The Effect of Multiple Modalities in Dialogue Act Annotation

Crystal Nakatsu and **Chris Brew**
Department of Linguistics
The Ohio State University
Columbus, OH 43210 USA
{cnakatsu,cbrew}@ling.ohio-state.edu

1 Introduction

In previous work, SWBD-DAMSL (Jurafsky et al., 1997) showed that interrater reliability could be improved by decreasing the set of possible tag combinations. However, this solution may not be an option for researchers wishing to study dialogue act phenomena at a lower level of detail. So how can we continue to improve interrater reliability scores without modifying the annotation scheme?

As an alternative to modifying the tagset, one might instead alter the process of annotation. Very few corpus developers mention whether they allow for audio playback of an utterance during the coding process. In this work, we investigate the effects of dual modality annotation on both annotation rate and interrater reliability.

2 Annotation Experiment

2.1 Annotation Procedure

Two native speakers of English independently labeled the Trains 93 corpus (Heeman and Allen, 1994), using a (re-)modified version of Doran et. al.'s (2001) modified C-Star dialogue act tagset.

The first 43 dialogues (2961 utterances) were annotated through reading the transcripts (i.e. text) only, while the remaining 52 dialogues (3875 utterances) were annotated by listening to the corresponding audio file while viewing the transcripts.

2.2 Annotation Rate

The text-only utterances were annotated at an average rate of 0.121 utt/s and the text-audio utterances at at rate of 0.157 utt/s. Initially, these rates imply that it is the use of audio that increases the annotation rate. However, since the utterances in the text-only condition are annotated before the utterances in text-audio condition, the increased rate could be attributed to increased familiarity with the tag set.

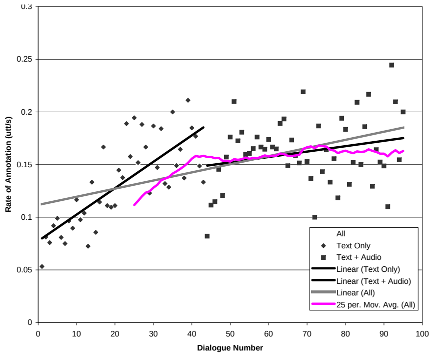

Figure 1: Rate of Annotation (seconds/utterance) with linear interpolation.

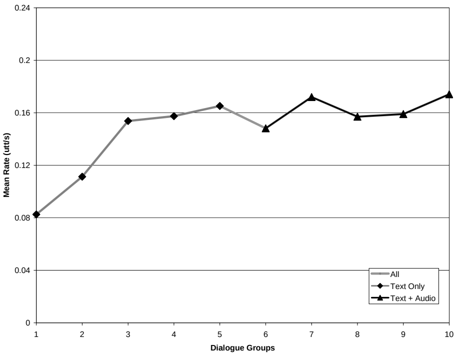

Figure 2: The Group Mean Annotation Rate (seconds/utterance)

The correlation analysis supports the influence of the familiarity effect, showing a significantly strong positive correlation ($r = .77, p < .001$) between dialogue number and annotation rate in the text-only condition, and a non-significant weak correlation ($r = .24, p < .1$) between the same two variables in the text-audio condition. Further analysis by a 2-factor ANOVA (F = 6.6, df = 8, $p <$ 1x10^{-6}), using dialogue number and modality as independent factors (depicted in Figure 2) more clearly indicates that the rising rate occurs mostly in the first two groups of the text-only dialogues

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 191–192.

and then flattens out in last 3 groups[1]. Furthermore, the rate from Groups 4 & 5 are maintained in groups 6-10 (with minor variance). This finding suggests that the addition of audio is not a factor in the increased annotation rate, but rather that annotation rate increases sharply at the onset of the annotation process as a result of some other factor that changes over time, such as an increase in familiarity of the tagset, and then flattens out, likely due to the annotators reaching maximum familiarity with the tagset.

Also, although the annotation rate is flattened in the later dialogues (Groups 4-10), it is maintained throughout the text-audio condition at about the same rate as the latter text-only dialogues. Thus, while annotation rate is not positively affected by the use of additional media, neither is it negatively affected.

2.3 Interrater Reliability

Raw agreement for all the utterances in the text-only modality is 66.7%, with $\kappa = 0.623$. This is slightly lower than the $\kappa = 0.66$ reported in (Doran et al., 2001) using their modified C-star tagset, but higher than the averaged $\kappa = 0.54$ achieved by the Trains 93 corpus annotators using the DAMSL scheme (Allen and Core, 1997) which allowed use of audio during annotation. In comparison, adding audio during our annotation resulted in an even higher raw agreement of 74.5% and $\kappa = 0.701$.

Again, at first glance these scores indicate that the increase in reliability is due to the use of the utterances' audio recordings during annotation. However, as before, due to the order of annotation, the increase in reliability could be due to increasing familiarity with the tagset.

The significant negative correlation (r = -0.45, $p <$ .005) in the text-only condition (in Figure 3) would seem to strongly indicate that κ did not improve as a result of familiarity, since we would expect a positive correlation in that case. This is further supported by a 2-factor ANOVA (F = 3.0, df = 8, $p <$ .005), which shows that κ decreases over time in the text-only condition, but is mostly level in the text-audio condition (Figure 4).

Having ruled out familiarity as a possibility for the improvement in interrater reliability, it seems that the improvement can indeed be correlated with the use of the corresponding audio recording of the utterance during annotation. In addition, this improvement comes at no obvious detriment to the annotation rate, since the annotation rate does not decline but rather remains somewhat steady throughout the text-audio condition.

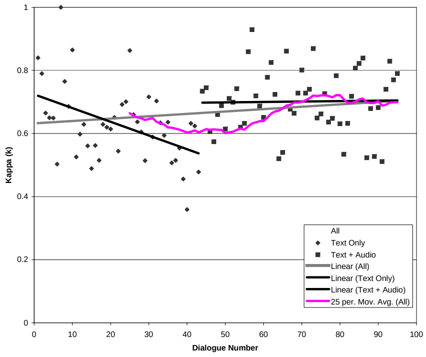

Figure 3: Interrater Reliability (κ) with linear interpolation and moving average.

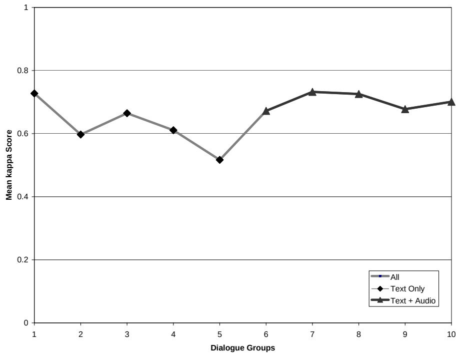

Figure 4: The Group Mean Interrater Reliability (kappa) Score

3 Acknowledgements

Thank you to Donna Byron, Michael White, and the Clippers Discussion group at OSU.

References

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers. Available at http://www.cs.rochester.edu/research/trains/annotation.

C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, pages 48–57.

Peter A. Heeman and James Allen. 1994. The TRAINS93 dialogues. Technical Report TRAINS TN 94-2, University of Rochester.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science, August.

[1] Groups 1-5 (text-only condition) were annotated first, and Groups 6-10 (text-audio condition) were annotated last.