Classification of Feedback Expressions in Multimodal Data

Costanza Navarretta
University of Copenhagen
Centre for Language Technology (CST)
Njalsgade 140, 2300-DK Copenhagen
costanza@hum.ku.dk

Patrizia Paggio
University of Copenhagen
Centre for Language Technology (CST)
Njalsgade 140, 2300-DK Copenhagen
paggio@hum.ku.dk

Abstract

This paper addresses the issue of how linguistic feedback expressions, prosody and head gestures, i.e. head movements and face expressions, relate to one another in a collection of eight video-recorded Danish map-task dialogues. The study shows that in these data, prosodic features and head gestures significantly improve automatic classification of dialogue act labels for linguistic expressions of feedback.

1 Introduction

Several authors in communication studies have pointed out that head movements are relevant to feedback phenomena (see McClave (2000) for an overview). Others have looked at the application of machine learning algorithms to annotated multimodal corpora. For example, Jokinen and Ragni (2007) and Jokinen et al. (2008) find that machine learning algorithms can be trained to recognise some of the functions of head movements, while Reidsma et al. (2009) show that there is a dependence between focus of attention and assignment of dialogue act labels. Related are also the studies by Rieks op den Akker and Schulz (2008) and Murray and Renals (2008): both achieve promising results in the automatic segmentation of dialogue acts using the annotations in a large multimodal corpus.

Work has also been done on prosody and gestures in the specific domain of map-task dialogues, also targeted in this paper. Sridhar et al. (2009) obtain promising results in dialogue act tagging of the Switchboard-DAMSL corpus using lexical, syntactic and prosodic cues, while Gravano and Hirschberg (2009) examine the relation between particular acoustic and prosodic turn-yielding cues and turn taking in a large corpus of task-oriented dialogues. Louwerse et al. (2006) and Louwerse et al. (2007) study the relation between eye gaze, facial expression, pauses and dialogue structure in annotated English map-task dialogues (Anderson et al., 1991) and find correlations between the various modalities both within and across speakers. Finally, feedback expressions (head nods and shakes) are successfully predicted from speech, prosody and eye gaze in interaction with Embodied Communication Agents as well as human communication (Fujie et al., 2004; Morency et al., 2005; Morency et al., 2007; Morency et al., 2009).

Our work is in line with these studies, all of which focus on the relation between linguistic expressions, prosody, dialogue content and gestures. In this paper, we investigate how feedback expressions can be classified into different dialogue act categories based on prosodic and gesture features. Our data are made up by a collection of eight video-recorded map-task dialogues in Danish, which were annotated with phonetic and prosodic information. We find that prosodic features improve the classification of dialogue acts and that head gestures, where they occur, contribute to the semantic interpretation of feedback expressions. The results, which partly confirm those obtained on a smaller dataset in Paggio and Navarretta (2010), must be seen in light of the fact that our gesture annotation scheme comprises more fine-grained categories than most of the studies mentioned earlier for both head movements and face expressions. The classification results improve, however, if similar categories such as head nods and jerks are collapsed into a more general category.

In Section 2 we describe the multimodal Danish corpus. In Section 3, we describe how the prosody of feedback expressions is annotated, how their content is coded in terms of dialogue act, turn and agreement labels, and we provide inter-coder agreement measures. In Section 4 we account for the annotation of head gestures, including inter-

Proceedings of the ACL 2010 Conference Short Papers, pages 318–324,
Uppsala, Sweden, 11-16 July 2010.

coder agreements results. Section 5 contains a description of the resulting datasets and a discussion of the results obtained in the classification experiments. Section 6 is the conclusion.

2 The multimodal corpus

The Danish map-task dialogues from the DanPASS corpus (Grønnum, 2006) are a collection of dialogues in which 11 speaker pairs cooperate on a map task. The dialogue participants are seated in different rooms and cannot see each other. They talk through headsets, and one of them is recorded with a video camera. Each pair goes through four different sets of maps, and changes roles each time, with one subject giving instructions and the other following them. The material is transcribed orthographically with an indication of stress, articulatory hesitations and pauses. In addition to this, the acoustic signals are segmented into words, syllables and prosodic phrases, and annotated with POS-tags, phonological and phonetic transcriptions, pitch and intonation contours.

Phonetic and prosodic segmentation and annotation were performed independently and in parallel by two annotators and then an agreed upon version was produced with the supervision of an expert annotator, for more information see Grønnum (2006). The Praat tool was used (Boersma and Weenink, 2009).

The feedback expressions we analyse here are *Yes* and *No* expressions, i.e. in Danish words like *ja* (yes), *jo* (yes in a negative context), *jamen* (yes but, well), *nej* (no), *næh* (no). They can be single words or multi-word expressions.

Yes and *No* feedback expressions represent about 9% of the approximately 47,000 running words in the corpus. This is a rather high proportion compared to other corpora, both spoken and written, and a reason why we decided to use the DanPASS videos in spite of the fact that the gesture behaviour is relatively limited given the fact that the two dialogue participants cannot see each other. Furthermore, the restricted contexts in which feedback expressions occur in these dialogues allow for a very fine-grained analysis of the relation of these expressions with prosody and gestures. Feedback behaviour, both in speech and gestures, can be observed especially in the person who is receiving the instructions (the *follower*). Therefore, we decided to focus our analysis only on the follower's part of the interaction. Because of time restrictions, we limited the study to four different subject pairs and two interactions per pair, for a total of about an hour of video-recorded interaction.

3 Annotation of feedback expressions

As already mentioned, all words in DanPASS are phonetically and prosodically annotated. In the subset of the corpus considered here, 82% of the feedback expressions bear stress or tone information, and 12% are unstressed; 7% of them are marked with onset or offset hesitation, or both. For this study, we added semantic labels – including dialogue acts – and gesture annotation. Both kinds of annotation were carried out using ANVIL (Kipp, 2004). To distinguish among the various functions that feedback expressions have in the dialogues, we selected a subset of the categories defined in the emerging ISO 24617-2 standard for semantic annotation of language resources. This subset comprises the categories *Accept*, *Decline*, *RepeatRephrase* and *Answer*. Moreover, all feedback expressions were annotated with an agreement feature (*Agree*, *NonAgree*) where relevant. Finally, the two turn management categories *TurnTake* and *TurnElicit* were also coded.

It should be noted that the same expression may be annotated with a label for each of the three semantic dimensions. For example, a *yes* can be an *Answer* to a question, an *Agree* and a *TurnElicit* at the same time, thus making the semantic classification very fine-grained. Table 1 shows how the various types are distributed across the 466 feedback expressions in our data.

Dialogue Act		
Answer	70	15%
RepeatRephrase	57	12%
Accept	127	27%
None	212	46%
Agreement		
Agree	166	36%
NonAgree	14	3%
None	286	61%
Turn Management		
TurnTake	113	24%
TurnElicit	85	18%
None	268	58%

Table 1: Distribution of semantic categories

3.1 Inter-coder agreement on feedback expression annotation

In general, dialogue act, agreement and turn annotations were coded by an expert annotator and the annotations were subsequently checked by a second expert annotator. However, one dialogue was coded independently and in parallel by two expert annotators to measure inter-coder agreement. A measure was derived for each annotated feature using the agreement analysis facility provided in ANVIL. Agreement between two annotation sets is calculated here in terms of Cohen's *kappa* (Cohen, 1960)[1] and corrected *kappa* (Brennan and Prediger, 1981)[2]. Anvil divides the annotations in slices and compares each slice. We used slices of 0.04 seconds. The inter-coder agreement figures obtained for the three types of annotation are given in Table 2.

feature	Cohen's *k*	corrected *k*
agreement	73.59	98.74
dial act	84.53	98.87
turn	73.52	99.16

Table 2: Inter-coder agreement on feedback expression annotation

Although researchers do not totally agree on how to measure agreement in various types of annotated data and on how to interpret the resulting figures, see Artstein and Poesio (2008), it is usually assumed that Cohen's *kappa* figures over 60 are good while those over 75 are excellent (Fleiss, 1971). Looking at the cases of disagreement we could see that many of these are due to the fact that the annotators had forgotten to remove some of the features automatically proposed by ANVIL from the latest annotated element.

4 Gesture annotation

All communicative head gestures in the videos were found and annotated with ANVIL using a subset of the attributes defined in the MUMIN annotation scheme (Allwood et al., 2007). The MUMIN scheme is a general framework for the study of gestures in interpersonal communication. In this study, we do not deal with functional classification of the gestures in themselves, but rather with how gestures contribute to the semantic interpretations of linguistic expressions. Therefore, only a subset of the MUMIN attributes has been used, i.e. *Smile, Laughter, Scowl, FaceOther* for facial expressions, and *Nod, Jerk, Tilt, SideTurn, Shake, Waggle, Other* for head movements.

A link was also established in ANVIL between the gesture under consideration and the relevant speech sequence where appropriate. The link was then used to extract gesture information together with the relevant linguistic annotations on which to apply machine learning.

The total number of head gestures annotated is 264. Of these, 114 (43%) co-occur with feedback expressions, with *Nod* as by far the most frequent type (70 occurrences) followed by *FaceOther* as the second most frequent (16). The other tokens are distributed more or less evenly, with a few occurrences (2-8) per type. The remaining 150 gestures, linked to different linguistic expressions or to no expression at all, comprise many face expressions and a number of tilts. A rough preliminary analysis shows that their main functions are related to focusing or to different emotional attitudes. They will be ignored in what follows.

4.1 Measuring inter-coder agreement on gesture annotation

The head gestures in the DanPASS data have been coded by non expert annotators (one annotator per video) and subsequently controlled by a second annotator, with the exception of one video which was annotated independently and in parallel by two annotators. The annotations of this video were then used to measure inter-coder agreement in ANVIL as it was the case for the annotations on feedback expressions. In the case of gestures we also measured agreement on gesture segmentation. The figures obtained are given in Table 3.

feature	Cohen's *k*	corrected *k*
face segment	69.89	91.37
face annotate	71.53	94.25
head mov segment	71.21	91.75
head mov annotate	71.65	95.14

Table 3: Inter-coder agreement on head gesture annotation

These results are slightly worse than those obtained in previous studies using the same annotation scheme (Jokinen et al., 2008), but are still sat-

[1] $(Pa - Pe)/(1 - Pe)$.

[2] $(Po - 1/c)/(1 - 1/c)$ where c is the number of categories.

isfactory given the high number of categories provided by the scheme.

A distinction that seemed particularly difficult was that between nods and jerks: although the direction of the two movement types is different (down-up and up-down, respectively), the movement quality is very similar, and makes it difficult to see the direction clearly. We return to this point below, in connection with our data analysis.

5 Analysis of the data

The multimodal data we obtained by combining the linguistic annotations from DanPASS with the gesture annotation created in ANVIL, resulted into two different groups of data, one containing all *Yes* and *No* expressions, and the other the subset of those that are accompanied by a face expression or a head movement, as shown in Table 4.

Expression	Count	%
Yes	420	90
No	46	10
Total	466	100
Yes with gestures	102	90
No with gestures	12	10
Total with gestures	114	100

Table 4: *Yes* and *No* datasets

These two sets of data were used for automatic dialogue act classification, which was run in the Weka system (Witten and Frank, 2005). We experimented with various Weka classifiers, comprising Hidden Naive Bayes, SMO, ID3, LADTree and Decision Table. The best results on most of our data were obtained using Hidden Naive Bayes (HNB) (Zhang et al., 2005). Therefore, here we show the results of this classifier. Ten-folds cross-validation was applied throughout.

In the first group of experiments we took into consideration all the Yes and No expressions (420 *Yes* and 46 *No*) without, however, considering gesture information. The purpose was to see how prosodic information contributes to the classification of dialogue acts. We started by totally leaving out prosody, i.e. only the orthographic transcription (Yes and No expressions) was considered; then we included information about stress (stressed or unstressed); in the third run we added tone attributes, and in the fourth information on hesitation. Agreement and turn attributes were used in all experiments, while Dialogue act annotation was only used in the training phase. The baseline for the evaluation are the results provided by Weka's ZeroR classifier, which always selects the most frequent nominal class.

In Table 5 we provide results in terms of precision (P), recall (R) and F-measure (F). These are calculated in Weka as weighted averages of the results obtained for each class.

dataset	Algor	P	R	F
YesNo	ZeroR	27.8	52.8	36.5
	HNB	47.2	53	46.4
+stress	HNB	47.5	54.1	47.1
+stress+tone	HNB	47.8	54.3	47.4
+stress+tone+hes	HNB	47.7	54.5	47.3

Table 5: Classification results with prosodic features

The results indicate that prosodic information improves the classification of dialogue acts with respect to the baseline in all four experiments with improvements of 10, 10.6, 10.9 and 10.8%, respectively. The best results are obtained using information on stress and tone, although the decrease in accuracy when hesitations are introduced is not significant. The confusion matrices show that the classifier is best at identifying *Accept*, while it is very bad at identifying *RepeatRephrase*. This result if not surprising since the former type is much more frequent in the data than the latter, and since prosodic information does not correlate with *RepeatRephrase* in any systematic way.

The second group of experiments was conducted on the dataset where feedback expressions are accompanied by gestures (102 *Yes* and 12 *No*). The purpose this time was to see whether gesture information improves dialogue act classification. We believe it makes sense to perform the test based on this restricted dataset, rather than the entire material, because the portion of data where gestures do accompany feedback expressions is rather small (about 20%). In a different domain, where subjects are less constrained by the technical setting, we expect gestures would make for a stronger and more widespread effect.

The Precision, Recall and F-measure of the ZeroR classifier on these data are 31.5, 56.1 and 40.4, respectively. For these experiments, however, we used as a baseline the results obtained based on stress, tone and hesitation information, the combination that gave the best results on the larger

dataset. Together with the prosodic information, Agreement and turn attributes were included just as earlier, while the dialogue act annotation was only used in the training phase. Face expression and head movement attributes were disregarded in the baseline. We then added face expression alone, head movement alone, and finally both gesture types together. The results are shown in Table 6.

dataset	Algor	P	R	F
YesNo	HNB	43.1	56.1	46.4
+face	HNB	43.7	56.1	46.9
+headm	HNB	44.7	55.3	48.2
+face+headm	HNB	49.9	57	50.3

Table 6: Classification results with head gesture features

These results indicate that adding head gesture information improves the classification of dialogue acts in this reduced dataset, although the improvement is not impressive. The best results are achieved when both face expressions and head movements are taken into consideration.

The confusion matrices show that although the recognition of both *Answer* and *None* improve, it is only the *None* class which is recognised quite reliably. We already explained that in our annotation a large number of feedback utterances have an agreement or turn label without necessarily having been assigned to one of our task-related dialogue act categories. This means that head gestures help distinguishing utterances with an agreement or turn function from other kinds. Looking closer at these utterances, we can see that nods and jerks often occur together with *TurnElicit*, while tilts, side turns and smiles tend to occur with *Agree*.

An issue that worries us is the granularity of the annotation categories. To investigate this, in a third group of experiments we collapsed *Nod* and *Jerk* into a more general category: the distinction had proven difficult for the annotators, and we don't have many jerks in the data. The results, displayed in Table 7, show as expected an improvement. The class which is recognised best is still *None*.

dataset	Algor	P	R	F
YesNo	HNB	43.1	56.1	46.4
+face	HNB	43.7	56.1	46.9
+headm	HNB	47	57.9	51
+face+headm	HNB	51.6	57.9	53.9

Table 7: Classification results with fewer head movements

6 Conclusion

In this study we have experimented with the automatic classification of feedback expressions into different dialogue acts in a multimodal corpus of Danish. We have conducted three sets of experiments, first looking at how prosodic features contribute to the classification, then testing whether the use of head gesture information improved the accuracy of the classifier, finally running the classification on a dataset in which the head movement types were slightly more general. The results indicate that prosodic features improve the classification, and that in those cases where feedback expressions are accompanied by head gestures, gesture information is also useful. The results also show that using a more coarse-grained distinction of head movements improves classification in these data.

Slightly more than half of the head gestures in our data co-occur with other linguistic utterances than those targeted in this study. Extending our investigation to those, as we plan to do, will provide us with a larger dataset and therefore presumably with even more interesting and reliable results.

The occurrence of gestures in the data studied here is undoubtedly limited by the technical setup, since the two speakers do not see each other. Therefore, we want to investigate the role played by head gestures in other types of video and larger materials. Extending the analysis to larger datasets will also shed more light on whether our gesture annotation categories are too fine-grained for automatic classification.

Acknowledgements

This research has been done under the project VKK (Verbal and Bodily Communication) funded by the Danish Council for Independent Research in the Humanities, and the NOMCO project, a collaborative Nordic project with participating research groups at the universities of Gothenburg, Copenhagen and Helsinki which is funded by the NOS-HS NORDCORP programme. We would also like to thank Nina Grønnum for allowing us to use the DanPASS corpus, and our gesture annotators Josephine Bødker Arrild and Sara Andersen.

References

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation*, 41(3–4):273–287.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Paul Boersma and David Weenink, 2009. *Praat: doing phonetics by computer*. Retrieved May 1, 2009, from http://www.praat.org/.

Robert L. Brennan and Dale J. Prediger. 1981. Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41:687–699.

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

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bullettin*, 76(5):378–382.

Shinya Fujie, Y. Ejiri, K. Nakajima, Y Matsusaka, and Tetsunor Kobayashi. 2004. A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, pages 159 – 164, september.

Agustin Gravano and Julia Hirschberg. 2009. Turn-yielding cues in task-oriented dialogue. In *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, September 2009*, pages 253–261, Queen Mary University of London.

Nina Grønnum. 2006. DanPASS - a Danish phonetically annotated spontaneous speech corpus. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proceedings of the 5th LREC*, pages 1578–1583, Genoa, May.

Kristiina Jokinen and Anton Ragni. 2007. Clustering experiments on the communicative prop- erties of gaze and gestures. In *Proceeding of the 3rd. Baltic Conference on Human Language Technologies*, Kaunas, Lithuania, October.

Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2008. Distinguishing the communicative functions of gestures. In *Proceedings of the 5th MLMI*, LNCS 5237, pages 38–49, Utrecht, The Netherlands, September. Springer.

Michael Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Ph.D. thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com.

Max M. Louwerse, Patrick Jeuniaux, Mohammed E. Hoque, Jie Wu, and Gwineth Lewis. 2006. Multimodal communication in computer-mediated map task scenarios. In R. Sun and N. Miyake, editors, *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1717–1722, Mahwah, NJ: Erlbaum.

Max M. Louwerse, Nick Benesh, Mohammed E. Hoque, Patrick Jeuniaux, Gwineth Lewis, Jie Wu, and Megan Zirnstein. 2007. Multimodal communication in face-to-face conversations. In R. Sun and N. Miyake, editors, *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 1235–1240, Mahwah, NJ: Erlbaum.

Evelyn McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. 2005. Contextual Recognition of Head Gestures. In *Proceedings of the International Conference on Multi-modal Interfaces*.

Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. 2007. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8–9):568–585.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2009. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84, Springer.

Gabriel Murray and Steve Renals. 2008. Detecting Action Meetings in Meetings. In *Proceedings of the 5th MLMI*, LNCS 5237, pages 208–213, Utrecht, The Netherlands, September. Springer.

Harm Rieks op den Akker and Christian Schulz. 2008. Exploring features and classifiers for dialogue act segmentation. In *Proceedings of the 5th MLMI*, pages 196–207.

Patrizia Paggio and Costanza Navarretta. 2010. Feedback in Head Gesture and Speech. To appear in *Proceedings of 7th Conference on Language Resources and Evaluation (LREC-2010)*, Malta, May.

Dennis Reidsma, Dirk Heylen, and Harm Rieks op den Akker. 2009. On the Contextual Analysis of Agreement Scores. In Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen, editors, *Multimodal Corpora From Models of Natural Interaction to Systems and Applications*, number 5509 in Lecture Notes in Artificial Intelligence, pages 122–137. Springer.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangaloreb, and Shrikanth Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition.

Harry Zhang, Liangxiao Jiang, and Jiang Su. 2005. Hidden Naive Bayes. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, pages 919–924.