In: *Language Resources and Evaluation,* 2016, 50(2):411–442. **POSTPRINT**
The final publication is available at Springer via http://doi.org/10.1007/s10579-016-9355-6.

The ALICO corpus: analysing the active listener

Zofia Malisz · Marcin Włodarczak · Hendrik Buschmeier · Joanna Skubisz · Stefan Kopp · Petra Wagner

Received: 2014-12-09 / Accepted: 2016-05-03 / Published online: 2016-05-21

Abstract The *Active Listening Corpus* (ALICO) is a multimodal data set of spontaneous dyadic conversations in German with diverse speech and gestural annotations of both dialogue partners. The annotations consist of short feedback expression transcriptions with corresponding communicative function interpretations as well as segmentations of interpausal units, words, rhythmic prominence intervals and vowel-to-vowel intervals. Additionally, ALICO contains head gesture annotations 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 briefly present the most important results obtained from analyses in recent years.

This article is an extended version of the paper 'ALICO: A multimodal corpus for the study of active listening' (Buschmeier et al 2014) presented at LREC 2014, the 9th Conference on Language Resources and Evaluation. Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier and Joanna Skubisz have contributed equally to this article.

Zofia Malisz
Department of Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
E-mail: malisz@kth.se

Marcin Włodarczak
Department of Linguistics, Stockholm University, Stockholm, Sweden
E-mail: wlodarczak@ling.su.se

Hendrik Buschmeier · Stefan Kopp
Faculty of Technology and CITEC, Bielefeld University, Bielefeld, Germany
E-mail: {hbuschme,skopp}@uni-bielefeld.de

Joanna Skubisz
Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisbon, Portugal
E-mail: joanna.skubisz@fcsh.unl.pt

Petra Wagner
Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany
E-mail: petra.wagner@uni-bielefeld.de

Keywords active listening · multimodal feedback · backchannels · head gestures · attention · multimodal corpus

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 mutual 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). Recent efforts in the collection and analysis of listener data can be found in de Kok and Heylen (2011) and Heldner et al (2013).

In the present paper, we report on the design and main results obtained from the *Active Listening Corpus* (ALICO), collected at Bielefeld University. ALICO is a multimodal corpus of German dialogues built to study spoken and gestural behaviour in face-to-face communication, with a special focus put on the listener. The corpus consists of 50 dialogues in which the communicative context – interaction with a storytelling partner – was set up to facilitate spontaneous, active listening behaviour. Short feedback expressions (henceforth SFEs) were extracted from the corpus (cf. Schegloff 1982; Ward and Tsukahara 2000; Edlund et al 2010) and classified using an inventory of communicative feedback functions (Buschmeier et al 2011). Short feedback expressions of the listeners have been annotated and analysed for 40 out of 50 ALICO dialogues so far. The corpus also includes head gesture annotations for both the listener (in 40 dialogues) and the storyteller (in 9 dialogues), along with gesture type tags with categories such as nod, shake, or tilt.

Additionally, the ALICO conversational sessions feature a task in which the listener's attention was manipulated experimentally. Previous studies reported that the listener's attentional state had 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. Both Bavelas et al (2000), and the results of a similar study by Kuhlen and Brennan (2010), demonstrated that the preoccupied listener produced less context-specific feedback, suggesting that listener distractedness affected the behaviour of the interlocutor and interfered with the speaker's speech. At the same time, the effect of distractedness on the verbal and non-verbal behaviour in the listener herself has so far received little attention. We adapted the task used by Bavelas et al (2000) in half of the ALICO conversational sessions to allow an investigation into how active listening behaviour changes when the attention level is varied in dialogue. In several analyses of the ALICO data (Buschmeier et al 2011; Malisz et al 2012; Włodarczak et al 2012), we managed to reveal some of the communicative strategies listeners used when distracted.

The corpus was also built for the purpose of studying temporal relations across modalities, within and between interlocutors. The rhythmic annotation layer in the speaker's speech (vocalic beat intervals and rhythmic prominence intervals) has been

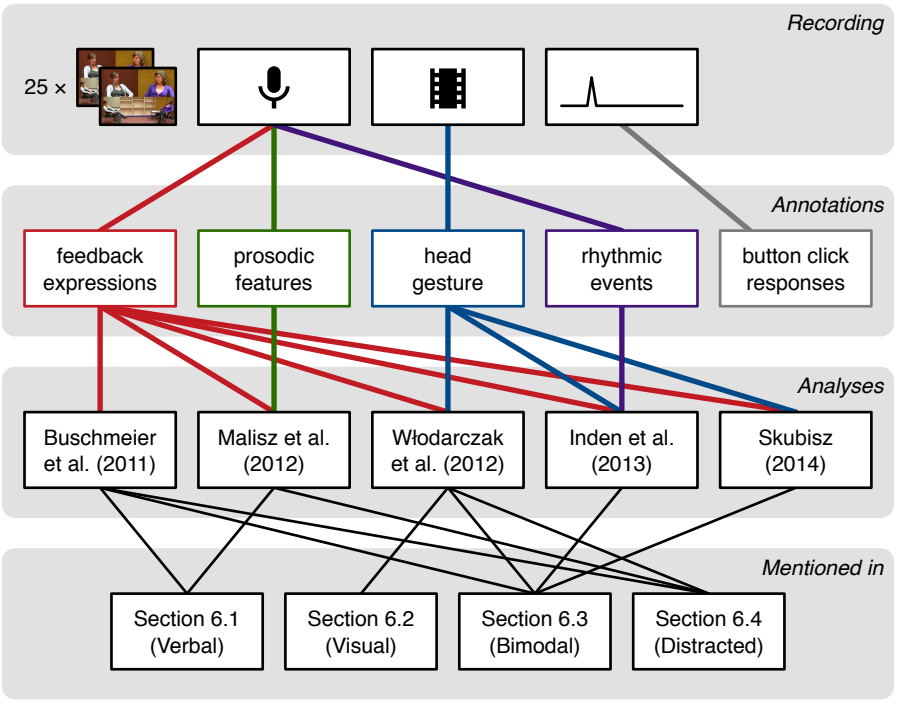

Figure 1 Overview of the work done on the ALICO corpus material. Connections show how annotations are related to raw data, which published analyses use which annotations and which sections in this article discuss these analyses. See also Table 1 and Table 13 for data overviews.

annotated in 20 dialogues and has served as input for coupled oscillator models providing an important testbed for hypotheses concerning interpersonal entrainment in dialogue (Wagner et al 2013). Inden et al (2013) reported on the first evaluations of entrained timing behaviour in two modalities implemented in an artificial agent.

The unique features of ALICO enable a targeted study of active listening with varying listener attention levels in the context of spontaneous interaction, thereby contributing to a better understanding of human discourse. The data is particularly well suited for studying temporal interactions between multimodal phenomena, effects of distractedness on communication and basic mechanisms of interaction, especially related to providing feedback and establishing common ground. In addition to informing basic research on human–human dialogue, the corpus can be useful for fields such as interaction quality monitoring, where detecting distracted users could assist the early prevention of communication problems. More generally, the design of ALICO resulted in conversations rich in feedback behaviour that could improve existing models of feedback implemented in dialogue systems and help build more human-like and sociable conversational agents. Indeed, analysis outcomes have already proven useful in these domains (Inden et al 2013).

The findings so far indicate a link between feedback form and its pragmatic function in both visual and verbal domains. Temporal and functional interdependencies

between the two domains further suggest complex patterns of mutual influence underlying multimodal interaction. For instance, while complex head gestures are more likely to be accompanied by verbal feedback, simple head movements exhibit tighter synchronisation with the verbal counterpart, and visual-only feedback is shorter than bimodal. In addition, distractedness was found to significantly influence both the amount and the type of feedback given by listeners. Specifically, distracted listeners produce less feedback overall and a decrease in communicating understanding of the interlocutor's message emerged as a consistent marker of distractedness across the modalities. Annotation and analysis of the remaining material is underway.

In the present paper, we describe the main corpus characteristics and summarise the most important results obtained from analyses done so far, demonstrating its utility to studies of spontaneous multimodal communication. In Sect. 2 we define the multimodal feedback phenomena that constitute the core of the ALICO corpus. Section 3 provides an overview of the corpus architecture. Sections 4 and 5 discuss annotations of verbal and non-verbal modalities in the data set. Section 6 summarises work to date on multimodal aspects of feedback in ALICO, followed by conclusions and plans for future work in Sect. 7.

A schematic guide to previous studies on ALICO and to the relevant discussion sections in this work is provided in Fig. 1.

2 Listener feedback in spontaneous interaction

Although the active speaker usually fulfils 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' constitute the majority of listener turns and form an integral part of face-to-face communication. Such verbal feedback expresses the ability and willingness to interact and understand, as well as conveys listener emotions and attitudes. The duties of grounding in dialogue fulfilled by verbal feedback (Clark 1996; Clark and Schaefer 1989) are shared with head gestures and other non-vocal behaviour. Additionally, head movements, by co-occurrence with mutual gaze (Peters et al 2005) and correlation with other active listening displays, emphasise the degree of listener involvement in conversation. By means of head gestures, the listener can also encourage the speaker to stay active during his or her speech at turn relevance places (Wagner et al 2014; Heldner et al 2013). Both kinds of feedback are discussed in more detail in the following sections.

2.1 Spoken feedback in active listeners

Short feedback expressions, as understood in the present paper, are located at the intersection of several other concepts found in literature defining dialogue phenomena, most significantly *backchannels* (Yngve 1970), *continuers* (Schegloff 1982), *discourse markers* (Grosz and Sidner 1986), or *accompaniment signals* (Kendon 1967). Unfortunately, distinctions between them are vague at best. In what was perhaps the most stringent attempt so far, Ward and Tsukahara (2000) give an overview of definitional criteria commonly adopted in literature, which include: responsive character to

the content of the interlocutor's utterance, optionality, brevity, expressing attitudinal meaning, non-turn-taking character and facilitating communication flow.

More generally, Schegloff (1982) groups theories of human interaction into two broad categories: those which describe feedback phenomena in terms of providing "evidence of attention, interest, and/or understanding on the listener's part" (e.g. Kendon 1967; Clark and Schaefer 1989) and those which treat them as a means of active listening and a mechanism of interspeaker coordination (e.g. Duncan and Fiske 1977).

In accordance with grounding theory (Clark and Schaefer 1989), we assume that SFEs are tightly linked to coordinative processes underlying successful communication, a position that motivated analyses of rhythmic feedback coordination in this data (Wagner et al 2013). At the same time, we explicitly manipulate attention levels in the recorded sessions by introducing distractors in order to describe the differences in feedback behaviour as a function of attention.

In the present study, we also take the middle ground on another distinction made in determining the concept of feedback, namely between semantic/functional (Allwood et al 1992; Schegloff 1982) and formal definitions relating to surface realisation (e.g. brevity, lexical form, position within the turn, Ward and Tsukahara 2000; Edlund et al 2010).

Hence, we adopt a predominantly functional definition of SFEs as indicating uptake of interlocutor's utterance, while at the same time incorporating certain surface criteria. In particular, following Koiso et al (1998), we require that, as the name implies, SFEs be short. Our approach thus stands in contrast to the original definition of backchannels by Yngve (1970), which also included longer stretches of speech, for instance, completions by the listener. Admittedly, this is a methodological compromise since longer feedback utterances do not yield themselves to the intended analyses for reasons of incomparable prosodic structure and timing properties (Wagner et al 2013; Heldner et al 2013; Włodarczak et al 2015). At the same time, surface criteria such as lexical type alone are not sufficient since, e.g., some occurrences of 'ja' function as answers or discourse markers rather than as feedback.

More importantly, in accordance with formal semantic accounts, we posit that the feedback category comprises a wide range of pragmatic meanings from merely conveying perception of interlocutor's communicative behaviour to higher-level evaluation of the message (acceptance, surprise, etc., Allwood et al 1992). However, as we show in Sect. 6.1, adopting a formal taxonomy leads to interpretative ambiguities when confronted with real data, suggesting that semantic definitions be augmented to allow optional interpretations related to sequential organisation of discourse (Gravano et al 2007) and affective stance (Kopp et al 2008; Buschmeier and Kopp 2012).

2.2 Head gesture feedback in active listeners

Functions of communicative head movements include marking contrast between topics, referring to objects by pointing with the head in space (McClave 2000; Heylen 2006) or eliciting feedback from dialogue partners (McClave 2000; Goodwin 1981). These functions are most frequently expressed by rotating the head vertically or laterally.

Here, we constrain our functional definition of head movements in a similar fashion as the SFE definition: we take that establishing common ground, displaying attention and facilitating dialogue coordination by providing feedback are the most prevalent reasons to use head gestures in communication, particularly in listeners (Włodarczak et al 2012; Heldner et al 2013).

Head movement forms are relatively free, since they are much less conventionalised than speech forms but more constrained than manual gestures. This freedom complicates the mapping of head gesture form and function. However, we observe that most annotation schemes developed so far describe simplified models of head movement forms that already carry a functional load. The load is due to the frequent re-occurrence and a certain coherence of specific head movement shapes in human dialogue, from which a group of prototypes emerges: nods, shakes, turns, etc. (see also Table 3). These prototypes are typically later analysed and matched with more detailed meanings.

Recent analyses of multimodal corpora report that head nodding predominates among head gestures in free conversation (56% in a multimodal corpus of Japanese in Ishi et al (2014)). When the listener role is assigned to the participants, 81.5% of head movements are nods, as previous analyses of the corpus featured in the present study revealed (Włodarczak et al 2012).

Several authors have attempted to further subclassify nodding into more finely grained feedback functions. However, surveys of the literature (Wagner et al 2014) reveal that the functional interpretation of a simple distinction such as single vs. multiple cycle nodding, varies from language to language (Hadar et al 1985 for English, Cerrato 2007 for Swedish, Włodarczak et al 2012 for German, Ishi et al 2014 for Japanese) and depending on the dialogue task and analysis.

Some general functional distinctions between gestures based on criteria such as the number of nodding cycles are possible. For example Hadar et al (1985) differentiated between linear and cyclic head movement forms and categorised between communicative and non-communicative head movement frequency ranges, equivalent to e.g. single and multiple nodding bouts. The approach in Hadar et al (1985) attempted at objectivising the description of head movement forms by using kinematic measurements. Modern motion tracking systems further enable the operationalisation of higher order derivatives of movement (velocity, jerk) and precise points of maximum extension. An attempt to map head movement prototypes to kinematic parameters derived from motion tracking was made in Kousidis et al (2012).

Qualitative descriptors such as jerkiness (Poggi et al 2010) or fluidity (Hartmann et al 2006) were also proposed, usually in order to find correlations between movement features and contextual or attitudinal information. Poggi et al (2010) suggest that simple vertical movements of the head share basic semantic cores that depend on speakership roles: in speaking turns, head nods communicate levels of "importance", while in listening turns, the main message conveyed is feedback (Poggi et al 2010 use the term "acceptance").

In the present work, we use the time-aligned annotations of head movement prototypes (Table 3) and report on several analyses relating the SFE function and the function of the co-occurring head movements. We also look at the temporal relations between the two modalities that might point to semantic coherence.

Figure 2 The ALICO recording session setup. Screenshot from a video file capturing the whole scene (*long shot*), and individual participants (*medium shots*). In the distracted listener condition, the listener is being distracted by counting words beginning with the letter 's' and pressing a button on a remote control hidden in her left hand.

3 Corpus architecture

ALICO consists of 50 same-sex conversations between 25 dyads of German native speakers (34 female and 16 male) recorded audio-visually. All participants were students at Bielefeld University and, apart from four 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 a payment of 4 euros. None of the participants reported any hearing impairments. The total length of the recorded material is 5 hours 12 minutes. Dialogues have a mean length of 6 minutes and 36 seconds (Min = 2:00 min, Max = 14:48 min, SD = 2:50 min). Table 13 in the Appendix summarises the data file types, sizes and annotation progress at the time of publication.

The sessions were recorded in a studio at Bielefeld University (MIntLab; Kousidis et al 2012). Dialogue partners were placed approximately three metres apart (to minimise cross-talk) in a comfortable setting. Participants wore high quality headset microphones (Sennheiser HSP 2 and Sennheiser ME 80) while another condenser microphone captured the whole scene. Camcorders (Sony VX 2000 E) recorded the interactions from three camera perspectives: medium shots showing the storyteller and the listener – enabling future fine-grained analysis of their head gestures – and a long shot showing the whole scene. Fig. 2 shows the setting and one of the dyads from all three perspectives.

Face-to-face dialogue forms the core of the corpus. The conversational scenario engaged one dialogue partner, the *storyteller*, in telling two holiday stories. The other participant, the *listener*, was instructed to listen actively to the stories told by the storyteller, make remarks and ask questions, if they pleased. Participants were assigned to their roles randomly and received their instructions separately.

Furthermore, similarly to Bavelas et al (2000), the listener was engaged in an ancillary task during one of the stories (the task order was counterbalanced across

Table 1 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). Speaker head gestures are annotated for nine speakers (see Sect. 5.2). Button press response-click annotations were extracted automatically from the audio.

		Annotation		Role		Condition	
	Tiers	Scheme	Examples	L	S	A	D
Speech	orthographic transcription	[kis]	'Reise'		✓	✓	
	phonetic transcription (SAMPA)	[kis]	`RaIz@`		✓	✓	
	phonemic segmentation	[kis]	`R`, `aI`, `z`, `@`		✓	✓	
	vowel-to-vowel interval		interval		✓	✓	✓
	rhythmic prominence interval	[bre]	interval		✓	✓	✓
IPU	interpausal units	[bre]	IPU, pause		✓	✓	✓
Feedback	feedback expressions	[bus]	'ja', 'm', 'okay'	✓		✓	✓
	feedback functions	[bus]	P1, P3A, N2	✓		✓	✓
Head	speaker head gesture units	[kou]	slide-1-right		✓	✓	
	listener head gesture units	[wło]	jerk-1+nod-2	✓		✓	✓
Click	button-click 's' responses		timestamps	✓			✓

[*bre*]—Breen et al (2012), [*kis*]—Kisler et al (2012), [*bus*]—Buschmeier et al (2011), [*kou*]—Kousidis et al (2013), [*wło*]—Włodarczak et al (2012)

dyads): he or she were to press a button on a hidden remote control (see Fig. 2) every time they heard a word beginning with a letter 's', which 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 to ensure they perform both tasks. The storyteller was made aware that the listener is going to search for something in the stories but no further information about the details of the listener's tasks was disclosed.

4 Speech annotation

Annotation of the interlocutors' speech was performed in Praat (Boersma and Weenink 2013) and post-processed using TextGridTools (Buschmeier and Włodarczak 2013), independently from head gesture annotation. As speech annotation differs for listener and speaker role in the corpus (see Table 1 for an overview of the respective annotation tiers), they are discussed in turn below.

4.1 The listener

We annotated short feedback expressions produced by the listener and interpreted the corresponding communicative feedback functions in 40 dialogues thus far, i.e. in 20 sessions involving the distraction task and 20 sessions with no distractions. We carried out listener SFE segmentation automatically on the basis of signal intensity in Praat and we subsequently checked and adjusted the segmentation manually. Another

annotator then transcribed the pre-segmented SFEs according to German orthographic conventions. Longer non-SFE listener turns were marked but not transcribed.

Three independent annotators assigned feedback functions to listener SFEs in each dialogue[1]. Each annotator carefully took the communicative context into account. A feedback function scheme was developed and first described in Buschmeier et al (2011), largely based on Allwood et al (1992). The inventory involves core feedback function categories 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. The levels can be treated as a hierarchy with an increasing value judgement of grounding 'depth' (Allwood et al 1992; Clark 1996; Włodarczak et al 2010). The negation of the respective functions was marked as N1–N3. Due to very low frequency of negative feedback, we discuss only positive feedback in the present paper.

The annotators had an option to extend listener feedback function categories with three modifiers. 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). Modifier A in turn referred to the listener's emotions or attitudes co-occurring with core functions, leading to categories such as P3A (Kopp et al 2008; Buschmeier and Kopp 2012). If an SFE functioned as a clear emotional or attitudinal display, the annotators were allowed to use the category A as a single, core category. Therefore, the attitudinal/emotional category is the only one in the ALICO scheme that can fulfil both a modifying and a core function. Table 2 summarises the resulting inventory of feedback functions.

Automatically derived majority labels determined the final feedback function interpretation. The annotators discussed the remaining disagreements, i.e. cases which could not be settled by majority labels (10% of all tagged feedback expressions) and resolved them manually. We discuss the procedure and its implications for interpretation of feedback functions in greater detail in Sect. 6.1.

4.2 The storyteller

Twenty sessions not involving a distraction task contain storyteller's speech annotations. In addition, annotators delimited the following rhythmic phenomena in storyteller speech: vowel-to-vowel intervals, rhythmic prominence intervals and minor intonational phrases (Breen et al 2012). First, vowel onsets were extracted semi-automatically from the data. Algorithms in Praat (Barbosa 2006) were used first, after which two annotators checked the resulting segmentation for accuracy by inspecting the spectrogram, formants and pitch curve in Praat. Rhythmic prominences e.g., perceptually salient syllables were judged perceptually. Whenever an annotator perceived a 'beat' on a given syllable, the syllable was marked as prominent, regardless of lexical or stress placement rules (Breen et al 2012). Similarly, every time a perceptually discernible gap occurred in the storyteller's speech, a phrase boundary was marked.

[1] Four annotators in total worked on the feedback function interpretation in ALICO, namely the first four authors of this paper, out of which JS, MW and ZM are competent but not native speakers of German. Annotation tasks were assigned in rotation to three annotators per recorded session.

Table 2 Feedback function inventory. Categories P1–P3, N1–N3 and the category and modifier A are based on Allwood et al (1992) and Kopp et al (2008). Modifiers C and E were adopted from Gravano et al (2007).

Label	Category	Modifier	Definition
P1	✓		The partner signals perception of the signal. *'I hear you and please continue.'*
N1	✓		The partner signals problems with perception. *'What are you saying?'*
P2	✓		The partner signals perception and understanding of the message content. *'I understand what you mean.'*
N2	✓		The partner signals perception of the message and problems with understanding the content. *'I do not understand what you mean.'*
P3	✓		The partner signals perception, understanding and acceptance of the message or agreement with the message. *'I accept/agree/believe what you say.'*
N3	✓		The partner signals perception and understanding but rejection or disagreement of the message. *'I disagree/do not accept what you say.'*
A	✓	✓	The partner expresses an attitude towards the message, e.g., surprise, excitement, admiration, anger, disgust.
C		✓	The partner introduces a new discourse segment or topic.
E		✓	The partner ends the current discourse segment or topic.
?	✓		Unresolved.

The resulting minimum pause length of 60 ms is comparable to pauses between so called 'interpausal units' used in other studies (e.g., Beňuš et al 2011).

Apart from manual rhythmic segmentation, the storyteller's speech was automatically segmented by forced alignment using the WebMAUS tool (Kisler et al 2012). WebMAUS produces a fairly accurately aligned and multi-layered annotation on small linguistic units, e.g. in segmented data. The output of the aligner provides tiers with word and phoneme segmentation along with SAMPA transcription. All automatically aligned tiers were checked and corrected manually.

5 Head gesture annotation

The corpus contains gestural annotation of both dialogue partners (see Table 1). Annotators performed the segmentation and labelling in ELAN (Wittenburg et al 2006) by close inspection of the muted video, stepping through the video frame-by-frame. Uninterrupted, communicative head movements were segmented as minimal annotation units. 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 was marked for head gesture type. The full inventory of gesture types is presented in Table 3. We illustrate prototypical movements along particular axes in Fig. 3, using mathematical conventions for 3D spatial coordinates and following bio-mechanical and physiological studies on head movements (Yoganandan et al 2009).

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

Head gesture type	Definition	Role	
		Listener	Speaker
nod	Rotation down–up	✓	✓
jerk	Reversed nod, up–down	✓	✓
tilt	Rotation left or right	✓	✓
bobble	Repeated tilting sideways		✓
turn	Turning left or right	✓	✓
shake	Repeated turning	✓	✓
protrusion	Pushing the head forward	✓	✓
retraction	Pulling the head back	✓	✓
slide	Translation left or right		✓
shift	Repeated horizontal slides		✓
waggle	Irregular connected movement		✓

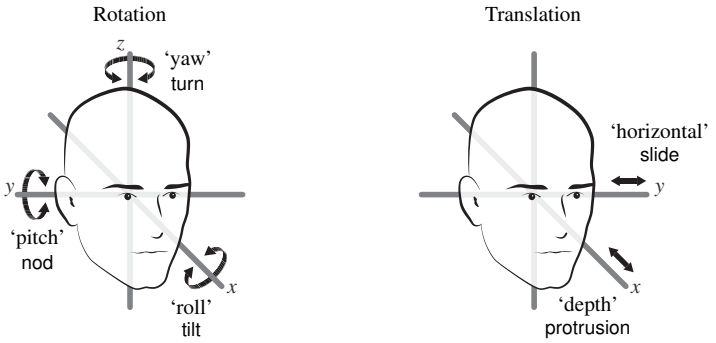

Figure 3 Schematic overview of rotations and translations along three axes (names in single quotes) as well as example movements most frequently used in communicative head gesturing (sans-serif).

The annotators identified constituent gestures in each HGU and also marked 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 is tilted to the right side of the screen.

The resulting head gesture labels describe *single*, *simple*, or *complex* gestural units. Single units refer to one head movement with one cycle, whereas complex HGUs denote multiple head movement types with different number of cycles. Simple head movement types consist of one movement type and at least two cycles (see Fig. 4). Where applicable, the annotated HGU labels also provide information about the following features: *complexity* (the number of subsequent gesture types in the phrase) and *cycle frequencies* of their component gestures.

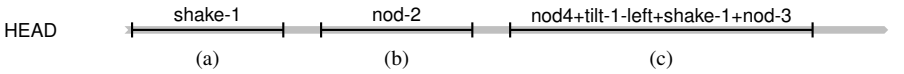

Figure 4 ALICO head gesture annotation categories that exemplify (a) single, (b) simple, and (c) complex head movement types.

Table 4 Absolute and relative frequency of listener HGU types found in 40 dialogues in ALICO.

	Frequency	
Listener HGU types	Absolute	Relative
nod	1818	0.69
jerk	108	0.04
tilt	98	0.04
shake	49	0.02
turn	44	0.02
retraction	28	0.01
protrusion	10	< 0.01
complex	443	0.17
Σ	2598	1.00

5.1 The listener

We annotated listener head gestures in 40 dialogues so far, i.e. in 20 sessions involving the distraction task and 20 sessions with no distractions. We found that listener head gesture type categories were limited to a subset of the inventory presented in Table 3, namely to nod, shake, tilt, turn, jerk, protrusion and retraction (Włodarczak et al 2012). Table 4 presents types of head gestures found for listeners in the corpus. Two annotators segmented, labelled and checked the listener HGUs for errors.

5.2 The storyteller

Co-speech head gestures produced by the storyteller are much more varied than those of the listener, which necessitated addition of several other categories, such as slide, shift and bobble. Consequently, we used the full inventory in Table 3, as described and evaluated on a similar German spontaneous dialogue corpus by Kousidis et al (2013). The inter-annotator agreement values found for the full inventory in Kousidis et al (2013) equalled 77% for event segmentation, 74% for labelling and 79% for duration. The annotation 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 in the storyteller than in the listener.

6 Multimodal feedback functions

6.1 Verbal

The listeners produced a total number of 1505 verbal feedback signals. The mean ratio of time spent producing feedback signals relative to other listener-produced turns, e.g. 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 dialogue 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: $p < 0.01$).

As the reader will recall from Sect. 4.1, three independent annotators assigned a pragmatic function from the inventory in Table 2 to each feedback expression in the listener's speech. Below we describe the observed disagreement patterns and the resolution procedure. We also consider their import on feedback function semantics.

We assess pairwise inter-annotator agreement on core verbal feedback function categories using Cohen's kappa (κ)[2]. The values range from $\kappa = 0.25$ to $\kappa = 0.43$ with a mean of $\kappa = 0.33$ (SD = 0.056; see Table 5). This degree of agreement is commonly regarded as 'fair' to 'moderate' (Landis and Koch 1977), and is usually considered to be insufficient for annotation of linguistic data (cf. Reidsma and Carletta 2008). It should be borne in mind, however, that feedback functions are not simple surface phenomena. Instead, specific dialogue acts are an expression of listener's non-observable internal states and intentions (Kopp et al 2008; Buschmeier and Kopp 2012), which makes the annotation task difficult. Importantly, comparable κ-values have been reported in existing work on feedback function annotation by naïve annotators using the DIT++ scheme (Geertzen et al 2008). While expert annotators in that study reached much higher agreement, the evaluation was based on a selection of dialogues from task-oriented corpora, which are likely to constitute a much narrower and semantically simpler domain. This is especially true of human–computer dialogues which made up the majority of the test material. By contrast, storytellers in our study constructed extended narratives pertaining to their personal experiences, which in turn elicited complex emphatic responses from listeners. Needless to say, these went far beyond the type of feedback required when interacting with a question answering railway information system or discussing a route on a map (Garrod and Anderson 1987).

More recently, Prévot et al (2015), reported a mean pair-wise κ of 0.6 for discrimination between basic feedback functions (*contact*, *acknowledgement*, *evaluation*). However, their tagset included functions related to question answering, feedback elicitation as well as *other* label indicating that an item had been erroneously classified as feedback at an earlier automatic processing stage. Arguably, these categories are much easier to identify, thus inflating the obtained κ values. For sub-classification of higher-order evaluative functions (*approval*, *expectation*, *amusement*, *confirmation/doubt*), they obtained an average agreement of 0.3, which is identical to our result.

[2] While it would be preferable to use a multi-annotator agreement measure, such as Fleiss's κ, this is somewhat problematic on the present dataset given that each dialogue was annotated by a different subset of annotators. For this reason, we resort to pairwise comparisons between individual annotators.

Table 5 Cohen's κ agreement values for feedback function annotation for all annotator pairs.

Annotator	HB	JS	MW
JS	0.25		
MW	0.33	0.29	
ZM	0.32	0.35	0.43

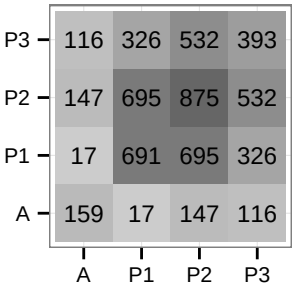

Figure 5 Confusion matrix for verbal feedback function categories. Cells show frequency of core category pairs (X,Y), with $X,Y \in \{\mathsf{P1},\mathsf{P2},\mathsf{P3},\mathsf{A}\}$, between any two annotators. Cells below the minor diagonal mirror values from above. See Table 6 for collated row-/columnwise agreement and disagreement frequencies.

Admittedly, however, the latter sub-classification task is more difficult than ours and would roughly correspond to interpreting various meanings subsumed by categories P3 and A in the ALICO inventory.

Additionally, the presented κ values were calculated using core categories obtained by stripping off any of the optional modifiers that actually formed the decision taken by the annotator. A crucial factor is the relatively subjective modifier A. It is possible that instead of committing to a core category, this modifier was attached to a core function category on occasion, especially if an interpretative overlap existed, e.g., in case of A vs. P3. Indeed, there were 164 cases in which one annotator used A as a modifier while another used it as a core category. This certainly points to a trade-off built into the inventory where interpretative flexibility given to the annotators impacts precision.

Fig. 5 presents the disagreement patterns between core feedback function categories in greater detail as a confusion matrix. The matrix reveals a substantial degree of disagreement. Only for categories P2 and A did two annotators most frequently choose the same label. By contrast, if one annotator chose P1, a second annotator most frequently chose P2. The same holds for P3. Given that feedback functions form a hierarchy of grounding strength, it seems that annotators tend to be drawn towards the middle category. Confusion matrices for individual annotator pairs can be found in the Appendix (Fig. 11).

We tabulate the proportion of agreements on a given feedback function relative to all judgements involving this function, $N(X,X)/N(X,?)$ in Table 6. This measure is equivalent to the conditional probability of an annotator choosing a particular category,

Table 6 Number of judgements in which any two annotators agreed, $N(X,X)$, and disagreed, $N(X,Y)$, on a given feedback function category. The proportion of agreements to all judgements involving this function, $N(X,X)/N(X,?)$, is compared against its empirical probability $P(X)$ and the baseline probability of any two annotators agreeing by chance, $P(X,X)$.

X	Agreement $N(X,X)$	Disagreement $N(X,Y)$	Agreement proportion $N(X,X)/N(X,?)$	Empirical prob. $P(X)$	Chance agreement $P(X,X)$
P1	691	1038	0.40	0.31	0.09
P2	875	1374	0.39	0.40	0.16
P3	393	974	0.29	0.22	0.05
A	159	280	0.36	0.08	0.01

given that another annotator chose the same category, $P((X,X)|(X,?))$. We compare the observed proportions against a baseline probability of two annotators selecting a given function category by chance, $P(X,X)$, given its empirical probability based on frequency in the data, $P(X)$. For all four core feedback function categories, if one annotator chose a particular category, the second annotator was more likely to choose any of the other categories. Here, P1 has the biggest likelihood (0.4) of another annotator choosing P1 as well, followed by P2 (0.39), A (0.36), and P3 (0.29). All of these values, however, are much higher than the probability of two annotators choosing the same category by chance.

To resolve the disagreements, the final SFE function labels used in subsequent analyses were determined algorithmically by calculating majority labels from individual annotations. If at least two out of three annotators agreed upon one of the core categories P1, P2, P3, or A, this category was chosen. If at least one of the annotators used a modifier (C, E, or A) in addition to a core category, the modifier was carried over to the final annotation label. This ensured that surface features represented by the modifiers (e.g., position relative to the full turn, an attitudinal component) as well as subtler semantic distinctions, especially related to emotional content of utterances, were preserved. Complete disagreements on core categories, i.e., cases in which all three annotators disagreed, were discussed and resolved manually.

This resolution method results in three classes of inter-annotator agreement-quality with (1) all three annotators agreeing, (2) two annotators agreeing, and (3) all annotators disagreeing on the assigned feedback function. We present the percentages of cases in each of these categories, additionally split on feedback function category as well as category type (complex vs. core) in Fig. 6. While disagreement was resolved predominantly by majority voting (57% of cases for both complex and core categories), full agreement between annotators is more frequent (33% vs. 26%) and full disagreement is less frequent (10% vs. 17%) when modifiers are stripped from listener's feedback function labels. For this reason, core categories were used in all subsequent analyses on ALICO. Concerning individual functions, all three annotators unanimously assigned P1 and P2 somewhat more frequently than P3 and A, suggesting that lower-level functions may be less ambiguous than higher-level ones. P3 in particular, showed a higher proportion of complete disagreements, which required manual intervention (12% for P3 against 8% in both P1 and P2).

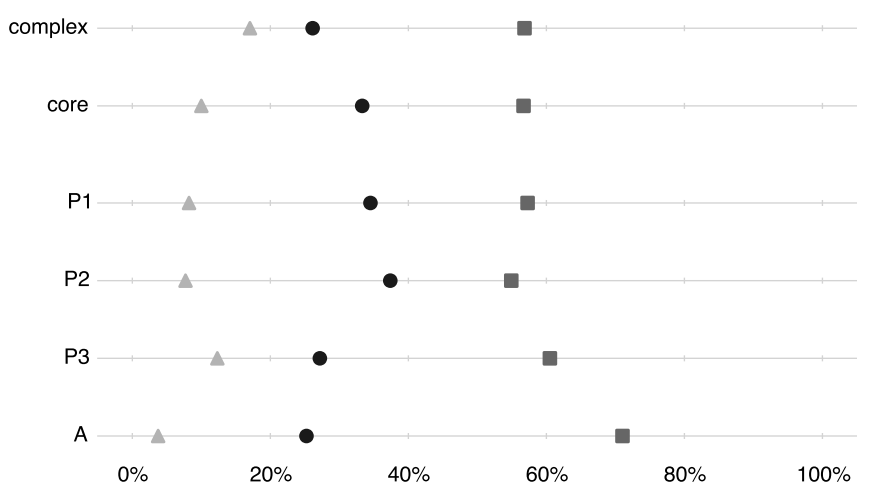

Figure 6 Distribution of inter-annotator agreement quality classes across items in general (complex vs. core feedback function) and across individual core feedback functions (P1, P2, P3, A). Markers indicate proportion of items decided upon: (●) by agreement of all three annotators; (■) by automatic majority vote taken when two of three annotators agreed; or (▲) by resolution through discussion when all annotators disagreed.

In Fig. 7, we further explore cases where the final label had to be resolved by majority voting. Since in these cases, two out of three annotators agreed on the assigned feedback function, the figure presents percentage-wise distribution of dissenting annotations given a majority category. The results are in line with the hierarchical dependencies between feedback functions implicit in our inventory. In particular, disagreements most likely involved neighbouring categories: P2 for the P1 and P3 majority labels and both P1 and P3 for the middle P2 majority label.

The observed disagreement is also consistent with the existence of entailment relations between preconditions of communicative functions (Bunt 2007; Włodarczak et al 2010). In our case, "acceptance", P3, logically implies "understanding", P2, which in turn implies "perception", P1. Fig. 7 reflects a similar tendency in that the most common minority category is logically entailed by the majority vote: P3 $\Rightarrow$ P2 and, to a lesser extent, P2 $\Rightarrow$ P1. Annotators are thus more likely to assign feedback functions with fewer rather than more preconditions, possibly because they either do not consider all higher-level function preconditions to be fully satisfied or because they choose a conservative strategy by favouring safe decisions.

We conclude the overview of listener's verbal feedback in ALICO by considering the relation of particular feedback expressions to their function interpretation. We present the most frequent German SFEs found in the corpus and their corresponding core feedback functions in Table 7. Each SFE lexical type listed, subsumes several variations that we interpret to reflect the same lexical form. For example, the lexical type 'ja' encompasses forms such as 'nja', 'na ja', 'ah ja' or 'ja ja ja', where the constituent 'ja' is repeated in quick succession. Similarly, the type 'klar' (*sure*) includes

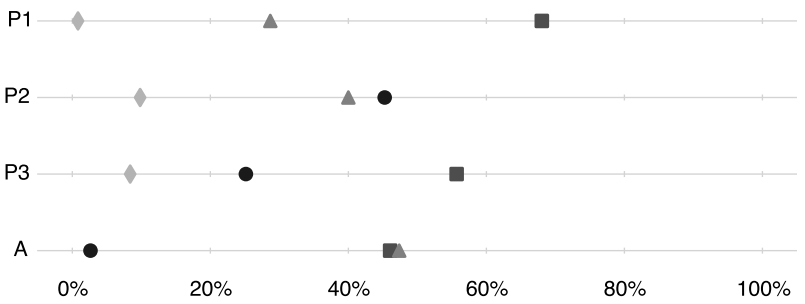

Figure 7 Distribution of minority core categories (● P1; ■ P2; ▲ P3; ♦ A) across those final individual core labels for which a majority vote was necessary. As an example, of the items that where automatically decided upon to be P1 through a majority vote (i.e., two of three annotators assigned P1), a third annotator assigned P2 in 68%, P3 in 29%, and A in 1% of the cases.

Table 7 Relative frequencies (expressed as percentages) of German short feedback expressions and their corresponding feedback functions produced by listeners in 40 ALICO dialogues. Less frequent expressions are grouped into into three semantic categories (in italics), a breakdown of which can be found in Table 14 in the Appendix.

SFE	P1	P2	P3	A	*other*	∑
'ja'	7.44	9.79	10.05	0.20	2.28	29.76
'm'	13.81	6.57	2.14	0.13	0.67	23.32
'mhm'	7.10	5.03	0.54	0.07	0.07	12.81
'okay'	0.20	7.10	3.69		0.47	11.46
excitement/enthusiasm		0.20	0.67	3.95		4.82
'ach so'		3.35			0.13	3.48
certainty		0.47	2.75	0.07		3.29
'ach'	0.20	2.21		0.60	0.13	3.14
'och'		1.27	0.20	1.41	0.07	2.95
other	0.20	0.74	0.60	0.87	0.54	2.95
'aha'	0.27	0.94			0.13	1.34
surprise/incredulity		0.13	0.07	0.40	0.07	0.67
∑	29.22	37.80	20.71	7.70	4.56	100.00

also 'na klar' or 'ja klar'. Values listed under 'ach', apart from the exclamation usually transcribed as /ax/, likewise represent all phonetic forms ending with an open, low vowel, e.g. transcribed as 'ah' or 'oah' but not subsumed under the bisyllabic 'aha' category, etc. We classify other SFEs found in the corpus into three semantic cores for the purposes of Table 7, namely *certainty* (e.g. 'genau', 'stimmt', 'richtig'), *excitement/enthusiasm* (e.g. 'wow', 'klasse', 'toll') and *surprise/incredulity* (e.g. 'echt?', 'krass', 'boah'). Table 14 in the Appendix provides a full list and glossary of these expressions.

From Table 7 it can be readily appreciated that some feedback expressions are used with certain functions more frequently than with others (on a related point see also Gravano et al 2007). The four most frequent German SFEs are 'ja', 'm', 'mhm' and 'okay', with 'ja' apparently exhibiting the most unspecific character in terms

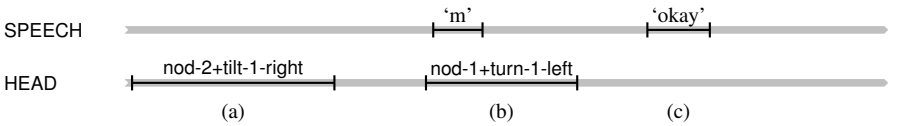

Figure 8 ALICO annotation tokens that exemplify (a) visual, (b) bimodal and (c) verbal feedback types.

of feedback functions it can be associated with. The SFEs 'm' and 'mhm' in turn are used infrequently as expressions of *agreement/acceptance* and affective displays while occurring relatively often in what can be generalised as a backchanneling capacity (P1 and P2). We find fairly good correlates of *understanding* (P2) in feedback forms such as 'okay', 'ach so' and other short exclamations in Table 7. Note that these SFEs are also potentially good 'news markers' as defined by Gardner (2001), where new information introduced by the partner into the discourse is signalled as received. The few specific lexical tokens listed in Table 7 and the Appendix under *excitement/enthusiasm*, *certainty*, *surprise/incredulity* predictably cluster within the categories P3 (especially words expressing *certainty*) and A (especially enthusiastic, surprised and disapproving attitudes).

Malisz et al (2012) also investigated the relation between the most common SFEs, 'm', 'mhm' and 'ja' and their function in a subset of 28 dialogues. The analysis revealed that 'ja' best classifies higher-level functions compared to P1. The opposite was true of 'mhm' and no statistically significant relation was found for 'm.' Additionally, feedback functions were linked to several prosodic features in Malisz et al (2012): feedback expressions communicating lower-levels of feedback were characterised by shorter durations, lower intensity, less pitch variation in their middle part and more pitch variation towards the end of the expression.

6.2 Visual

Non-verbal behaviour is an important source of information in the study of active listening. Head gestures, produced by the listener conjointly with or separately from spoken feedback expressions, significantly contribute to feedback in dialogue (Wagner et al 2014). In fact, the number of visual feedback expressions found in ALICO exceeds the number of verbal expressions: the ratio of head gesture units ($N = 2598$) to SFEs ($N = 1505$, including bimodals) equals 1.7 in 40 dialogues.

In the present section, we concentrate on a subset of 20 ALICO dialogues with complete gestural annotations for listeners not engaged in a distraction task. We describe head gesture units with respect to their modality, that is whether they overlap with SFEs or constitute 'pure' visual feedback without co-occurring speech (see Fig. 8 for an illustration). We also differentiate among HGU structures: simple, where only one gesture movement type occurs vs. complex. We additionally look at HGU cyclicity, where single repetitions of the same movement are juxtaposed with those of $N > 1$ repetitions (Fig. 4). The dependencies of complex HGUs with overlapping speech are discussed in greater detail in the following section.

Table 8 Relative and absolute frequencies of HGU types in 20 ALICO dialogues with an *attentive listener*. The data is split according to cyclicity and modality. Absolute frequencies are provided in parentheses.

		Cyclicity		Modality	
HGU type	Frequency	single	multiple	visual	bimodal
nod	786	0.22 (172)	0.78 (614)	0.66 (517)	0.34 (269)
jerk	38	1.00 (38)		0.42 (16)	0.58 (22)
tilt	33	0.94 (31)	0.06 (2)	0.70 (23)	0.30 (10)
retr	15	0.93 (14)	0.07 (1)	0.27 (4)	0.73 (11)
shake	14	0.57 (8)	0.43 (6)	0.71 (10)	0.29 (4)
turn	12	1.00 (12)		0.92 (11)	0.08 (1)
pro	3	1.00 (3)		0.33 (1)	0.62 (2)
complex	102		1.00 (102)	0.36 (37)	0.64 (65)
$\sum$	1003	0.28 (278)	0.72 (725)	0.62 (619)	0.38 (384)

Table 9 Relative and absolute frequencies of HGU types in 20 ALICO dialogues with a *distracted listener*. The data is split according to cyclicity and modality. Absolute frequencies are provided in parentheses.

		Cyclicity		Modality	
HGU type	Frequency	single	multiple	visual	bimodal
nod	1032	0.23 (236)	0.77 (796)	0.81 (836)	0.19 (196)
jerk	70	0.99 (69)	0.01 (1)	0.79 (55)	0.21 (15)
tilt	65	0.95 (62)	0.05 (3)	0.77 (50)	0.23 (15)
retr	13	1.00 (13)		0.85 (11)	0.15 (2)
shake	35	0.26 (9)	0.74 (26)	0.89 (31)	0.11 (4)
turn	32	0.94 (30)	0.06 (2)	0.84 (27)	0.16 (5)
pro	7	1.00 (7)		0.71 (5)	0.29 (2)
complex	341		1.00 (341)	0.60 (206)	0.40 (135)
$\sum$	1595	0.27 (426)	0.73 (1169)	0.77 (1221)	0.23 (374)

Table 8 lists the relative and absolute frequencies of specific head movement categories for this subset. We found a total of 1003 HGUs, with more instances of pure visual feedback (62%, $N = 619$) compared to bimodal feedback (38%, $N = 384$). Nodding constitutes the most frequent movement category in attentive gestural feedback (78.4%, $N = 786$). In contrast, the next two most frequent head movement types occur fewer than forty times each (for jerk $N = 38$, for tilt $N = 33$).

In single head movements, listeners produced twice as many multiple as single HGUs (72% vs. 28%). In accordance with previous outcomes (Włodarczak et al 2012), we found that the multiple nod is used more often than single nod labels (78% vs. 22%). In general, labels nod-2 ($N = 248$), nod-3 ($N = 154$) and nod-4 ($N = 70$) constitute the majority (60%) of all nod labels, with nod-2 being the most frequent nodding gesture (31.5%). The higher relative prevalence of polycyclical head movements in listeners, especially multiple nods, was also evidenced in other studies and languages, e.g. in English (Hadar et al 1985), in Swedish (Allwood et al 2007) and in Japanese (Ishi et al 2014).

The investigation of visual feedback in ALICO also shows a strong tendency for listeners to produce simple head movements ($N = 901$) relative to complex head ges-

tures ($N = 102$), that is, listeners generally disprefer stringing several head movement types within one head gesture unit.

We also provide an analogous summary of visual feedback frequencies for *distracted* listeners in Table 9. The table indicates that distracted listeners, while also generally preferring simple gestures, produce complex gestures relatively more frequently than attentive listeners (20% vs. 10 %). Additionally, a greater proportion of these complex gestures is realised in the purely visual modality rather than as part of a bimodal feedback expression. We discuss further details of an analysis of multimodal feedback in distracted listeners in Sect. 6.4.

6.3 Bimodal

Given the availability of verbal and visual annotations, ALICO facilitates studying mutual influences between the two communicative modalities. Above all, we investigate the possibility that those listener head movements that overlap with spoken feedback are semantically and temporally constrained by the verbal expression (see also Wagner et al 2014 for a discussion of speech-gesture interdependencies).

We again turn to attentive listeners in ALICO who produced 384 bimodal expressions, that is 38% of the head gesture units in 20 dialogues. The majority of complex HGUs, that is those containing at least two distinct head gesture types one after another, occur while in overlap with a verbal SFE. This relation is presented in Fig. 9. A χ^2 - test was applied to investigate whether the relationship between modality and HGU complexity is independent. The test revealed a statistically significant relationship ($\chi^2 = 22.7$, $p < 0.001$) between the internal complexity of an HGU and whether it overlaps with an SFE. In other words, more complex gestures, e.g. beyond the most frequent multiple nod class, are more likely to accompany verbal feedback, possibly achieving means to express relatively more specific and complementary meanings in concert.

As in Włodarczak et al (2012), we first assume that the feedback function assigned to listener's verbal expression in bimodal feedback defines the function of the overlapping head movement. That is, e.g., a double nod overlapping with a short feedback expression bearing the majority label P1 ("perception") likewise expresses perception.

Table 10 presents the frequency of listener's HGU types across feedback function categories assigned to SFEs the head gestures overlap with. The results show that the most frequent and 'generic' gestural response, i.e. the nod, happens almost equally often in overlap with P1 as P2, less so with P3. The jerk, similarly to results in Włodarczak et al (2012), seems to correlate with verbal feedback expressing 'understanding', likewise complex HGU gestures overlapping with SFEs. The label tilt is characteristic for higher level functions, increasing in frequency from P2 through to A. Summarising, nodding almost single-handedly performs in the backchanneling function P1, while other frequent head gestures such as jerk and tilt correlate with higher feedback categories.

The assumption that the function of head gestures is identical with the overlapping SFE functional category is most likely only partially viable, since communicative gestures can add complementary information to the co-occurring verbal expression, as

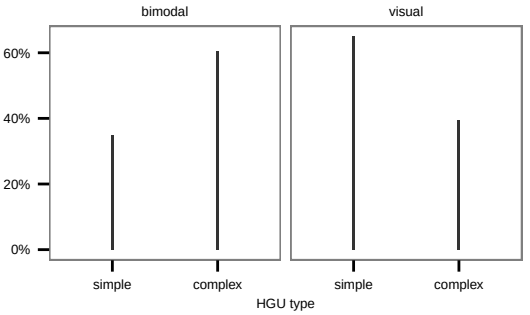

Figure 9 Proportions of simple and complex HGUs in the visual and the bimodal domains.

Table 10 Relative and absolute frequency of HGU types in bimodal feedback and their corresponding verbal feedback function category observed in 20 ALICO dialogues with an *attentive listener*.

HGU type	P1	P2	P3	A	*other*	∑
nod	29.69 (114)	27.10 (104)	8.85 (34)	1.56 (6)	2.86 (11)	70.06 (269)
jerk	0.78 (3)	4.43 (17)	0.52 (2)			5.73 (22)
tilt		0.78 (3)	0.78 (3)	1.04 (4)		2.60 (10)
retr	0.26 (1)	1.82 (7)	0.78 (3)			2.86 (11)
other	0.26 (1)	0.26 (1)	0.78 (3)	0.26 (1)	0.26 (1)	1.82 (7)
complex	3.13 (12)	8.85 (34)	3.12 (12)	1.82 (7)		16.93 (65)
∑	34.11 (131)	43.23 (166)	14.84 (57)	4.69 (18)	3.13 (12)	100 (384)

Table 11 Relative and absolute frequencies of HGU types from 10 one-minute extracts taken from ALICO and used in the rating study in Skubisz (2014). The data is split according to cyclicity and modality. Absolute frequencies are provided in parentheses.

		Cyclicity		Modality	
HGU type	Frequency	single	multiple	visual	bimodal
nod	128	0.33 (42)	0.67 (86)	0.70 (90)	0.30 (38)
jerk	11	1.00 (11)		0.55 (6)	0.45 (5)
tilt	5	1.00 (5)		0.60 (3)	0.40 (2)
retr	1	1.00 (1)			1.00 (1)
shake	4	0.50 (2)	0.50 (2)	1.00 (4)	
turn	1	1.00 (1)		1.00 (1)	
pro	1	1.00 (1)		1.00 (1)	
complex HGUs	8		1.00 (8)	0.63 (5)	0.37 (3)
∑	159	0.40 (63)	0.60 (96)	0.69 (110)	0.31 (49)

well as carry redundant meaning (Goldin-Meadow et al 1993; Bergmann and Kopp 2006), not to mention attitudinal and affective nuancing.

Consequently, Skubisz (2014) asked if communicative feedback functions of head gestures could be identified independently from speech. She applied the feedback function inventory from Buschmeier et al (2011) in the non-verbal modality. Ten

Table 12 Inter-rater agreement results (%). Head movement type majority agreement among 11 annotators for most frequent head gesture types. The no agreement category includes cases where fewer than 6 annotators agreed.

HGU type	P1	P2	P3	*no agreement*	Σ
nod	45.31	36.72	5.47	12.50	100
jerk	27.27	36.36	27.27	9.10	100
complex	45.00	50.00		5.00	100

one-minute long clips from different ALICO dialogues involving an attentive listener were used in a rating task. Clips with at least 10 HGUs and the smallest number of listener's verbal interruptions were selected. The resulting excerpts contained 159 listener's HGUs and exhibited similar distributions of visual (69%) and bimodal feedback (31%), single (40%), multiple (60%) and complex gestures as the ones found in the whole ALICO corpus (compare Table 8). For instance, nods amounted to 80.5% of all gestures in the test sample vs. 78% in ALICO. For jerk the respective proportions were 7% and 3%, and for tilt equalled 3.2% in both datasets, yielding a comparable set of gestures types to be rated.

Participants assigned three positive feedback functions, namely, perception (P1), understanding (P2), and agreement (P3) to the head movement gestures of the listener by inspecting the video clips carefully. Modifier categories listed in Table 2 were not available to the raters in this study. The video material was presented without the sound track to avoid the impact of verbal feedback expressions on the annotators' judgement.

Agreement among 11 annotators was first estimated by Fleiss' κ for 11 raters. The resulting value shows low agreement among all the raters: $\kappa = 0.22$. Skubisz (2014) further analysed the level of agreement in this task by calculating percentage agreement on feedback function ratings per head movement type. Table 12 presents majority based categories determined by a vote of at least six raters and their relative frequencies per gesture type. Head gesture types that appeared at least eight times in the watched and rated clips of the listener are shown in Table 11. The results revealed that the majority of raters could not agree on the function of ca. 11% of all HGUs. The majority agreement has shown that nods were mostly assigned to the P1 and P2-function. It seems that only jerk elicited a fair amount of classifications (27%) into the 'acceptance' category (P3) with most instances agreed upon for P2.

The results of the first study on independent feedback function assignment to head gestures that we are aware of indicate difficulties in judging the meaning of the movement, at least with the scheme proposed in Table 2. The scheme was originally developed for spoken feedback. A restricted number of categories was used in the HGU rating task, the inter-rater agreement values are nonetheless low. It was possible however, to establish majority labels for specific head gesture types where up to 90% of the most frequent tokens were classified. The patterns observed follow in part the conclusions on the functional interpretation of bimodals, e.g. especially in case of the jerk label, classified as 'understanding'.

Linguistic constraints on movement may also manifest themselves in the location of the head gesture onset relative to the verbal expression and the relative duration

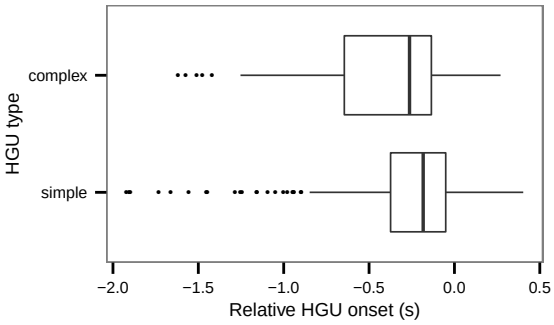

Figure 10 Distribution of HGU onset relative to utterance onset in simple and complex HGUs. Positive values signify gesture lag and negative values signify gesture lead.

of the two components. Studies have shown that the onset of communicative manual gestures tends to precede their lexical affiliate by a few hundred milliseconds (Kendon 1980; Schegloff 1984; Nobe 2000; de Ruiter 2000). In case of multimodal feedback, the SFE corresponds to the lexical affiliate that the head movement overlaps with. In Włodarczak et al (2012), the median distance from the head movement onset to the overlapping verbal expression onset, i.e. the lead of the head gesture, equalled 220 milliseconds.

In the present study, based on a superset of the data analysed in Włodarczak et al (2012), the gesture onset predominantly precedes listener's verbal expression by a median of 190 ms (min. lead= -1920 ms, max. lag= 400 ms; cf. Dittmann and Llewellyn 1968). We plot the distribution of head movement onset in bimodal feedback in Fig. 10 for the dialogue sessions with no listener distraction. The HGU onset is presented relative to the onset of the verbal element located at $t = 0$s, with positive values signifying gesture lag and negative values signifying gesture lead. We differentiate the data according to whether the bimodal expression in question involves a simple or a complex HGU. The figure suggests that simple head gestures are more tightly synchronised with their verbal affiliate (median HGU lead $= -0.18$s) than complex gestures (median HGU lead $= -0.26$s). The difference in the complex head gesture lead compared to the simple head gesture lead is statistically significant (Wilcoxon rank sum test, $p < 0.01$).

Morrel-Samuels and Krauss (1992) found that the degree by which gesture precedes speech, as well as the duration of the gesture, appeared to be a function of how familiar the lexical affiliate was to the speaker. SFEs in German generally involve a lexically and phonologically restricted set of forms (see Table 7) that are used repeatedly (but probably not randomly) in the course of a dialogue. These forms are therefore readily available in terms of lexical access. Head movements expressing feedback on the other hand, predominantly involve simple bi-phasic movements such as nods, jerks and tilts. Given the two facts, the short search times for feedback representations (especially in backchanneling) and the uncomplicated motoric programmes of most communicative head gestures, a close temporal coordination between head gestures

and their verbal affiliates can be expected. The shorter lead for simple gestures found in the present study confirms this mechanism, while the fronting of complex overlapping gestures suggests that coordination of bimodal expressions might be planned.

We further look at the dependence of listener's HGU duration on the number of movement cycles it contains, and the movement complexity it exhibits. We add a factor of modality to test for the possible effect of the overlapping verbal expression on head gesture unit duration. In order to account for all these factors, a Linear Mixed Model was formulated with the abovementioned independent variables, as well as log(HGU duration) as the dependent variable. The exact head movement label and dialogue ID were entered as random factors. We report the model coefficients and p-values here that resulted from the best model fit, the quality was determined by log-likelihood.

The model coefficients show that the duration of head gesture feedback significantly depends on movement cyclicity. The patterns are predictable: multiple gestures (e.g., nod-10) are longer than complex gestures (e.g., nod-2+jerk-1) ($b = 0.48$, $p < 0.001$), while single gestures (e.g., nod-1) are shorter than complex gestures ($b = -0.627$, $p < 0.001$). The interaction between modality and cyclicity was not significant, meaning that there is no interdependent impact of the two factors on head gesture feedback duration. The listener however, is significantly influenced by feedback modality in that visual gestures are produced in less time than bimodal gestures ($b = -0.11$, $p < 0.001$).

Finally, Inden et al (2013) reported on timing of multimodal feedback in ALICO based on a study implemented in an artificial agent. 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 well established, the former was not strongly attested in the literature: the specific nature of the conversational situation in ALICO, specifically concentrated on active listening, provided a sufficiently constrained setting, and revealed the function of the visual modality in this discourse context.

6.4 Feedback in distracted listeners

One of the fundamental design decisions behind ALICO involved introducing an ancillary task with a view to studying variation in feedback produced by distracted listeners. Thus, studies of ALICO complement earlier results which demonstrated the effect a distracted listener has on the speaker (Bavelas et al 2000; Kuhlen and Brennan 2010) but omitted changes in the listener's behaviour itself.

Analyses so far show that distracted listeners produce less feedback overall. In particular, a decreased rate of feedback communicating understanding (i.e., P2) was consistently associated with distractedness (Buschmeier et al 2011; Włodarczak et al 2012), which can be interpreted either in terms of listeners' inability to retrieve semantic content of the message in conditions of attention deficits or an intentional strategy aimed at avoiding confusion caused by explicitly feigning understanding. Notably, distracted listeners continue to produce *reactive feedback*, possibly based

on shallow processing of the utterance surface features, as well as communicate acceptance of the interlocutor's message, thereby conveying *implied* understanding.

Subsequently, changes in prosodic realisation of SFEs in distracted and non-distracted listeners were investigated in Malisz et al (2012). Significant differences were found in the intensity and pitch domains: feedback expressions produced by distracted listeners were on average quieter and showed less variability in F_0. At the same time, SFEs in the distracted condition varied more in intensity. In addition, some of the features were sensitive to segmental realisation of specific feedback expression.

Regarding the interaction between modalities and feedback functions in the corpus, Włodarczak et al (2012) found that in HGUs overlapping with verbal feedback expressions 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.

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). Information in Table 9 also suggests that complex head gesturing is avoided when providing feedback concurrent with an SFE, unlike in the attentive listener.

7 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. It includes extensive and detailed annotations of dialogue partners' verbal and head movement behaviour: from segmental transcription and local timing phenomena to entire talkspurts and gestural units. In addition, corpus design based on a storytelling scenario makes ALICO a particularly rich resource for both fundamental and applied research on communicative feedback across modalities. An inventory of feedback functions, based on existing standards, was used to assign pragmatic functions to short feedback expressions in a multi-annotator setup. Subsequent analyses revealed a number of regularities between feedback functions and their realisations, pertaining, among others, to their lexical form, prosodic and temporal properties, as well as the structure of accompanying head gestures. Finally, the inclusion of a distraction task allowed to identify decreased attention effects on surface feedback features.

The results outlined in the previous sections highlight ALICO's relevance for both fundamental and applied research. On the one hand, they provide an insight into the inner workings of such fundamental mechanisms in interpersonal communication as grounding, communicating agreement and expressing involvement. On the other hand, they pave the way towards more human-like and sociable dialogue systems. Above all,

however, they underscore the importance of studying communicative feedback and, more generally, listener-specific behaviours, in spontaneous, naturalistic conversation.

Work on additional tiers containing lexical, morphological information, turn segmentations and further prosodic labels is ongoing and the annotations are being continuously updated. A corpus extension is planned with recordings using motion capture and gaze tracking available in the MIntLab (Kousidis et al 2012).

Due to legal restrictions on dissemination of multimodal data, we are unable to make the audio and video recordings accessible to third parties. However, secondary ALICO data, such as transcriptions, annotations or extracted prosodic profiles of feedback expressions are available on request. They provide information on temporal and prosodic organisation and feedback signal frequency in both verbal and non-verbal domains, as well as on interdependencies between modalities. Since this information relies exclusively on surface forms, annotation reliability poses no significant problem even without access to the signal itself. While reliability is more of an issue for the functional classification, the annotations are available as well.

Acknowledgements This research was 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), as well as the Swedish Research Council (VR) projects "Samtalets rytm" (2009–1766) and "Andning i samtal" (2014–1072).

References

Allwood J, Nivre J, Ahlsén E (1992) On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9:1–26, DOI 10.1093/jos/9.1.1

Allwood J, Cerrato L, Jokinen K, Navarretta C, Paggio P (2007) The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation 41:273–287, DOI 10.1007/s10579-007-9061-5

Barbosa PA (2006) Incursões em torno do ritmo da fala *[Incursions into speech rhythm]*. Pontes, Campinas, Brasil

Bavelas JB, Coates L, Johnson T (2000) Listeners as co-narrators. Journal of Personality and Social Psychology 79:941–952, DOI 10.1037/0022-3514.79.6.941

Beňuš Š, Gravano A, Hirschberg J (2011) Pragmatic aspects of temporal accommodation in turn-taking. Journal of Pragmatics 43:3001–3027, DOI 10.1016/j.pragma.2011.05.011

Bergmann K, Kopp S (2006) Verbal or visual? how information is distributed across speech and gesture in spatial dialogue. In: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, Potsdam, Germany, pp 90–97

Boersma P, Weenink D (2013) Praat: Doing phonetics by computer [computer program]. Version 5.3.68, `http://www.praat.org/`

Breen M, Dilley LC, Kraemer J, Edward G (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, DOI 10.1515/cllt-2012-0011

Bunt H (2007) Multifunctionality and multidimensional dialogue act annotation. In: Ahlsén E, Henrichsen PJ, Hirsch R, Nivre J, Abelin Å, Strömqvist S, Nicholson S (eds) Communication – Action – Meaning. A Festschrift to Jens Allwood, Gothenburg University Press, Gothenburg, pp 237–259

Buschmeier H, Włodarczak M (2013) TextGridTools: A TextGrid processing and analysis toolkit for Python. In: Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung, Bielefeld, Germany, pp 152–157

Buschmeier H, Kopp S (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, Paris, France, pp 12–20

Buschmeier H, Malisz Z, Włodarczak M, Kopp S, Wagner P (2011) 'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness. In: Proceedings of Interspeech 2011, Florence, Italy, pp 2057–2060

Buschmeier H, Malisz Z, Skubisz J, Włodarczak M, Wachsmuth I, Kopp S, Wagner P (2014) ALICO: A multimodal corpus for the study of active listening. In: Proceedings of the 9th Conference on Language Resources and Evaluation, Reykjavík, Iceland, pp 3638–3643

Cerrato L (2007) Investigating communicative feedback phenomena across languages and modalities. PhD thesis, KTH Stockholm, Department of Speech, Music and Hearing, Stockholm, Sweden

Clark HH (1996) Using Language. Cambridge University Press, Cambridge, UK, DOI 10.1017/CBO9780511620539

Clark HH, Schaefer EF (1989) Contributing to discourse. Cognitive Science 13:259–294, DOI 10.1207/s15516709cog1302_7

Dittmann AT, Llewellyn LG (1968) Relationship between vocalizations and head nods as listener responses. Journal of Personality and Social Psychology 9:79–84, DOI 10.1037/h0025722

Duncan S, Fiske DW (1977) Face-to-face interaction: Research, methods, and theory. Erlbaum, Hillsdale, NJ

Edlund J, Heldner M, Al Moubayed S, Gravano A, Hirschberg J (2010) Very short utterances in conversation. In: Proceedings Fonetik 2010, Lund, Sweden, pp 11–16

Gardner R (2001) When Listeners Talk. Response Tokens and Listener Stance. John Benjamins Publishing Company, Amsterdam, The Netherlands, DOI 10.1075/pbns.92

Garrod S, Anderson A (1987) Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. Cognition 27:181–218, DOI 10.1016/0010-0277(87)90018-7

Geertzen J, Petukhova V, Bunt H (2008) Evaluating dialogue act tagging with naive and expert annotators. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, pp 1076–1082

Goldin-Meadow S, Alibali M, Church S (1993) Transitions in concept acquisition: Using the hand to read the mind. Psychological Review 100:279–297, DOI 10.1037/0033-295X.100.2.279

Goodwin C (1981) Conversational organization: Interaction between speakers and hearers. Academic Press, New York, NY, USA

Gravano A, Beňuš Š, Hirschberg J, Mitchell S, Vovsha I (2007) Classification of discourse functions of affirmative words in spoken dialogue. In: Proceedings of Interspeech 2007, Antwerp, Belgium, pp 1613–1616

Grosz BJ, Sidner CL (1986) Attention, intentions, and the structure of discourse. Computational Linguistics 12:175–204

Hadar U, Steiner T, Rose CF (1985) Head movement during listening turns in conversation. Journal of Nonverbal Behavior 9:214–228, DOI 10.1007/BF00986881

Hartmann B, Mancini M, Pelachaud C (2006) Implementing expressive gesture synthesis for embodied conversational agents. In: Proceedings of the 6th International Gesture Workshop, Berder Island, France, pp 188–199, DOI 10.1007/11678816_22

Heldner M, Hjalmarsson A, Edlund J (2013) Backchannel relevance spaces. In: Nordic Prosody XI, Tartu, Estonia, Peter Lang Publishing Group, pp 137–146

Heylen D (2006) Head gestures, gaze and the principle of conversational structure. International Journal of Humanoid Robotics 3:241–267, DOI 10.1142/S0219843606000746

Heylen D, Bevacqua E, Pelachaud C, Poggi I, Gratch J, Schröder M (2011) Generating listening behaviour. In: Petta P, Pelachaud C, Cowie R (eds) Emotion-Oriented Systems: The Humaine Handbook, Springer-Verlag, Berlin, Germany, DOI 10.1007/978-3-642-15184-2_17

Inden B, Malisz Z, Wagner P, Wachsmuth I (2013) Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In: Proceedings of the 15th International Conference on Multimodal Interaction, Sydney, Australia, pp 181–188, DOI 10.1145/2522848.2522890

Ishi CT, Ishiguro H, Hagita N (2014) Analysis of relationship between head motion events and speech in dialogue conversation. Speech Communication 57:233–243, DOI 10.1016/j.specom.2013.06.008

Kane J, Gobl C (2011) Identifying regions of non-modal phonation using features of the wavelet transform. In: Proceedings of INTERSPEECH 2011, Florence, Italy, pp 177–180

Kendon A (1967) Some functions of gaze-direction in social interaction. Acta Psychologica 26:22–63, DOI 10.1016/0001-6918(67)90005-4

Kendon A (1980) Gesture and speech: Two aspects of the process of utterance. In: Key MR (ed) Nonverbal communication and language, The Hague: Mouton, pp 207–227

Kisler T, Schiel F, Sloetjes H (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, Hamburg, Germany, pp 30–34

Koiso H, Horiuchi Y, Tutiya S, Ichikawa A, Den Y (1998) An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Language and Speech 41:295–321, DOI 10.1177/002383099804100404

de Kok I, Heylen D (2011) The MultiLis corpus – Dealing with individual differences in nonverbal listening behavior. In: Proceedings of the 3rd COST 2102 International Training School, Caserta, Italy, pp 362–375, DOI 10.1007/978-3-642-18184-9_32

Kopp S, Allwood J, Grammar K, Ahlsén E, Stocksmeier T (2008) Modeling embodied feedback with virtual humans. In: Wachsmuth I, Knoblich G (eds) Modeling Communication with Robots and Virtual Humans, Springer-Verlag, Berlin, Germany, pp 18–37, DOI 10.1007/978-3-540-79037-2_2

Kousidis S, Pfeiffer T, Malisz Z, Wagner P, Schlangen D (2012) Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In: Proceedings of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue, Stevenson, WA, USA, pp 39–42

Kousidis S, Malisz Z, Wagner P, Schlangen D (2013) Exploring annotation of head gesture forms in spontaneous human interaction. In: Proceedings of the Tilburg Gesture Meeting (TiGeR 2013), Tilburg, The Netherlands

Kuhlen AK, Brennan SE (2010) Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling. Discourse Processes 47:567–587, DOI 10.1080/01638530903441339

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174, DOI 10.2307/2529310

Malisz Z, Włodarczak M, Buschmeier H, Kopp S, Wagner P (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, Edinburgh, UK, pp 36–39

McClave EZ (2000) Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32:855–878, DOI 10.1016/s0378-2166(99)00079-x

Morrel-Samuels P, Krauss RM (1992) Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Human Learning and Memory 18:615–622, DOI 10.1037/0278-7393.18.3.615

Nobe S (2000) Where do most spontaneous representational gestures actually occur with respect to speech? In: McNeill D (ed) Language and Gesture, Cambridge University Press, Cambridge, UK, pp 186–198, DOI 10.1017/CBO9780511620850.012

Oertel C, Cummins F, Edlund J, Wagner P, Campbell N (2013) D64: A corpus of richly recorded conversational interaction. Journal on Multimodal User Interfaces 7:19–28, DOI 10.1007/s12193-012-0108-6

Peters C, Pelachaud C, Bevacqua E, Mancini M, Poggi I (2005) A model of attention and interest using gaze behavior. In: Proceedings of the 5th International Working Conference on Intelligent Virtual Agents, Kos, Greece, pp 229–240, DOI 10.1007/11550617_20

Poggi I, D'Errico F, Vincze L (2010) Types of nods. the polysemy of a social signal. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valletta, Malta

Prévot L, Gorish J, Mukherjee S (2015) Annotation and classification of french feedback communicative functions. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29), pp 302–310

Reidsma D, Carletta J (2008) Reliability measurement without limits. Computational Linguistics 34:319–326, DOI 10.1162/coli.2008.34.3.319

de Ruiter JP (2000) The production of gesture and speech. In: McNeill D (ed) Language and Gesture, Cambridge University Press, Cambridge, UK, pp 284–311, DOI 10.1017/CBO9780511620850.018

Schegloff EA (1982) Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In: Tannen D (ed) Analyzing Discourse: Text and Talk, Georgetown University Press, Washington, DC, USA, pp 71–93

Schegloff EA (1984) On some gestures' relation to talk. In: Atkinson J, Heritage J (eds) Structures of Social Action, Cambridge University Press, Cambridge, pp 266–296, DOI 10.1017/CBO9780511665868.018

Sidner CL, Kidd CD, Lee C, Lesh N (2004) Where to look: A study of human-robot engagement. In: Proceedings of the 9th International Conference on Intelligent User Interfaces, Funchal, Madeira, Portugal, pp 78–84, DOI 10.1145/964442.964458

Skubisz J (2014) Multimodale Feedbackäußerungen im Deutschen. Eine korpusbasierte Analyse zu nonverbalen Feedbackfunktionen am am Beispiel einer Beurteilungsstudie. Master's thesis, Fakultät für Linguistik und Literaturwissenschaft, Bielefeld University, Bielefeld, Germany

Truong KP, Poppe R, de Kok I, Dirk H (2011) A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In: Proceedings of Interspeech 2011, Florence, Italy, pp 2973–2976

Wagner P, Malisz Z, Inden B, Wachsmuth I (2013) Interaction phonology – A temporal co-ordination component enabling representational alignment within a model of communication. In: Wachsmuth I, de Ruiter J, Jaecks P, Kopp S (eds) Alignment in Communication. Towards a new theory of communication, John Benjamins Publishing Company, Amsterdam, The Netherlands, pp 109–132, DOI 10.1075/ais.6.06wag

Wagner P, Malisz Z, Kopp S (2014) Gesture and speech in interaction: An overview. Speech Communication 57:209–232, DOI 10.1016/j.specom.2013.09.008

Ward N, Tsukahara W (2000) Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 38:1177–1207, DOI 10.1016/S0378-2166(99)00109-5

Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H (2006) ELAN: A professional framework for multimodality research. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, pp 1556–1559

Włodarczak M, Bunt H, Petukhova V (2010) Entailed feedback: evidence from a ranking experiment. In: Łupkowski P, Purver M (eds) Aspects of Semantic and Pragmatics of Dialogue, Poznań, Poland, pp 159–162

Włodarczak M, Buschmeier H, Malisz Z, Kopp S, Wagner P (2012) Listener head gestures and verbal feedback expressions in a distraction task. In: Proceedings of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue, Stevenson, WA, USA, pp 93–96

Włodarczak M, Heldner M, Edlund J (2015) Communicative needs and respiratory constraints. In: Proceedings of Interspeech 2015, Dresden, Germany

Yngve VH (1970) On getting a word in edgewise. In: Campbell MA, et al (eds) Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, Chicago Linguistic Society, Chicago, IL, USA, pp 567–577

Yoganandan N, Pintar FA, Zhang J, Baisden JL (2009) Physical properties of the human head: Mass, center of gravity and moment of inertia. Journal of Biomechanics 42:1177–1192, DOI 10.1016/j.jbiomech.2009.03.029

Appendix

See Fig. 11 and Tables 13, 14.

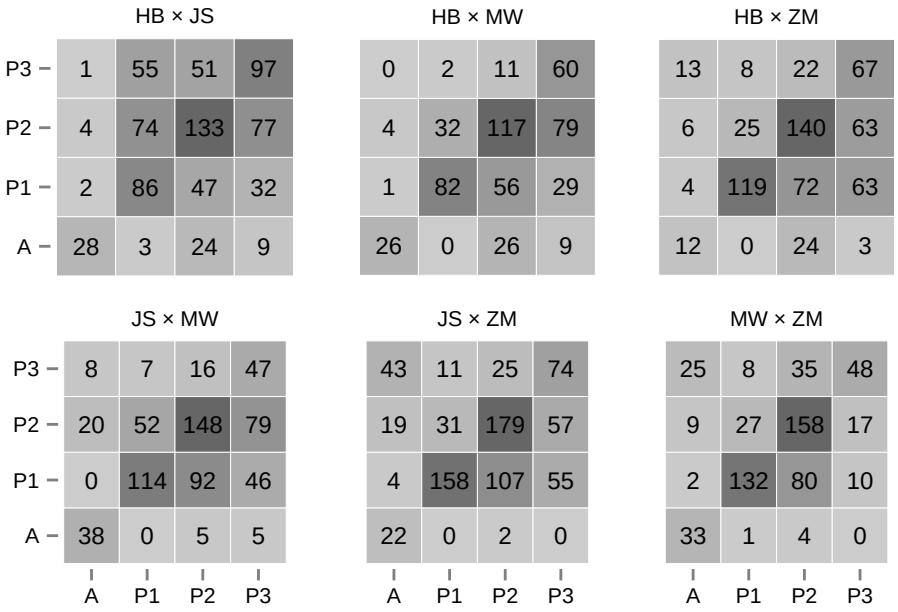

Figure 11 Confusion matrices for each annotator pair annotating core feedback functions categories: P1, P2, P3, and A. Labels were stripped off all modifiers (e.g., C or E or A in modifier role). The shades of the cells indicate relative frequency for each label combination and can be compared across confusion matrices. The numbers in each cell show absolute frequencies and are not comparable across confusion matrices.

Table 13 ALICO data overview. (L) denotes the listener and (S) the storyteller roles. (A) denotes the attentive listener condition, (D) the distracted listener condition.

	Condition		File format	Quality	File size
	A	D			
audio	25	25	AIFF	16 bit, 48 kHz	4.73 GB
video (no audio)	25	25	MPEG Video 2	320p, 25 fps	18.85 GB
video	25	25	MPEG-4, AAC	576p, 25 fps	2.1 GB
total duration	2:30h	2:42h			
completed speech annot.	20	20	Praat TextGrid	—	1.8 MB
completed L HGU annot.	20	20	ELAN eaf	—	3.2 MB
completed S HGU annot.	9	–	ELAN eaf	—	1.8 MB
prosodic features∗	20	20	CSV table	—	5.9 MB

∗duration (msec); mean, sd, slope [pitch (Hz), intensity (dB), peak slope, NAQ (Kane and Gobl 2011)] for a) the whole SFE and b) SFE initial, middle and final chunks of equal duration (Malisz et al 2012).

Table 14 Frequency of specific short feedback expressions (SFEs) found in ALICO as classified into three semantic categories (see Table 7). Words in parentheses were attached in some of the cases. English glosses after a usage based dictionary aggregator 'dict.cc'

Semantic category	German SFE	English gloss	Frequency
excitement/enthusiasm	'(ist) cool'	*cool*	25
	'(na / och) schön'	*nice, pretty*	24
	'wow'	*wow*	6
	'(ja) gut'	*good*	6
	'krass'	*sick, wicked*	4
	'super'	*super*	3
	'geil'	*awesome*	2
	'(och wie) süß'	*sweet, cute*	2
	'toll'	*terrific*	2
	'klasse'	*neat, marvellous*	1
	'sauber'	*neat*	1
	'spannend'	*exciting*	1
certainty	'(ja / na) klar'	*sure*	22
	'(ja) genau'	*exactly*	10
	'(das / ja) stimmt'	*true*	8
	'natürlich'	*of course*	2
	'richtig'	*that's right*	2
	'bestimmt'	*surely*	1
	'(ja) eben'	*exactly*	1
	'sicher'	*sure*	1
	'sicherlich'	*surely*	1
surprise/incredulity	'boah'	*wow*	1
	'echt'	*really, seriously*	1
	'heftig'	*fierce*	1
	'komisch'	*strange*	1
	'schrecklich'	*terrible*	1
	'wirklich'	*really, seriously*	1