct	*Go ahead and try that card, okay?*
Clarification	Reformulation /summarization /specification	A: *Find the card with the dog.* B: *The yellow dog?*
	Suggest possible correction or intention	A: *I like Murakami's style, he's sort of a...* B: *Surrealist?*
	Confirmation	*You've got it, right?*
	Signal non-understanding: Acoustic	A: *Excuse me! I'm looking for a bathroom.* *B: Pardon?*
	Signal non-understanding: Semantic/referential	A: *Over there.* B: *Where is 'there'?*
Rhetorical question	Agreement	A: *Do you want to do that then?* B: *Sure, why not?*
	Point	A: *He married his adopted daughter!* B: *Who would do such a thing?*
	Backchannel	A: *She totally had it out with him!* B: *Oh, really?*
Information request	Factual	*What card are you looking at?*
	Comment	*What do you think?*
	Suggest	*Which card do you think we should match?*

Table B.2: Question function types

Appendix C

ACW Labeling Guidelines

These guidelines for labeling the discourse/pragmatic functions of affirmative cue words were developed by Julia Hirschberg, Stefan Benus, 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**."*

[**Ack**] **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. 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 **Ack**.)

[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: (Ack+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: (Ack+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

- *"okay so"* / *"okay now"* / *"okay then"* / etc., where both words are uttered together, *okay* seems to convey **Ack**, and *so* / *now* / *then* seems to convey **CBeg**. Since we do not label words like *so*, *now* or *then*, we label *okay* as **PBeg**.

- 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*:Ack *yeah*:Ack *yeah*:? *yeah*:?".

Appendix D

Turn-taking Labeling Guidelines

These guidelines for labeling turn-taking phenomena were developed by Julia Hirschberg, Stefan Benus, Agustín Gravano, Héctor Chávez and Enrique Henestroza at Columbia University, and were based on the labeling scheme proposed in Beattie (1982).

———

Turn exchanges, 'turns' tier

Label only the turn intervals inside tasks (tasks are marked by intervals that start with "Images:" in the 'tasks' tier).

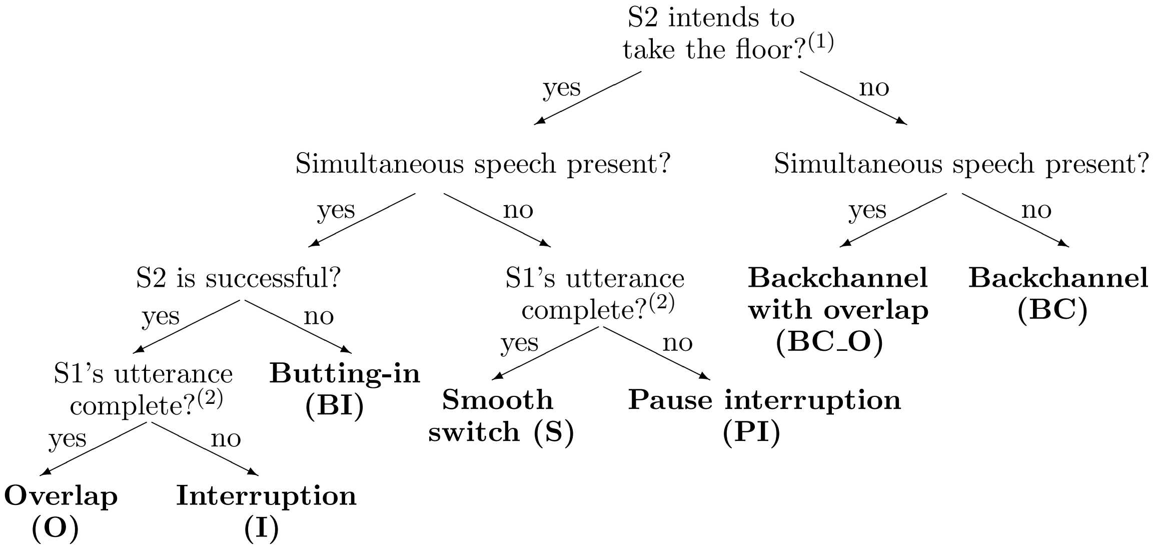
For each turn interval by S2, where S1 is the other speaker, label S2's turn interval as follows:

(1) Backchannels were identified by three annotators for the *affirmative cue words* project, who were provided with the following definition:

> *Backchannel: The function of 'okay' [or 'alright', 'mm-hm', 'yeah', etc.] in response to another speaker's utterance that indicates only "I'm still here / I hear you and please continue".*

When a simple majority of annotators (i.e., at least two out of three) considered an utterance to be a backchannel, it was labeled **BC** or **BC_O**.

(2) We use Beattie's informal definition of utterance completeness: "Completeness was

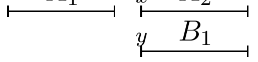

S2 intends to take the floor?(1)
yes
no
Simultaneous speech present?
yes
no
Simultaneous speech present?
yes
no
S2 is successful?
yes
no
S1's utterance complete?(2)
yes
no
Backchannel with overlap (BC_O)
Backchannel (BC)
S1's utterance complete?(2)
yes
no
Butting-in (BI)
Smooth switch (S)
Pause interruption (PI)
Overlap (O)
Interruption (I)

judged intuitively, taking into account the intonation, syntax, and meaning of the utterance" (Beattie, 1982, page 100).

Special cases

We identified three common cases in which no turn exchange occurs, and the corresponding turn interval receives a special label **X[1-3]**.

- **Task beginnings:** If a turn interval begins a new task, then label it **X1**.
- **Continuation after a backchannel:** If a turn interval t is a continuation from the previous turn by the same speaker after a **BC** or **BC_O**, then label it **X2_O** if t overlaps the backchannel, or **X2** if not.
- **Simultaneous start:** If two turn intervals begin almost simultaneously — formally, within 210 ms of each other (Fry, 1975) — then the speakers are most probably reacting to the preceding turn interval:

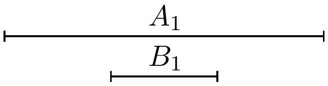
  

  $0 < |y - x| < 210\text{ms}$

  In the figure, A_2 and B_1 occur most likely in response to A_1. Thus, B_1 should be labeled with respect to A_1 (not A_2); A_2 should be labeled **X3**.

Notes

- The figure below shows a frequent pattern consisting of a complete short utterance (B_1) fully contained within a longer utterance (A_1) by the other speaker, such that the floor is briefly *shared* by both speakers, and A_1 is not disrupted by B_1. In such cases, the most appropriate label for B_1, according to our labeling scheme, is **O**; it is neither **I** nor **BI** because both utterances are complete.



Miscellaneous tier

Collaborative contributions

If a speaker completes, or attempts to complete, an utterance from their interlocutor, as if trying to help them, add a **'Help'** label in the misc tier.

Other

Mark in the misc tier any other situation not contemplated in these guidelines.

Appendix E

Turn-Taking Results Per Speaker

E.1 Evidence of turn-yielding cues per speaker

	Abs pitch slope			Syllables per sec			Phonemes per sec		
Speaker	S	H	*p*	S	H	*p*	S	H	*p*
101	298.0	134.6	∼0	5.10	4.02	∼0	11.45	8.69	∼0
102	237.8	167.2	∼0	7.33	5.94	∼0	16.76	12.35	∼0
103	224.3	157.9	∼0	5.00	4.28	∼0	11.57	9.60	∼0
104	180.4	94.8	0.02	4.77	4.12	∼0	11.15	9.71	∼0
105	222.6	161.8	∼0	5.75	4.99	∼0	12.60	10.87	∼0
106	295.0	227.8	∼0	5.27	4.91	∼0	12.21	10.88	∼0
107	154.1	105.0	0.03	5.04	4.28	∼0	11.06	8.78	∼0
108	215.7	155.5	0.01	5.36	3.99	∼0	12.66	9.00	∼0
109	210.2	121.6	∼0	5.50	4.08	∼0	12.83	9.14	∼0
110	255.8	209.1	0.06	5.40	4.93	∼0	12.28	11.42	0.04
111	214.8	163.5	∼0	5.16	4.28	∼0	11.68	9.39	∼0
112	188.8	115.4	∼0	4.85	4.42	∼0	11.49	9.68	∼0
113	242.0	177.4	0.03	5.00	4.49	∼0	11.62	9.84	∼0

Table E.1: Absolute pitch slope over the final 300ms of the IPU, and syllables and phonemes per second over the whole IPU, for IPUs preceding **S** and **H**. The *p*-values correspond to ANOVA tests between the two groups.

	Mean intensity			Mean pitch			Number of words		
Speaker	S	H	p	S	H	p	S	H	p
101	64.3	64.7	0.60	110.1	116.1	0.07	6.51	4.90	∼0
102	69.3	72.4	∼0	119.4	134.3	∼0	5.44	3.79	∼0
103	67.3	70.6	∼0	131.1	134.3	0.22	6.42	3.98	∼0
104	72.5	74.8	∼0	98.5	99.9	0.34	4.63	3.71	0.01
105	65.1	67.6	∼0	115.6	122.3	∼0	4.94	3.47	∼0
106	63.5	66.7	∼0	113.2	112.2	0.55	6.79	4.91	∼0
107	59.9	64.8	∼0	85.1	90.6	∼0	5.32	3.53	∼0
108	67.1	68.4	0.02	101.1	104.1	0.11	6.58	4.97	∼0
109	60.0	63.3	∼0	95.4	101.1	0.01	4.53	3.58	∼0
110	63.5	65.4	∼0	120.3	127.1	∼0	4.87	3.52	∼0
111	64.6	66.4	∼0	112.3	112.6	0.85	6.76	4.38	∼0
112	63.5	66.5	∼0	117.6	126.8	∼0	5.78	3.54	∼0
113	64.3	66.3	∼0	124.5	127.3	0.24	5.78	3.46	∼0

Table E.2: Mean intensity and pitch levels over the final 500ms of the IPU, and number of words in the entire IPU, for IPUs preceding **S** and **H**.

	Jitter			Shimmer			NHR		
Speaker	S	H	p	S	H	p	S	H	p
101	0.020	0.011	∼0	0.108	0.073	∼0	0.324	0.188	∼0
102	0.015	0.011	∼0	0.120	0.091	∼0	0.314	0.200	∼0
103	0.011	0.007	∼0	0.090	0.071	∼0	0.163	0.116	∼0
104	0.017	0.012	∼0	0.083	0.066	∼0	0.145	0.096	∼0
105	0.013	0.010	∼0	0.104	0.081	∼0	0.186	0.120	∼0
106	0.021	0.016	∼0	0.127	0.101	∼0	0.326	0.261	∼0
107	0.020	0.015	∼0	0.110	0.091	∼0	0.307	0.190	∼0
108	0.016	0.014	0.01	0.088	0.076	∼0	0.243	0.189	∼0
109	0.015	0.010	∼0	0.091	0.065	∼0	0.211	0.121	∼0
110	0.012	0.011	∼0	0.103	0.087	∼0	0.177	0.147	∼0
111	0.013	0.010	∼0	0.089	0.077	∼0	0.155	0.127	∼0
112	0.011	0.007	∼0	0.095	0.069	∼0	0.160	0.095	∼0
113	0.014	0.012	0.27	0.099	0.089	0.05	0.202	0.163	∼0

Table E.3: Jitter, shimmer and noise-to-harmonics ratio, computed over the final 500ms of the IPU, for IPUs preceding **S** and **H**.

Speaker ID	S		H	
101	104	(77.6%)	233	(55.9%)
102	196	(77.8%)	244	(46.5%)
103	244	(85.9%)	346	(57.0%)
104	95	(75.4%)	193	(53.9%)
105	337	(88.5%)	371	(51.6%)
106	314	(80.3%)	486	(51.8%)
107	214	(85.3%)	348	(46.6%)
108	144	(77.4%)	402	(60.9%)
109	130	(71.4%)	357	(48.0%)
110	212	(83.5%)	455	(53.5%)
111	306	(82.9%)	323	(51.4%)
112	227	(81.7%)	283	(54.3%)
113	126	(79.7%)	231	(56.8%)
Total	2649	(81.6%)	4272	(52.6%)

Table E.4: Number and proportion of complete IPUs preceding **S** and **H** per speaker, as predicted by our SVM-based automatic classifier.

E.2 Evidence of backchannel-inviting cues per speaker

	Pitch slope			Mean intensity			Mean pitch		
Speaker	S	H	*p*	S	H	*p*	S	H	*p*
102	208.2	29.1	$\sim$0	71.3	72.4	0.48	140.6	134.3	0.34
103	173.7	58.8	$\sim$0	72.6	70.6	0.09	138.3	134.3	0.48
105	163.1	-8.8	$\sim$0	68.5	67.6	0.13	124.7	122.3	0.43
106	153.5	45.0	0.02	68.9	66.7	$\sim$0	115.8	112.2	0.29
108	109.7	56.1	0.28	71.3	68.4	0.01	105.0	104.1	0.80
110	217.9	-4.9	$\sim$0	65.3	65.4	0.93	131.8	127.1	0.33
111	67.0	12.2	0.09	70.3	66.4	$\sim$0	129.8	112.6	$\sim$0
112	217.3	3.3	$\sim$0	68.9	66.5	$\sim$0	144.0	126.8	$\sim$0
113	119.7	6.4	0.01	69.7	66.3	$\sim$0	141.3	127.3	$\sim$0

Table E.5: Pitch slope over the final 300ms of the IPU, and mean intensity and pitch levels over the final 500ms of the IPU, for IPUs preceding **BC** and **H**. The *p*-values correspond to ANOVA tests between the two groups.

	Number of words			NHR		
Speaker	S	H	p	S	H	p
102	4.654	3.790	0.16	0.207	0.200	0.84
103	6.612	3.975	~ 0	0.075	0.116	0.04
105	5.050	3.466	~ 0	0.087	0.120	~ 0
106	7.169	4.912	~ 0	0.210	0.261	0.01
108	7.727	4.965	~ 0	0.176	0.189	0.57
110	4.638	3.524	0.02	0.080	0.147	~ 0
111	7.294	4.382	~ 0	0.074	0.127	~ 0
112	5.400	3.539	~ 0	0.058	0.095	0.01
113	6.100	3.459	~ 0	0.086	0.163	~ 0

Table E.6: Number of words in the entire IPU, and noise-to-harmonics ratio over the final 500ms of the IPU, for IPUs preceding **BC** and **H**.

102		103		105		106		108	
NN NN	7	DT NN	27	DT NN	39	DT NN	25	DT NN	16
DT NN	7	JJ NN	6	JJ NN	20	NN NN	10	DT JJ	2
PRP VBP	2	VBZ VBG	5	NN NN	10	IN NN	9	# NN	2
IN NN	2	DT JJ	2	DT NNP	3	JJ NN	6	IN NN	1
NNS NN	2	UH NN	1	# NN	3	DT JJ	3	NN VB	1
JJ NN	2	IN PRP	1	NN IN	1	# NN	3	IN PRP	1
# IN	1	CD NNS	1	# RB	1	RB VB	1	NN NN	1
	...		...		...		...		...
Total	26	Total	49	Total	80	Total	65	Total	33

110		111		112		113	
DT NN	18	DT NN	35	DT NN	21	DT NN	20
JJ NN	8	JJ NN	17	JJ NN	14	JJ NN	8
NN NN	7	NN NN	8	NN NN	11	# NN	7
# NN	3	NN VBZ	3	# NN	5	NN NN	5
NN NNS	2	IN DT	2	IN PRP	3	NN RB	2
DT JJ	1	NN RB	2	DT NNP	1	DT NNP	1
CD NNS	1	NNS VBP	2	DT CD	1	DT JJ	1
	...		...		...		...
Total	47	Total	85	Total	65	Total	60

Table E.7: Counts of the most frequent final POS bigrams in IPUs preceding **BC**, per speaker.

Appendix F

ACWs Results By Word

	alright		*mm-hm*		*okay*				
	Ack	CBeg	Ack	BC	Ack	BC	CBeg	PEnd	PBeg
H-H%	5	0	21	145	38	24	1	5	0
[!]H-L%	9	3	12	76	163	13	64	23	9
L-H%	8	2	8	51	121	57	17	9	3
L-L%	29	31	1	17	303	13	132	40	29
other	8	4	1	3	118	1	70	14	22

	right			*uh-huh*		*yeah*		
	Ack	Chk	Mod	Ack	BC	Ack	BC	PEnd
H-H%	0	19	42	6	18	3	4	0
[!]H-L%	8	4	30	3	25	59	8	0
L-H%	4	8	35	5	37	90	31	2
L-L%	43	2	131	1	11	257	12	9
other	5	1	363	0	0	137	1	1

Table F.1: ToBI phrase accents and boundary tones per ACW. The ‘other’ category consists of cases with no phrase accent and/or boundary tone present at the target word.

	IPU initial		IPU medial		IPU final		Single-word IPU	
alright	Total	63	Total	6	Total	10	Total	77
	Ack	30	Ack	1	Ack	6	Ack	39
	BTsk	1	CBeg	4	FEnd	1	BTsk	4
	CBeg	32	Mod	1	Mod	3	CBeg	25
							FEnd	9
okay	Total	655	Total	74	Total	154	Total	1224
	Ack	248	Ack	41	Ack	113	Ack	690
	BTsk	1	CBeg	20	BC	1	BC	119
	CBeg	365	Mod	5	CBeg	11	BTsk	27
	Chk	1	PBeg	4	CEnd	1	CBeg	147
	PBeg	39	Stl	4	FEnd	12	CEnd	3
	Stl	1			Mod	12	Chk	4
					PBeg	1	FEnd	206
					Stl	3	Mod	1
							PBeg	20
							Stl	7
yeah	Total	251	Total	60	Total	70	Total	449
	Ack	251	Ack	60	Ack	68	Ack	375
					FEnd	2	BC	58
							FEnd	16
mm-hm	Total	6	Total	0	Total	1	Total	450
	Ack	5			Ack	1	Ack	52
	BC	1					BC	394
							FEnd	4
uh-huh	Total	1	Total	0	Total	2	Total	114
	Ack	1			Ack	2	Ack	13
							BC	101
right	Total	63	Total	485	Total	573	Total	71
	Ack	11	Ack	7	Ack	11	Ack	45
	Mod	52	Chk	7	Chk	36	Chk	6
			Mod	471	Mod	526	Mod	20

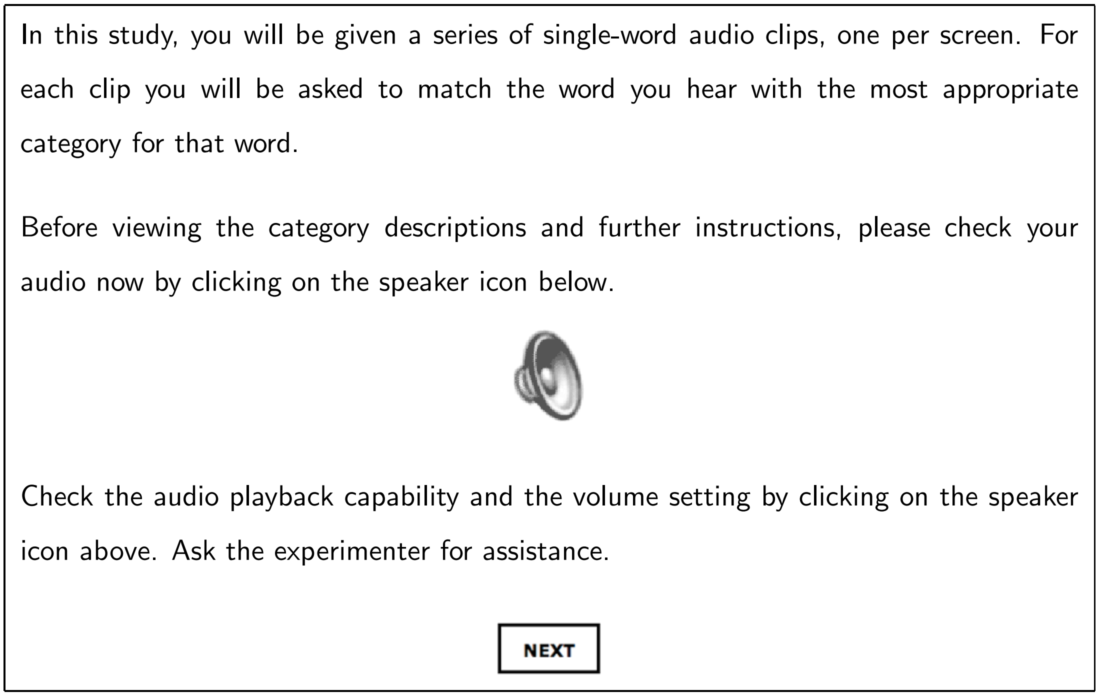
Table F.2: Distribution of ACWs and discourse/pragmatic functions per position in the inter-pausal unit (IPU). See Figure 13.1 on page 104.

	Turn initial		Turn medial		Turn final		Single-word turn	
alright	Total	88	Total	35	Total	8	Total	25
	Ack	39	Ack	15	Ack	4	Ack	18
	BTsk	1	BTsk	1	BTsk	1	BTsk	2
	CBeg	45	CBeg	16	FEnd	2	FEnd	5
	FEnd	3	Mod	3	Mod	1		
okay	Total	985	Total	210	Total	139	Total	773
	Ack	436	Ack	105	Ack	102	Ack	449
	BTsk	3	BC	1	BC	1	BC	118
	CBeg	471	BTsk	1	BTsk	3	BTsk	21
	Chk	1	CBeg	64	CBeg	3	CBeg	5
	FEnd	23	CEnd	1	CEnd	2	CEnd	1
	PBeg	50	Chk	1	Chk	1	Chk	2
	Stl	1	FEnd	3	FEnd	19	FEnd	173
			Mod	9	Mod	8	Mod	1
			PBeg	13			PBeg	1
			Stl	12			Stl	2
yeah	Total	269	Total	118	Total	71	Total	372
	Ack	268	Ack	118	Ack	67	Ack	301
	FEnd	1			FEnd	4	BC	58
							FEnd	13
mm-hm	Total	12	Total	0	Total	1	Total	444
	Ack	9			Ack	1	Ack	48
	BC	2					BC	393
	FEnd	1					FEnd	3
uh-huh	Total	4	Total	0	Total	4	Total	109
	Ack	2			Ack	3	Ack	11
	BC	2			BC	1	BC	98
right	Total	31	Total	639	Total	485	Total	37
	Ack	19	Ack	10	Ack	13	Ack	32
	Mod	12	Chk	11	Chk	33	Chk	5
			Mod	618	Mod	439		

Table F.3: Distribution of ACWs and discourse/pragmatic functions per position in the conversational turn. See Figure 13.2 on page 105.

Appendix G

Instructions for the Perception Study of *Okay*

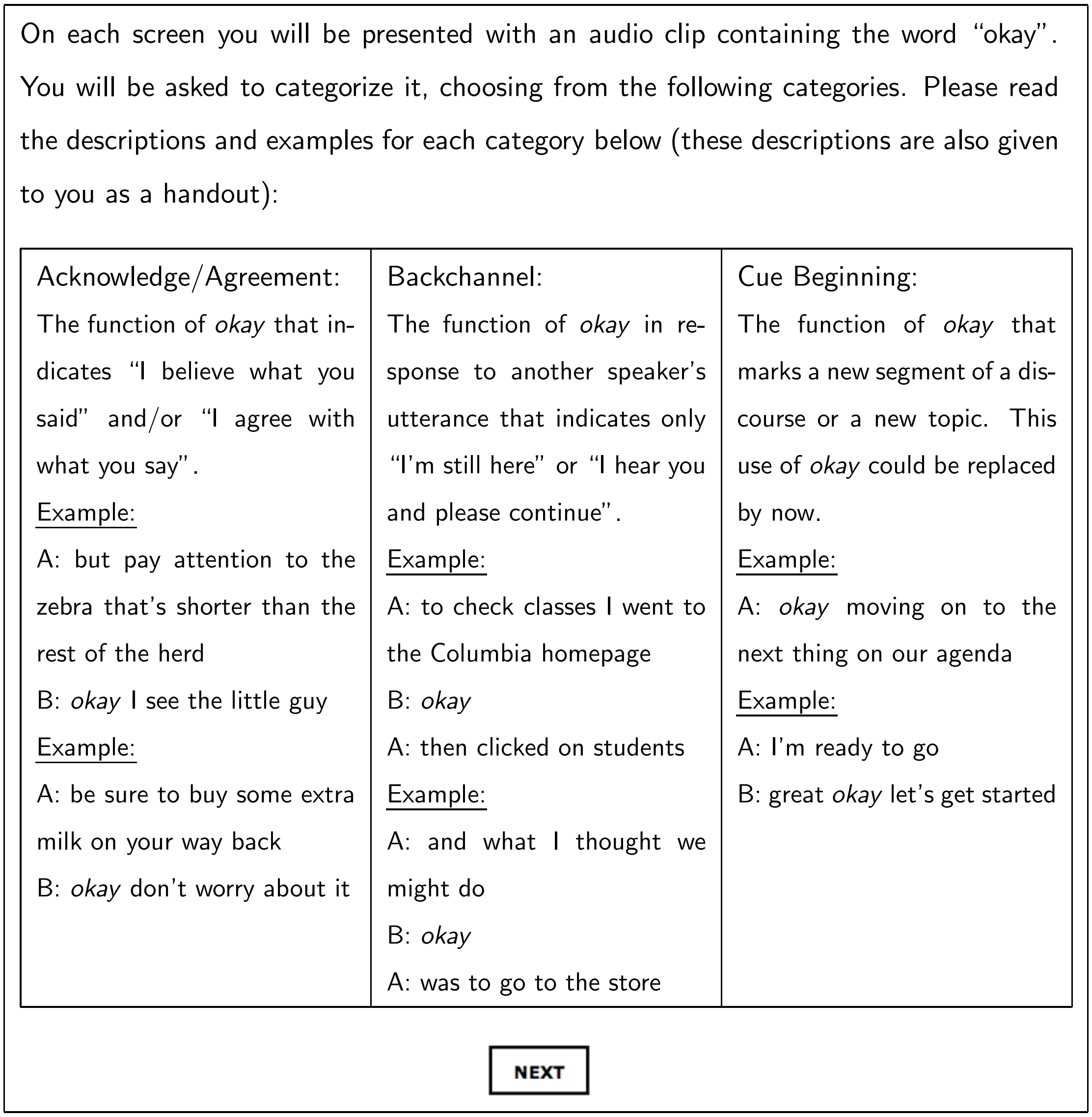


In this study, you will be given a series of single-word audio clips, one per screen. For each clip you will be asked to match the word you hear with the most appropriate category for that word.

Before viewing the category descriptions and further instructions, please check your audio now by clicking on the speaker icon below.

Check the audio playback capability and the volume setting by clicking on the speaker icon above. Ask the experimenter for assistance.

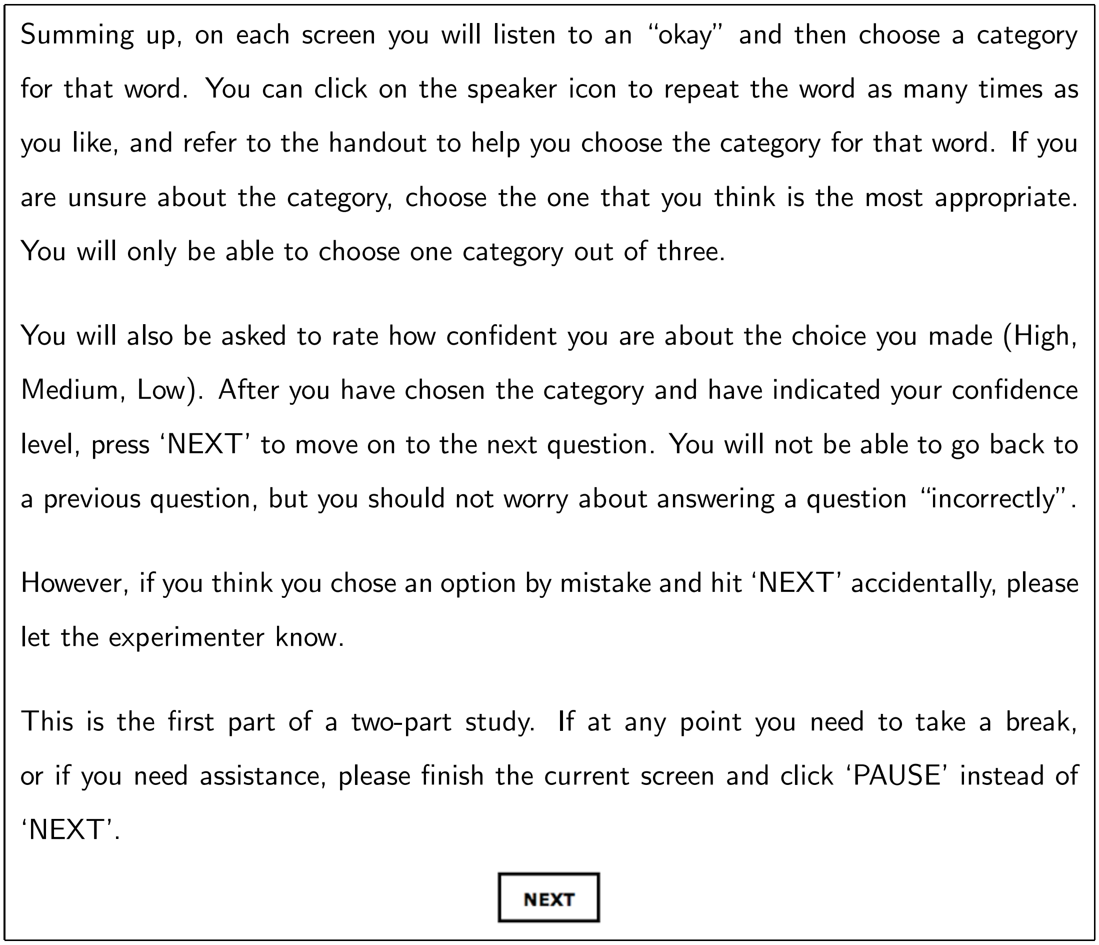
Figure G.1: First instructions screen for the first part (isolated condition) of the perception study of *okay*.



On each screen you will be presented with an audio clip containing the word "okay". You will be asked to categorize it, choosing from the following categories. Please read the descriptions and examples for each category below (these descriptions are also given to you as a handout):

Acknowledge/Agreement:	Backchannel:	Cue Beginning:
The function of *okay* that indicates "I believe what you said" and/or "I agree with what you say". Example: A: but pay attention to the zebra that's shorter than the rest of the herd B: *okay* I see the little guy Example: A: be sure to buy some extra milk on your way back B: *okay* don't worry about it	The function of *okay* in response to another speaker's utterance that indicates only "I'm still here" or "I hear you and please continue". Example: A: to check classes I went to the Columbia homepage B: *okay* A: then clicked on students Example: A: and what I thought we might do B: *okay* A: was to go to the store	The function of *okay* that marks a new segment of a discourse or a new topic. This use of *okay* could be replaced by now. Example: A: *okay* moving on to the next thing on our agenda Example: A: I'm ready to go B: great *okay* let's get started

Figure G.2: Second instructions screen for the first part (isolated condition) of the perception study of *okay*.



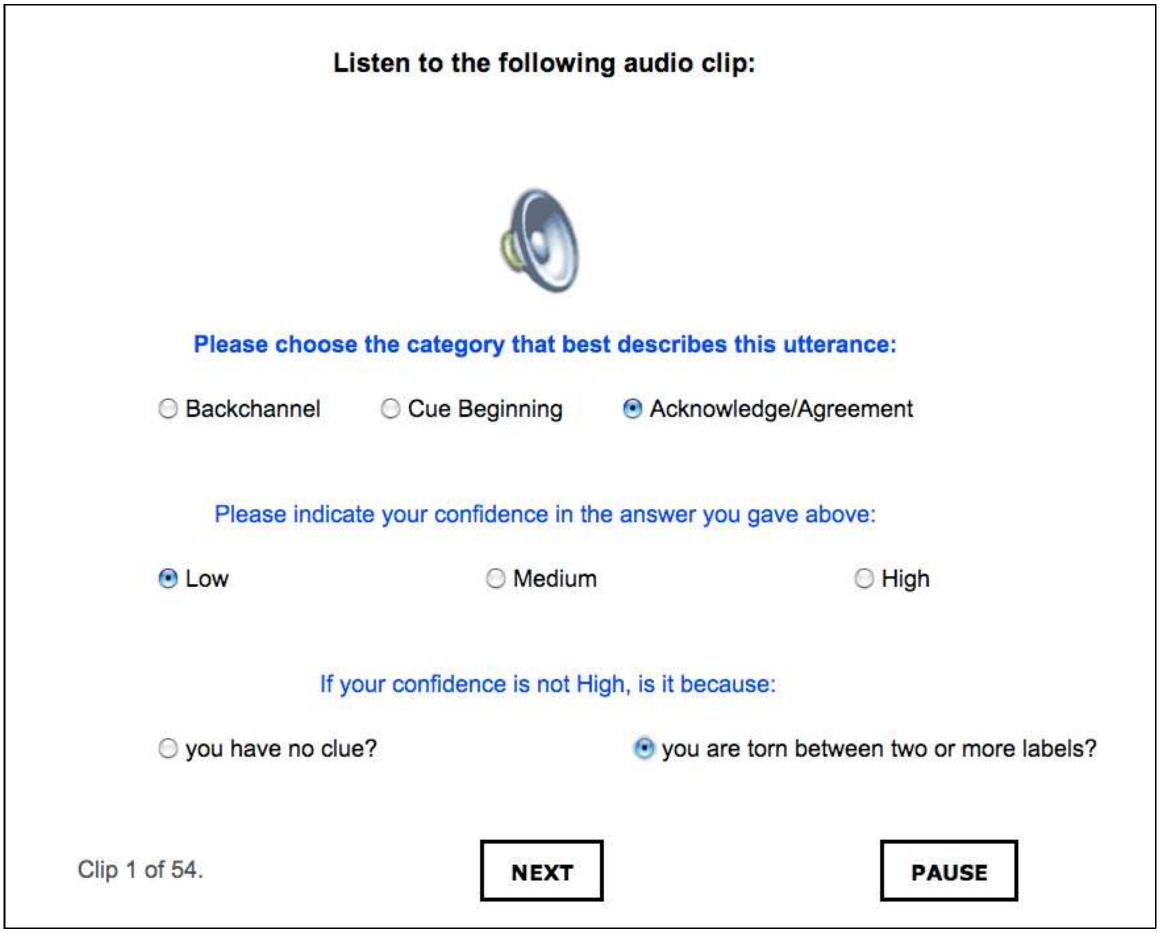
Summing up, on each screen you will listen to an "okay" and then choose a category for that word. You can click on the speaker icon to repeat the word as many times as you like, and refer to the handout to help you choose the category for that word. If you are unsure about the category, choose the one that you think is the most appropriate. You will only be able to choose one category out of three.

You will also be asked to rate how confident you are about the choice you made (High, Medium, Low). After you have chosen the category and have indicated your confidence level, press 'NEXT' to move on to the next question. You will not be able to go back to a previous question, but you should not worry about answering a question "incorrectly".

However, if you think you chose an option by mistake and hit 'NEXT' accidentally, please let the experimenter know.

This is the first part of a two-part study. If at any point you need to take a break, or if you need assistance, please finish the current screen and click 'PAUSE' instead of 'NEXT'.

Figure G.3: Third instructions screen for the first part (isolated condition) of the perception study of *okay*.



Listen to the following audio clip:

Please choose the category that best describes this utterance:

○ Backchannel ○ Cue Beginning ● Acknowledge/Agreement

Please indicate your confidence in the answer you gave above:

● Low ○ Medium ○ High

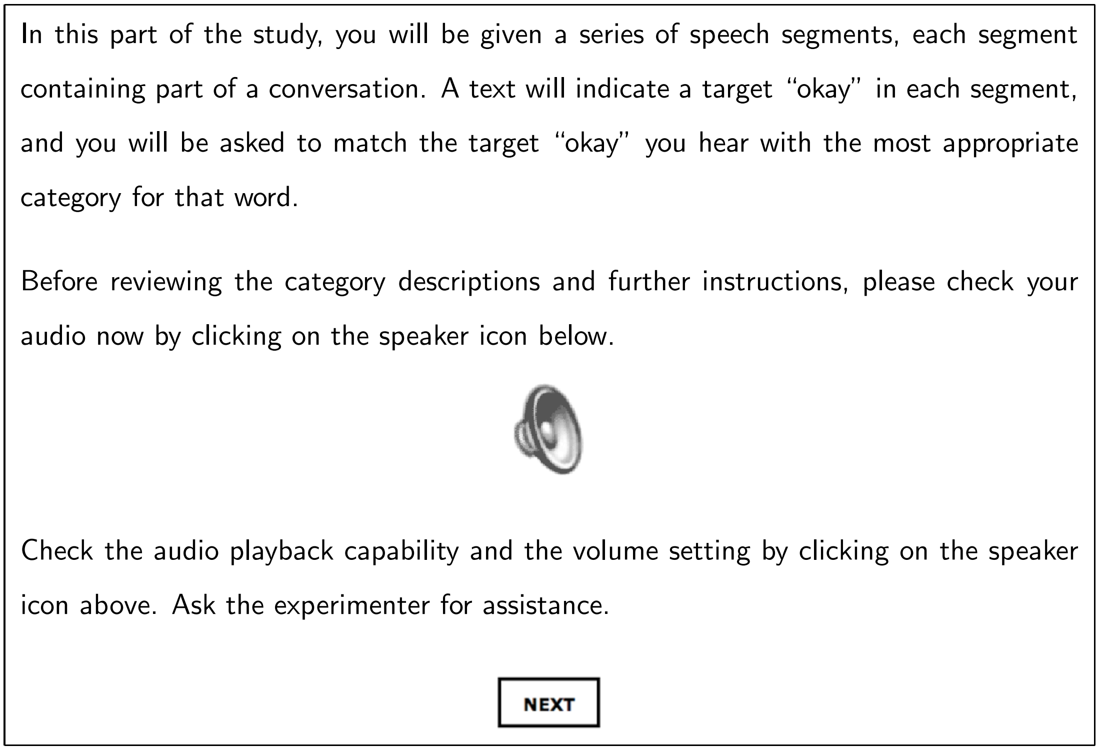
If your confidence is not High, is it because:

○ you have no clue? ● you are torn between two or more labels?

Clip 1 of 54.

Figure G.4: Sample screen of the first part (isolated condition) of the perception study of *okay*.

Note: The confidence rates were not used in the studies presented in this thesis.

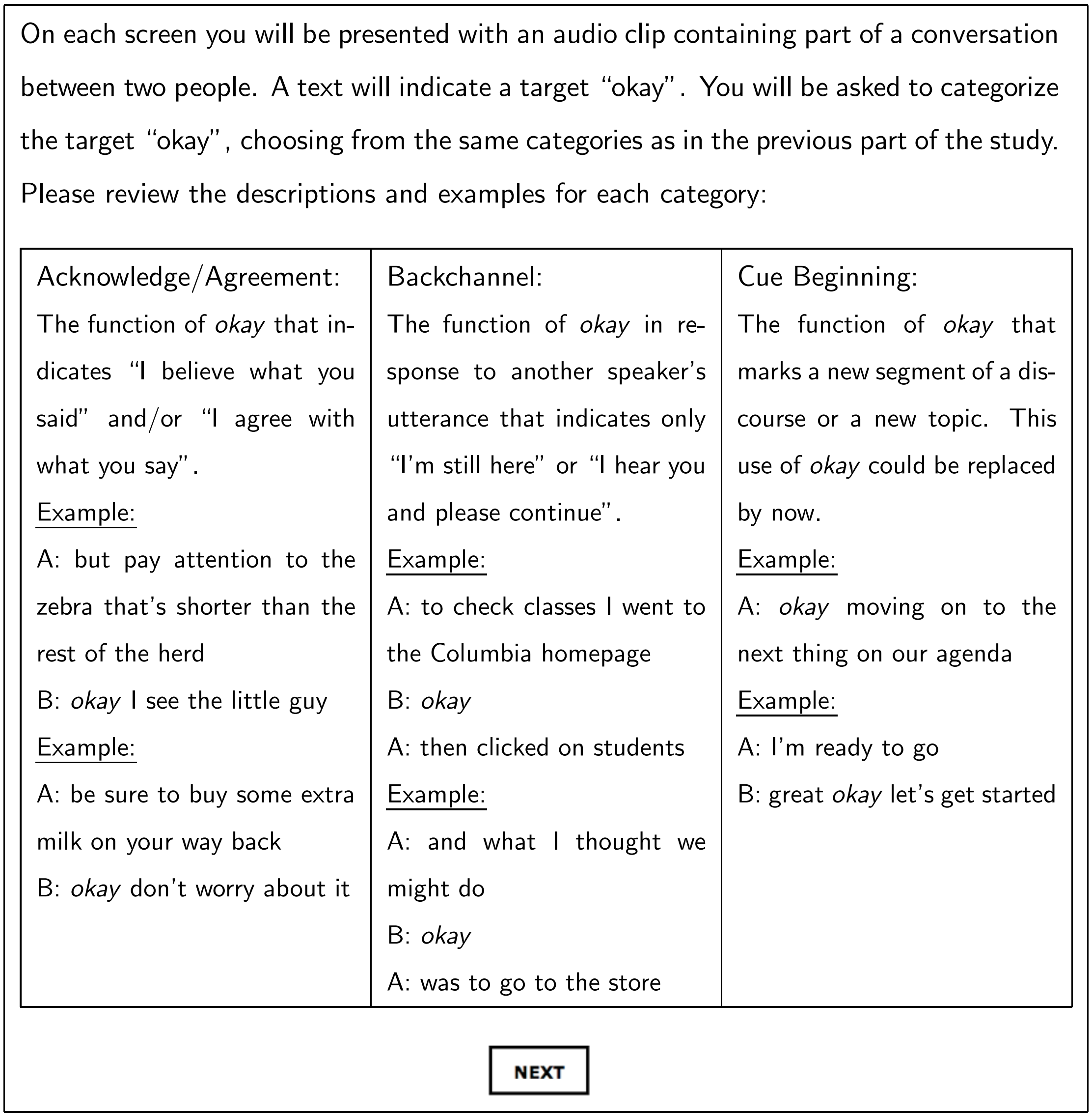


In this part of the study, you will be given a series of speech segments, each segment containing part of a conversation. A text will indicate a target "okay" in each segment, and you will be asked to match the target "okay" you hear with the most appropriate category for that word.

Before reviewing the category descriptions and further instructions, please check your audio now by clicking on the speaker icon below.

Check the audio playback capability and the volume setting by clicking on the speaker icon above. Ask the experimenter for assistance.

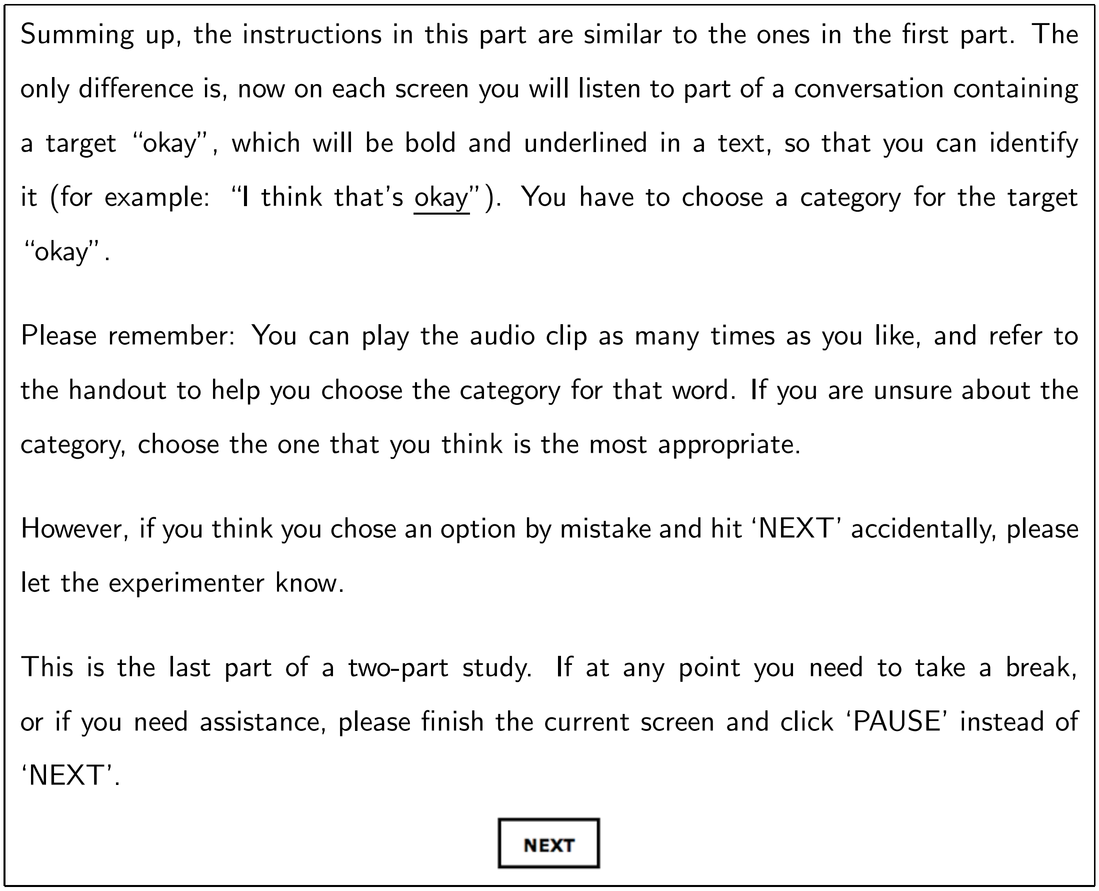
Figure G.5: First instructions screen for the second part (contextualized condition) of the perception study of *okay*.



On each screen you will be presented with an audio clip containing part of a conversation between two people. A text will indicate a target "okay". You will be asked to categorize the target "okay", choosing from the same categories as in the previous part of the study. Please review the descriptions and examples for each category:

Acknowledge/Agreement:	Backchannel:	Cue Beginning:
The function of *okay* that indicates "I believe what you said" and/or "I agree with what you say". Example: A: but pay attention to the zebra that's shorter than the rest of the herd B: *okay* I see the little guy Example: A: be sure to buy some extra milk on your way back B: *okay* don't worry about it	The function of *okay* in response to another speaker's utterance that indicates only "I'm still here" or "I hear you and please continue". Example: A: to check classes I went to the Columbia homepage B: *okay* A: then clicked on students Example: A: and what I thought we might do B: *okay* A: was to go to the store	The function of *okay* that marks a new segment of a discourse or a new topic. This use of *okay* could be replaced by now. Example: A: *okay* moving on to the next thing on our agenda Example: A: I'm ready to go B: great *okay* let's get started

Figure G.6: Second instructions screen for the second part (contextualized condition) of the perception study of *okay*.



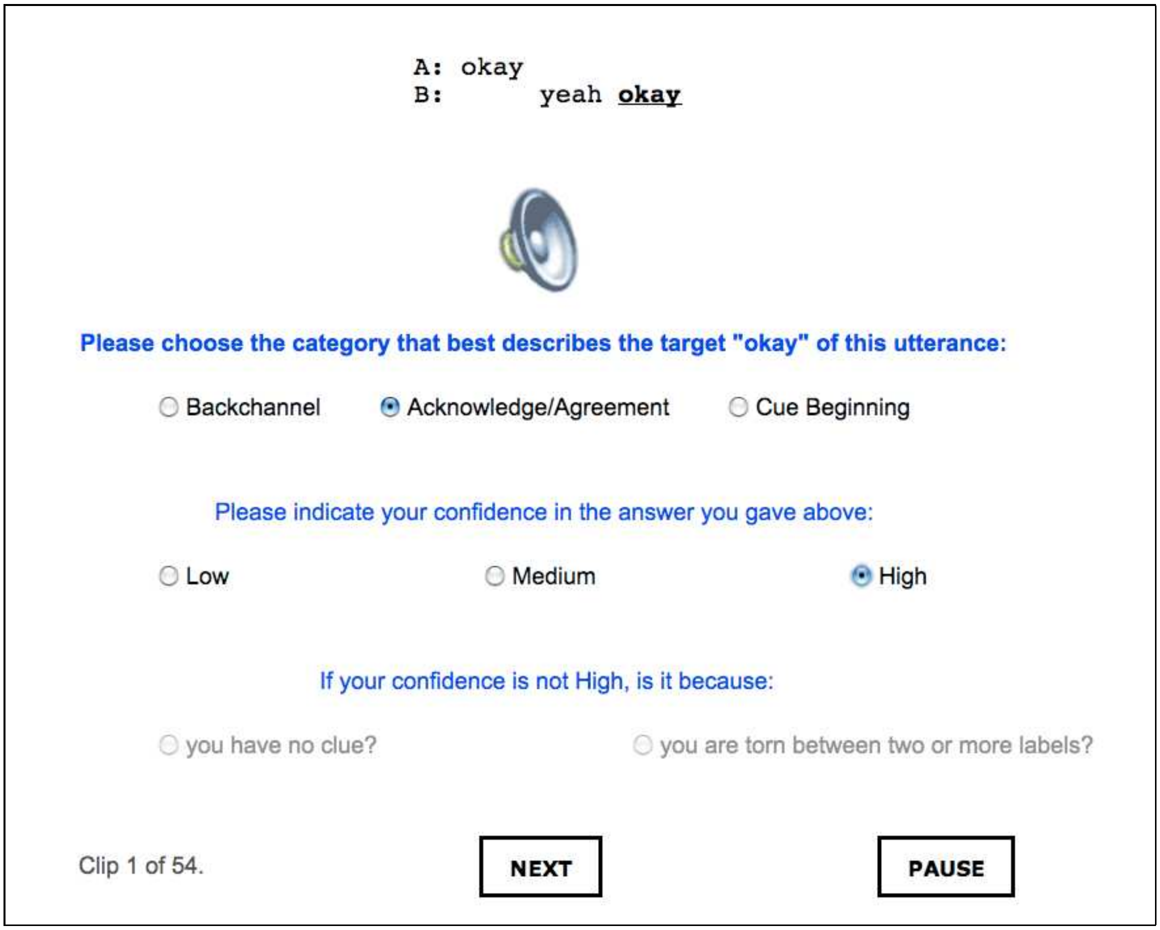
Summing up, the instructions in this part are similar to the ones in the first part. The only difference is, now on each screen you will listen to part of a conversation containing a target "okay", which will be bold and underlined in a text, so that you can identify it (for example: "I think that's okay"). You have to choose a category for the target "okay".

Please remember: You can play the audio clip as many times as you like, and refer to the handout to help you choose the category for that word. If you are unsure about the category, choose the one that you think is the most appropriate.

However, if you think you chose an option by mistake and hit 'NEXT' accidentally, please let the experimenter know.

This is the last part of a two-part study. If at any point you need to take a break, or if you need assistance, please finish the current screen and click 'PAUSE' instead of 'NEXT'.

Figure G.7: Third instructions screen for the second part (contextualized condition) of the perception study of *okay*.



```
A: okay
B:      yeah okay
```

Please choose the category that best describes the target "okay" of this utterance:

○ Backchannel ● Acknowledge/Agreement ○ Cue Beginning

Please indicate your confidence in the answer you gave above:

○ Low ○ Medium ● High

If your confidence is not High, is it because:

○ you have no clue? ○ you are torn between two or more labels?

Clip 1 of 54.

Figure G.8: Sample screen of the second part (contextualized condition) of the perception study of *okay*.

Note: The confidence rates were not used in the studies presented in this thesis.

Part VI

Bibliography

Bibliography

[Abney, 1996] Steven Abney. Partial parsing via finite-state cascades. *Journal of Natural Language Engineering*, 2(4):337–344, 1996.

[Atterer et al., 2008] Michaela Atterer, Timo Baumann, and David Schlangen. Towards incremental end-of-utterance detection in dialogue systems. In *Coling*, Manchester, UK, 2008.

[Baumann, 2008] Timo Baumann. Simulating spoken dialogue with a focus on realistic turn-taking. In *Proceedings of the 13th ESSLLI Student Session*, Hamburg, Germany, 2008.

[Beattie, 1981] G. Beattie. The regulation of speaker turns in face-to-face conversation; some implications for conversation in soundonly communication channels. *Semiotica*, 34:55–70, 1981.

[Beattie, 1982] Geoffrey W. Beattie. Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted. *Semiotica*, 39(1/2):93–114, 1982.

[Beckman and Edwards, 1990] M. E. Beckman and J. Edwards. Lengthening and shortening and the nature of prosodic constituency. In J. Kingston and M. E. Beckman, editors, *Laboratory Phonology I*, pages 152–178. Cambridge University Press, 1990.

[Beckman and Hirschberg, 1994] Mary E. Beckman and Julia Hirschberg. The ToBI annotation conventions. *Ohio State University*, 1994.

[Benus et al., 2007] Stefan Benus, Agustín Gravano, and Julia Hirschberg. The prosody of backchannels in American English. In *ICPhS*, 2007.

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

[Boersma and Weenink, 2001] Paul Boersma and David Weenink. Praat: Doing phonetics by computer. *http://www.praat.org*, 2001.

[Bohus and Rudnicky, 2003] D. Bohus and A.I. Rudnicky. RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. In *Eighth European Conference on Speech Communication and Technology*. ISCA, 2003.

[Bosch et al., 2005] Louis ten Bosch, Nelleke Oostdijk, and Lou Boves. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47:80–86, 2005.

[Brennan, 1996] S.E. Brennan. Lexical entrainment in spontaneous dialog. In *ISSD*, pages 41–44, 1996.

[Brown et al., 1980] G. Brown, K.L. Currie, and J. Kenworthy. *Questions of Intonation*. Routledge, 1980.

[Bull and Aylett, 1998] Matthew Bull and Matthew Aylett. An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In *ICSLP*, 1998.

[Charniak and Johnson, 2001] Eugene Charniak and Mark Johnson. Edit detection and parsing for transcribed speech. In *Proceedings of NAACL*, 2001.

[Chartrand and Bargh, 1999] TL Chartrand and JA Bargh. The chameleon effect: the perception-behavior link and social interaction. *J Pers Soc Psychol*, 76(6):893–910, 1999.

[Cohen, 1960] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[Cohen, 1984] Robin Cohen. A computational theory of the function of clue words in argument understanding. In *Proceedings of ACL*, 1984.

[Cohen, 1995] William C. Cohen. Fast effective rule induction. In *Machine Learning: Proceedings of the Twelfth International Conference*, 1995.

[Collins, 2003] Michael Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support vector networks. In *Machine Learning*, pages 273–297, 1995.

[Coulston et al., 2002] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *ICSLP'02*, 2002.

[Cutler and Pearson, 1986] E. A. Cutler and M. Pearson. On the analysis of prosodic turn-taking cues. In C. Johns-Lewis, editor, *Intonation in Discourse*, pages 139–156. College-Hill, San Diego, CA, 1986.

[Du Bois et al., 1993] J.W. Du Bois, S. Schuetze-Coburn, S. Cumming, and D. Paolino. Outline of discourse transcription. *Talking Data: Transcription and Coding in Discourse Research*, pages 45–89, 1993.

[Duncan and Fiske, 1977] S. Duncan and D.W. Fiske. *Face-To-Face Interaction: Research, Methods, and Theory.* Lawrence Erlbaum Associates, 1977.

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

[Duncan, 1973] S. Duncan. Toward a grammar for dyadic conversation. *Semiotica*, 9(1):29–46, 1973.

[Duncan, 1974] S. Duncan. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2):161–180, 1974.

[Duncan, 1975] S. Duncan. Interaction units during speaking turns in dyadic, face-to-face conversations. *Organization of Behavior in Face-to-Face Interaction, Mouton Publishers, Den Hague*, pages 199–213, 1975.

[Edlund et al., 2005] J. Edlund, M. Heldner, and J. Gustafson. Utterance segmentation and turn-taking in spoken dialogue systems. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, pages 576–587, 2005.

[Eskenazi et al., 1990] L. Eskenazi, DG Childers, and DM Hicks. Acoustic correlates of vocal quality. *Journal of Speech, Language and Hearing Research*, 33(2):298–306, 1990.

[Ferguson, 1977] N. Ferguson. Simultaneous speech, interruptions and dominance. *British Journal of Social and Clinical Psychology*, 16:295–302, 1977.

[Ferrer et al., 2002] L. Ferrer, E. Shriberg, and A. Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Proceedings of ICSLP*, pages 2061–2064, 2002.

[Ferrer et al., 2003] L. Ferrer, E. Shriberg, and A. Stolcke. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *Proceedings of ICASSP*, 2003.

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

[Ford and Thompson, 1996] C.E. Ford and S.A. Thompson. Interactional units in conversation: Syntactic, intonational and pragmatic resources for the management of turns. In E.A. Schegloff, E. Ochs, E.A. Schegloff, and S.A. Thompson, editors, *Interaction and Grammar*, pages 134–184. Cambridge University Press, 1996.

[Fraser, 1999] B. Fraser. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952, 1999.

[Fry, 1975] DB Fry. Simple reaction-times to speech and non-speech stimuli. *Cortex*, 11(4):355–60, 1975.

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

[Godfrey et al., 1992] JJ Godfrey, EC Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 1992.

[Goleman, 2006] Daniel Goleman. *Social Intelligence: The New Science of Human Relationships.* Bantam, 2006.

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

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

[Heckerman et al., 1995] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[Hirschberg and Litman, 1987] Julia Hirschberg and Diane Litman. Now let's talk about now: Identifying cue phrases intonationally. In *Proceedings of ACL*, 1987.

[Hirschberg and Litman, 1993] Julia Hirschberg and Diane Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19:3, 1993.

[Hirschberg and Nakatani, 1996] Julia Hirschberg and Christine Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of ACL*, pages 286–293, 1996.

[Hirschberg, 1990] J. Hirschberg. Accent and discourse context: Assigning pitch accent in synthetic speech. *Proceedings of the Eighth National Conference on Artificial Intelligence*, 2:952–957, 1990.

[Hockey, 1991] Beth Ann Hockey. Prosody and the interpretation of 'okay'. In *Working Notes of the AAAI Fall Symposium*, 1991.

[Hockey, 1992] Beth Ann Hockey. Prosody and the role of okay and uh-huh in discourse. In M. Bernstein, editor, *Eastern States Conference on Linguistics*, pages 128–136, 1992.

[Jensen, 1996] F.V. Jensen. *Introduction to Bayesian Networks*. Springer-Verlag, New York, 1996.

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

[Kendon, 1972] A. Kendon. Some relationships between body motion and speech. In A. W. Siegman and B. Pope, editors, *Studies in Dyadic Communication*, pages 177–210. Pergamon Press, Elmsford, NY, 1972.

[Kitch et al., 1996] J.A. Kitch, J. Oates, and K. Greenwood. Performance effects on the voices of 10 choral tenors: Acoustic and perceptual findings. *Journal of Voice*, 10(3):217–227, 1996.

[Koehn et al., 2000] Philipp Koehn, Steven Abney, Julia Hirschberg, and Michael Collins. Improving intonational phrasing with syntactic information. In *Proceedings of ICASSP*, volume 3, pages 1289–1290, 2000.

[Kowtko, 1997] Jacqueline C. Kowtko. *The function of intonation in task-oriented dialogue*. PhD thesis, University of Edinburgh, 1997.

[Lafferty et al., 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning*, pages 282–289, San Francisco, CA, 2001. Morgan Kaufmann.

[Lai, 2008] Catherine Lai. Prosodic cues for backchannels and short questions: Really? In *Proceedings of the Fourth Conference on Speech Prosody*, 2008.

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

[Litman and Hirschberg, 1990] Diane Litman and Julia Hirschberg. Disambiguating cue phrases in text and speech. In *Proceedings of COLING*, 1990.

[Litman and Passonneau, 1995] D.J. Litman and R.J. Passonneau. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of ACL*, pages 108–115, 1995.

[Litman, 1994] Diane Litman. Classifying cue phrases in text and speech using machine learning. In *Eleventh National Conference on Artificial Intelligence - AAAI*, 1994.

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

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

[McNeill, 1992] David McNeill. *Hand and Mind: what gestures reveal about thought.* University of Chicago Press, 1992.

[Nenkova et al., 2008] A. Nenkova, A. Gravano, and J. Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL/HLT*, 2008.

[Ogden, 2002] Richard Ogden. Creaky voice and turn-taking in finnish. In *Colloquium of the British Association of Audiological Physicians*, 2002.

[Ogden, 2004] R. Ogden. Non-modal voice quality and turn-taking in Finnish. *Sound Patterns In Interaction: Cross-Linguistic Studies From Conversation*, 2004.

[Pickering and Garrod, 2004] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

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

[Pierrehumbert, 1980] J.B. Pierrehumbert. *The phonology and phonetics of English intonation.* PhD thesis, Massachusetts Institute of Technology, 1980.

[Pitrelli et al., 1994] John F. Pitrelli, Mary E. Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of ICSLP*, pages 123–126, 1994.

[Price et al., 1991] P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90:2956, 1991.

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

[Rabiner, 1989] L.R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Ratnaparkhi et al., 1996] A. Ratnaparkhi, E. Brill, and K. Church. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, 1996.

[Raux and Eskenazi, 2008] Antoine Raux and Maxine Eskenazi. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *SIGdial*, Columbus, OH, 2008.

[Raux et al., 2006] Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In *Proceedings of Interspeech*, 2006.

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

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

[Reitter and Moore, 2007] D. Reitter and J. Moore. Predicting success in dialogue. In *Proceedings of ACL*, 2007.

[Reitter et al., 2006] D. Reitter, F. Keller, and J.D. Moore. Computational modelling of structural priming in dialogue. In *Proceedings of HLT/NAACL*, 2006.

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

[Schaffer, 1983] Deborah Schaffer. The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, 11:243–257, 1983.

[Schiffrin, 1987] Deborah Schiffrin. *Discourse Markers*. Cambridge University Press, Cambridge, England, 1987.

[Schlangen, 2006] David Schlangen. From reaction to prediction: Experiments with computational models of turn-taking. In *Proceedings of Interspeech*, 2006.

[Shriberg et al., 2001] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. EUROSPEECH*, 2:1359–1362, 2001.

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

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

[Ward and Litman, 2007] Arthur Ward and Diane Litman. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *SLaTE Workshop on Speech and Language Technology in Education*, Farmington, PA, 2007. ISCA.

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

[Ward et al., 2005] N.G. Ward, A.G. Rivera, K. Ward, and D.G. Novick. Root causes of lost time and user stress in a simple dialog system. In *Interspeech*, 2005.

[Wennerstrom and Siegel, 2003] A. Wennerstrom and A. F. Siegel. Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36(2):77–107, 2003.

[Wichmann and Caspers, 2001] Anne Wichmann and Johanneke Caspers. Melodic cues to turn-taking in English: evidence from perception. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.

[Wightman et al., 1992] C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91:1707, 1992.

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

[Yngve, 1970] V.H. Yngve. On getting a word in edgewise. *Sixth Regional Meeting of the Chicago Linguistic Society*, 6:657–677, 1970.

[Young et al., 2006] S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK Book, version 3.4, 2006.

[Yuan et al., 2007] Jiahong Yuan, Mark Liberman, and Christopher Cieri. Towards an integrated understanding of speech overlaps in conversation. In *ICPhS XVI*, Saarbrücken, Germany, 2007.

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