ALICO: A multimodal corpus for the study of active listening

Hendrik Buschmeier[1,2], Zofia Malisz[1,3], Joanna Skubisz[3], Marcin Włodarczak[3,4], Ipke Wachsmuth[1,2], Stefan Kopp[1,2], Petra Wagner[1,3]

[1]CITEC, [2]Faculty of Technology, [3]Faculty of Linguistics and Literary Studies,
Bielefeld University, Bielefeld, Germany
{firstname.lastname}@uni-bielefeld.de
[4]Department of Linguistics, Stockholm University, Stockholm, Sweden
wlodarczak@ling.su.se

Abstract

The *Active Listening Corpus* (ALICO) is a multimodal database of spontaneous dyadic conversations with diverse speech and gestural annotations of both dialogue partners. The annotations consist of short feedback expression transcription with corresponding communicative function interpretation as well as segmentation of interpausal units, words, rhythmic prominence intervals and vowel-to-vowel intervals. Additionally, ALICO contains head gesture annotation of both interlocutors. The corpus contributes to research on spontaneous human–human interaction, on functional relations between modalities, and timing variability in dialogue. It also provides data that differentiates between distracted and attentive listeners. We describe the main characteristics of the corpus and present the most important results obtained from analyses in recent years.

Keywords: active listening; multimodal feedback; head gestures; attention

1. Introduction

Multimodal corpora are a crucial part of scientific research investigating human–human interaction. Recent developments in data collection of spontaneous communication emphasise the co-influence of verbal and non-verbal behaviour between dialogue partners (Oertel et al., 2013). In particular, the listener's role during interaction has attracted attention in both fundamental research and technical implementations (Sidner et al., 2004; Kopp et al., 2008; Truong et al., 2011; Heylen et al., 2011; de Kok and Heylen, 2011; Buschmeier and Kopp, 2012).

The *Active Listening Corpus* (ALICO) collected at Bielefeld University is a multimodal corpus built to study verbal/vocal and gestural behaviour in face-to-face communication, with a special focus on the listener. The communicative situation in ALICO, interacting with a storytelling partner, was designed to facilitate active and spontaneous listening behaviour. Although the active speaker usually fulfills the more dynamic role in dialogue, the listener contributes to successful grounding by giving verbal and non-verbal feedback. Short vocalisations like 'mhm', 'okay', 'm' that constitute listener's turns express the ability and willingness to interact, understand, convey emotions and attitudes and constitute an integral part of face-to-face communication. We use the term *short feedback expressions* (SFE; cf. Schegloff, 1982; Ward and Tsukahara, 2000; Edlund et al., 2010) and classify SFEs using an inventory of communicative feedback functions (Buschmeier et al., 2011). Both SFE transcriptions and feedback function labels are annotated and included in the ALICO database.

Apart from vocal feedback, listeners show their engagement in conversation by means of non-vocal behaviour such as head gestures. Visual feedback emphasises the degree of listener involvement in conversation and encourages the speaker to stay active during his or her speech at turn relevance places (Wagner et al., 2014; Heldner et al., 2013). Head movements also co-occur with mutual gaze (Peters et al., 2005) and correlate with active listening displays. ALICO contains head gesture annotations, including gesture type labeling such as nod, shake or tilt, for both interlocutors. First evaluations of the head gesture inventory can be found in (Kousidis et al., 2013).

Additionally, the ALICO conversational sessions included a task in which the listener's attention was experimentally manipulated, with a view to revealing communicative strategies listeners use when distracted. Previous studies have reported that the listener's attentional state has an influence on the quality of speaker's narration and the number of feedback occurrences in dialogue. Bavelas et al. (2000) carried out a study in which the listener was distracted by an ancillary task during a conversational session. The findings have shown the preoccupied listener to produce less context-specific feedback. These findings are in accordance with the results of Kuhlen and Brennan (2010). All the above authors confirm that distractedness of the listener affects the behaviour of the interlocutor and interferes with the speaker's speech. Several analyses performed so far on the ALICO corpus deal with the question of how active listening behaviour changes when the attention level is varied in dialogue (Buschmeier et al., 2011; Malisz et al., 2012; Włodarczak et al., 2012).

The corpus was also annotated for the purpose of studying temporal relations across modalities, within and between interlocutors. The rhythmic annotation layer (vocalic beat intervals and rhythmic prominence intervals) has served as input for coupled oscillator models providing an important

The first four authors are listed in alphabetical order.

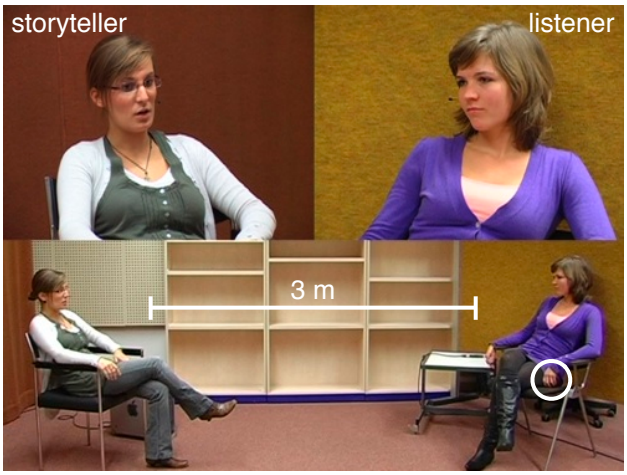

Figure 1: Screenshot from a video file capturing the whole scene (*long camera shot*), and perspectives of each participant (*medium camera shots*). The listener is being distracted by counting words beginning with letter 's' and pressing a button on a remote control hidden in her left hand.

testbed for hypotheses concerning interpersonal entrainment in dialogue (Wagner et al., 2013). First evaluations of entrained timing behaviour in two modalities implemented in an artificial agent are reported on by Inden et al. (2013).

By enabling a targeted study of active listening that includes varying listener attention levels, the ALICO corpus contributes to better understanding of human discourse. Analysis outcomes have proven useful in applications such as artificial listening agents (Inden et al., 2013). The corpus also provides a unique environment for studying temporal interactions between multimodal phenomena. In the present report we describe the main corpus characteristics and summarise the most important results obtained from analyses done so far.

2. Corpus architecture

ALICO's audiovisual dataset consists of 50 same-sex conversations between 25 German native speaker dyads (34 female and 16 male). All the participants were students at Bielefeld University and, apart from 4 dialogue partners, did not know each other before the study. Participants were randomly assigned to dialogue pairs and rewarded for their effort with credit points or 4 euros. No hearing impairments were reported by the participants. The total length of the recorded material is 5 hours 31 minutes. Each dialogue has a mean length of 6 minutes and 36 seconds (Min = 2:00 min, Max = 14:48 min, SD = 2:50 min).

A face-to-face dialogue study forms the core of the corpus. The study was carried out in a recording studio (MintLab; Kousidis et al., 2012) at Bielefeld University. Dialogue partners were placed approximately three metres apart in a comfortable setting (see Figure 1). Participants wore high quality headset microphones (Sennheiser HSP 2 and Sennheiser ME 80), another condenser microphone captured the whole scene and three Sony VX 2000 E camcorders recorded the video.

One of the dialogue partners (the 'storyteller') told two holiday stories to the other participant (the 'listener'), who was instructed to listen actively, make remarks and ask questions, if appropriate. Participants were assigned to their roles randomly and received their instructions separately. Furthermore, similar to Bavelas et al. (2000), the listener was engaged in an ancillary task during one of the stories (the order was counterbalanced across dyads): he or she was instructed to press a button on a hidden remote control (see Figure 1) every time they heard the letter 's' at the beginning of a word. The letter 's' is the second most common word-initial letter in German and often corresponds to perceptually salient sibilant sounds. A fourth audio channel was used to record the 'clicks' synthesised by a computer when listeners pressed the button on the remote control. The listeners were also required to retell the stories after the study and to report on the number of 's' words. The storyteller was aware that the listener is going to search for something in the stories; no further information about the details of the listener's tasks was disclosed to the storyteller.

3. Speech annotation

Annotation of the interlocutors' speech was performed in Praat (Boersma and Weenink, 2013), independently from head gesture annotation. Speech annotation tiers differ for listener and speaker role (see Table 1 for an overview of the annotation tiers).

3.1. The listener

The listener's SFEs with corresponding communicative feedback functions have been annotated in 40 dialogues thus far, i.e. in 20 sessions involving the distraction task and 20 sessions with no distractions. Segmentation of the listener SFEs was carried out automatically in Praat based on signal intensity and was subsequently checked manually. After that, another annotator transcribed the pre-segmented SFEs according to German orthographic conventions. Longer listener turns were marked but not transcribed.

A total number of 1505 feedback signals was identified. The mean ratio of time spent producing feedback signals to other listener turns ("questions and remarks", normalised by their respective mean duration per dialogue) equals 65% (Min = 32%; Max = 100%), suggesting that the corpus contains a high density of spoken feedback phenomena. The mean feedback rate is 10 signals per minute, mean turn rate is 5 turns per minute, with a significantly higher turn rate in the attentive listener (6 turns/min) than in the distracted listener (4 turns/min, two-sample Wilcoxon rank sum test: $p < .01$).

Three labelers independently assigned feedback functions to listener SFEs in each dialogue. A feedback function inventory was developed and first described in Buschmeier et al. (2011), largely based on Allwood et al. (1992). The inventory involves core feedback functions that signal *perception* of the speaker's message (category P1), *understanding* (category P2) of what is being said, *acceptance/agreement* (category P3) with the speaker's message. These levels can be treated as a hierarchy with an increasing value judgement of grounding 'depth'. The negation of the respective functions was marked as N1–N3. An option to extend listener

Table 1: Overview of the annotation tiers in ALICO. Speech and gesture annotation tiers differ between listener (L) and speaker (S) roles. All annotation tiers are available in the attentive listener condition (A) but not in the distracted listener condition as yet (D).

				Role		Condition	
	Tiers	Annotation examples	Annotation scheme	L	S	A	D
IPU	interpausal units	utterance, pause	Breen et al. (2012)	—	✓	✓	✓
Speech	words	*Reise*	Kisler et al. (2012)	—	✓	✓	—
	pronounciation (SAMPA)	RaIz@	Kisler et al. (2012)	—	✓	✓	—
	phonemic segmentation	R, aI, z, @	Kisler et al. (2012)	—	✓	✓	—
	vowel-to-vowel interval	interval		—	✓	✓	✓
	rhythmic prominence interval	interval	Breen et al. (2012)	—	✓	✓	✓
Feedback	feedback expressions	*ja, m, okay*	Buschmeier et al. (2011)	✓	—	✓	✓
	feedback functions	P1, P3A, N2	Buschmeier et al. (2011)	✓	—	✓	✓
Head	speaker head gesture units	slide-1-right	Kousidis et al. (2013)	—	✓	✓	—
	listener head gesture units	jerk-1+nod-2	Włodarczak et al. (2012)	✓	—	✓	✓

Table 2: Proportions of the most frequent German SFEs (short feedback expressions) and their corresponding feedback functions (P1: *perception*, P2: *understanding*, P3: *acceptance/agreement* and *other*) produced by listeners in forty ALICO dialogues.

%	P1	P2	P3	*other*	∑
ja	6.9	6.4	5.4	7.6	26.3
m	13.2	5.5	1.5	2.6	22.8
others	0.2	2.2	2.5	15.1	19.9
mhm	6.6	4.2	0.4	1.5	12.7
okay	0.2	5.4	2.5	2.7	10.8
achso	0	1.4	0	1.8	3.2
cool	0	0	0	1.5	1.5
klar	0	0.1	0.9	0.5	1.4
ah	0.2	0.1	0	1.1	1.4
∑	27.2	25.2	13.2	34.4	100.0

feedback function labels by three modifiers was available to the annotators, where modifier A referred to the listener's emotions/attitudes co-occurring with SFEs, leading to labels such as P3A (Kopp et al., 2008). Modifier A was also appended to the resulting majority label if it was used by at least one annotator so that subtle (especially emotion-related) distinctions were preserved. Modifiers C and E referred to feedback expressions occurring at the beginning or the end of a discourse segment initiated by the listener (Gravano et al., 2007). The most frequent SFEs with corresponding feedback functions found in the corpus are presented in Table 2. Communicative context was carefully and independently taken into account by each annotator during feedback function interpretation.

Majority labels between annotators determined the feedback functions in the final version of the listener's speech annotation. Disagreements in the labeling, i.e. cases which could not be settled by majority labels, corresponding to 10% of all feedback expressions, were discussed and resolved.

3.2. The storyteller

The storyteller's speech was annotated in 20 sessions involving no distractions. The following rhythmic phenomena were delimited in the storyteller's speech: vowel-to-vowel intervals, rhythmic prominence intervals and minor phrases (Breen et al., 2012). Vowel onsets were extracted semi-automatically from the data. Algorithms in Praat (Barbosa, 2006) were used first, after which the resulting segmentation was checked for accuracy by two annotators who inspected the spectrogram, formants and pitch curve in Praat as well as verified each other's corrections. Rhythmic prominences, judged perceptually, were marked whenever a 'beat' on a given syllable was perceived, regardless of lexical or stress placement rules (Breen et al., 2012). Phrase boundaries were marked manually every time a perceptually discernible gap in the storyteller's speech occurred. The resulting minimum pause length of 60 msec is comparable to pauses between so called Interpausal Units as segmented automatically, in e.g., Beňus et al. (2011). Interannotator agreement measurements regarding prominence and phrase annotations are forthcoming. In the study by Inden et al. (2013), the prosodic annotation carried out on storyteller's speech served as input to the modeling of local timing for an embodied converstational agent.

Apart from manual rhythmic segmentation, forced alignment was carried out on the storytellers' speech, using the WebMAUS tool (Kisler et al., 2012). Automatic segmentation and labeling facilitates work with large speech data and is less time-consuming, expensive and error prone than manual annotation. It produces a fairly accurately aligned and multi-layered annotation on small linguistic units, in e.g. segmented data. WebMAUS output provides tiers with *word segmentation*, *SAMPA transcription* and *vowel-consonant segmentation*).

4. Head gesture annotation

The corpus contains gestural annotation of both dialogue partners (see Table 1). Annotations were performed in ELAN (Wittenburg et al., 2006) by close inspection of the muted

Table 3: Head gesture type inventory (adapted from Kousidis et al. (2013).

Label	Description
nod	Rotation down–up
jerk	'Inverted nod', head upwards
tilt	'Sideways nod'
shake	Rotation left–right horizontally
protrusion	Pushing the head forward
retraction	Pulling the head back
turn	Rotation left OR right
bobble	Shaking by tilting left–right
slide	Sideways movement(no rotation)
shift	Repeated slides left–right
waggle	Irregular connected movement

Table 4: Frequency table of listener's head movement types found in 40 dialogues in the *Active Listening Corpus*.

Listener's head movement types	count	%
nod	1685	69.06
jerk	105	4.30
shake	89	3.65
turn	48	1.97
retraction	30	1.23
protrusion	6	0.25
complex HGUs	385	15.78
other	92	3.76
Σ	2440	100

video (stepping through the video frame-by-frame). Uninterrupted, communicative head movements were segmented as annotation events. Movements resulting from inertia, slow body posture shifts, ticks, etc. were excluded from the annotation. Thus obtained *head gesture units (HGUs)* contain perceptually coherent, communicative head movement sequences, without perceivable gaps.

Each constituent gesture in an HGU label was marked for head gesture type. The full inventory of gesture types is presented in Table 3. Prototypical movements along particular axes are presented in Figure 2. Mathematical conventions for 3D spatial coordinates are used in Figure 2, as done in biomechanical and physiological studies (Yoganandan et al., 2009) on head movements.

The identified constituent gestures in each HGU were also annotated for the number of gesture cycles and, where applicable, the direction of the gesture (left or right, from the perspective of the annotator). For example, the label nod-2+tilt-1-right describes a sequence consisting of two different movement types with two- and one cycle, respectively, where the head tilting is performed to the right side of the screen.

The resulting head gesture labels describe simple, complex or single gestural units. Complex HGUs denote multiple head movement types with different number of cycles, whereas single units refer to one head movement with one repetition. Simple head movement types consist of one movement type and at least two cycles (see Figure 3). The annotated HGU labels may provide information about the following features: *complexity* (the number of subsequent gesture types in the phrase) or *cycle frequencies* of all HGUs and both dialogue partners.

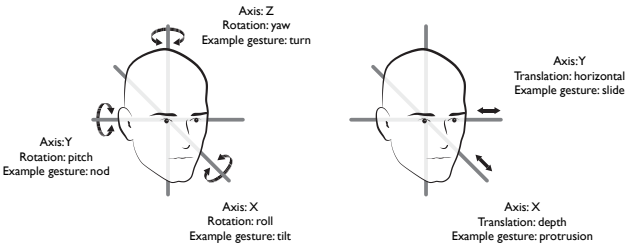

Figure 2: Schematic overview of rotations and translations along three axes as well as example movements most frequently used in communicative head gesturing (reprinted from Wagner et al. (2014) with permission from Elsevier).

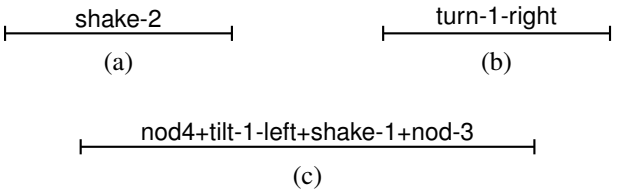

Figure 3: Examples of (a) simple, (b) single and (c) complex head movement types found in the ALICO inventory.

4.1. The listener

Listener head gestures were annotated in 40 dialogues so far, i.e. in 20 sessions involving the distraction task and 20 sessions with no distractions. Listener head gesture type categories were found to be limited to the subset of the inventory presented in Table 3, namely to nod, shake, tilt, turn, jerk, protrusion and retraction (Włodarczak et al., 2012). The most frequent head gestures found for listeners in the corpus are presented in Table 4. Listener HGUs were labeled and checked for errors by two annotators, however no inter-annotator agreement was calculated.

4.2. The storyteller

Co-speech head gestures produced by the storyteller are much more differentiated than those of the listener. Consequently, we used an extended inventory, as described and evaluated on a different German spontaneous dialogue corpus by Kousidis et al. (2013) and presented in Table 3. Several additional categories were necessary to fully describe the variety of head movements in the storyteller, e.g. slide, shift and bobble. The inter-annotator agreement values found for the full inventory in Kousidis et al. (2013) equaled 77% for event segmentation, 74% for labeling and 79% for duration. The labeling of storyteller's head gestures has been completed in 9 conversations in the no-distraction subset so far, as the density and complexity of gestural phenomena is much greater than in the listener.

5. Analysis and results

5.1. Analysis toolchain

Typically, ALICO annotations prepared in Praat and ELAN are combined and processed using TextGridTools (Buschmeier and Włodarczak, 2013, `http://purl.org/net/tgt`), a Python toolkit for manipulating and querying annotations stored in Praat's TextGrid format. Data analyses are then carried out in a Python-based scientific computing environment (IPython, NumPy, pandas, SciPy, matplotlib; McKinney, 2012) as well as in R when more complex statistical methods are needed.

5.2. Results

Analyses on the ALICO corpus so far show that distracted listeners communicate understanding by feedback significantly less frequently than attentive listeners (Buschmeier et al., 2011). They do however, communicate acceptance of the interlocutor's message, thereby conveying *implied* understanding. We discuss this strategy in a few possible pragmatic scenarios in Buschmeier et al. (2011).

Furthermore, the ratio of non-verbal to verbal feedback significantly increases in the distracted condition, suggesting that distracted listeners choose a more basic modality of expressing feedback, i.e. with head gestures rather than verbally (Włodarczak et al., 2012). We also found that spoken feedback expressions of distracted listeners have a different prosodic profile than those produced by attentive listeners (Malisz et al., 2012). Significant differences were found in the intensity and pitch domain.

Regarding the interaction between modalities and feedback functions in the corpus, Włodarczak et al. (2012) found that in HGUs overlapping with verbal feedback expressions (bimodal feedback), nods, especially multiple ones, predominate. However, the tilt was found to be more characteristic of higher feedback categories in general, while the jerk was found to express understanding. A significant variation shown in the use of the jerk, between distracted and attentive listeners (Włodarczak et al., 2012) is in accordance with the previous result in Buschmeier et al. (2011). Hitherto ALICO provided two converging sources of evidence confirming the hypothesis that communicating *understanding* is a marker of attentiveness.

Beyond the analysis of correlates of distractedness and multimodal feedback function, Inden et al. (2013) report on timing analyses of multimodal feedback in ALICO. The analysis, conducted on attentive listeners only, was implemented in an artificial agent by Inden et al. (2013). The results indicate that listeners distribute head gestures uniformly across the interlocutor's utterances, while the probability of verbal and bimodal feedback increases sharply towards the end of the storyteller's turn and into the following pause. While the latter hypothesis is established, the former was not strongly attested in the literature: the specific nature of the conversational situation in ALICO, strongly concentrated on active listening, provided a sufficiently constrained setting, revealing the function of the visual modality in this discourse context.

Most recent results suggest that onsets of Head Gesture Units in attentive listeners are timed with the *interlocutor's* vowel onsets, providing evidence that listeners are entrained to the vocalic rhythms of the dialogue partner (Malisz and Wagner, under review).

6. Conclusions and future work

The Active Listening Corpus offers an opportunity to study multimodal and cognitive phenomena that characterise listeners in spontaneous dialogue and to observe mutual influences between dialogue partners. The annotations are being continuously updated. Work on additional tiers containing lexical, morphological information, turn segmentations and further prosodic labels is ongoing. A corpus extension is planned with recordings using motion capture and gaze tracking available in the MintLab (Kousidis et al., 2012).

7. Acknowledgements

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence EXC 277 "Cognitive Interaction Technology" (CITEC).

8. References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Plínio A. Barbosa. 2006. *Incursões em torno do ritmo da fala*. Pontes, Campinas, Brasil.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79:941–952.

Stefan Beňus, Augustín Gravano, and Julia Hirschberg. 2011. Pragmatic aspects of temporal accomodation in turn-taking. *Journal of Pragmatics*, 43:3001–3027.

Paul Boersma and David Weenink. 2013. Praat: Doing phonetics by computer [computer program]. Version 5.3.68, `http://www.praat.org/`.

Mara Breen, Laura C. Dilley, John Kraemer, and Gibson Edward. 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (tones and break indices) and RaP (rhythm and pitch). *Corpus Linguistics and Linguistic Theory*, 8:277–312.

Hendrik Buschmeier and Marcin Włodarczak. 2013. TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung*, pages 152–157, Bielefeld, Germany.

Hendrik Buschmeier and Stefan Kopp. 2012. Using a Bayesian model of the listener to unveil the dialogue information state. In *SemDial 2012: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, pages 12–20, Paris, France.

Hendrik Buschmeier, Zofia Malisz, Marcin Włodarczak, Stefan Kopp, and Petra Wagner. 2011. 'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness. In *Proceedings of Interspeech 2011*, pages 2057–2060, Florence, Italy.

Iwan de Kok and Dirk Heylen. 2011. The MultiLis corpus – Dealing with individual differences in nonverbal listening behavior. In Anna Esposito, Antonietta M. Esposito, Raffaele Martone, Vincent C. Müller, and Gaetano

Scarpetta, editors, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, pages 362–375. Springer-Verlag, Berlin, Germany.

Jens Edlund, Matthias Heldner, Al Moubayed Samer, Agustín Gravano, and Hirschberg Julia. 2010. Very short utterances in conversation. In *Proceedings Fonetik 2010*, pages 11–16, Lund, Sweden.

Augustín Gravano, Stefan Beňus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Proceedings of Interspeech 2007*, pages 1613–1616, Antwerp, Belgium.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Proceedings of Nordic Prosody XI*, pages 137–146, Tartu, Estonia.

Dirk Heylen, Elisabetta Bevacqua, Catherine Pelachaud, Isabella Poggi, Jonathan Gratch, and Marc Schröder. 2011. Generating listening behaviour. In Paolo Petta, Catherine Pelachaud, and Roddy Cowie, editors, *Emotion-Oriented Systems: The Humaine Handbook*. Springer-Verlag, Berlin, Germany.

Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. 2013. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th International Conference on Multimodal Interaction*, pages 181–188, Sydney, Australia.

Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal processing via web services: The use case WebMAUS. In *Proceedings of the Workshop on Service-oriented Architectures for the Humanities: Solutions and Impacts*, pages 30–34, Hamburg, Germany.

Stefan Kopp, Jens Allwood, Karl Grammar, Elisabeth Ahlsén, and Thorsten Stocksmeier. 2008. Modeling embodied feedback with virtual humans. In Ipke Wachsmuth and Günther Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, pages 18–37. Springer-Verlag, Berlin, Germany.

Spyros Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. 2012. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue*, pages 39–42, Stevenson, WA, USA.

Spyridon Kousidis, Zofia Malisz, Petra Wagner, and David Schlangen. 2013. Exploring annotation of head gesture forms in spontaneous human interaction. Proceedings of the Tilburg Gesture Meeting (TiGeR 2013), Tilburg, The Netherlands.

Anna K. Kuhlen and Susan E. Brennan. 2010. Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling. *Discourse Processes*, 47:567–587.

Zofia Malisz and Petra Wagner. Under review. Listener rhythms. Manuscript.

Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Stefan Kopp, and Petra Wagner. 2012. Prosodic characteristics of feedback expressions in distracted and non-distracted listeners. In *Proceedings of The Listening Talker. An Interdisciplinary Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions*, pages 36–39, Edinburgh, UK.

Wes McKinney. 2012. *Python for Data Analysis*. O'Reilly, Sebastopol, CA, USA.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2013. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7:19–28.

Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. A model of attention and interest using gaze behavior. In *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents*, pages 229–240, Kos, Greece.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown University Press, Washington, DC, USA.

Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: A study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pages 78–84, Funchal, Madeira, Portugal.

Khiet P. Truong, Poppe Ronald, Iwan de Kok, and Heylen Dirk. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proceedings of Interspeech 2011*, pages 2973–2976, Florence, Italy.

Petra Wagner, Zofia Malisz, Benjamin Inden, and Ipke Wachsmuth. 2013. Interaction phonology – A temporal co-ordination component enabling representational alignment within a model of communication. In Ipke Wachsmuth, Jan de Ruiter, Petra Jaecks, and Stefan Kopp, editors, *Alignment in Communication. Towards a new theory of communication*, pages 109–132. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32:1177–1207.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1556–1559, Genoa, Italy.

Marcin Włodarczak, Hendrik Buschmeier, Zofia Malisz, Stefan Kopp, and Petra Wagner. 2012. Listener head gestures and verbal feedback expressions in a distraction task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue*, pages 93–96, Stevenson, WA, USA.

Narayan Yoganandan, Frank A. Pintar, Jiangyue Zhang, and Jamie L. Baisden. 2009. Physical properties of the human head: Mass, center of gravity and moment of inertia. *Journal of Biomechanics*, 42:1177–1192.