On the role of context and prosody in the interpretation of ‘okay’

Abstract

We examine the effect of contextual and acoustic cues in the disambiguation of three discourse-pragmatic functions of the word okay. Results of a perception study show that contextual cues are stronger predictors of discourse function than acoustic cues. However, acoustic features capturing the pitch excursion at the right edge of okay feature prominently in disambiguation, whether other contextual cues are present or not.

1 Introduction

CUE PHRASES (also known as DISCOURSE MARKERS) are linguistic expressions that can be used to convey explicit information about the structure of a discourse or to convey a semantic contribution (Grosz and Sidner, 1986; Reichman, 1985; Cohen, 1984). For example, the word okay can be used to convey a ‘satisfactory’ evaluation of some entity in the discourse (the movie was okay); as a backchannel in a dialogue to indicate that one interlocutor is still attending to another; to convey acknowledgment or agreement; or, in its ‘cue’ use, to start or finish a discourse segment (Jefferson, 1972; Schegloff and Sacks, 1973; Kowtko, 1997; Ward and Tsukahara, 2000). A major question is how speakers indicate and listeners interpret such variation in meaning. From a practical perspective, understanding how speakers and listeners disambiguate cue phrases is important to spoken dialogue systems, so that systems can convey potentially ambiguous terms with their intended meaning and can interpret user input correctly.

There is considerable evidence that the different uses of individual cue phrases can be distinguished by variation in the prosody with which they are realized. For example, (Hirschberg and Litman, 1993) found that cue phrases in general could be disambiguated between their ‘semantic’ and their ‘dis- course marker’ uses in terms of the type of pitch accent borne by the cue phrase, the position of the phrase in the intonational phrase, and the amount of additional information in the phrase. Despite the frequency of the word okay in natural dialogues, relatively little attention has been paid to the relationship between its use and its prosodic realization. (Hockey, 1993) did find that okay differs in terms of the pitch contour speakers use in uttering it, suggesting that a final rising pitch contour “categorically marks a turn change,” while a downstepped falling pitch contour usually indicates a discourse segment boundary. However, it is not clear which, if any, of the prosodic differences identified in this study are actually used by listeners in interpreting these potentially ambiguous items.

In this study, we address the question of how hearers disambiguate the interpretation of okay. Our goal is to identify the acoustic, prosodic and phonetic features of okay tokens for which listeners assign different meanings. Additionally, we want to determine the role that discourse context plays in this classification: i.e., can subjects classify okay tokens reliably from the word alone or do they require additional context?

Below we describe a perception study in which listeners were presented with a number of spoken productions of okay, taken from a corpus of dialogues between subjects playing a computer game. The tokens were presented both in isolation and in context. Users were asked to select the mean-
ing of each token from three of the meanings that okay can take on: ACKNOWLEDGEMENT / AGREEMENT, BACKCHANNEL, and CUE OF AN INITIAL DISCOURSE SEGMENT. Subsequently, we examined the acoustic, prosodic and phonetic correlates of these classifications to try to infer what cues listeners used to interpret the tokens, and how these varied by context condition. Section 2 describes our corpus. Section 3 describes the perception experiment. In Section 4 we analyze inter-subject agreement, introduce a novel representation of subject judgments, and examine the acoustic, prosodic, phonetic and contextual correlates of subject classification of okays. In Section 5 we discuss our results and future work.

2 Corpus

The materials for our perception study were selected from a portion of the X Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from speakers of Standard American English. The corpus was collected and annotated jointly by the Spoken Language Group at X University and the Department of Linguistics at Y University.

Subjects were paid to play two series of computer games (the CARDS GAMES and the OBJECTS GAMES), requiring collaboration between partners to achieve a common goal. Participants sat in front of laptops in a soundproof booth with a curtain between them, so that all communication would be verbal. Each player played with two different partners in two different sessions. On average, each session took 45m 39s, totalling 9h 8m of dialogue for the whole corpus. All interactions were recorded, digitized, and downsampled to 16K.

The recordings were orthographically transcribed and words were aligned by hand by trained annotators in a ToBI (Beckman and Hirschberg, 1994) orthographic tier using Praat (Boersma and Weenink, 2001) to manipulate waveforms. The corpus contains 2239 unique words, with 73,831 words in total. Nearly all of the Objects Games part of the corpus has been intonationally transcribed, using the ToBI conventions. Pitch, energy and duration information has been extracted for the entire corpus automatically, using Praat.

In the Objects Games portion of the corpus each player’s laptop displayed a gameboard containing 5–7 objects (Figure 1). In each segment of the game, both players saw the same set of objects at the same position on each screen, except for one object (the TARGET). For one player (the DESCRIBER), this target appeared in a random location among other objects on the screen. For the other player (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 their screen so that the follower could move their representation of the target to the same location on their own screen. After the players had negotiated what they determined to be the best location match, they were awarded up to 100 points based on the actual match of the target location on the two screens. The game proceeded in this way through 14 tasks, with describer and follower alternating roles. On average, the Objects Games portion of each session took 21m 36s, resulting in 4h 19m of dialogue for the twelve sessions in the corpus. There are 1484 unique words in this portion of the corpus, and 36,503 words in total.

Figure 1: Sample screen of the Objects Games.

Throughout the Objects Games, we noted that subjects made frequent use of cue phrases, such as okay, yeah, alright, which appeared to vary in meaning. To investigate the discourse functions of affirmative words, we first asked three labelers to independently classify all occurrences of alright, gotcha, huh, mmhm, okay, right, uhhuh, yeah, yep, yes, yup in the entire Games Corpus into one of 10 categories:
Labelers were asked to choose the most appropriate category for each token, or indicate with ‘?’ if they could not make a decision. They were allowed to read the transcripts and listen to the speech as they labeled.

For our perception experiment we chose materials from the tokens of the most frequent of our labeled affirmative words, okay, from the Objects Games, which contained most of these tokens. Altogether, there are 1151 instances of okay in this part of the corpus; it is the third most frequent word, following the, with 4565 instances, and of, with 1534. At least two labelers agreed on the functional category of 902 (78%) okay tokens. Of those tokens, 286 (32%) were classified as BACKCHANNEL, 255 (28%) as ACKNOWLEDGE / AGREEMENT, 141 (16%) as CUE BEGINNING, 116 (13%) as PIVOT BEGINNING, and 104 (11%) as one of the other functions. We sampled from tokens the annotators had labeled as Cue beginning discourse segment, Backchannel, and Acknowledgement / agreement, the most frequent categories in the corpus; we will refer to these below simply as ‘C’, ‘B’, and ‘A’ classes, respectively.

3 Experiment

We next designed a perception experiment to examine naive subjects’ perception of these tokens of okay. To obtain good coverage both of the (labeled) A, B, and C classes, as well as the degrees of potential ambiguity among these classes, we identified 9 categories of okay tokens to include in the experiment: 3 classes (A, B, C) x 3 levels of labeler agreement (UNANIMOUS, MAJORITY, NO-AGREEMENT). ‘Unanimous’ refers to tokens assigned to a particular class label by all 3 labelers, ‘majority’ to tokens assigned to this class by 2 of the 3 labelers, and ‘no-agreement’ to tokens assigned to this class by only 1 labeler. To decrease variability in the stimuli, we selected tokens only from speakers who produced at least one token for each of the 9 conditions. There were 6 such speakers (3 female, 3 male), which gave us a total of 54 tokens.

To see whether subjects’ classifications of okay were dependent upon contextual information or not, we prepared two versions of each token. The isolated versions consisted of only the word okay extracted from the waveform. For the contextualized versions, we extracted two full speaker turns for each okay including the full turn\(^1\) containing the target okay plus the full turn of the previous speaker. The isolated okay tokens were single channel audio files; the contextualized okay tokens were formatted so that each speaker was presented to subjects on a different channel, with the speaker uttering the target okay consistently on the same channel.

The perception study was divided into two parts. In the first part, each subject was presented with the 54 isolated okay tokens, in a different random ordering for each subject. They were given a forced choice task to classify them as A, B, or C, with the corresponding labels (Acknowledgement / agreement, Backchannel, and Cue beginning) also presented in a random order for each token. In the second part, the same subject was 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 were native speakers of Standard American English, except for one subject who was born in Jamaica but a native speaker of English. All subjects reported no hearing problems. 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 sections of the study.

For the study itself, for each token in the isolated condition, subjects were shown a screen with the three randomly ordered classes and a link to the token’s waveform. They could listen to the sound files as many times as they wished but were instructed

\(^1\)We define a TURN as a maximal sequence of words spoken by the same speaker during which the speaker holds the floor.
not to be concerned with answering the questions “correctly”, but to answer with their immediate re-
sponse if possible. However, they were allowed 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 ortho-
graphic transcription of part of the contextualized to-
ken, to help them identify the target okay. The mean
duration of the first part of the study was 25 minutes,
and of the second part, 27 minutes.

4 Results

4.1 Subject ratings

The distribution of class labels in each experimental
condition is shown in Table 1. While this distribu-
tion 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 A
(acknowledgment / agreement) than as B (backchan-
nel) or C (cue to topic beginning). This supports the
hypothesis that acknowledgment / agreement may
function as the default interpretation of okay.

<table>
<thead>
<tr>
<th>Isolated</th>
<th>Contextualized</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 426 (39%)</td>
<td>452 (42%)</td>
</tr>
<tr>
<td>B 324 (30%)</td>
<td>306 (28%)</td>
</tr>
<tr>
<td>C 330 (31%)</td>
<td>322 (30%)</td>
</tr>
<tr>
<td>Total 1080 (100%)</td>
<td>1080 (100%)</td>
</tr>
</tbody>
</table>

Table 1: Distribution of label classes in each study condition.

We examined inter-subject agreement using Fleiss’ $\kappa$ measure of inter-rater agreement for mul-
tiple raters (Fleiss, 1971). Table 2 shows Fleiss’ $\kappa$
calculated for each individual label vs. the other two
labels and for all three labels, in both study condi-
tions. From this table we see that, while there is very
little overall agreement among subjects about how
to classify tokens in the isolated condition, agree-
ment is higher in the contextualized condition, with
a moderate agreement for class C ($\kappa$ score of .497).
This suggests that context helps distinguish the cue
beginning discourse segment function more than the
other two functions of okay.

<table>
<thead>
<tr>
<th></th>
<th>Isolated</th>
<th>Contextualized</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs. rest</td>
<td>.089</td>
<td>.227</td>
</tr>
<tr>
<td>B vs. rest</td>
<td>.118</td>
<td>.164</td>
</tr>
<tr>
<td>C vs. rest</td>
<td>.157</td>
<td>.497</td>
</tr>
<tr>
<td>all</td>
<td>.120</td>
<td>.293</td>
</tr>
</tbody>
</table>

Table 2: Fleiss’ $\kappa$ for each label class
in each study condition.

Recall from Section 3 that the okay tokens were
chosen in equal numbers from three classes accord-
ing to the level of agreement of our three original
labelers (unanimous, majority, and no-agreement),
who had the full dialogue context to use in making
their decisions. Table 3 shows Fleiss’ $\kappa$
measure now grouped by amount of agreement of the orig-
inal labelers, again presented for each context con-
dition. We see here that the inter-subject agreement

<table>
<thead>
<tr>
<th></th>
<th>Isolated</th>
<th>Context.</th>
<th>OL</th>
</tr>
</thead>
<tbody>
<tr>
<td>no-agreement</td>
<td>.085</td>
<td>.104</td>
<td>-</td>
</tr>
<tr>
<td>majority</td>
<td>.092</td>
<td>.299</td>
<td>-</td>
</tr>
<tr>
<td>unanimous</td>
<td>.158</td>
<td>.452</td>
<td>-</td>
</tr>
<tr>
<td>all</td>
<td>.120</td>
<td>.293</td>
<td>.312</td>
</tr>
</tbody>
</table>

Table 3: Fleiss’ $\kappa$ in each study condition, grouped
by agreement of the three original labelers (‘OL’).

also mirrors the agreement of the three original la-
belers. In both study conditions, tokens which the
original labelers agreed on also had the highest $\kappa$
scores, followed by tokens in the majority and no-
agreement classes, in that order. In all cases, tokens
which subjects heard in context showed more agree-
ment than those they heard in isolation.

The overall $\kappa$ is small at .120 for the isolated con-
dition, and fair at .293 for the contextualized con-
dition. The three original labelers also achieved fair
agreement at .312. The similarity between the lat-
ter two $\kappa$ scores suggests that the full context avail-
able to the original labelers and the limited context
presented to the experiment subjects offer compar-
able amounts of information to disambiguate between
the three functions, although lack of any context
clearly affected subjects’ decisions. We conclude

\footnote{This measure of agreement above chance is interpreted as
follows: 0 = None, 0.2 - 0.4 = Small, 0.2 - 0.4 = Fair, 0.4 - 0.6 =
Moderate, 0.6 - 0.8 = Substantial, 0.8 - 1 = Almost perfect.}

\footnote{For the calculation of this $\kappa$, we considered four label
classes: A, B, C, and a fourth class ‘other’ that comprises the
remaining 7 word functions mentioned in Section 2. In conse-
quence, these $\kappa$ scores should be compared with caution.}
from these results that context is of considerable importance in the interpretation of the word okay, although even a very limited context appears to suffice.

4.2 Representing subject judgments

In this section, we present a graphical representation of subject decisions, useful for interpreting, visualizing, and comparing the way our subjects interpreted the different tokens of okay. For each individual okay in the study, we define an associated three-dimensional vote vector, whose components are the proportions of subjects that classified the token as A, B or C. For example, if a particular okay was labeled as A by 5 subjects, as B by 3, and as C by 12, then its associated vote vector is \((\frac{5}{20}, \frac{3}{20}, \frac{12}{20}) = (0.25, 0.15, 0.6)\). Following this definition, the vectors \(A = (1, 0, 0)\), \(B = (0, 1, 0)\) and \(C = (0, 0, 1)\) correspond to the ideal situations in which all 20 subjects agreed on the label. We call these vectors the **unanimous-vote vectors**.

Figure 2.i shows a two-dimensional representation that illustrates these definitions. The black dot represents the vote vector for our example okay, the vertices of the triangle correspond to the three unanimous-vote vectors (A, B and C), and the cross in the center of the triangle represents the vote vector of a three-way tie between the labelers \((\frac{1}{3}, \frac{1}{3}, \frac{1}{3})\).

We are thus able to calculate the Euclidean distance of a vote vector to each of the unanimous-vote vectors. The shortest of these distances corresponds to the label assigned by the plurality of subjects.

Also, the smaller that distance, the higher the inter-subject agreement for that particular token. For our example okay, the distances to \(A\), \(B\) and \(C\) are 0.972, 1.070 and 0.495, respectively; its plurality label is C.

In our experiment, each okay has two associated vote vectors, one for each context condition. To illustrate the relationship between decisions in the isolated and the contextualized conditions, we first grouped each condition’s 54 vote vectors into three clusters, according to their plurality label. Figure 2.ii shows the clustering centroids in a two-dimensional representation of vote vectors. The filled dots correspond to the cluster centroids of the isolated condition, and the empty dots, to the centroids of the contextualized condition. Table 4 shows the distances in each condition from the cluster centroids (denoted \(A_c\), \(B_c\), \(C_c\)) to the respective unanimous-vote vectors (\(A\), \(B\), \(C\)), and also the distance between each pair of cluster centroids.

<table>
<thead>
<tr>
<th></th>
<th>Isolated</th>
<th>Contextualized</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d(A_c, A))</td>
<td>.54</td>
<td>.44 (–18%)</td>
</tr>
<tr>
<td>(d(B_c, B))</td>
<td>.57</td>
<td>.52 (–10%)</td>
</tr>
<tr>
<td>(d(C_c, C))</td>
<td>.52</td>
<td>.28 (–47%)</td>
</tr>
<tr>
<td>(d(A_c, B_c))</td>
<td>.41</td>
<td>.48 (+17%)</td>
</tr>
<tr>
<td>(d(A_c, C_c))</td>
<td>.49</td>
<td>.86 (+75%)</td>
</tr>
<tr>
<td>(d(B_c, C_c))</td>
<td>.54</td>
<td>.91 (+69%)</td>
</tr>
</tbody>
</table>

Table 4: Distances from the cluster centroids \((A_c, B_c, C_c)\) to the unanimous-vote vectors \((A, B, C)\) and between cluster centroids, in each condition.

In the isolated condition, the three cluster centroids are approximately equidistant from each other—that is, the three word functions appear to be equally confusable. In the contextualized condition, while \(C_c\) is further apart from the other two centroids, the distance between \(A_c\) and \(B_c\) remains practically the same. This suggests that, with some context available, A and B tokens are still fairly confusable, while both are more easily distinguished from C tokens. We posit two possible explanations for this: First, C is the only function for which the speaker uttering the okay necessarily continues speaking; thus the role of context in disambiguating seems quite clear. Second, both A and B have a common element of ‘acknowledgement’ that might affect inter-subject agreement.
4.3 Features of the okay tokens

In this section, we describe a set of acoustic, prosodic, phonetic and contextual features which may help to explain why subjects interpret okay differently. Acoustic features were extracted automatically using Praat. Phonetic and prosodic features were hand-labeled by expert annotators. Contextual features were considered only in the analysis of the contextualized condition, since they were not available to subjects in the isolated condition.

We examined a number of phonetic features to determine whether these correlated with subject classifications. We first looked at the production of the three phonemes in the target okay (/oul/, /kl/, /el/), noting the following possible variations:

- /oul/: [l, [a], [y], [o], [u], [m], [n], [e], [o].
- /kl/: [y], [k], [kx], [q], [x].
- /el/: [e], [et], [f], [e].

We also calculated the duration of each phone and of the velar closure. Whether the target okay was at least partially whispered or not, and whether there was glottalization in the target okay were also noted.

For each target okay, we also examined its duration and its maximum, mean and minimum pitch and intensity, as well as the speaker-normalized versions of these values. We considered its pitch slope, intensity slope, and stylized pitch slope, calculated over the whole target okay, its last 50, 80 and 100 milliseconds, its second half, its second syllable, and the second half of its second syllable, as well.

We used the ToBI labeling scheme (Pitrelli et al., 1994) to label the prosody of the target okays and their surrounding context.

- Pitch accent, if any, of the target okay (e.g., H*, H+!H*, L*).
- Break index after the target okay (0-4).
- Phrase accent and boundary tone, if any, following the target okay (e.g., L-L%, !H-H%).

For contextualized tokens, we included several features related to the exchange between the speaker uttering the target okay (Speaker A) and the other speaker (Speaker B).

- Number of words uttered by Speaker A in the context, before and after the target okay. Same for Speaker B.
- Latency of Speaker A before Speaker B’s turn.
- Duration of silence of Speaker B before and after the target okay.
- Duration of speech by Speaker B immediately before and after the target okay and up to a silence.

4.4 Cues to interpretation

We conducted a series of Pearson’s tests to look for correlations between the proportion of subjects that chose each label and the numeric features described in Section 4.3, together with two-sided t-tests to find whether such correlations differed significantly from zero. Tables 5 and 6 show the significant results (two-sided t-test, p < 0.05) for the isolated and contextualized conditions, respectively.

<table>
<thead>
<tr>
<th>Acknowledgement / agreement</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>length of realization of /kl/</td>
<td>-0.299</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backchannel</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>stylized pitch slope over 2nd half 2nd syllable</td>
<td>0.752</td>
</tr>
<tr>
<td>pitch slope over 2nd half of 2nd syllable</td>
<td>0.409</td>
</tr>
<tr>
<td>speaker-normalized maximum intensity</td>
<td>-0.372</td>
</tr>
<tr>
<td>pitch slope over last 80 ms</td>
<td>0.349</td>
</tr>
<tr>
<td>speaker-normalized mean intensity</td>
<td>-0.327</td>
</tr>
<tr>
<td>length of realization of /el/</td>
<td>0.278</td>
</tr>
<tr>
<td>length</td>
<td>0.277</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cue to discourse segment beginning</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>stylized pitch slope over the whole word</td>
<td>-0.380</td>
</tr>
<tr>
<td>pitch slope over the whole word</td>
<td>-0.342</td>
</tr>
<tr>
<td>pitch slope over 2nd half of 2nd syllable</td>
<td>-0.319</td>
</tr>
</tbody>
</table>

Table 5: Features correlated to the proportion of votes for each label. Isolated condition.

Table 5 shows that in the isolated condition, subjects tended to classify tokens of /okay/ as Acknowledgment / agreement (A) which had a longer realization of the /k/ phoneme (/kl/). They tended to classify tokens as Backchannels (B) which had a lower intensity, a longer duration, a longer realization of the /el/ phoneme (/el/), and a final rising pitch. They tended to classify tokens as C (cue to topic beginning) that ended with falling pitch.
Acknowledgement / agreement | $r$
---|---
latency of Spkr A before Spkr B’s turn | -0.528
length of silence by Spkr B before okay | -0.404
number of words by Spkr B after okay | -0.277

Backchannel | $r$
---|---
pitch slope over 2nd half of 2nd syllable | 0.520
pitch slope over last 50 ms | 0.500
pitch slope over 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
length of speech by Spkr B after okay | -0.413
latency of Spkr A before Spkr B’s turn | -0.385
length of silence by Spkr B before okay | 0.295
intensity slope over 2nd syllable | -0.279

Cue to discourse segment beginning | $r$
---|---
latency of Spkr A before Spkr B’s turn | 0.645
pitch slope over last 50 ms | -0.564
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
length of speech by Spkr B after okay | 0.338

Table 6: Features correlated to the proportion of votes for each label. Contextualized condition.

In the contextualized condition, we do find very different correlations. Table 6 shows that nearly all of the strong correlations in this condition involve 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 that shows strong correlations in the isolated condition shows the same strong correlation in the contextualized condition: the pitch slope at the end of the word. In both conditions subjects tended to label tokens with a final rising pitch contour as B, and tokens with a final falling pitch contour as C. This supports (Hockey, 1993)’s findings on the role of pitch contour in disambiguating okay.

We next conducted a series of two-sided Fisher’s exact tests to find correlations between subjects’ labelings of okay and the nominal features described in Section 4.3. We found significant associations between the realization of the /əʊ/ phoneme and the okay function in the isolated condition ($p < 0.005$).

Table 7 shows that, in particular, [m] seems to be the preferred realization for B okays, while [ə] seems to be the preferred one for A okays, and [ɔə] and [ɔ] for A and C okays.

<table>
<thead>
<tr>
<th></th>
<th>[ə]</th>
<th>[ʊ]</th>
<th>[ɔ]</th>
<th>[ŋ]</th>
<th>[ɔə]</th>
<th>[ɔ]</th>
<th>[m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7: Realization of the /ou/ phoneme, grouped by subject plurality label. Isolated condition only.

Notably, we did not find such significant associations in the contextualized condition. We did find significant correlations in both conditions, however, between okay classifications and the type of phrase accent and boundary tone following the target (Fisher’s Exact Test, $p < 0.05$ for the isolated condition, $p < 0.005$ for the contextualized condition). Table 8 shows that L-L% tends to be associated with A and C classes, H-H% with B classes, and L-H% with A and B classes. In this case, such correlations are present in the isolated condition, and sustained or enhanced in the contextualized condition.

<table>
<thead>
<tr>
<th></th>
<th>H-H%</th>
<th>H-L%</th>
<th>L-H%</th>
<th>L-L%</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Context.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 8: Phrase accent and boundary tone, grouped by subject plurality label.

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

In the second part of the study, when subjects listened to contextualized versions of the same to-
kens of okay, most of the strong correlations of word
function with acoustic, prosodic and phonetic fea-
tures were replaced by correlations with contextual
features, like latency and turn length. In other words,
these results suggest that contextual features might
override the effect of most acoustic, prosodic and
phonetic features of okay. There is nonetheless one
notable exception: word final intonation —captured
by the pitch slope and the ToBI labels for phrase ac-
cent and boundary tone— seems to play a central
role in the interpretation of both isolated and con-
textualized okays.

5 Conclusion and future work

In this study, we have presented evidence of differ-
ences in the interpretation of the function of isolated
and contextualized okays. We have shown the im-
portance of word final intonation for disambiguating
between the three studied functions in both condi-
tions. Contextualized features, when available, also
seem to play a central role.

We have also presented results suggesting that ac-
knowledgment / agreement acts as a default function
for both isolated an contextualized okays. Further-
more, while that function remains confusable with
backchannel in both conditions, the availability of
some context helps in distinguishing those two func-
tions from cue to topic beginning.

These results are relevant to spoken dialogue sys-
tems in suggesting how systems can convey the cue
word okay with the intended meaning and can inter-
pret users’ productions of okay correctly. How these
results extend to other cue words and to other word
functions remains an open question.

In this paper we have also introduced the vote
vectors, a novel representation of multi-subject clas-
sifications of tokens into a set of categories. This
technique, originally devised for three label classes
but easily generalizable to any number, proved very
useful for visualizing and interpreting our data.

As future work, we will extend this study to in-
clude the over 5800 occurrences of alright, gotcha,
huh, mmhm, okay, right, uhhuh, yeah, yep, yes, yup
in the entire Games Corpus, and all 10 discourse
functions described in Section 2, as annotated by
our three original labelers. Since we have observed
considerable differences in conversation style in the
two parts of the corpus (the Objects Games elicted
more ‘dynamic’ conversations, with more overlaps
and interruptions than the Cards Games), we will
compare cue phrase usage in these two settings. Fi-
ally, we are also interested in examining speaker
entrainment in cue phrase usage, or how subjects
adapt their choice and production of cue phrases to
their conversation partner’s.

References

Mary E. Beckman and Julia Hirschberg. 1994. The ToBI
annotation conventions. Ohio State University.


Robin Cohen. 1984. A computational theory of the func-
tion of clue words in argument understanding. Proceed-
ing of the 22nd conference on Association for
Computational Linguistics, pages 251–258.

Joseph L. Fleiss. 1971. Measuring nominal scale agree-
ment among many raters. Psychological Bulletin,

Barbara J. Grosz and Candace L. Sidner. 1986. Atten-
tion, Intentions, and the Structure of Discourse. Com-
putational Linguistics, 12(3):175–204.

Studies on the Disambiguation of Cue Phrases. Com-

Beth Ann Hockey. 1993. Prosody and the role of okay
and uh-huh in discourse. Proceedings of the Eastern

interaction, 294:338.

Jacqueline C. Kowtko. 1997. The function of intonation
in task-oriented dialogue. Ph.D. thesis, University of
Edinburgh.

John Pitrelli, Mary Beckman, and Julia Hirschberg.
1994. Evaluation of prosodic transcription labeling re-
liability in the ToBI framework. In icslp94, volume 2,
pages 123–126, Yokohama.

Rachel Reichman. 1985. Getting Computers to Talk Like
You and Me: Discourse Context, Focus, and Seman-
tics: (an ATN Model). MIT Press.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening

Nigel Ward and Wataru Tsukahara. 2000. Prosodic fea-
tures which cue back-channel responses in English and