The ICSI Meeting Recorder Dialog Act (MRDA) Corpus

Elizabeth Shriberg[1,2], Raj Dhillon[1], Sonali Bhagat[1], Jeremy Ang[1], Hannah Carvey[1,3]

[1]International Computer Science Institute
[2]SRI International
[3] CSU Hayward
{ees,rdhillon,sonalivb,jca,hmcarvey}@icsi.berkeley.edu

Abstract

We describe a new corpus of over 180,000 hand-annotated dialog act tags and accompanying adjacency pair annotations for roughly 72 hours of speech from 75 naturally-occurring meetings. We provide a brief summary of the annotation system and labeling procedure, inter-annotator reliability statistics, overall distributional statistics, a description of auxiliary files distributed with the corpus, and information on how to obtain the data.

1 Introduction

Natural meetings offer rich opportunities for studying a variety of complex discourse phenomena. Meetings contain regions of high speaker overlap, affective variation, complicated interaction structures, abandoned or interrupted utterances, and other interesting turn-taking and discourse-level phenomena. In addition, meetings that occur naturally involve real topics, debates, issues, and social dynamics that should generalize more readily to other real meetings than might data collected using artificial scenarios. Thus meetings pose interesting challenges to descriptive and theoretical models of discourse, as well as to researchers in the speech recognition community [4,7,9,13,14,15].

We describe a new corpus of hand-annotated dialog acts and adjacency pairs for roughly 72 hours of naturally occurring multi-party meetings. The meetings were recorded at the International Computer Science Institute (ICSI) as part of the ICSI Meeting Recorder Project [9]. Word transcripts and audio files from that corpus are available through the Linguistic Data Consortium (LDC). In this paper, we provide a first description of the meeting recorder dialog act (MRDA) corpus, a companion set of annotations that augment the word transcriptions with discourse-level segmentations, dialog act (DA) information, and adjacency pair information. The corpus is currently available online for research purposes [16], and we plan a future release through the LDC.

2 Data

The ICSI Meeting Corpus data is described in detail in [9]. It consists of 75 meetings, each roughly an hour in length. There are 53 unique speakers in the corpus, and an average of about 6 speakers per meeting. Reflecting the makeup of the Institute, there are more male than female speakers (40 and 13, respectively). There are a28 native English speakers, although many of the nonnative English speakers are quite fluent. Of the 75 meetings, 29 are meetings of the ICSI meeting recorder project itself, 23 are meetings of a research group focused on robustness in automatic speech recognition, 15 involve a group discussing natural language processing and neural theories of language, and 8 are miscellaneous meeting types. The last set includes 2 very interesting meetings involving the corpus transcribers as participants (example included in [16]).

3 Annotation

Annotation involved three types of information: marking of DA segment boundaries, marking of DAs themselves, and marking of correspondences between DAs (adjacency pairs, [12]). Each type of annotation is described in detail in [7]. Segmentation methods were developed based on separating out speech regions having different discourse functions, but also paying attention to pauses and intonational grouping. To distinguish utterances that are prosodically one unit but which contain multiple DAs, we use a pipe bar (|) in the annotations. This allows the researcher to either split or not split at the bar, depending on the research goals.

We examined existing annotation systems, including [1,2,5,6,8,10,11], for similarity to the style of interaction in the ICSI meetings. We found that SWBD-DAMSL [11], a system adapted from DAMSL [6], provided a fairly good fit. Although our meetings were natural, and thus had real agenda items, the dialog was less like human-human or human-machine task-oriented dialog

TAG TITLE	SWBD-DAMSL	MRDA
Indecipherable	*%*	*%*
Abandoned	%-	%--
Interruption		**%-**
Nonspeech	x	x
Self-Talk	t1	t1
3rd-Party Talk	t3	t3
Task-Management	t	t
Communication-Management	c	
Statement	*sd*	*s*
Subjective Statement	*sv*	*s*
Wh- Question	qw	qw
Y/N Question	qy	qy
Open-Ended Question	qo	qo
Or Question	qr	qr
Or Clause After Y/N Question	qrr	qrr
Rhetorical Question	qh	qh
Declarative- Question	d	d
Tag Question	g	g
Open-Option	oo	
Command	*ad*	*co*
Suggestion	*co*	*cs*
Commit (self-inclusive)	cc	cc

TAG TITLE	SWBD-DAMSL	MRDA
Conventional-Opening	fp	
Conventional-Closing	fc	
Topic Change		**tc**
Explicit-Performative	fx	
Exclamation	fe	fe
Other-Forward-Function	fo	
Thanks	ft	ft
Welcome	fw	fw
Apology	fa	fa
Floor-Holder		**fh**
Floor-Grabber		**fg**
Accept, Yes Answers	*ny, aa*	*aa*
Partial Accept	aap	aap
Partial Reject	arp	arp
Maybe	am	am
Reject, No Answers	*nn, ar*	*ar*
Hold	h	h
Collaborative-Completion	2	2
Backchannel	b	b
Acknowledgment	bk	bk
Mimic	m	m
Repeat		**r**

TAG TITLE	SWBD-DAMSL	MRDA
Reformulation	bf	bs
Appreciation	ba	ba
Sympathy	by	by
Downplayer	bd	bd
Misspeak Correction	bc	bc
Rhetorical-Question Backchannel	bh	bh
Signal Non understanding	br	br
Understanding Check		**bu**
Defending/Explanation		**df**
Misspeak Self-Correction		**bsc**
"Follow Me"		**f**
Expansion/Supporting addition	*e*	*e*
Narrative-affirmative answers	na	na
Narrative-negative answers	ng	ng
No knowledge answers	no	no
Dispreferred answers	nd	nd
Quoted Material	q	
Humorous Material		**j**
Continued from previous line	+	
Hedge	h	
Nonlabeled		**z**

Figure 1: Mapping of MRDA tags to SWBD-DAMSL tags. Tags in boldface are not present in SWBD-DAMSL and were added in MRDA. Tags in italics are based on the SWBD-DAMSL version but have had meanings modified for MRDA. The ordering of tags in the table is explained as follows: In the mapping of DAMSL tags to SWBD-DAMSL tags in the SWBD-DAMSL manual, tags were ordered in categories such as "Communication Status", "Information Requests", and so on. In the mapping of MRDA tags to SWBD-DAMSL tags here, we have retained the same overall ordering of tags within the table, but we do not explicitly mark the higher-level SWBD-DAMSL categories in order to avoid confusion, since categorical structure differs in the two systems (see [7]).

(e.g., [1,2,10]) and more like human-human casual conversation ([5,6,8,11]). Since we were working with English rather than Spanish, and did not view a large tag set as a problem, we preferred [6,11] over [5,8] for this work. We modified the system in [11] a number of ways, as indicated in Figure 1 and as explained further in [7]. The MRDA system requires one "general tag" per DA, and attaches a variable number of following "specific tags". Excluding nonlabelable cases, there are 11 general tags and 39 specific tags. There are two disruption forms (%-, %--), two types of indecipherable utterances (x, %) and a non-DA tag to denote rising tone (rt).

An interface allowed annotators to play regions of speech, modify transcripts, and enter DA and adjacency pair information, as well as other comments. Meetings were divided into 10 minute chunks; labeling time averaged about 3 hours per chunk, although this varied considerably depending on the complexity of the dialog.

4 Annotated Example

An example from one of the meetings is shown in Figure 2 as an illustration of some of the types of interactions we observe in the corpus. Audio files and additional sample excerpts are available from [16]. In addition to the obvious high degree of overlap—roughly one third of all words are overlapped—note the explicit struggle for the floor indicated by the two failed floor grabbers (fg) by speakers c5 and c6. Furthermore, 6 of the 19 total utterances express some form of agreement or disagreement (arp, aa, and nd) with previous utterances. Also, of the 19 utterances within the excerpt, 9 are incomplete due to interruption by another talker, as is typical of many regions in the corpus showing high speaker overlap. We find in related work that regions of high overlap correlate with high speaker involvement, or "hot spots" [15]. The example also provides a taste of the frequency and complexity of adjacency pair information. For example, within only half a minute, speaker c5 has interacted with speakers c3 and c6, and speaker c6 has interacted with speakers c2 and c5.

5 Reliability

We computed interlabeler reliability among the three labelers for both segmentation (into DA units) and DA labeling, using randomly selected excerpts from the 75 labeled meetings. Since agreement on DA segmentation does not appear to have standard associated metrics in the literature, we developed our own approach. The philosophy is that any difference in words at the beginning and/or end of a DA could result in a different label for that DA, and the more words that are mismatched, the more likely the difference in label. As a very strict measure of reliability, we used the

Time	Chan	DA	AP	Transcript
2804-2810	c3	s^df^e.%-	34a	i mean you can't just like print the - the vaues out in ascii and you know look at them to see if they're ==
2810-2811	c6	fg		well ==
2810-2811	c5	s^arp^j	34b	not unless you had a lot of time .
2811-2812	c5	%-		and ==
2811-2814	c6	s^bu	35a	uh and also they're not - i mean as i understand it you – you don't have a way to optimize the features for the final word error .
2814-2817	c6	qy^d^g^rt	35a+	right ?
2818-2818	c2	s^aa	35b	right .
2818-2820	c6	s^bd		i mean these are just discriminative .
2820-2823	c6	s.%-	36a	but they're not um optimized for the final ==
2822-2823	c2	s^nd	36b	they're optimized for phone discrimination
2823-2825	c2	s^e.%-		not for ==
2823-2835	c6	s^bk|s.%-	37a	right | so it - there's always this question of whether you might do better with those features if there was a way to train it for the word error metric that you're actually - that you're actually ==
2824-2825	c5	s^aa		that's right .
2829-2830	c5	s.%-		well the other ==
2831-2832	c5	fg|%-		yeah | th- - the ==
2833-2835	c2	%-		huh- - huh ==
2834-2835	c5	s^nd	37b.38a	well you actually are .
2835-2837	c5	s^e	37b+.38a+	but – but it – but in an indirect way .
2837-2840	c6	s^aa|s^df.%-		well right | it's indirect so you don't know ==

Figure 2: Example from meeting Bmr023. Time marks are truncated here; actual resolution is 10 msec. "Chan": channel (speaker); "DA": full dialog act label (multiple tags are separated by "^"); "==": incomplete DA; "xx - xx": disfluency interruption point between words; "xx-": incomplete word; "AP": adjacency pairs (use arbitrary identifiers). For purposes of illustration, overlapped speech regions are indicated in the figure by reverse font color. Audio and other samples available from [16].

following approach: (1) Take one labeler's transcript as a reference. (2) Look at each other labeler's words. For each word, look at the utterance it comes from and see if the reference has the exact same utterance. (3) If it does, there is a match. Match every word in the utterance, and then mark the matched utterance in the reference so it cannot be matched again (this prevents felicitous matches due to identical repeated words). (4) Repeat this process for each word in each reference-labeler pair, and rotate to the next labeler as the reference. Note that this metric requires perfect matching of the full utterance a word is in for that word to be matched. For example in the following case, labelers agree on 3 segmentation locations, but the agreement on our metric is only 0.14, since only 1 of 7 words is matched:

. yeah . I agree it's a hard decision .
. yeah . I agree . it's a hard decision .

Overall segmentation results on this metric are provided by labeler pair in Table 1.

We examined agreement on DA labels using the Kappa statistic [3], which adjusts for chance agreement. Because of the large number of unique full label combinations, we report Kappa values in Table 2 using various class mappings distributed with the corpus. Values are shown by labeler pair.

Table 1: Results for strict segmentation agreement metric

Reference Labeler	Comparison Labeler	Agree	Total	Agree %
1	2	3004	4915	61.1
1	3	2797	3820	73.2
2	1	3004	4908	61.2
2	3	5253	7906	66.4
3	1	2797	3808	73.5
3	2	5253	7889	66.6
Overall		22108	33246	66.5

Table 2: Kappa values for DAs using different class mappings. Map 1: Disruptions vs. backchannels vs. fillers vs. statements vs. questions vs. unlabelable; does not break at the "|". Map 2: Same as Map 1 but breaks at the "|". Map 3: Same as Map 2 but breaks down fillers and questions into further subclasses. See [16] for further details.

Labeler	**Labeler**	**Map 1**	**Map 2**	**Map 3**
1	2	.75	.73	.72
1	3	.82	.81	.80
2	3	.82	.77	.75

The overall value of Kappa for our basic, six-way classmap (Map1) is 0.80, representing good agreement for this type of task.

6 Distributional Statistics

We provide basic statistics based on the dialog act labels for the 75 meetings. If we ignore the tag marking rising intonation (rt), since this is not a DA tag, we find 180,218 total tags. Table 3 shows the distribution of the tags in more detail.

Table 3: Distribution of tags. Tags are listed in order of descending frequency; values are percentages of the 180,218 total tags.

s 42.85	b 8.42	fh 4.65	%--4.39	bk 4.05
aa 3.38	%- 3.33	qy 3.10	df 2.29	e 2.02
d 1.74	fg 1.73	cs 1.69	ba 1.37	z 1.36
bu 1.28	qw 1.15	na 0.97	g 0.89	% 0.69
no 0.57	ar 0.53	j 0.49	2 0.48	co 0.46
h 0.44	f 0.41	m 0.40	nd 0.39	tc 0.38
r 0.34	t 0.33	fe 0.29	ng 0.28	bd 0.25
cc 0.24	qh 0.23	qrr 0.22	am 0.21	t3 0.20
x 0.18	t1 0.16	fa 0.16	aap 0.15	br 0.14
qr 0.12	qo 0.11	arp 0.10	bsc 0.09	bs 0.09
bh 0.09	ft 0.08	bc 0.03	by 0.01	

If instead we look at only the 11 obligatory general tags, for which there is one per DA, and if we split labels at the pipe bar, the total is 113,560 (excluding tags that only include a disruption label). The distribution of general tags is shown in Table 4.

Table 4: Distribution of general tags; values are percentages of 113,560 total general tags.

s 68.00	b 13.37	fh 7.38	qy 4.91
fg 2.74	qw 1.82	h 0.70	qh 0.36
qrr 0.35	qr 0.20	qo 0.17	

7 Auxiliary Information

We include other useful information with the corpus. Word-level time information is available, based on alignments from an automatic speech recognizer. Annotator comments are also provided. We suggest various ways to group the large set of labels into a smaller set of classes, depending on the research focus. Finally, the corpus contains information that may be useful in for developing automatic modeling of prosody, such as hand-marked annotation of rising intonation.

8 Acknowledgments

We thank Chuck Wooters, Don Baron, Chris Oei, and Andreas Stolcke for software assistance, Ashley Krupski for contributions to the annotation scheme, Andrei Popescu-Belis for analysis and comments on a release of the 50 meetings, and Barbara Peskin and Jane Edwards for general advice and feedback. This work was supported by an ICSI subcontract to the University of Washington on a DARPA Communicator project, ICSI NSF ITR Award IIS-0121396, SRI NASA Award NCC2-1256, SRI NSF IRI-9619921, an SRI DARPA ROAR project, an ICSI award from the Swiss National Science Foundation through the research network IM2, and by the EU Framework 6 project on Augmented Multi-party Interaction (AMI). The views are those of the authors and do not represent the views of the funding agencies.

References

[1] Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., et al. Dia logue Acts in VERBMOBIL-2 Second Edition. *VM-Report 226*, DFKI Saarbrücken, Germany, July 1998.

[2] Anderson, A. H., Bader, M., Bard, E. G., et al. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*(4), 351-366.

[3] Carletta, J., 1996. Assessing agreement on classification tasks: The Kappa Statistic. *Computational Linguistics*, 22:2, 249-254.

[4] Cieri, C., Miller, D. & Walker, K., 2002. Research methodologies, observations, and outcomes in conversational speech data collection. *Proc. HLT 2002*.

[5] Clark, A. & Popescu-Belis, A., 2004. Multi-level Dialogue Act Tags. *In Proceedings of SIGDIAL '04 (5th SIGDIAL Workshop on Discourse and Dialog)*. Cambridge, MA.

[6] Core, M. & Allen, J., 1997. Coding dialogs with the DAMSL annotation scheme. *Working Notes: AAAI Fall Symposium*, AAAI, Menlo Park, CA, pp. 28-35.

[7] Dhillon, R., Bhagat, S., Carvey, H., & Shriberg, E., 2004. Meeting Recorder Project: Dialog Act Labeling Guide. ICSI Technical Report TR-04-002, International Computer Science Institute.

[8] Finke, M., Lapata, M., Lavie, A., et al., 1998. CLARITY: Inferring discourse structure from speech. *AAAI '98 Spring Symposium Series*, March 23-25, 1998, Stanford University, California.

[9] Janin, A. et al., 2003. The ICSI Meeting Corpus. *Proc. ICASSP-2003*.

[10] Jekat, S., Klein, A., Maier, E., et al. Dialogue Acts in Verbmobil, Verbmobil-Report No. 65, April 1995.

[11] Jurafsky, D., Shriberg, E., & Biasca, D., 1997. Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Technical Report 97-02, Univ. of Colorado Institute of Cognitive Science.

[12] Levinson, S., 1983. *Pragmatics*. Cambridge: Cambridge University Press.

[13] NIST meeting transcription project, www.nist.gov/speech/test_beds

[14] Waibel, A., et al., 2001. Advances in automatic meeting record creation and access. *Proc. ICASSP-2001*.

[15] Wrede, B. & Shriberg, E., 2003. The relationship between dialogue acts and hot spots in meetings. *Proc. IEEE Speech Recognition and Understanding Workshop*, St. Thomas.

[16] www.icsi.berkeley.edu/~ees/dadb contains the annotation corpus and sample (audio + annotations) excerpts.