Classification of Discourse Functions of Affirmative Words in Spoken Dialogue

Agustín Gravano [1], *Stefan Benus* [2], *Julia Hirschberg* [1], *Shira Mitchell* [3], *Ilia Vovsha* [1]

[1] Department of Computer Science, Columbia University, New York, NY, USA
[2] Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA
[3] Department of Mathematics, Harvard University, Boston, MA, USA

agus@cs.columbia.edu, sb513@nyu.edu, julia@cs.columbia.edu,
mitchel4@fas.harvard.edu, iv2121@columbia.edu

Abstract

We present results of a series of machine learning experiments that address the classification of the discourse function of single affirmative cue words such as *alright*, *okay* and *mm-hm* in a spoken dialogue corpus. We suggest that a simple discourse/sentential distinction is not sufficient for such words and propose two additional classification sub-tasks: identifying (a) whether such words convey acknowledgment or agreement, and (b) whether they cue the beginning or end of a discourse segment. We also study the classification of each individual word into its most common discourse functions. We show that models based on contextual features extracted from the time-aligned transcripts approach the error rate of trained human aligners.

Index Terms: cue words, discourse markers, spoken dialogue systems.

1. Introduction

Cue Phrases (or, Discourse Markers) are linguistic expressions that can be used to convey explicit information about the structure of a discourse or to convey a more literal, semantic contribution ([1][2][3]). 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 dialogue to indicate that one interlocutor is still attending to another; to convey acknowledgment or agreement; or, in its 'cue' use, to begin or end a discourse segment ([4][5][6][7]).

The ability to correctly determine the function of cue phrases is critical for important natural language processing tasks, including anaphora resolution ([1]), argument understanding ([3]), plan recognition ([8][1]), and discourse segmentation ([9]). 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 ([10]).

Prior work on the automatic classification of cue phrases includes studies by Litman and Hirschberg ([11][12] inter alia), which focused on differentiating between the discourse and sentential senses of cue phrases in spoken monologue. [13] presented a method for incorporating cue phrase identification into the process of part-of-speech tagging, for spoken dialogue. More recently, [14] studied the automatic classification of *like* and *well* into their discourse and sentential senses, achieving a performance close to that of human annotators.

In this paper we extend previous research by focusing on Single Affirmative Cue Words such as *alright*, *okay* and *mm-hm,* and their discourse functions in task-oriented spoken dialogues. For these words, sentential uses are rare (with the exception of *right*) and, thus, distinguishing among different discourse functions is more important than disambiguating between discourse and sentential uses. We employ machine learning (ML) techniques to automate the construction of models for classifying these affirmative cue words from empirical data. We present a series of experiments that induce classification models from sets of hand-annotated cue phrases and their features. Finally, we discuss the contribution of contextual, acoustic, and prosodic features, and compare the performance of the automatic classifiers to that of trained human annotators.

2. Material

The material for our study comes from the Columbia Games Corpus, a collection of 12 spontaneous task-oriented dyadic conversations elicited from speakers of Standard American English. Subjects were paid to play two types of collaborative games (Cards and Objects) on laptops, while seated in a soundproof booth divided by a curtain to ensure that all communication was verbal.

In the Cards games, subjects received points for finding cards depicting the same objects on their respective screens. One player described a card on her board, and the other searched for a full or partial match on his board. In the Objects games, one player described the position of a target object with respect to other fixed objects on her screen, while the other tried to move his representation of the target object to the same position on his screen. Players were given points based on the proximity of the target object to its correct location. Both games were designed to encourage discussion, and the subjects switched roles repeatedly.

13 subjects (7 males and 6 females) participated in the games; 11 played with two different partners in two different sessions and 2 played a single session. On average, each session took 45m, totaling 9h of dialogue for the whole corpus. There are 2245 unique words, and 73,844 words in total. All interactions were recorded, digitized, and downsampled to 16K. The recordings were orthographically transcribed, and words were aligned to the source by hand. Nearly all of the Objects part of the corpus has also been intonationally transcribed, using the ToBI conventions ([15]).

2.1. Labeling discourse functions

We asked three labelers to independently classify all occurrences of the single affirmative words *alright, gotcha, huh, mm-hm, okay, right, uh-huh, yeah, yep, yes, yup* in the entire Games Corpus into one of 11 categories, as shown in Table 1.

Labelers were given examples of each category, and labeled using both transcripts and speech together. Interlabeler reliability was measured by Fleiss' κ ([16]) at 0.69, where values between 0.6 and 0.8 correspond to substantial agreement. In this study we use Majority Labels, where at least two labelers assigned a token to the same class, as the gold standard. We assign the '?' label to a token when either its majority label is '?', or when it was assigned a different label by each labeler.

Table 1. Labeled discourse functions

A1	**Acknowledgment/agreement.** Indicates *"I believe what you said"*, and/or *"I agree with what you say"*.
A2	**Backchannel.** Indicates only *"I hear you and please continue"*, in response to another speaker's utterance.
C	**Cue beginning discourse segment.** Marks a new segment of a discourse or a new topic.
E	**Cue ending discourse segment.** Marks the end of a current segment of a discourse or a current topic.
P	**Pivot beginning (A1+C).** Functions both to acknowledge/agree and to cue a beginning segment.
F	**Pivot ending (A1 + E).** Functions both to acknowledge /agree and to cue the end of the current segment.
N	**Literal modifier.** Example: "*I think that's okay***"**.
B	**Back from a task.** Indicates *"I've just finished what I was doing and I'm back"*.
K	**Check.** Used with the meaning *"Is that okay?"*
S	**Stall.** Used to stall for time while keeping the floor.
?	Cannot decide.

Table 2 shows the distribution of each affirmative word and label in the Games Corpus. Note that *okay* is the most frequent affirmative word, as well as the only one conveying all ten functions as defined in Table 1. The remaining words have a small number of predominant functions, with A1 (acknowledgment/agreement) always among them. A1 is the most common discourse function overall, followed by A2 (backchannel) and C (cue beginning). Even though N (literal modifier) has a high frequency, 97% of its occurrences correspond to the word *right*, fact that is explained by the spatial descriptions involved in the OBJECTS games, e.g., "*it's to the right of the mirror*," or "*the ear is right on top of the nail*." The remaining words are rarely used in its literal sense (1% for *okay*; 3% for *alright*), or do not have a literal sense.

Table 2. Distribution of each word and label. 'Rest' = {*gotcha, huh, yep, yes, yup*}

	alright	*mm-hm*	*okay*	*right*	*uh-huh*	*yeah*	Rest	**Total**
A1	**99**	**61**	**1137**	**114**	18	**808**	133	2370
A2	6	**402**	**121**	14	**143**	**72**	5	763
C	**89**	0	**548**	2	0	2	0	641
E	8	0	10	0	0	0	0	18
P	5	0	**68**	0	0	0	0	73
F	13	12	**232**	2	0	22	17	298
N	9	0	29	**1079**	0	0	1	1118
B	9	1	33	0	0	0	0	43
K	0	0	6	**53**	0	1	8	68
S	1	0	15	1	0	2	0	19
?	56	27	235	10	3	65	11	407
Total	295	503	2434	1275	164	972	175	5818

3. Method

3.1. Features

For our ML experiments, we extracted a number of contextual, acoustic and prosodic features from the affirmative words uttered by the reference speaker, and from the surrounding context uttered by either the reference speaker or their interlocutor. Continuous acoustic/prosodic and durational features were extracted automatically with Praat ([17]). Normalizations were computed using *z*-score: $z = (X - \text{mean}) / \text{standard deviation}$.

Text-based features (TX) include the identity of the word; the part-of-speech tag (labeled automatically using Ratnaparkhi's maxent tagger, [18]) and simplified POS tag (silence, noun, verb, adjective, adverb, contraction, other) of the target word and its preceding and following words; position of the target word in its INTER-PAUSAL UNIT (or IPU, maximal sequence of words surrounded by pause longer than 50 ms); position of the target word in its TURN (maximal sequence of IPUs separated by silences shorter than 5 sec and including no speech from the interlocutor); IPU and turn length in words; and whether the previous and following turn were uttered by same speaker.

Timing features (TM) include duration of the target word, in raw ms and normalized by speaker; duration of the IPU and turn in which the target word occurred; duration and type of overlap (if any) of the target word, and its containing IPU and turn with the interlocutor (none, complete, at the beginning, at the end); time elapsed to the target word from the beginning of its IPU and turn, in ms and as a percentage of the duration of its IPU and turn.

Word acoustic/prosodic features (WA) consist of the ratio of voiced to unvoiced frames; minimum, maximum, mean, standard deviation of pitch and intensity (raw values, and normalized for the IPU and for the speaker in the current session); pitch slope, intensity slope, and stylized pitch slope, each calculated over the whole word, or over its last 100, 200 and 300 ms. We approximated the location of the syllable boundary (using Mermelstein's algorithm, [19]) and of the accented syllable (based on the maximum intensity). A manual check showed that the results were reasonably accurate for *okay*, but not for the other two-syllable words, so we included these features only for this word.

Previous-turn acoustic/prosodic features (PA) include**,** for target words occurring in turn-initial IPUs, the same pitch and intensity features described above, but this time calculated over the last IPU of the interlocutor's preceding turn.

Note that the word transcriptions and their time alignments are the only features annotated by hand — all other features are computed automatically. We combined the feature sets presented above into the following larger feature sets, to assess performance on several application tasks described below.

- **Text only (TX):** Features in this set were extracted solely from the dialogue transcripts. If a text-to-speech system can correctly classify cue words in the input text, it can then synthesize them using different intonational models according to their functions ([10]).
- **Contextual (TX+TM):** These features are extracted from the text transcriptions with their corresponding time alignment, but without access to acoustic information. This set would apply in situations where the transcription alignment is reliable, but the acoustics are not, due e.g. to excessive noise or overlap.
- **Acoustic (WA+PA):** Conversely to TX+TM, this feature set includes information which does ***not*** rely on a text transcription of the conversation. This set is used to assess the amount of information available in the signal alone.
- **Full Set (TX+TM+WA+PA):** This set includes all available features.

3.2. Classification Tasks

We performed a number of classification tasks using the JRIP machine learning algorithm, WEKA's ([20]) version of RIPPER, a propositional rule learner presented in [21]. We used 10-fold cross validation in all experiments.

The first task was a simple binary classification of affirmative words into their sentential vs. discourse uses, similar to experiments performed in previous studies. For this purpose, the sentential sense corresponds to our N label, and the discourse sense includes all others. Our second task was to identify the words used to signal the beginning (C, P in our labeling scheme) or end (E, F) of a discourse segment vs. all others. The goal of this experiment is to see how feasible it might be to use such distinctions in automatic discourse segmentation. Our third experiment involved identifying tokens bearing an acknowledgment or agreement function (in our labeling scheme, {A1, A2, P, F} vs. all other labels). Here we address the important goal in spoken dialogue systems of recognizing which information has been understood by the user.

In our final set of experiments we attempted to disambiguate the function of each affirmative word separately. We considered only classes with at least 50 tokens, since fewer would not be suitable to perform 10-fold cross validation. Note that only five affirmative words in our corpus have more than one function meeting such a requirement (highlighted in Table 2). We collapsed the functions with lower counts into an 'other' category, but only *alright* and *okay* have high enough counts in it. Therefore, the final functions for *alright* are {A1, C, other}; for *mm-hm,* {A1, A2}; for *okay,* {A1, A2, C, F, P, other}; for *right,* {A1, K, N}; for *yeah,* {A1, A2}.

4. Results

4.1. Classification Tasks

Tables 3 and 4 summarize the performance of the ML algorithm on the various classification tasks, using the feature sets defined in 3.1. We consider two types of baseline, one a majority class baseline, and one that employs a simple rule based on word identity (e.g. in the first task, *right* → sentential sense, other words → discourse sense). The rule used in each baseline is indicated in the tables. The error rate and F-measure for each human annotator were computed by comparing the labels assigned by each of them with the MAJORITY LABELS as defined in 2.1. Tables 3 and 4 include the mean error rate and mean F-measure for the three labelers together for comparison with automatic predictions.

In the discourse/sentential classification task, the low error rate (4.2%) of the word-based baseline is explained by the fact that, of the 1118 cue words labeled as N, 1079 correspond to tokens of *right* (see Table 2). The text-only, contextual, and full models achieve approximately a 50% reduction in error rate over this baseline. The acoustic model, which does not use word identity as a feature, improves the error rate about 60% over the majority-class baseline, but does not reach the level of the word-based baseline.

For the classification of cue words according to their discourse function, the word-based and majority class baselines are identical. While the contextual and full feature set models reduce the error rate by 50% over the baseline, the other models achieve slightly inferior improvement. The F-measures for this task show that identifying the 'cue end' discourse function is much harder than the 'cue beginning' one, both for human annotators and for the ML algorithm.

When detecting the acknowledgment function of affirmative cue words, the contextual and full models accomplish improvements of about 55% over the word-based baseline, while the acoustic model improves roughly 50% over the majority-class baseline.

The performance of the trained models for classifying individual words, although better than the baseline in most cases, is much lower than for the previous three tasks. One possible explanation is the smaller amount of training data available: while there are 5411 tokens available for the three general tasks, there are only 239 tokens of *alright*, 476 of *mm-hm*, and 907 of *yeah*. Another explanation might be the greater ambiguity of the tasks, which is also reflected in the higher mean error rate by the human labelers — as high as 14.9% for *alright* and 14% for *okay*. Nonetheless, the F-measure values reveal that the classification of particular functions for some of these words approaches the performance of human labelers, e.g. C (cue-beginning) for *alright* and *okay*, A2 (backchannel) for *mm-hm*, A1 (acknowledgment) for *okay* and *yeah*, and N (literal sense) for *right*.

4.2. Performance of Features

Looking at the performance of the different feature sets, we observe that, in all cases, the text-only and contextual feature sets outperform the acoustic features, approaching the mean error rate achieved by human labelers. Even though it uses only features extracted from the dialogue transcripts, with no timing or acoustic information, the text-only model achieves a remarkably low error rate, and, in one case (discourse vs. sentential sense), even outperforms the full set of features.

To estimate the importance of individual features in our classification tasks, we ranked them according to a information-gain metric. Results show that, in all tasks, contextual features dominate. In both the acknowledgment detection and the discourse/sentential classification tasks, the highest ranked features are word identity, POS tag of the previous word, IPU and turn length, and number and proportion of preceding words in the turn. In the discourse boundary classification task the highest-ranked features are word identity, POS tag of the following word, number and proportion of succeeding words in the turn, and context-normalized mean intensity.

In the classification of individual words, IPU and turn length appear among the highest predictive features for the five words. Speaker-normalized maximum intensity ranks first in the classification of *alright*. Pause length after the target word, and number and proportion of succeeding words in the turn are among the highest ranked features for classifying *alright, mm-hm, okay* and *yeah*. In the case of *mm-hm*, the length of the speech by the other speaker before and after the current turn is also ranked high. Finally, for classifying *right*, POS tag of the preceding word, and number and proportion of preceding words are the highest ranked features.

In addition to the features described above, we looked at a set of nominal prosodic features extracted from the ToBI-labeled portion of the corpus. These features included type of pitch accent, phrase accent, boundary tone, and break index, both of the target word and of the final part of the previous turn. We built models from this feature set alone and also added the ToBI features to the full feature set. ToBI features alone produced worse results than any other feature set alone. When ToBI features were added to the full set, we found no significant improvement. However, the small size of the data sets used in these experiments might be the cause for this, and therefore they should be repeated with larger training sets. Finally, we also investigated the effect of including gender

Table 3. Error rate and F-measure for JRip with different feature sets, for the baselines, and for the human labelers. First tasks.

	Discourse vs. Sentential Sense			**Discourse Segment Boundaries**			**Acknowledgment**	
	Error Rate	F-Measure		Error Rate	F-Measure		Error Rate	F-Measure
		discourse	sentence		begin	end		
Text only	**1.9 %**	.99	.95	**11.6 %**	.77	.30	**8.3 %**	.94
Contextual	**2.1 %**	.99	.95	**9.8 %**	.81	.53	**6.2 %**	.95
Acoustic	**8.1 %**	.95	.81	**14.2 %**	.66	.19	**17.2 %**	.87
Full set	**2.2 %**	.99	.95	**9.6 %**	.81	.57	**6.5 %**	.95
Majority class baseline	**20.7 %**	.89	.00	**19.0 %**	.00	.00	**35.2 %**	.79
	majority class: discourse sense			majority class: no boundary			majority class: acknowledgment	
Word-based baseline	**4.2 %**	.97	.91	**19.0 %**	.00	.00	**16.7 %**	.88
	rule: *right* → sentential sense others → discourse sense			rule: all words→ no boundary			rule: {*huh, right* } → no ack. others → ack.	
Human labelers	**1.8 %**	.99	.98	**5.7 %**	.94	.71	**5.5 %**	.98

Table 4. Same as Table 3 for the classification of each separate affirmative cue word.

	***alright* A1, C, other**			***mm-hm* A1, A2**			***okay* A1, A2, C, F, P, other**						***right* A1, K, N**				***yeah* A1, A2**		
	ER	F-Measure		ER	F-Measure		ER	F-Measure					ER	F-Measure			ER	F-Measure	
		A1	C		A1	A2		A1	A2	C	F	P		A1	K	N		A1	A2
Text only	**36.4**	.63	.72	**12.7**	.42	.93	**31.7**	.76	.16	.77	.33	.09	**8.1**	.80	.26	.96	**8.9**	.95	.23
Contextual	**34.3**	.63	.74	**9.9**	.53	.94	**25.6**	.79	.31	.82	.67	.18	**8.3**	.79	.24	.96	**7.7**	.96	.49
Acoustic	**38.5**	.65	.75	**14.0**	.06	.92	**40.2**	.69	.24	.64	.25	.03	**10.4**	.63	.00	.94	**9.4**	.95	.28
Full set	**33.9**	.65	.76	**10.2**	.56	.94	**25.5**	.80	.46	.83	.66	.21	**8.7**	.77	.28	.95	**8.2**	.96	.48
Majority class baseline	**58.6**	.59	.00	**13.2**	.00	.93	**48.3**	.68	.00	.00	.00	.00	**13.4**	.00	.00	.93	**8.2**	.96	.00
	majority class: A1			majority class: A2			majority class: A1						majority class: N				majority class: A1		
Human labelers	**14.9**	.86	.93	**6.3**	.81	.97	**14.0**	.89	.78	.94	.73	.56	**2.8**	.94	.66	.99	**7.5**	.96	.82

and identity of both speakers as features in our classification tasks, but found no significant improvement in performance.

5. Conclusion

We have presented the results of machine learning experiments classifying the discourse function of single affirmative words such as *alright*, *okay* and *mm-hm*. We find that, for spoken dialogue, the simple discourse/sentential distinction is insufficient. Thus, we added two additional classification tasks, the detection of acknowledgment and of discourse segment boundary functions — as well as the classification of each individual cue word into its most frequent functions. We showed that models based on contextual features extracted from the time-aligned transcripts approach the error rate of trained human aligners in all tasks, while our acoustic features offered little improvement. Future work will incorporate nominal prosodic features in the analysis, and evaluate the performance of clustering techniques for cue phrase sense disambiguation.

6. Acknowledgments

This work was funded in part by NSF IIS-0307905. We thank Gregory Ward, Elisa Sneed, and Michael Mulley for their valuable help in collecting and labeling the data, and the anonymous reviewers for helpful comments and suggestions.

7. References

[1] Grosz, B.J. & Sidner, C.L., "Attention, intentions, and the structure of discourse", Comp. Ling., 12(3):175. 1986.

[2] Reichman, R., "Getting Computers to Talk like You and Me", MIT Press, Cambridge, MA, 1985.

[3] Cohen, R., "A computational theory of the function of clue words in argument understanding", ACL, 251, 1984.

[4] Jefferson, G., "Side sequences", Studies in Social Interaction, 294:338, 1972.

[5] Schegloff, E. A. & Sacks, H., "Opening up closings", Semiotica, 8(4):289-327, 1973.

[6] Kowtko, J. C., "The function of intonation in task-oriented dialogue", Ph.D. Thesis, U. of Edinburgh, 1997.

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

[8] Litman, D.J. & Allen, J.F., "A plan recognition model for subdialogues in conversation", Cognitive Sc,11:163, 1987.

[9] Litman, D.J. & Passonneau, R.J., "Combining multiple knowledge sources for disc. segmentation", ACL, 1995.

[10] Hirschberg, J., "Accent and discourse context: Assigning pitch accent in synthetic speech", AAAI, 1990

[11] Hirschberg, J. & Litman, D.J., "Empirical studies on the disambiguation of cue phrases", Computational Linguistics, 19(3):501-530, 1993.

[12] Litman, D.J., "Cue phrase classification using ML", J. of Artificial Intelligence Research, 5:53-94, 1996.

[13] Heeman, P.A., Byron, D. & Allen, J.F., "Identifying discourse markers in spoken dialog", AAAI Spring Symposium on Applying ML and Discourse Proc., 1998.

[14] Zufferey, S. & Popescu-Belis A., "Towards automatic disambiguation of discourse markers: The case of 'like'", SIGDIAL, p.63-71, 2004.

[15] Beckman, M.E. & Hirschberg, J., "The ToBI annotation conventions", Ohio State University, 1994.

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

[17] Boersma, P. & Weenink, D., "Praat: Doing phonetics by computer", http://www.praat.org, 2001.

[18] Ratnaparkhi, A., "A maximum entropy model for part-of-speech tagging", Conf. on EMNLP, 133-142, 1996.

[19] Mermelstein, P., "Automatic segmentation of speech into syllabic units", J. Acoustical Soc. America, 58:880, 1975.

[20] Witten, I.H. & Frank, E., "Data Mining: Practical ML tools and techniques", 2nd Ed., Morgan Kaufmann, 2005.

[21] Cohen, W.W., "Fast effective rule induction", 12th International Conf. on Machine Learning, 115-123, 1995.