University of Cologne

Faculty of Arts and Humanities

Bachelor's thesis

Multimodal feedback signals in dyadic conversations

submitted for the degree Bachelor of Arts

by

Matteo Schmelzer

matteo.schmelzer@uni-koeln.de

Two-subject B.A. Linguistics & Phonetics, and English Studies

Supervisor: Prof. Dr. Martine Grice

Cologne, September 30, 2025

Contents

1 Introduction **1**

2 Background **1**

2.1 Conversation Analysis . . . 1
2.1.1 Turn-taking . . . 1
2.1.2 Floor management . . . 2
2.1.3 Communication in the back channel . . . 2
2.2 Conversational feedback . . . 3
2.2.1 Functions of feedback . . . 3
2.2.2 Relation of embodied gesture to speech . . . 4
2.2.3 Head gesture feedback . . . 4
2.2.4 Clarification of terminology . . . 5
2.3 Head nods . . . 6
2.3.1 Cultural and language-based differences . . . 6
2.3.2 Functions of nods and other head movements . . . 6
2.3.3 Nods in response to polar questions . . . 7
2.3.4 Difference between nods produced by listeners and speakers . . . 8
2.3.5 Cyclicity of nods . . . 8
2.3.6 Head nods compared to head jerks . . . 9
2.3.7 Working definition head nod gesture unit and components . . . 10
2.4 Manual- and automatic classification of nods . . . 10
2.4.1 Manual nod classification . . . 10
2.4.2 Automatic nod classification . . . 12
2.5 Backchannels . . . 13
2.5.1 Communicative functions of BCs . . . 13
2.5.2 BCs compared to filled pauses and continuers . . . 14
2.5.3 BC pitch . . . 15
2.5.4 Cultural variation . . . 15
2.5.5 Working definition BC . . . 15
2.6 Overlapping spoken nods and BCs . . . 16
2.6.1 Function of overlapping nods and BCs . . . 16
2.6.2 Temporal placement of nods and BCs . . . 17
2.6.3 Cueing of nods and BCs . . . 17

3 Data **18**

3.1 Details on data origin . . . 18
3.2 Annotations . . . 19
3.2.1 Nod annotations . . . 19
3.2.2 Interlocutor intonation phrase boundary tones . . . 20
3.2.3 BC annotations . . . 20

4 Methods **21**

4.1 Software and tools . . . 21
4.2 Time alignment of multimodal data . . . 21

5 Analysis **21**

5.1 Description of nods and BCs in the data . . . 22
5.1.1 Nods . . . 22
5.1.2 BCs . . . 24

5.2 Annotated nods overlapping with annotated BCs 27
5.2.1 Categories of overlap 27
5.2.2 Overlap between PR BCs and nods 28
5.2.3 Overlapping nods cross-correlated with features of BCs 32
5.2.4 Overlap between IS BCs and nods 33
5.2.5 Nods and PR BCs not overlapping 34
5.3 Window concept and effect on overlaps 35
5.4 Interlocutor's IP end pitch direction (DIMA boundary tones) 37
5.5 Summary of findings 39

6 Considerations for generalisability **40**
6.1 Validity of Lab Speech 40
6.2 Ecological validity 41
6.3 WEIRD samples 42
6.4 Limitations of the dataset 43

7 Future directions **43**

References **45**

List of Figures

1 Nod rate/min by speaker . . . 22
2 Nod duration by speaker . . . 23
3 Duration of PR BCs by speaker . . . 25
4 PR BC intonation in ST (positive values indicate rising pitch) and by category . . . 26
5 BC types distingiushed by (non-)lexical category . . . 27
6 Types of overlap . . . 28
7 Comparison of BCs, nods, and overlaps (BC and nod) by speaker . . . 29
8 Distribution of overlap size by speaker (in % to facilitate direct comparison of annotations of different length) . . . 30
9 Overlap category by speaker . . . 30
10 Overlap category by speaker correlated with BC type . . . 31
11 Duration of overlapping nod and BC by overlap category . . . 32
12 Duration of overlapping nod and BC by BC type (collapsed categories) 33
13 Comparison of IS BCs, nods, and overlaps (BC and nod) by speaker . 34
14 Schematic representation how the window added to the interval of the BC establishes association (through overlap) between annotations 35
15 Effect of added window size on amount of overlapping annotations (in %), window size on x-axis denotes individual interval added to start and end of BC . . . 36
16 Overlap category by speaker, with a 2 s window applied . . . 37
17 Distance between speaker boundary tone and listener BC, distinguishing between (non-)overlapping annotations . . . 38
18 Distance between speaker boundary tone and listener nod, distinguishing between overlapping (circles) and non-overlapping (triangles) annotations . . . 38

List of Tables

1 Dyad balance based on visual feedback (larger in-dyad count in bold) 24
2 Dyad balance based on vocal feedback (larger in-dyad count in bold) 25
3 Mean offset between start time of overlapping nod and BC . . . 31
4 BC type by speaker . . . 33
5 Count and mean duration of nods that do not overlap with any BC, by speaker . . . 34

1 Introduction

Language is more than words, especially in face-to-face interaction. Humans use a variety of multimodal signals to exchange information with their interlocutors. Information exchanged may contain statements, inquiries, or feedback, as part of a co-operative process resulting in conversation.

For a long time, seemingly inconspicuous gestures such as head nods have been broadly neglected in favour of word-based representations of language. Particularly the interplay between modalities has been underserved. But linguistic research using naturalistic, interactive, and multimodal data is becoming increasingly important, as part of an "interactive turn" (Kendrick, 2017, p. 7) that has long been needed (Cowley, 1998, p. 542). Therefore, this thesis investigates the relation of feedback signals across modalities in spontaneous German conversation. The analysis focusses specifically on head nods and vocal backchannels and how they relate in the production of feedback.

2 Background

The next section starts with a summary of the management of conversation (2.1) and the notion of conversational feedback (2.2). Then feedback in the form of head nods (2.3) and their classification (2.4), backchannels (2.5), as well as their combined use (2.6) are described.

2.1 Conversation Analysis

Face to face conversation between humans is very efficient. It unfolds in turns of on average 2 seconds, and gaps between speaker turns of around 200 milliseconds (Levinson, 2016). This offset is shorter than the time required to construct a response, which indicates that planning of a response needs to happen while the interlocutor's turn is unfolding. Turn-taking is universal across languages, at least in those investigated by Stivers et al. (2009).

2.1.1 Turn-taking

The turn-taking framework as described by Sacks et al. (1974) is a foundational theory to the emerging field of conversation analysis in the 1970s. According to the framework, conversation unfolds in sequentially alternating turns produced by interlocutors. It attempts to mechanically describe the norms according to which language is produced interactively between interlocutors via turn allocation managed between interlocutors at the end of turns.

The specifics of the turn-taking framework as put forward in Sacks et al. (1974) is contentious. Cowley (1998) for example decisively critiques it as a reductive, structuralist metaphor that ought to be abandoned when trying to understand conversation. Cowley (1998) criticises the fact that conversation is transcribed in turns, which are more akin to sentences than utterances since they do not maintain the timing of constituent words. Further, the exclusive interpretation of conversation as content encoded through word-based language, rather than taking timing, prosody or embodied gesture into account is viewed as reductive. The turn-taking device hypothesized by Sacks et al. remains elusive, despite the "grossly apparent facts" touted (1974, p. 700). While the framework as presented by Sacks et al. is not suitable to understand the intricacies of conversation, as a metaphor it has its uses. For example to conceptualise the roles of speaker and listener and their contributions to conversation.

2.1.2 Floor management

Edelsky (1981) proposes a distinction between the concepts of floor and turn. The floor consists of an interactionally established topic of conversation, negotiated between interlocutors and is demonstrative of the current function of the conversation. It is possible to take a turn without contributing to the floor as well as that no floor currently exists as part of a break in conversation. A turn in comparison consists of a vocal- or gestural utterance intended to convey a relevant message. Edelsky crucially distinguishes between content – which is produced in-turn, and feedback – which is produced out-of-turn.

Duncan (1972) (see also Duncan & Fiske (1985), ch.3) describes three signals that interlocutors can use to negotiate turn- and floor-management. First, turn-yielding signals are performed by the turn-holder, indicating the end of their turn and giving the listener the opportunity to take the floor. Second, attempt-suppressing signals are performed by the turn-holder to communicate intent to continue their turn. Third, listeners can use feedbacks signals – described as back-channel communication by Duncan – to decline taking a turn in support of their interlocutor's turn.

2.1.3 Communication in the back channel

Preceding much of the aforementioned work in conversation analysis, the term "back channel" (or contemporarily "backchannel") was coined in an influential paper by Yngve (1970). According to Yngve, while speakers and listeners or speaking and listening are clearly distinct activities, information is exchanged concurrently by both interlocutors. While a speaker produces a turn, a listener may use the backchannel to provide feedback. This is done while "speak[ing] out of turn" (1970, p. 958),

i.e. while the listener neither has the floor nor produces a turn (cf. similarly Duncan, 1972; Edelsky, 1981). Yngve illustrates the necessity of feedback signals by comparing it to a conversation over the telephone in which a speaker might ask for continued attention by the listener, when they do not provide feedback signals. The feedback in Yngve's conception is produced using both vocal and gestural modalities, and is often (but not always) produced concurrently with the speaker's turn.

Yngve was not the first to describe feedback, Dittmann & Llewellyn (1968) proposed the concept as listener response, and according to Duncan (1972), the term turn-taking first used by Yngve (later made widely known through articles such as Sacks et al. (1974)) had in parallel been coined by Goffman. It must be noted that while Yngve used the term "back channel" to describe a parallel stream of communicative signals to the primary one (a "front channel", produced by the turn-holding speaker), contemporarily the term is used to denote the content produced in it. This modern notion is described in detail in 2.3 and –Section 2.5.

2.2 Conversational feedback

Conversational feedback can be provided as vocal- or gestural backchannel responses. Other descriptions of feedback term it listener response (Dittmann & Llewellyn, 1968) or continuer (Schegloff, 1982).[1] Goffman terms it "response" and describes four criteria: the response is produced in response to a prior speaker; the response indicates the responder's stance on the object of the response; the response indicates what it responds to; and the response is elicited or understood to be recognised by the addressee (1976, p. 280). Backchannel feedback is considered as one of three mechanisms for creating common ground (Fusaroli et al., 2017).

2.2.1 Functions of feedback

Other-oriented feedback in both vocal- (backchannels, laughter) and gestural- (head nods, etc.) modalities has been shown to support the speaker. Canadian participants tasked with telling a narrative, told it better when a listener provided feedback (Bavelas et al., 2000). There, feedback was distinguished into two categories: generic – which was consistently provided by listeners and specific – which required more attention and buy-in to the story by the listener. As part of a collaborative model of conversation proposed by Bavelas et al., both interlocutors actively contribute to conversations. The listener plays an important role by providing the speaker with feedback, allowing them to adjust their communicative output. Feedback can be used as part of what Sacks et al. describe as recipient design (1974, p. 727) and is made possible by other-monitoring (cf. Clark & Krych, 2004).

[1] For a more detailed dissection of the use of feedback, see De Stefani (2021).

The modality used to deliver feedback influences its function. In American English (AmE) storytelling, vocal backchannel feedback was found to display alignment with the storyteller, while gestural nods additionally indicated the listener's understanding of the speaker's viewpoint (Stivers, 2008). The use of feedback in the form of head nods increases the likeability and trust in (virtual) interlocutors. Aburumman et al. (2022) tested the effect of fast nods and mimicry implemented in virtual humans and found that study participants rated them more likeable and trustworthy when they provided feedback in the form of nods.

2.2.2 Relation of embodied gesture to speech

McNeill (1992) categorises gestures on a continuum[2] according to their dependence on a vocal component to convey meaning. On the one end are co-speech gestures[3], i.e. gestures fully dependent on speech produced alongside the gesture for interpretation. On the other end McNeill poses sign language, which functions fully independently of vocal signals. These gestures have since been described as pro-speech gestures, as they are fully independent of spoken language (cf. Schlenker, 2019).

Head gestures and more specifically head nods can fit in multiple places on this continuum. Yes-confirmatory nods (and their diametrically opposite head shakes) could be considered close to what Kendon and McNeill describe as emblems due to their conventionally established form-function mapping. Emblems are largely independent of speech and closest to sign language in McNeill's classification and can be used as pro-speech gestures. More general feedback nods (those that signal e.g. attention or agreement) however, are harder to place on the continuum. This is because their intended function is context- and to some degree form dependent. Due to their context dependency, these gestures are closer to gesticulation on the language-dependent end of McNeill's continuum. These nods may be described as co-speech gestures, if they occur with speech that enhances their communicative function. This may appear contentious due to the iconic gesture form, but additional context in the form of speech can, in some contexts, give additional insight into the intended function of the nod. See Knight (2009, p. 79f) for additional discussion on the matter.

2.2.3 Head gesture feedback

Interlocutors use a variety of embodied gestures to non-verbally produce feedback signals. Embodied feedback signals may occur alone or in conjunction with vocal backchannels. Other than head gestures, specifically head nods – the central focus

[2]This continuum was, according to McNeill, first proposed by Kendon (1988).

[3]"Gesticulation" in McNeill's terms (1992, p. 37). Co-speech gestures have also been described as speech focussed movements (SFM) (Butterworth & Beattie, 1978).

of this thesis – a variety of other articulators associated with the head are used to produce feedback.

Eyebrow raises and furrows have been found to signal problems in understanding a speaker, other-initiating a repair sequence (Hömke et al., 2025). The intentionality of these gestures remains debated despite a form-function mapping. This is because they also occur outside of communication as a function of input processing. Moments of mutual gaze initiated by the speaker can be used to elicit feedback from the listener (Bavelas et al., 2002; Duncan & Fiske, 1985, p. 46). By using gaze, the turn holder can collaboratively elicit feedback from the listener without effecting a turn-exchange. The co-occurrence of smiles with head nods has been observed and interpreted as a feedback signal, rather than an expression of pleasure (Dittmann & Llewellyn, 1968).

While head nods are the most frequent type of head gesture to signal feedback (Cerrato, 2007, p. 15), their opposite – the head shake – is in certain western European cultures (e.g. German or English) a conventionalised gesture to (most frequently) signal negation (Kendon, 2002).[4] A head shake consists of horizontal rotation of the head moving into either direction and returning to the starting position at least once. Variations of amplitude and repetitions are used to perform different functions. Head shakes are used by both speakers and listeners, either on their own before and after speech or in conjunction with vocal components.

In a study by Bavelas et al. (2008), participants produced gestures in a conversation over the telephone (although to a lesser degree than in face-to-face interaction). Vocal feedback signals in the form of backchannels (and filled pauses for that matter) are typically produced subconsciously (Castello & Gesuato, 2019; Wehrle, 2023, pp. 96, 117). Note however that some – particularly neurodivergent individuals – have to consciously use feedback signals. This indicates that embodied feedback too is produced subconsciously, although in accordance with the communicative situation. Additionally, embodied feedback signals may serve further functions for the producer.

2.2.4 Clarification of terminology

In the following sections, two specific forms of feedback investigated in this thesis will be defined in terms of their form and function, preparing their analysis in 5. Gestural (sometimes described as visual) feedback in the form of head nods are described in 2.3. They will be referred to here as (head) nods, descriptive of their characteristic vertical up and down movement. Vocal feedback in the form of backchannels are described in 2.5. They will be referred to here as backchannels (BCs) as is customary in broader research, but explicitly only pertain to vocal feedback signals. This is

[4] There is however notable cultural variation, cf. 2.3.1

done for ease of terminology, since feedback head nods also function as backchannels. The combined occurrence of both nods and BCs is described in 2.6. Other modes of feedback such as manual gestures, gaze, or other head gestures are not considered from this point on.

2.3 Head nods

While there is a variety of head gestures produced by both listener and speaker, the focus here will be specifically on nods produced by the listener. Head nods are movements of the head up and down (or down and up) often repeated, conventionally established as a signal of affirmation, particularly in European languages.

2.3.1 Cultural and language-based differences

One big caveat for the generalised description of head nods and other embodied feedback is language- or cultural specificity. This is why, whenever research is referred to throughout this thesis, the language analysed is explicitly mentioned to avoid overgeneralisation.

There are, at times marked, differences in the use of head movements between different language communities. Regarding function of head gestures, Jakobson (1972) reports that in Bulgarian, Greek, and southern varieties of Italian an upward- and back movement is used as a marker of negation (see also Andonova & Taylor, 2012). This movement, especially when repeated for emphasis, may appear similar to a nod to speakers of German or Jakobson's native Russian. Besides function, frequency of feedback differs between languages. In Japanese for example, embodied feedback is notably more frequent in comparison to AmE (Kita & Ide, 2007; Maynard, 1987; Ward & Tsukahara, 2000). Ishi et al. (2014) found that in Japanese dyads the interpersonal relationship between interlocutors has an impact on the amount of nods produced. The more distant the interlocutors, the more nods – especially multiple nods – were produced.

2.3.2 Functions of nods and other head movements

Backchannel feedback is one of the primary functions of head nods that is assumed by nearly all approaches distinguishing different functions of head nods in spoken (cf. Wagner et al., 2014) and signed languages (cf. Bauer et al., 2024).

Functions of head nods (and head gestures generally) are sometimes grouped by interlocutor role in a conversation (e.g. Poggi et al. (2010)). Some researchers describe overall function categories, without strictly pairing interlocutor role with function. Otsuka & Tsumori (2020) for example distinguish between head gestures that are speech-related (i.a. emphasis, turn management) and those that function

as reactions (i.a. BC, assessment). For their analysis Otsuka & Tsumori do not assume that reactions are only produced by listeners and vice versa.

A variety of functions are distinguished, often with terms unique to the researcher and varying in number according to the scope of inquiry. Gurion et al. (2020) consider three primary functions of head nods in their investigation of BrE: backchannel feedback, mimicry, and listener response to speaker trouble. They find that the three major functions considered cannot account for a substantial amount of head nods, implying more complexity.

In the aforementioned study by Otsuka & Tsumori, the authors distinguished 32 functions of head gestures. Poggi et al. (2010) similarly distinguish a breadth of functions in their conceptionalisation of the head nod as a polysemic signal. What makes their categorisation unique is that it distinguishes between speakers, listeners, and third-party listeners. The functions of head nods distinguished for third party listeners constitute a subset of functions available to the directly addressed listener.

Hadar et al. (1985) distinguishes four functions of listener head movements in British English (BrE) based on context in other modalities. Thus head movements accompanied by (dis-)confirming verbalisations (or semantically interpretable as such) were labelled "yes" and "no". Movements that occurred within 100 ms of a speaking interlocutor's stressed syllable were labelled as synchronous, and movements that occurred within 2 s prior to the listener's own vocalisation were labelled as anticipatory.

In their study on head nods in German sign-language (DGS), Bauer et al. (2024) distinguish between affirming and feedback nods.

The function of head gesture feedback has also been investigated from a conversation analysis perspective. McClave (2000) found that speakers employ head nods to request feedback from listeners in AmE. De Stefani (2021) analyses head nods that occur on their own in the backchannel as responses to polar questions in French and Italian. This rare phenomenon is an interesting edge-case between feedback and turn-based responses, since it occurs in the middle of the speaker's turn without provoking a turn-exchange, while also completing an adjacency pair in the way that a full response would.

2.3.3 Nods in response to polar questions

Affirmation and negation are conventionally expressed through emblematic head nods and -shakes in a variety of European languages (Darwin, 1899; Jakobson, 1972; Kendon, 2002). Additionally, head gestures have been found to indicate listener engagement (Bavelas & Gerwing, 2011). Emblematic head nods produced by the listener could be described as (yes-)confirmatory nods when they occur shortly after the speaker posed a polar question. That is, these nods function as a replacement

or reinforcement of a vocal reply of the same affirming semantic content. Otsuka & Tsumori (2020) differentiate yes-confirmatory nods as a category of head movement, along with a negative pendant for negation. Ferré & Renaudier (2017) explicitly exclude nods occurring following questions because they might be replies to the question rather than feedback signals.

2.3.4 Difference between nods produced by listeners and speakers

Speakers, similar to listeners, also produce head nods during conversation. The function of head nods typically differs by role of the interlocutor, however. Speakers' nods are generally not used to produce feedback.

Maynard (1987) and Otsuka & Tsumori (2020) both distinguish a variety of functions of nods produced by speakers in Japanese. During speech, nods mark emphasis and agreement. Further, nods are used to elicit listener response, either in the form of a full turn or feedback. At the end of their turn, speakers may mark the end of their phrase and potential turn transition, while the incipient speaker may mark the start of their turn with a nod. Otsuka & Tsumori further describe that speakers use nods to select the next speaker. It ought to be noted that nods at turn transition points occur rarely in their data (2020, p. 217178, Table 2). The greater variety of speaker nods may be a function of an overall greater amount of feedback produced in Japanese conversation.

Gurion et al. (2020) take mimicry into account as a source of both speaker- and listener head nods in their work on BrE. According to mimicry theory, either interlocutor may copy their interlocutor's gestures (both vocal and gestural), which is thought to strengthen the bond between interlocutors, increase persuasiveness, likeability, and empathy.

Research in sign-language, which has previously strongly focussed on signs produced by the turn-producing interlocutor, finds that nods produced with varying cyclicity and speed fulfil differing functions (cf. Bauer et al., 2024). Signing addressees also use nods, primarily to provide feedback signals.

2.3.5 Cyclicity of nods

Head gesture form has been a focus of the earliest research on head nods. Papers by Birdwhistell (1970) and Hadar et al. (1983, 1985) analyse speed and amplitude of nods. While these are less prevalent in more recent research (likely because they are only feasible with mechanised methods of data collection), the cyclicity, i.e. whether or not, and how often nods repeat is of particular interest in many papers.

Nod gestures are thus distinguished into two major categories: single and cyclical (or multiple) nods (Cerrato, 2007; Knight, 2009; Malisz et al., 2016; Oomen et

al., 2023). Malisz et al. (2016) distinguish three types: single (one cycle), simple (more than two cycles of the same type) and complex (combination of multiple different movements) gesture units (Knight (2009) takes a similar approach). In terms of function Malisz et al. (2016) – in reviewing previous analyses – conclude that distinctions between single and cyclical nods vary from language to language. An exception to this generally applied distinction is Stivers (2008), who based on their review of (CA) literature considers nods – independent of speed, amplitude or cyclicity – to be the same gesture type.

Hadar et al. find that cyclical nods and shakes in BrE typically consist of three cycles produced with declining amplitude (1985, p. 224). Cyclical repetition of head gestures is reported to be used as an emphasizer (Jakobson, 1972). Based on limited data Cerrato (2007) finds that in Swedish listeners single nods are used as continuer signals, whereas cyclic nods are used to signal acceptance (i.e. understanding). In Cerrato's data, politeness i.e. reactions to courtesies offered by the speaker are exclusively associated with single, frequently slow nods.

2.3.6 Head nods compared to head jerks

Two distinct types of vertical head movement can be distinguished as head nods and head jerks. The distinguishing feature is the direction of movement, at least in German. For repeated movements the initial direction is decisive. Nods are described as downstrokes (i.e. a down movement followed by an upward movement), while jerks are described as upstrokes or inverted nods (Allwood et al., 2007; Cerrato, 2007; Kousidis et al., 2013; Wagner et al., 2014). According to Cerrato (2007; also cf. Allwood et al., 2007) a jerk consists of either a single or cyclical backward movement of the head produced at increased velocity (as implied by the term). Cerrato further takes gesture speed into account, classifying a slow up-down movement not as a jerk (2007, p. 47).

Włodarczak et al. (2012) found that jerks function as markers of understanding in German. In contrast to nods which are described as a default, jerks are more marked. Malisz et al. similarly found jerks to express understanding, and report that only 5 % of vertical head movements are jerks according to their data (2016, p. 429, Table 8). Paggio & Navarretta (2013) describe jerks as "up-nods" and collapse both categories because of the low frequency and difficulty in manually distinguishing them. Otsuka & Tsumori (2020), who adopt the definition by Włodarczak et al. (2012), similarly merge nods and jerks into one category.

2.3.7 Working definition head nod gesture unit and components

For the purpose of this thesis, head nods consist of vertical movements of the head that change direction at least once (i.e., down-up or up-down). While both turn-holder and listener produce vertical head movements, only head nods produced by interlocutors that are not currently holding the floor (i.e. listeners) are considered. Despite concerns about the term "nod" expressed by Maynard (1987, p. 592), arguing that it implies conventional pragmatic function of nods to signal agreement, the term is used here to describe all gestures following the given form dimensions.

Assumptions about where a nod gesture starts and ends have an impact on overlap and comparisons with other interval-based annotations, such as those of spoken backchannels analysed in this paper. The start of a nod annotation is marked at the beginning of what is considered in gesture research the stroke phase of a gesture unit (cf. Kendon, 1980, p. 212; Rohrer et al., 2023, p. 25). The stroke phase was chosen as the start, since it is the first part of a nod gesture that identifies it as such. The end of a nod is marked when the vertical movement (that is perceived as a nod) ends and the head returns to a rest position, or another (non-nod) head gesture is initiated. In this conception of the end of a nod gesture, cyclical nods typically recede in amplitude before the end of the gesture.

Włodarczak et al. use a similar definition when proposing a "head gesture unit" (HGU) as a "a perceptually coherent and continuous movement sequence" (2012, p. 94). Their approach is adopted here: single- and multiple (or cyclical) nods are annotated as one single gesture of nod, without distinction of the amount of cycles. This choice was supported by the impression that the majority of nods in the data are cyclical. While duration is recorded, it should not be used to infer cyclicity of the gesture, due to inter- and intra-speaker variation of nodding speed. A minimum gap of 300 ms between annotations as proposed for gesture units in the M3D scheme (Rohrer et al., 2023, p. 15) is adopted here. The gap adopted is arbitrary, yet conventional.[5] The difference between nods (downstrokes) and jerks (upstrokes) is not distinguished here, due to low frequencies.[6]

2.4 Manual- and automatic classification of nods

2.4.1 Manual nod classification

Which features of nods or more broadly head gestures to annotate is not a universally answered question. A review of different gesture annotation systems by Wagner et al. (2014) highlights a limited interest in the analysis of head- or nod gestures specifically. The only systems indicated to take head movements into account in

[5] In practice, few nod annotations were collapsed into single, long gesture annotations.
[6] As done in Paggio & Navarretta (2013), Otsuka & Tsumori (2020).

Wagner et al.'s list are *Multimodal Score* (Caldognetto et al., 2004) and *MUMIN* (Allwood et al., 2007).

Multimodal Score focusses on a variety of articulators but only annotates head nods, without further consideration what form this gesture takes. MUMIN focusses on the function of feedback signals used for turn-management and sequencing. By virtue of their focus on function, only communicative gestures are considered. In their thesis Cerrato (2007) (one of the co-authors of MUMIN) uses a similar but not identical scheme to categorise head gestures. Cerrato distinguishes between the following gestures based on form: single/repeated nod, single/repeated jerk, single slow backwards up, move forward, move backward, single/repeated tilt (sideways), side-turn, shake (repeated) (2007, p. 47).

Włodarczak et al. (2012) describe head gestures as superordinate head gesture units (HGU) to investigate function and timing of head gestures and verbal[7] feedback. A single HGU is considered to be a continuous movement delimited by pauses in movement. HGUs are distinguished by form: nod, jerk, tilt, turn, protrusion, and retraction and the number of cycles is annotated. Kousidis et al. (2013) present a superset of Włodarczak et al.'s scheme, additionally categorising turns (different from Włodarczak et al., movement in a single direction is described here, but shakes (termed turns by Włodarczak et al.) are also distinguished), bobbles, slides, shifts (repeated slides), and waggles (repeated irregular movement).

The M3D system (Rohrer et al., 2023) suggests the annotation of head gestures as complete gesture units without distinguishing gesture phases (preparation, stroke, hold (pre-/post stroke), and recovery). For head gestures Rohrer et al. only annotate start, end, and apex (point in time at which the articulator changes direction) in addition to the direction of the movement along its axis. They adopt head gesture types from Wagner et al. (2014, p. 212)[8], who distinguish nod, turn (i.e. shake), tilt, slide, and protrusion. For all gesture annotations Rohrer et al. use a minimum gap between gestures of 300 ms (2023, p. 15).

Esselink et al. (2024) distinguish between the following head movements (separately from head poses) in their annotation system[9] for non-manual markers (NMM) in sign-language research: " 'nod' (single nod), 'nodding' (multiple nods), 'shake' (single shake), 'shaking' (multiple shakes), 'sideways' (single sideways movement of the head), and 'neutral' " (2024, p. 71). In assessing inter-rater agreement, Esselink et al. (2024) find that annotators reliably distinguish between single and multiple nods. This overview of existing schemes for the description and classification of head nods showed a bias towards form- rather than function-based categorisations

[7] sic! Because *verbal* relates to propositional content, it is better to think of it as *vocal*, i.e. spoken

[8] See Wagner et al. (2014, p. 212 Fig. 1) for a visualisation of axes used for specific head movements.

[9] See also Oomen et al. (2023) introducing the system.

2.4.2 Automatic nod classification

Researchers have subjected participants to a variety of contraptions for more than 50 years in an effort to mechanically recorded head gestures. Dittmann & Llewellyn (1968) for example used the signal produced by friction of a record player needle against a chair to record movement. The needle was hung from participant's heads using a headband and would move when the participant produced a head gesture. Hadar et al. (1983, 1985) recorded participants using a polarised-light goniometer, a setup built to capture kinematic information about head movements. During the experiment participants wore two head mounted photosensors (positioned at the top and temple of the head) at which a light source was pointed. Using data recorded by the photosensors, the experimenters were able to capture position of the head. Hadar et al. describe the mechanism as being able to produce "very accurate record of amplitude, timing, and cyclicity of movement, but somewhat crude information about its direction" (1985, p. 216). Cerrato (2007) and Ishi et al. (2014) used similar setups by recording head gestures using reflective markers placed on the participant's face and captured by a number of infrared-cameras. The system sampled the position of the markers at 60 Hz.

The majority of the studies using specialised mechanical instruments suffer from small sample sizes, influence of the equipment on the participants' natural behaviour. Contemporarily, there is a variety of open source computer-vision (CV) tools for the tracking of human body movements. In line with the existing interest in manual- and facial gestures, these focus on providing an overall framework (Lugaresi et al., 2019), the body (Cao et al., 2018), the hands (Pouw et al., 2025), the face (Baltrusaitis et al., 2018; Cheong et al., 2023), and the head (Yung, 2022).

All of these tools estimate position of defined parts of articulators from videos or static images. The position and movement of articulators is either estimated as a 2D pixel based difference over time or a 3D inferred spacial position of the articulator. An advantage of the CV approach to gesture tracking is that it is less invasive, only requiring a camera to record the participants' movements. Some tools are limited in the type of video that can be used for analysis, *EnvisionHGdetector* (Pouw et al., 2025) for example is designed to only track a single person from a frontal perspective.

Different tools produce different output, given their respective focus on different aspects of gesture. Most produce time series kinematic data that describes the approximated movement of individually tracked positions of parts of articulators (Cao et al., 2018; Cheong et al., 2023; Lugaresi et al., 2019; Pouw et al., 2025). Tools focussing on the face provide categorised facial gestures in the form of facial action units (AU) (Ekman & Friesen, 1977), (cf. Baltrusaitis et al., 2018; Cheong et al., 2023). *EnvisionHGdetector* (Pouw et al., 2025) provides estimated types of manual gestures. Yung's *nodding pigeon* (2022) classifies head movement into nods

and rotations. All tools described (excluding *nodding pigeon*) provide some form of visualisation of the tracked movement.

As the above survey indicates, there are few tools suited to the analysis of head movements specifically. Labels of nods or rotations as provided by *nodding pigeon* (Yung, 2022) are too coarse for linguistic purposes. No other tool surveyed here offers deeper classification of head gestures than distinguishing head nods and shakes.

A tentative test of an adaptation of *Mediapipe* (Lugaresi et al., 2019; adapted by Pouw & Akamine, 2025) showed that the tracking is too unstable temporally to be of any use for the analysis of head movements. Agirrezabal et al. (2023) tested detection and classification of head movements with a variety of tools (including OpenPose (Cao et al., 2018)) on the Danish NOMCO Corpus. While the accuracy of their tool was state of the art the time of writing, Akamine et al. (Forthcoming) propose that while automated systems can be useful, they cannot be a total replacement for manual annotation and investigation of one's data. This is to say that even with an automatically annotated corpus, it is highly advisable to proofread these annotations, to correct inaccurate annotations and become familiar with the data on a low level.

2.5 Backchannels

Backchannels (BC) are short vocal utterances serving as feedback signals. They are produced as reactions to a speaker most typically signalling attention and agreement, facilitating the co-operative development of conversation. In the initial description of backchannels, Yngve included short comments, questions, and completions as possible backchannels (Yngve, 1970, p. 574f). In contrast to Yngve, Malisz et al. (2016) specifically describe BC as short-feedback expressions (SFE) and exclude them. The case has been made for multi-unit backchannels (MUB). These consist of combinations of multiple, otherwise independently used BCs, such as "mmh okay". A third of BCs in Mereu et al.'s (2024) Italian corpus consisted of MUBs marking agreement.

2.5.1 Communicative functions of BCs

In the literature, two primary types of BC are distinguished by their respective function. Passive Recipiency (PR) or turn-yielding BCs (sometimes continuer) are signals produced by the listener to indicate to the speaker that they should continue their turn (Jefferson, 1984). This requires that the listener yields their opportunity to start a turn, as indicated in the term turn-*yielding*. PR BCs are the most frequent type of BC used. By producing a PR BC, the listener assures their *passive* recipiency, i.e. is attentive to the floor-holder's turn and does not attempt to

initiate a turn. Gardner (2001) specifies PR BCs further by distinguishing three functions: basic attention signals (continuers), signals that indicate understanding or agreement (acknowledgements), and signals highlighting discourse-new information (news markers).

Incipient Speakership (IS) BCs are produced by the listener shortly before they start a turn of their own. In AmE, incipient speakership BCs are typically produced using "yeah", whereas PR is signalled using "Mm mh" (Drummond & Hopper, 1993; Jefferson, 1984). Further, BCs have been distinguished based on the producer's investment in the conversation. While generic BCs only signal attention and participation, specific BCs additionally indicate the listener's stance on the speaker's turn (Bavelas et al., 2000, ; Tolins & Fox Tree, 2014).

Tongue clicks are an adjacent – perhaps paravocal – mode to produce backchannel signals. In English, tongue clicks can fulfil three functions: incipient speakership (just before a turn, often produced as a by-product of opening the mouth to start articulating a word), sequence management (e.g. word search), and the display of negative stance (typically deliberately loud and directed at the current speaker) (Ogden, 2013).

2.5.2 BCs compared to filled pauses and continuers

Filled pauses (FP) constitute functionally opposite to backchannels (Wehrle, 2023, p. 116; see O'Connell & Kowal, 2004 for FP). Typical examples in English are "uh" and "um". These particular examples have been shown to announce short- and long delays respectively, when produced by a floor-holder (Clark & Fox Tree, 2002). While they are primarily employed by the active speaker to keep the floor and continue speaking, they are also used by listeners to interrupt the active speaker. When used to interrupt, they function opposite to passive recipiency backchannels, which are considered to be turn-yielding.

In conversation analysis the term *continuer* is used to refer to a backchannel. The term is unfortunate, since at face value it can be understood as a signal produced by the listener for the speaker to continue, as a signal of intent by the speaker to keep the floor and continue (continuation rise), a phenomenon that is also used for feedback requests (uptalk). In Schegloff's terms, a continuer is equivalent to a backchannel. It is used to signal understanding that the floor-holder intends to continue speaking, and that the producer of the signal (the listener) yields their opportunity to initiate a turn (1982, p. 81).

2.5.3 BC pitch

Möking surveys studies investigating BC pitch that show that it correlates with the function of the BC and the nature of the conversation (2025, p. 11f). Sbranna et al. (2022) found a relationship between rising intonation in PR BCs compared to falling intonation in IS BCs in speakers of German. Janz (2022), also analysing German, reports more rising intonation in task oriented dialogue (in this case a map task) compared to more falling intonation in spontaneous conversation.

2.5.4 Cultural variation

Comparing Spanish[10] and AmE, Berry (1994) finds that Spanish speakers prefer longer, more elaborate feedback produced in the backchannel concurrently with a speaker's turn, compared to speakers of AmE. Li (2006) compared mixed- and matched dyads of Chinese and Canadian English. They find that matched Chinese dyads produce more backchannels than Canadian ones and further that the use of backchannels in mixed dyads led to less successful information transmission compared to matched dyads. Besides the difference in BC rate, this indicates that there are relevant cultural differences in BC behaviour.

2.5.5 Working definition BC

Backchannels are short vocal signals that consist of either lexical- (e.g. "yes", "okay") or non-lexical (e.g. "mhm", "ah") vocalisations. Backchannels are produced by a non-floor holding interlocutor (in other words a listener), during the active speaker's turn, typically co-occurring with the speaker's speech. Their primary purpose is the facilitation of the flow of turn-taking, and thus conversation, by signalling most frequently attention and agreement. Two primary functions of BC are distinguished: passive recipiency (PR) and incipient speakership (IS). PR BCs are feedback signals that most frequently convey agreement or understanding, but do not take over the floor. This is relevant because PR BCs often overlap with the interlocutor's turn. IS BCs occur before a speakers' turn begins, indicating the listener's intention to transition and take the floor as the speaker. The distinction of those categories precludes the analysis of responses to polar questions, since these function to answer a question, rather than signal feedback. Combinations of multiple individual backchannel signals occurring in immediate sequence are considered and classified as multi-unit backchannels (MUB) (e.g. "mhm okay"). The definition given here is in line with previous work on BCs, in part analysing the same data (i.a. Janz, 2022; Möking, 2025; Sbranna et al., 2022; Spaniol et al., 2023; 2024; Spaniol, Forthcoming; Wehrle, 2023).

[10] It is unclear what variety of Spanish is analysed.

2.6 Overlapping spoken nods and BCs

The following section considers the combined use of feedback signals and the distribution of feedback across multiple channels. The term overlap used throughout this thesis describes the intersection or co-occurrence of vocal and gestural feedback. Some authors describe overlapping feedback signals as bi- or multimodal.

While some research in the early 1970s already considered language as a multimodal system – especially taking into account embodied gestures for the production of conversational feedback (e.g. Duncan, 1972; Kendon, 1972; Yngve, 1970) – only since the early 2000s have researchers resumed the active investigation of the relationship between vocal and embodied feedback signals.

In a study of Swedish conversation, Allwood & Cerrato (2003) find that vocal feedback (in the form of BCs) and gestural feedback are frequently produced together. The most frequent head movements are nods and jerks. This is compatible with previous findings by Shattuck-Hufnagel & Ren (2018), according to whom 80 % of manual gestures are produced aligned with pitch accents. Further research has also shown that participants judge simulated and real speakers as more natural when they used nods rather than BCs (Poppe et al., 2011).

2.6.1 Function of overlapping nods and BCs

Dittmann & Llewellyn describe that nods and BC (in AmE) overlap more than chance would suggest and find that 70 % of bimodal feedback signals contain IS BCs (1968, p. 81). Ferré & Renaudier (2017) come to a different conclusion (for BrE), describing bimodal feedback to function as markers of agreement.

McNeill (1992) proposed a pragmatic synchrony rule, according to which co-occurring vocal- (BC) and embodied- (nod) gestures fulfil the same function, i.e. only one function can be produced at the same time (1992, p. 29). Włodarczak et al. (2012) adhere to this rule to determine the function of bimodal feedback and exclude instances of overlap of different functions. According to Wagner et al. (2014) this is apparent by a relatively closer match in function between head gesture and speech, than e.g. gaze and speech. Malisz et al. (including all authors of Wagner et al. (2014)) however relativise this claim consequently and describe that the rule cannot exhaustively explain the function of combined feedback signals. This is because the combination of feedback from both modes can serve to add nuanced information about listener stance (2016, p. 431).

Paggio & Navarretta (2013) found that in both task-oriented and spontaneous Danish conversation head gestures and facial expressions are used in conjunction with BCs to signal feedback.

Since overlapping speech is conversationally dispreferred in accordance with Sacks

et al.'s (1974) model of turn-taking, researchers have considered a different use of embodied compared to vocal BCs. In Dittmann et al.'s (1968) data, nods precede BCs – which is interpreted to serve politeness – since it is less interruptive if feedback is sent through, at least initially in a non-overlapping channel. In addition to nods, Dittmann & Llewellyn (1968) also consider smiles and gaze as embodied backchannels that can be used politely without interrupting the speaker. This is compatible with Clark and Krych's (2004) description of a principle of least joint effort. According to this principle, conversationalists will choose the mode that will effect the least strain on both interlocutors' attentional resources. The use of multiple channels for the transmission of feedback is advantageous, since it allows the concurrent exchange of information, while avoiding miscommunication and clashes within a single (vocal) channel (Swerts & Krahmer, 2020). This preference is reflected in Ferré and Renaudier's (2017, p. 27) data, in which unimodal nod feedback is most frequently used while the speaker speaks, in comparison to bimodal- and unimodal vocal feedback.

2.6.2 Temporal placement of nods and BCs

Regarding the timing of nods and BCs, Dittmann & Llewellyn (1968) and Włodarczak et al. (2012) find that nod starts precede BCs in AmE and German respectively by approximately the mean duration of a BC (200 ms). This confirms a generally accepted assumption that gesture onset precedes speech onset (Wagner et al., 2014, p. 218). In Japanese, Ishi et al. (2014) finds that single nods are produced synchronously with BCs. Ferré (2010) finds that gestures generally precede speech in French, while Karpiński et al. (2009) and Chui (2005) find more complicated relationships between gestures and speech in Polish and Chinese respectively.

Comparing the use of nods, BCs, and combinations of both in BrE, Ferré & Renaudier (2017) find that BCs are primarily placed during pauses – so as not to interrupt the speaker, nods primarily during speech – since the gestural signal is less interruptive, and bimodal feedback, i.e. co-occurring or overlapping nods and BCs are produced anywhere in the speakers turn. Regarding the functions fulfilled by these three types of feedback, they find that nods are used as continuers, BCs as assessment markers, while combinations of both are used to signal agreement.

2.6.3 Cueing of nods and BCs

Feedback in both the form of vocal BCs and gestural feedback (i.a.) head nods can be cued by speakers. These opportunities in conversation when listeners can contribute a BC have been described as i.a. backchannel relevant spaces (BRS in Heldner et al., 2013), backchannel opportunity points (BOP coined by Gratch et al.

(2006) in reference to Ward & Tsukahara (2000), and analysed as such in Blomsma et al. (2024)), or backchannel inviting cues (Gravano & Hirschberg, 2009, 2011). Heldner et al. propose it as an analogy to transition relevance places (TRP – cf. Sacks et al., 1974), at which a new speaker takes the floor or remains a listener producing feedback in the form of a continuer (Schegloff, 1982).

Speakers use windows of brief mutual gaze not only to co-ordinate turn-taking, but also to elicit BCs from the listener (Bavelas et al., 2002). Further cues involve the modulation of pitch (Gravano & Hirschberg, 2011; Ward & Tsukahara, 2000), intensity (2011), syntactic completion (Duncan, 1972; Koiso et al., 1998) as well as pauses (Cathcart et al., 2003). The combination of multiple different cues has been shown to decrease listener response time in producing feedback (Gravano & Hirschberg, 2011; Hjalmarsson, 2011).[11]

Not all opportunities for BC feedback are used by listeners (Blomsma et al., 2024; Gravano & Hirschberg, 2011; Heldner et al., 2013). About a third of BOPs are used for backchannels (2024; 2013). Blomsma et al. describe the use of BCs as idiosyncratic, marking it as a source of individual variability. BOPs lead to increased BC production, with greater variability in Blomsma et al.'s data. Listeners' awareness of the state of their interlocutor's turn in the form of turn-end projection (cf. De Ruiter et al., 2006) further underscores the active role of the listener in conversation.

3 Data

3.1 Details on data origin

The data analysed here was originally collected as part of a Tangram experiment. The experiment consisted of three consecutive ten minute segments, two spontaneous interaction, one task-oriented. 28 (13f, 15m) dyadic interactions were recorded in 14 gender- and age-matched dyads (with one exception). In addition, nine dyads of participants diagnosed with autism spectrum disorder (ASD) were recorded and analysed by Spaniol (Forthcoming). The data has been used previously for analyses (Lorenzen et al., 2025; Möking, 2025; Spaniol et al., 2023; 2024).

In the *Introduction* participants had the opportunity to get acquainted with one another. During the *Tangram* segment participants were asked to perform a matching task. The task consisted of Tangram shapes presented to participants, out of which they were asked to find matching pairs. In the *Discussion*, participants were asked to reflect on their experience solving the Tangram task.

For the recording, participants were seated alone in a room, opposite each other at a ~30° angle toward a camera framing both participants. Both participants wore

[11] The analyses of cues listed here are based on an overview provided in Blomsma et al. (2024, p. 1160).

Pupil Invisible eye-tracking glasses (cf. *Invisible*, n.d.; Tonsen et al., n.d. for a detailed description) with additional microphones mounted on the frame. These provided additional perspective video recordings, gaze data, telemetry on spatial movement of the glasses and audio recordings.[12] A camera angle framing both participants was recorded at 1920x1080p (Full HD), 30 frames per second (i.e. 30 Hz). Audio was recorded in stereo, one channel per speaker, at 44.1 kHz.

For the analysis of the interplay of multimodal feedback strategies – specifically head nods and vocal backchannels, recordings of three dyads were selected. Dyads 2 and 3 were chosen due to existing gesture annotations produced for another research project, and dyad 5 was added because it matched speaker gender and age. The analysis here focusses on the *Discussion* segment, since it represents the most naturalistic interaction in comparison to the *Introduction* and *Tangram* segments. The interlocutors were more acquainted in comparison to the *Introduction* and had a shared experience to discuss. All speakers in the subset (N=6) were female aged 22-27 (mean = 24). Except for one speaker, all speakers learned German as their L1.

Speaker identities were coded numerically in addition with an index for their dyad. For the present analysis, data from dyads 2, 3, and 5 (i.e. speakers 3-6 and 9, 10) were selected.

3.2 Annotations

The analyses in this thesis are in part enabled by pre-existing annotations of the dataset produced as part of SFB 1252 *Prominence in Language*. These include gesture annotations; BC pitch, function, and lexical (and non-lexical) type annotations; and intonation phrase boundary tone annotations.

3.2.1 Nod annotations

Head movements were annotated manually as part of gesture annotation using the M3D framework (Rohrer et al., 2023). Annotation was carried out as part of a collaboration between A02 *Individual Specificity* and A07 *Metrical Prominence* of SFB 1252. The annotations were used in analyses first presented by Lorenzen et al. (2025). While all categories of head movement proposed in M3D (nod, turn, tilt, slide, protrusion (2023, pp. 22–23)) were annotated, only nod annotations were extracted for analyses here. Gesture annotation was carried out, as customary in gesture research, without sound. Only gesture events that occurred while no experimenter was in the room were annotated.

[12] Accelerometer data was initially considered to be used to extract speed and amplitude of head movements as a method to generate nod annotations. A comparison with manual annotation is currently planned for future work (cf. 7).

The existing M3D annotations were proofread, focussing exclusively on head nods. During this process a few missing nods were added and some annotations extended, to contain the tail-end of prolonged cyclical nods. Further, it ensured that as many nods as possible are captured in the annotation, since the M3D annotation takes all gesturing into account, with limited attention to head gestures. Proofreading also established parity with three additional speakers annotated by the author, for a total of six speakers across two dyads. For these annotations the focus was exclusively on what is defined here as a head nod (2.3.7). As nods were transcribed exclusively on the basis of visual information, yes-confirmatory nods (i.e. nods that occur immediately after questions) were not specially tagged or excluded from analysis.

3.2.2 Interlocutor intonation phrase boundary tones

The intonation of the final boundary of an intonational phrase (IP) (Selkirk, 1984) used by a speaker had been collected previously as part of a DIMA annotation. The annotation was carried out by the author and proofread by another rater and was first presented as part of a multimodal analysis by Lorenzen et al. (2025). DIMA (Kügler et al., 2022) is a system for the analysis of German intonation, similar to other systems such as GToBI (Grice et al., 2005). A major difference is that accent- and boundary tones are transcribed decomposed rather than combined into pitch trajectories. The annotations of minor and major intonation phrases distinguish high (`H`) and low (`L`) pitch boundary tones based on auditory perception. For analysis, tones marked with diacritics for up- or downsteps were merged into the two primary categories by perceptual similarity.

3.2.3 BC annotations

Vocal feedback in the form of BCs were annotated by Möking (2025) for their analysis and (i.a.) include distinctions for discourse function, (non-/)lexical type and pitch. See 2.5.5 for a comprehensive definition of BCs. Two functions were distinguished: passive recipiency (PR) and incipient speakership (IS). IS BCs were taken to occur within less than 400 ms before the incipient speaker's turn. BC types were categorised by lemmatising the transcribed form. They consist of primarily lexical types, such as "ja", "okay", "genau". One non-lexical type is distinguished as "mmhm". Categories relevant to the subset are given in 5.1.2.

BC pitch was extracted using Praat (Boersma & Weenink, 2024) and smoothed using mausmooth (Cangemi, 2015). Pitch data is represented in two separate data types: numerically depicting the difference of pitch height between start and end in semitones as well as categorically describing pitch direction as rise, level, and fall. BCs are considered level if the difference in height from start to end is in the range

between ± 1 ST.

4 Methods

4.1 Software and tools

The following software was used for preparation and analysis of the data: vocal backchannels and intonation boundaries were annotated manually using Praat (Boersma & Weenink, 2024). Head movements were manually annotated in ELAN (Wittenburg et al., 2023). For video cutting and precise video timecodes LosslessCut (Finstad, 2021) was used. Further handling, pre-processing, and analysis of the annotations was performed in Python (2023). Visualisations were created in R (2024), using RStudio (2024), and tweaked using GIMP (2023). The following R packages were used for visualisation: reshape2, ggplot2, tidyverse, & dplyr (Wickham, 2007, 2016; Wickham et al., 2019, 2025), cowplot (Wilke, 2025), see (Lüdecke et al., 2021), ggforce & patchwork (Pedersen, 2025a, 2025b), and ggpubr (Kassambara, 2025). All code written for analysis and visualisation, figures created, along with the annotations are available online in the project's GitHub repository upon request.

4.2 Time alignment of multimodal data

For analysis of timings of feedback signals, all annotations needed to be manually aligned. The clapper board used during the recording sessions (once at the start of each of the three segments) was used, since it is observable in video and audio signals. Using this as an anchor, offsets from the start of each annotation were taken. For gesture annotations based on video (sampled at 30 FPS, i.e. 30Hz) the frame was chosen at which the clapper appears fully shut using LosslessCut. For audio based annotations (sampled at 44.1kHz) the timecode at which the noise of the clapper closing first peaks on the spectrogram was noted using Praat. Due to the greatly differing sampling rates, the synchronisation between both media types can only be considered an approximation.

5 Analysis

The analysis and discussion section is grouped as follows: in 5.1 properties of nods and BCs are presented. In 5.2 the notion of overlap between nods and BCs is described and differences between overlapping and non-overlapping nods and BCs are analysed. In 5.3 testing of a window applied to the BC annotations to increase overlap is detailed. In 5.4 the effect of a speaker's intonation phrase's final tone on

a listener's feedback behaviour in BC and nod is described. Results are presented indicating dyad-relation, due to the interactive link between participants. The discussion of findings is given interspersed and summarised in 5.5.

5.1 Description of nods and BCs in the data

5.1.1 Nods

A total of 307 nod gestures were identified across all six speakers for an overall mean of 4,67 nods/min. The number of nods varies strongly between dyads and speakers (Figure 1). The variance between speakers is greater than suggested in previous findings by Hadar et al. (1985), whose individual means varied ± 2 nods/min on average. For the entire dataset analysed, nod counts range from 25 to 90 nods per speaker. Within all dyads one interlocutor produces notably nods more than their partner. Nod gestures are primarily cyclical – that is they consist of more than one up-and-down cycle – similar to previous findings by Malisz et al. (2016), whose data identified 78 % of nods identified as cyclical.

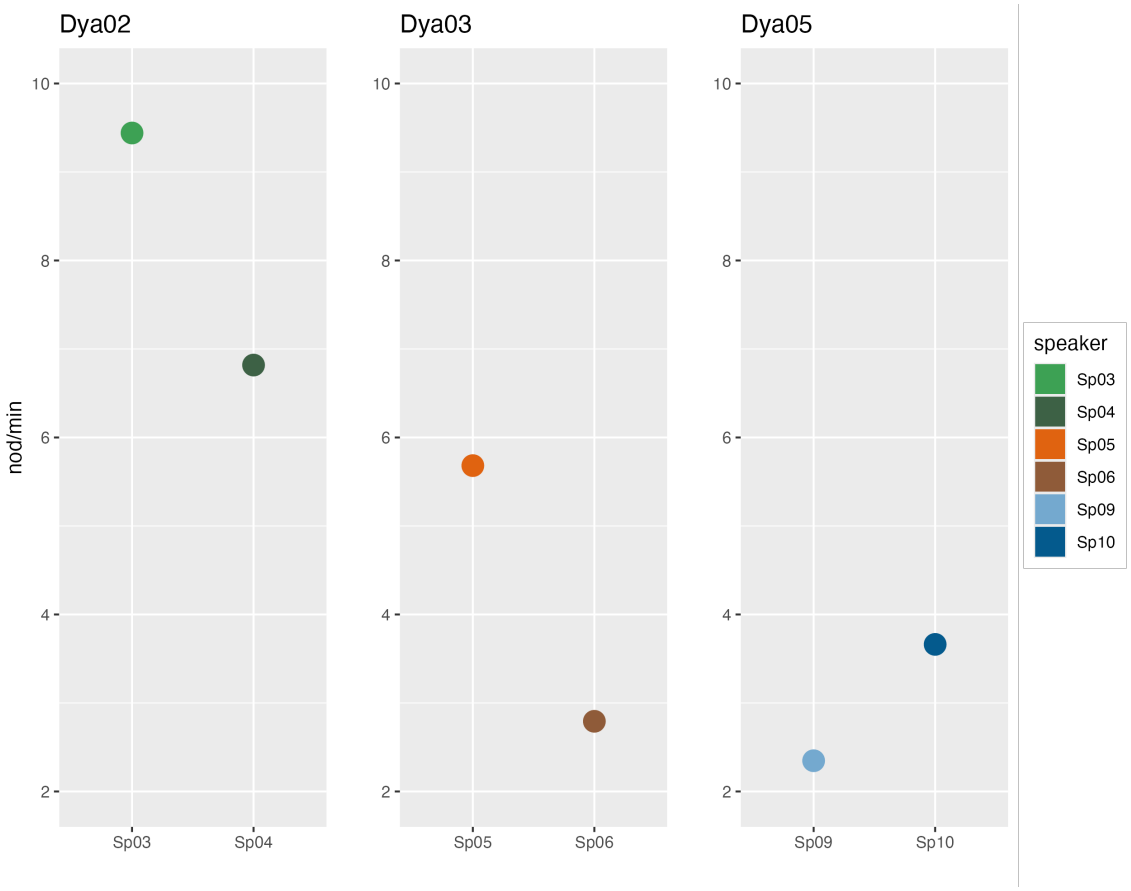

Fig. 1: Nod rate/min by speaker

The overall mean duration of all annotated nod gesture units is 1.14 s. There is notable individual variation between speakers, ranging between 0.77 s and 1.9 s (with outliers, Figure 2). Three outliers were excluded from Figure 2 (including

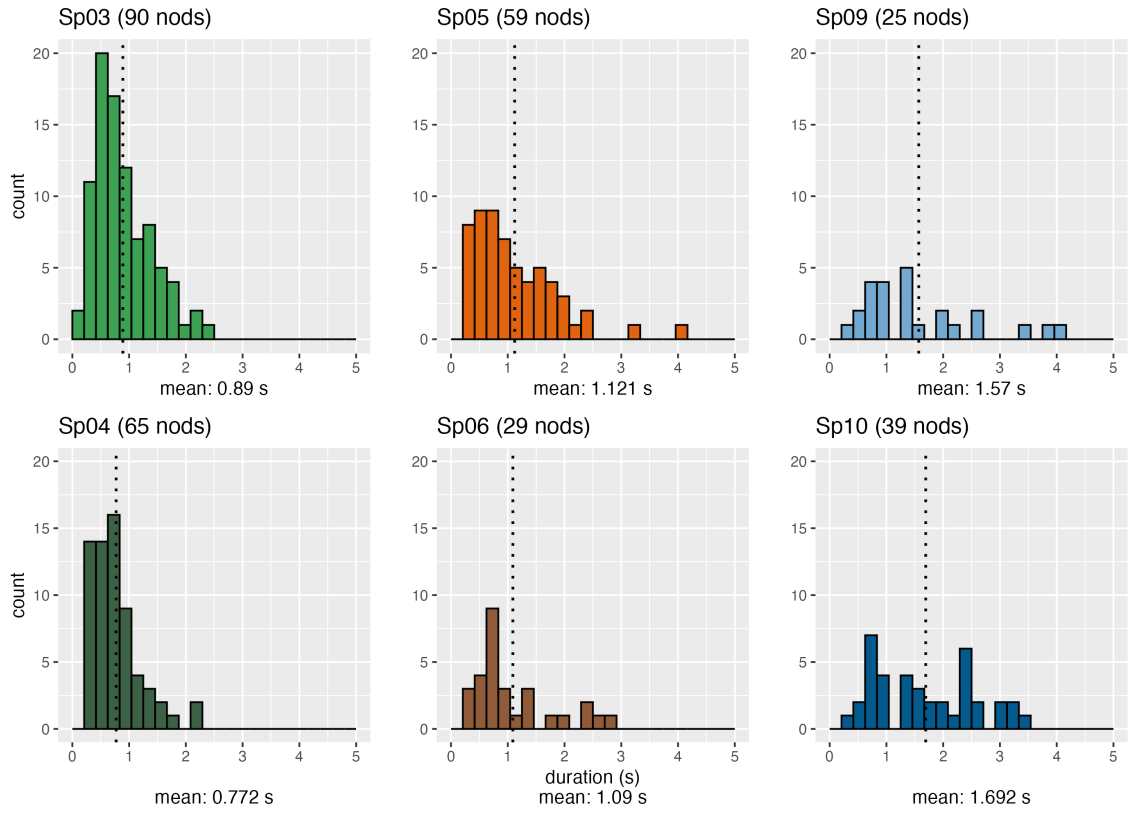

Fig. 2: Nod duration by speaker

derived data depicted in it) for visual clarity[13]. The means of speakers 9 and 10 increase by 121 and 333 ms respectively while outliers are included. The extreme duration of these annotations has two possible reasons. First the cyclical nature of nod gestures. The amplitude of these individual nods typically decreases over time which contributed to the outliers noted above. In those cases participants took longer to stop the oscillatory movements of the cyclical gesture. A second reason is the arbitrary 300 ms minimum gap between gesture units adopted from M3D (Rohrer et al., 2023).

The mean duration of nods appears to be coupled between speakers within each dyad, indicating that both interlocutors may be accomodating each other. Speaker 10 is the only speaker that might distinguish two types of nod gestures by duration, indicated by two peaks. At 0.7 s (present in other speakers) in addition to another peak at 2.4 s duration. Overall, however, the distribution of nod durations does not indicate identifiable categories based on duration alone. Speakers 3, 4, and 5 produced the most nods, which might contribute to the coherent distribution of nod durations.

Table 1 shows that each dyad has an interlocutor that produces notably more head nods. The time spent nodding varies between 31.6 and 80.09 s (mean 55.54 s). This leads to overall 9.01 % of time spent nodding, only half of what Włodarczak

[13]Excluded outliers (duration > 4.5 s): Sp09: 7.919 s; Sp10: 4.943, 7.1 s.

Table 1: Dyad balance based on visual feedback (larger in-dyad count in bold)

dyad	Dya02		Dya03		Dya05		mean
speaker	Sp03	Sp04	Sp05	Sp06	Sp09	Sp10	
nod count	**90**	65	**59**	29	25	**39**	51.16
nod/min	**9.44**	6.82	**5.68**	2.79	2.35	**3.66**	4.67
nod duration (s)	**80.09**	50.15	**66.15**	31.6	39.24	**66**	55.54
time spent nodding (%)	**14**	8.77	**10.62**	5.07	6.14	**10.33**	9.55

et al. (2012, p. 94) report for the mean proportion of head gestures (i.e. nods and other head gestures, such as shakes) in their data. The mean found here is somewhat low considering that previous research has shown a clear dominance of nods within the larger category of head gestures (cf. Malisz et al., 2016, p. 432). The finding is closer to a related finding by Hadar et al. (1983) for BrE, who find that participants' heads were moving for 12.8 % of the time while listening.

5.1.2 BCs

The number of passive recipiency (PR) BCs produced in the *Discussion* segment varies widely between both dyads and speakers, from 66 to 27 BCs, see Figure 3. The overall mean duration of all annotated BCs categorised as (N=260) is 299 ms, varying between 211 and 459 ms for individual speakers.

Table 2 shows the balance of the dyads in terms of speech produced, using tokens[14] and PR BCs produced, and total time spent speaking (sum of the duration of annotated spoken tokens) as measures. All three dyads appear to have a dominant interlocutor who produces more speech than the other. Counts in bold indicate the larger number comparing both speakers in a dyad. The data available here does not entirely confirm the intuition that within a dyad the more one speaker contributes to the conversation (operationalised here through the amount of tokens), the more their interlocutor – the listener – backchannels, as this is only apparent in dyads 2 and 3. Still, all three dyads show that the person that spends less time speaking in a dyadic conversation produces about twice as many backchannels.

For the full dataset Möking (2025, p. 30) reports a similar median of 301 ms during the *Discussion* section of the experiment. For all three sections of the experiment (*Introduction*, *Tangram*, and *Discussion*) the mean is higher, at 318 ms. Möking explains this difference by a difference in lexical (and non-lexical) BC types used, rather than a difference in production speed.

The intonation of BCs was measured in semitones (ST) as the difference between

[14]The term token is here used to denote the number of instances of "words", as in e.g. corpus linguistics – in opposition to types.

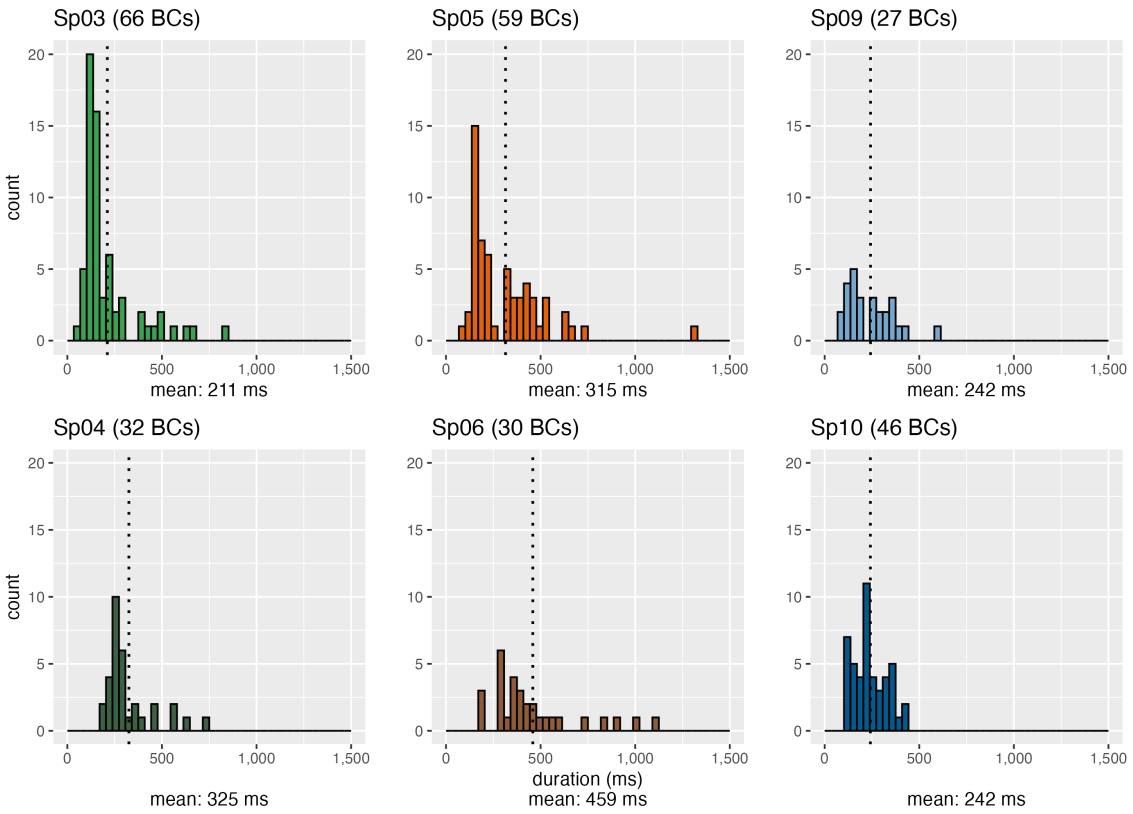

Fig. 3: Duration of PR BCs by speaker

Table 2: Dyad balance based on vocal feedback (larger in-dyad count in bold)

dyad	**Dya02**		**Dya03**		**Dya05**		**total**
speaker	Sp03	Sp04	Sp05	Sp06	Sp09	Sp10	
BC (PR) count	**66**	32	**59**	30	27	**46**	260
BC (PR) mean (ms)	211	**325**	315	**459**	242	242	299
token count	868	**1337**	1142	**1269**	904	**1137**	6657
token duration (s)	177.86	**193.01**	207.90	**210.52**	**223.16**	205.63	1218

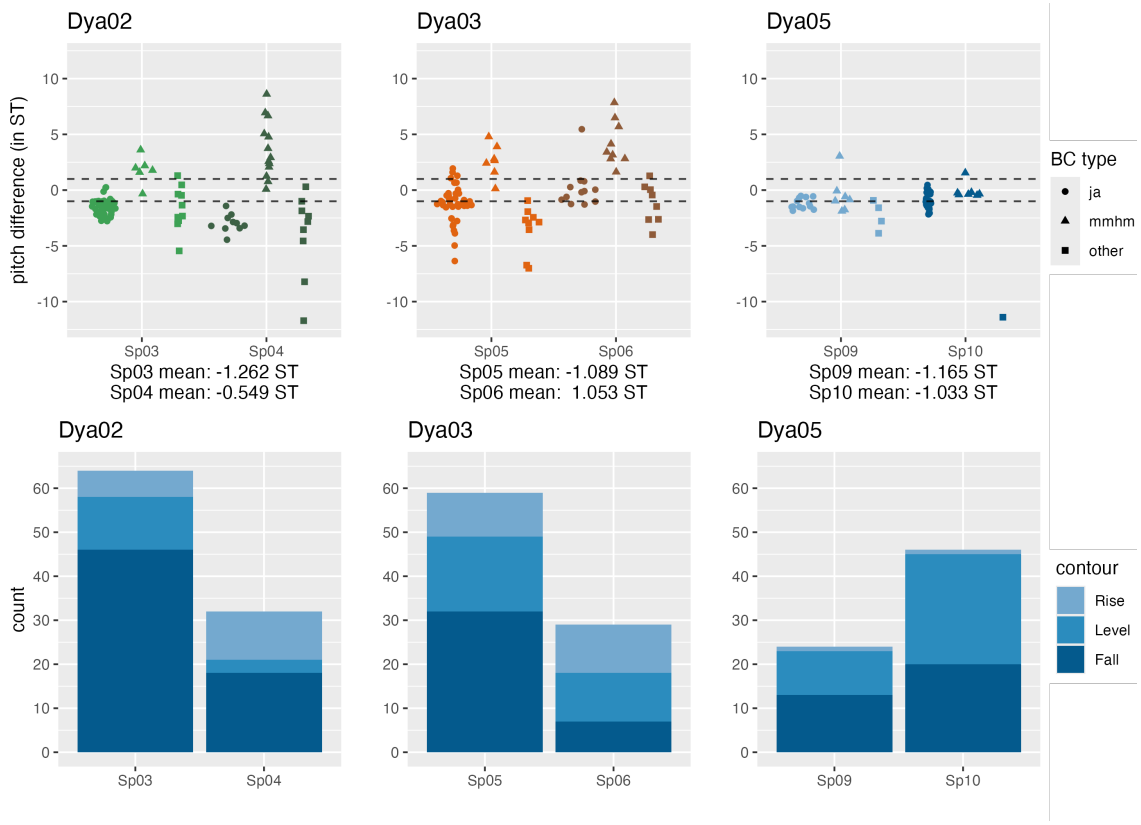

Fig. 4: PR BC intonation in ST (positive values indicate rising pitch) and by category

pitch at the start and end of the BC. Using these values, three categories of BC are distinguished: rising, level (range of ± 1 ST from 0), and falling intonation. Speakers produced BCs with predominantly level and falling contours, as indicated by the means of all but one speaker ranging from -1.262 to -0.549 ST (overall mean: -0,674) in Figure 4[15]. Lexical BCs are overall level to falling. "ja" BCs cluster closer together than "other" BCs and are positioned toward the level category. Non-lexical "mmhm" is produced level to rising across all speakers, except for speaker 9 who produces "mmhm" with a falling contour.

Speaker 6 is an outlier with a mean of 1.053 ST difference. In comparison to the full dataset reported in Möking (2025, p. 41f), the mean in the subset here is slightly lower than the overall *Discussion* mean at -0.83 ST. Despite the closely grouped mean values, the three assigned categories show individual variability. Both speakers 9 and 10 (dyad 5) diverge from the other speakers by producing less BCs overall across a narrower intonational range.

Figure 5 shows the distribution of lexical (ja, other) and non-lexical (mmhm) BC types by speaker. Annotated types were collected into three major categories. Due to low frequency, BCs such as okay, genau, and multi-unit backchannels (MUB – instances of more than one BC, if they occur within 200 ms of one another) are

[15]Six BC annotations were excluded because no measurement could be extracted using Praat. They are distributed as follows: Sp03: 2, Sp06: 1, Sp09: 3.

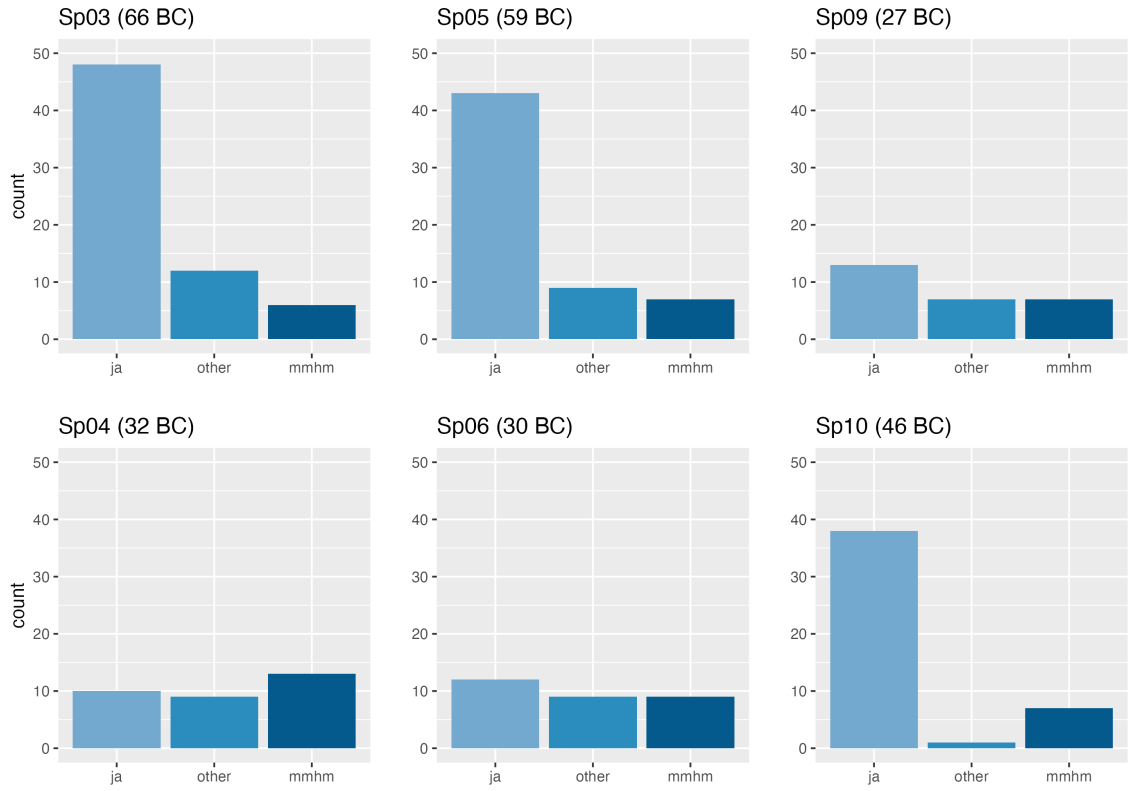

Fig. 5: BC types distingiushed by (non-)lexical category

presented here collected in the "other" column.

As previously reported in Möking (2025, p. 36) for the full dataset, "ja" is also the most frequent BC type in the subset. The relative preference for this lexical type varies however by speaker: speakers 3, 5, and 10 show a clear preference for it, while their dyad partners (who also produce less BCs overall) do not have as strong a preference for it. The distribution of their partners' (speakers 4, 6, and 9) BC types are representative of the relative proportions between types reported by Möking (2025, p. 38, Fig 6) for the full dataset during the *Discussion* segment. The second most frequent unique BC type is the non-lexical "mmhm" (except for speaker 4). For the full dataset Möking reports more variety in BC tokens, particularly during the *Introduction* and *Discussion* segments (2025, p. 40).

5.2 Annotated nods overlapping with annotated BCs

5.2.1 Categories of overlap

Three types of overlap between BC and nod are distinguished in Figure 6. The term overlap is used here to describe the simultaneous co-occurrence of feedback signals across modalities. The size of overlap is not taken into consideration. *Nod-BC* – nod begins before- and ends at some point during the annotated BC. *Nod-BC-nod* – the nod similarly starts before the BC, but continues past the end of the BC,

encompassing the BC. *BC-nod* – the inverse of *nod-BC* overlaps, in that the nod begins at some point during the BC and continues after the BC has ended. A category in which the nod is encompassed by the BC has not been considered given the mean nod duration being nearly four times as long as that of BCs. Cases of no overlap are not distinguished in terms of relative position of BC to nod.

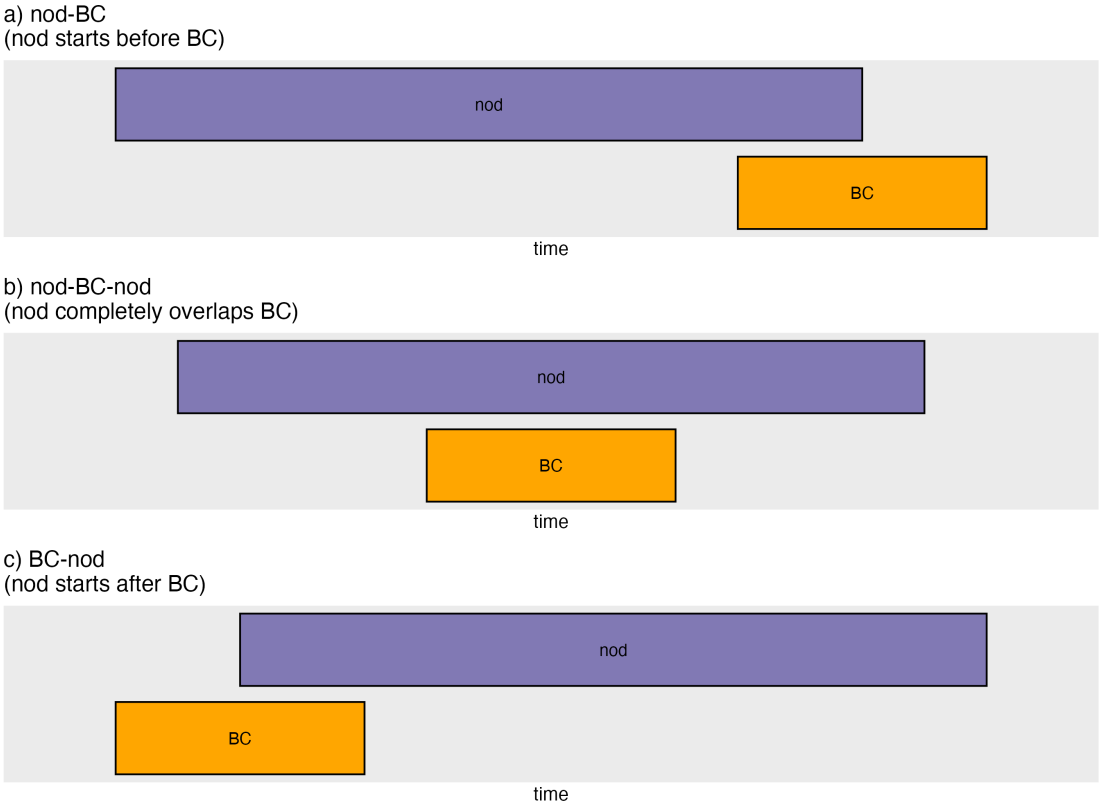

Fig. 6: Types of overlap

5.2.2 Overlap between PR BCs and nods

The following section analyses overlapping feedback, produced as a joint feedback signal. Figure 7 shows the amount of overlapping annotations relative to nods and BCs produced by individual speakers. The number of overlaps varies drastically between speakers without indication of an overall trend, reflecting reports in Allwood & Cerrato (2003) and Cerrato (2007). Within dyads however, speakers appear to match one-another, in terms of the relation between BCs and overlapping feedback. The number of overlaps ranges between six and 46.

Malisz et al.'s (2016, p. 429, Table 8) finding that a third of nods overlap with BCs in attentive German listeners is approximately reflected in this data set. Cerrato's (2007, p. 104, Fig. 6.4) study of spontaneous Swedish conversations finds a similar distribution (with some inter-dyad variability), where on average more than a third of feedback signals are produced in cross-modal overlap.

The finding for the amount of overlap sets head gestures apart from manual

gestures, based on Shattuck-Hufnagel & Ren's (2018, p. 6 Table 1) finding (for AmE) that about 80 % of manual gestures occur aligned with pitch-accented syllables.

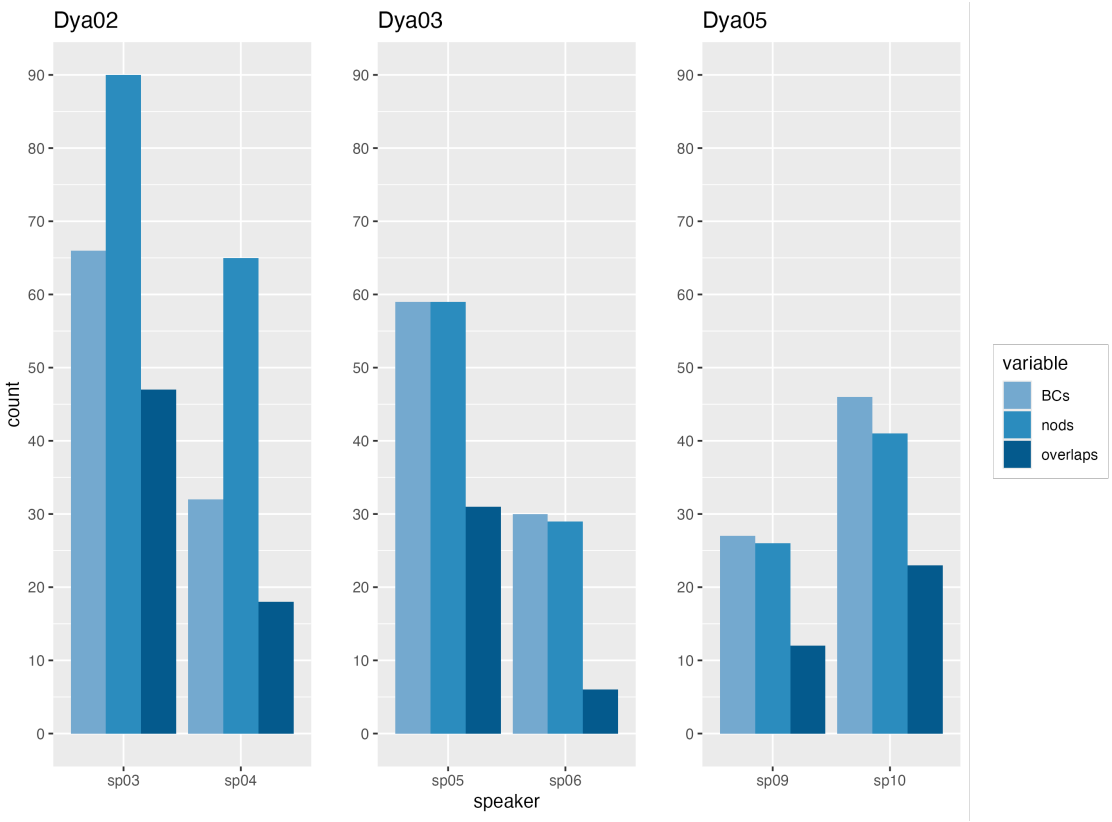

Fig. 7: Comparison of BCs, nods, and overlaps (BC and nod) by speaker

Figure 8 shows that individual speakers' values range between 18.83 % and 41.81 %. The mean of the amount of overlap is 31.17 %. Overlap here describes the percentage of a nod interval that is produced in overlap with a BC. In all dyads the interlocutor that produces more feedback signals, produces it with shorter overlap in comparison to their interlocutor.

Figure 9 shows that overlaps in the dataset consist predominantly of nods starting before an overlapping BC and extending beyond the end of it (type nod-BC-nod). This is consistent with findings for single nods in Japanese dyads (Ishi et al. (2014), 239, Fig. 5). And overall findings that feedback gesture onset precedes that of voice (Dittmann & Llewellyn, 1968; Włodarczak et al., 2012). Overlaps of types nod-BC and BC-nod occur with a uniformly low frequency. The finding that overlaps predominantly consist of nods encompassing BCs supports findings by Ferré & Renaudier (2017, p. 23) who found the same for head gestures in BrE.

Figure 10 shows the distribution of overlap types according to BC type. "Ja" is predominantly embedded in a nod. "Other" lexical BCs are also most frequently associated with this overlap, but show a tendency to start during, and extend after the nod. For nonlexical "mmhm" the converse is true, here the BC appears to frequently precede the nod gesture. While nod-BC-nod overlaps form the majority

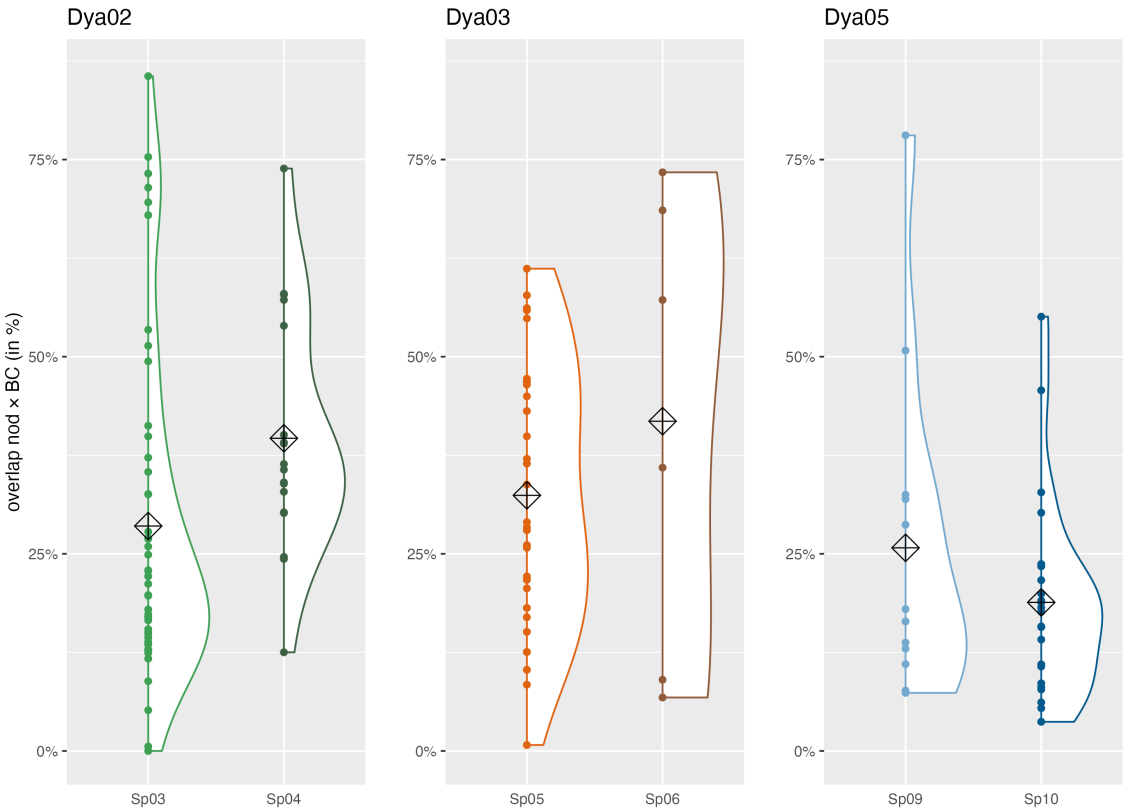

Fig. 8: Distribution of overlap size by speaker (in % to facilitate direct comparison of annotations of different length)

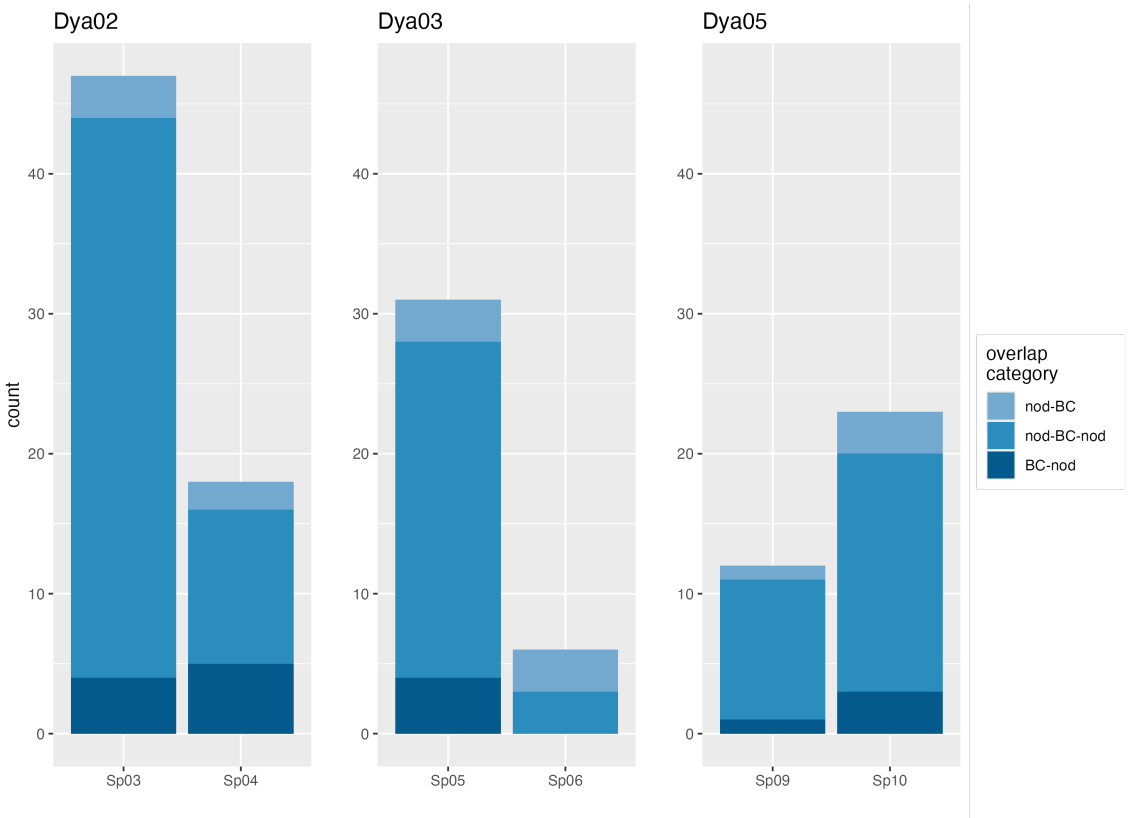

Fig. 9: Overlap category by speaker

Fig. 10: Overlap category by speaker correlated with BC type

Table 3: Mean offset between start time of overlapping nod and BC

dyad	Dyad 2		Dyad 3		Dyad 5		total
speaker	Sp03	Sp04	Sp05	Sp06	Sp09	Sp10	
mean (s)	0.395	0.16	0.607	0.686	1.72	1.515	0.847

of overlaps, the position of nonlexical "mmhm" before a nod may be used to signal growing understanding in the listener, reinforced by the nod gesture. The same may be true for "other" BCs. In their case the BC may be used as an emphasizer of the nod gesture.

The mean offset between overlapping nods and BCs is 847 ms. This is drastically higher than what Dittmann & Llewellynn (1968, p. 82) reported for English (175 ms mean) and what Włodarczak et al. (2012, p. 95) found for the overlap of head gesture units (a superset of nods) and BCs in German (202 ms). This may be due to unique experimental conditions, measurements or individual differences. Table 3 shows that speakers in dyad 2 are below (160 ms) and closer (395 ms) to the means reported by Włodarczak et al. (2012) and Dittmann & Llewellyn (1968). The overall mean given above is strongly influenced by the means computed for the offset in dyad 5, exceeding 1.5 s. Comparing within-dyad offsets, it appears that the offset is established in a dyad's individual dynamic and applies to both speakers.

5.2.3 Overlapping nods cross-correlated with features of BCs

Fig.s 11 and 12 map the duration of nods (x-axis) corresponding to the duration of the overlapping BC (y-axis), in combination with overlap category and lexical type. Nod outliers longer than 4.5 s are again excluded for visual clarity (as in Figure 2). Note that it is possible for BC annotations to occur multiple times in overlap with a BC due to the duration and proximity of nods.

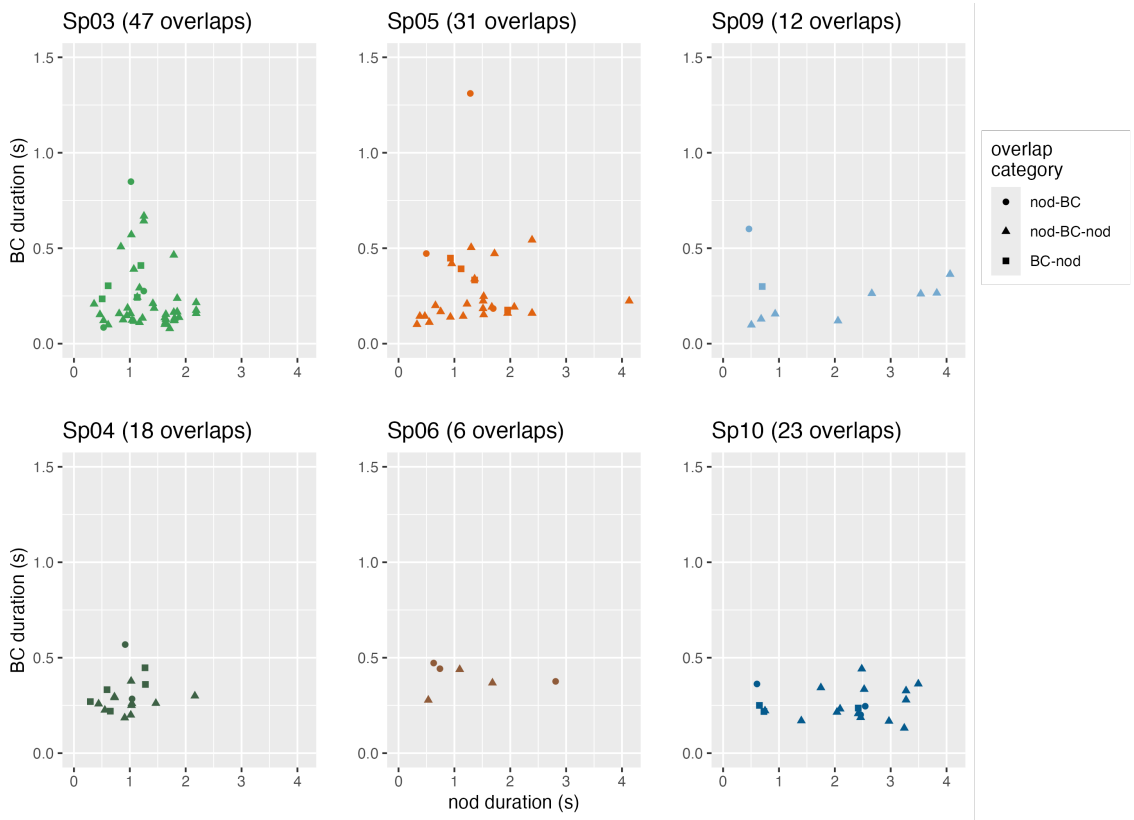

Fig. 11: Duration of overlapping nod and BC by overlap category

In Figure 11, the majority of overlaps (nod-BC-nod) do not appear to pattern with specific durations of neither nods nor BCs. Both Nod-BC and BC-nod overlaps are characterised by BCs longer than 200 ms, while nod duration varies freely, similar to nod-BC-nod overlaps.

The choice of neither lexical ("ja", "other") nor nonlexical ("mmhm") appear to have a relation to the duration of an overlapping nod Figure 12. The strong preference for "ja" as lexical BC type overlapping with nods is apparent, which is expected since "ja" is the most frequent BC type overall.

Overall both figures illustrate the individual variability in duration of overlapping nods and BCs. The type of overlap, lexical category, or pitch category (which has also been tested) do not show notable correlation with BC or nod duration.

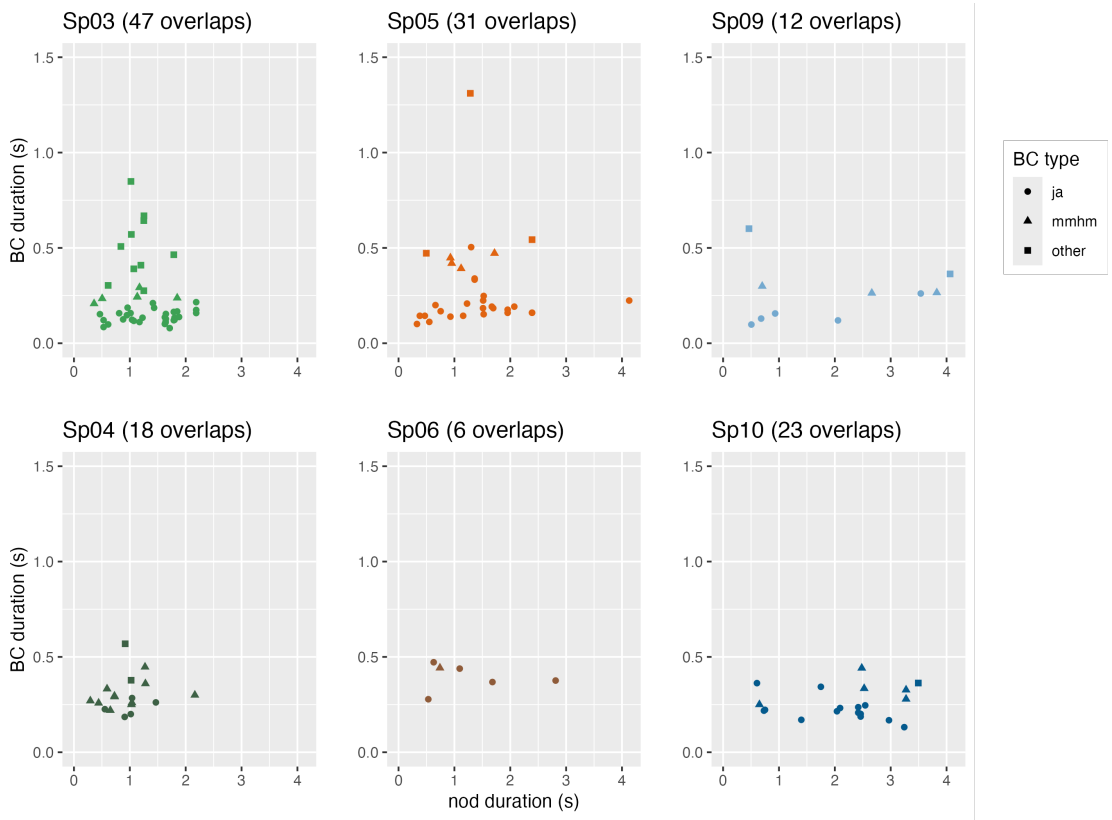

Fig. 12: Duration of overlapping nod and BC by BC type (collapsed categories)

Table 4: BC type by speaker

dyad	Dya02		Dya03		Dya05		total
speaker	Sp03	Sp04	Sp05	Sp06	Sp09	Sp10	
BC (PR)	66	32	59	30	27	46	260
BC (IS)	17	17	6	14	4	10	68

5.2.4 Overlap between IS BCs and nods

After focussing heavily on PR BCs, the annotated BCs categorised to function as markers of incipient speakership (IS) overlapping with nods are considered. Overlap of a nod with a BC fulfilling a differing function such as IS may have an impact on the function of the nod. Table 4 compares counts of both BC functions. IS BCs occur less than PR BCs with a speaker specific variance.

A total of 15 IS annotations overlap with nods as shown in Figure 13. IS BCs do not overlap with nods in speakers 6 and 9. While there are significantly less IS BCs than PR BCs, IS BCs also overlap less with nods than PR BCs. The overlaps observed can be categorised as seven in which the nod precedes the BC, five in which the nod encompasses the BC, and three in which the BC precedes the nod. Based on these limited datapoints it appears that IS BCs that are produced in overlap with a nod start later than PR BCs. Due to very limited data it is impossible to draw conclusions with any claim to generality.

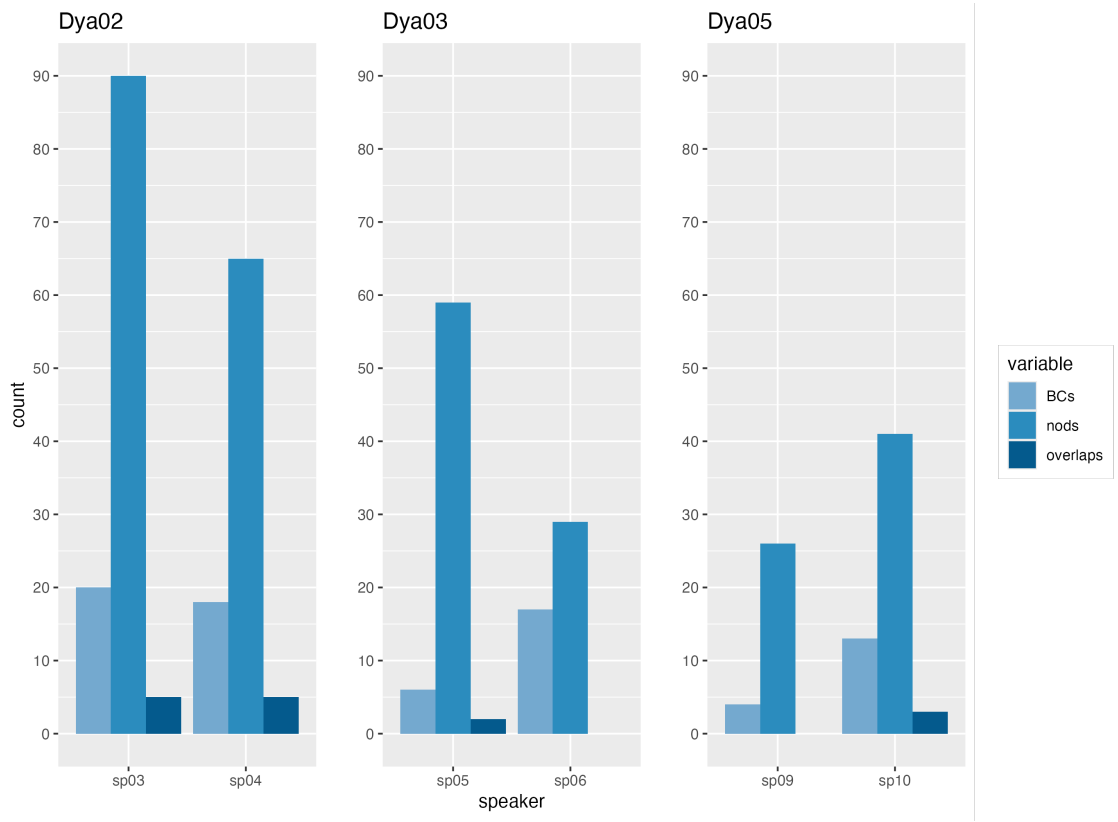

Fig. 13: Comparison of IS BCs, nods, and overlaps (BC and nod) by speaker

Table 5: Count and mean duration of nods that do not overlap with any BC, by speaker

dyad	Dyad 2		Dyad 3		Dyad 5		total
speaker	Sp03	Sp04	Sp05	Sp06	Sp09	Sp10	
count/total	49/90	42/65	31/59	23/29	15/25	20/39	180/307
mean duration (s)	0.664	0.732	0.966	1.045	1.320	1.406	1.022

5.2.5 Nods and PR BCs not overlapping

Table 5 indicates counts[16] and mean duration of annotated nods that do not overlap with PR BCs. Mean duration is lower in all non-overlapping nods. The meaningfulness of this difference is questionable however, since the longer the annotated nod the higher the chance that a BC overlaps with it.

PR BCs that do not overlap with nods are overall similar to those that overlap with nods. The mean duration is equivalent. In terms of intonation, non-overlapping BCs are more frequently level, this is evident in a higher overall mean comparatively (-0.885 ST). Lexical category is distributed similary, with regard to the most frequent category of "ja". However, the frequency of lexical "other" and non-lexical "mmhm"

[16]The number of nods not overlapping with BCs may appear incongruent with the total number of nods and that of nods overlapping with BCs. This incongruity is caused by the way that overlaps are counted, from the perspective of the BC rather than the nod. Because some nods overlap with multiple BCs, the number of overlaps exceed the difference between total nod count and that of non-overlapping nods.

is flipped, in comparison to overlapping BCs. The larger amount of "other" PR BCs not overlapping with nods may be explained by the diversity of lexical forms collected in the category, or the fact that semantically more propositional content is already present in the form, compared to non-lexical "mmhm".

5.3 Window concept and effect on overlaps

The application of a window in the form of a variably sized interval was tested. The basic idea is to increase the number of associated feedback through overlap. A window was applied to both ends of the BC annotation interval as indicated in Figure 14. The second and third window in Figure 14 schematically establish a BC-nod type overlap. The BC annotation was chosen instead of the nod since it is better understood and had previously been used for applying a window (cf. Türk et al. (2025)). In the present dataset the number of overlaps does not change when the window is applied to the nod annotations.

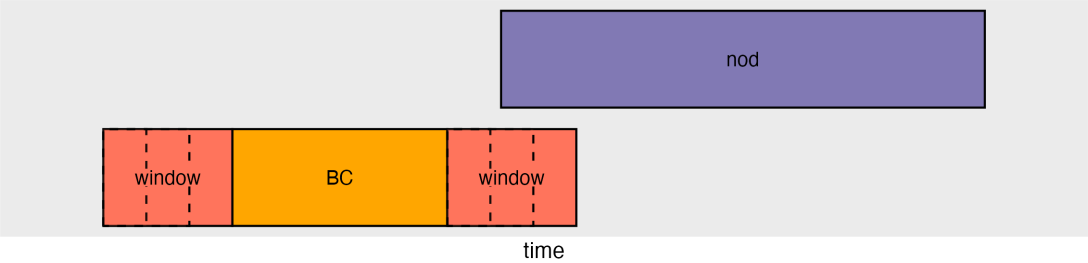

Fig. 14: Schematic representation how the window added to the interval of the BC establishes association (through overlap) between annotations

This approach can, generally speaking, aid in the establishment of associations (through overlap) between temporally sequential events. There may be systematic delays between signals that are easiest to associate with one another through the application of a buffer or "window". Testing was carried out in 100 ms increments (effectively 200 ms = 100 ms pre-BC + 100 ms post-BC) up to 5 seconds.

The amount of overlaps compared to the total number of BCs resulting from testing is given in Figure 15. The number of overlaps is expressed in percent to ease comparison between speakers. Except for speaker 6, ca. 50-60 % of all speakers' nods overlap with BCs without the use of a window artificially facilitating more overlap. The more feedback a speaker produces, the more relative overlap exists. The difference between speakers paired in a dyad is maintained in dyads 2 and 3 when applying increasingly large windows. In dyad 5 the difference in relative overlap count starts comparatively closer and the application of a window causes intersections of the trend lines.

When the number of overlaps reaches 100 %, all BCs are associated with at least one nod. Total overlap (i.e. when each BC is associated with a nod) is reached for

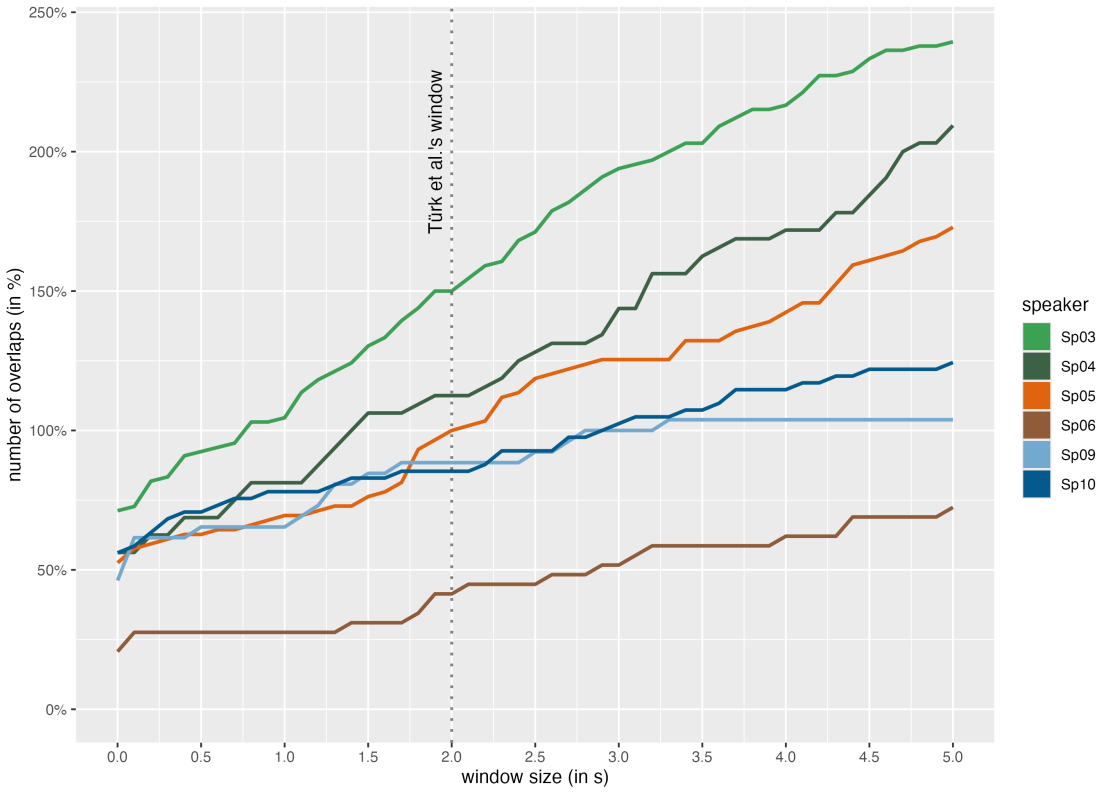

Fig. 15: Effect of added window size on amount of overlapping annotations (in %), window size on x-axis denotes individual interval added to start and end of BC

five (and exceeded for four) out of six speakers at 3.25 s.

When a 2 s window (i.e. effectively 2+2 s, as in Türk et al. (2025)) is applied, the initial amount of overlap approximately doubles in all speakers. Overlap categories shift drastically with this window applied, from the majority consisting of nods encompassing BCs (*nod-BC-nod*) (cf. Figure 9), to the majority consisting of a nod that precedes the overlapping BC (*nod-BC*) (cf. Figure 16). The amount of overlaps that are of the type nod-BC with a 2 s window implies that overlaps without a window are closer to the end of the nod than to the start. Türk et al. (2024) cite psychological research by Pöppel (2009) to motivate their chosen window size of 2 s, as they argue that a 2-3 s range is what is perceived as the current moment. At the 2 s window size, speakers 3 to 6 (those that produce the most feedback in both modalities) have reached 100 % overlap.

It was expected that if there were a systemic gap between nod and BC, the development graph would plateau within a certain window range. This is not the case. Instead, it appears that multiple BCs are associated with nods with increasing window size due to proximity of feedback signals, resulting in a linear increase of overlapping annotations. Based on the lack of a plateauing or other effect in the data and sizeable amount of overlap without a window, further analysis of the data using a window is foregone. When a window is used to facilitate overlap between

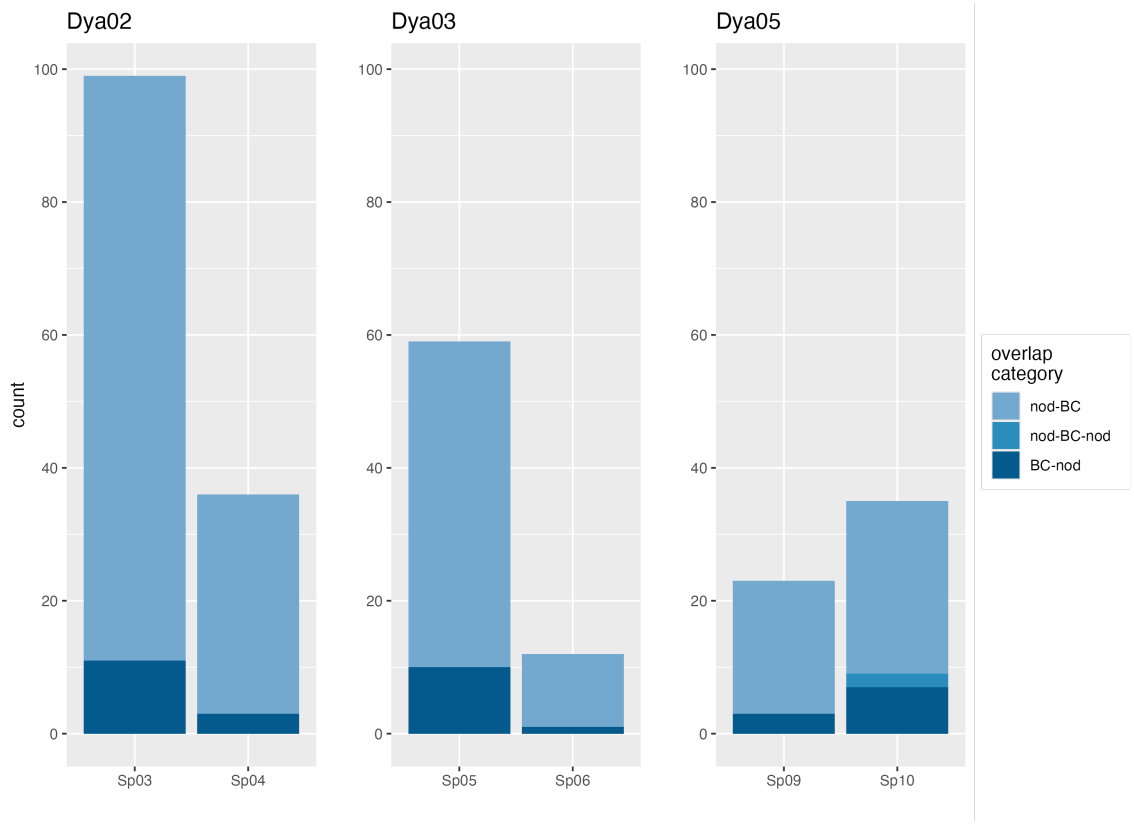

Fig. 16: Overlap category by speaker, with a 2 s window applied

interval-based data, testing of the impact of the window appears generally advisable.

5.4 Interlocutor's IP end pitch direction (DIMA boundary tones)

Figure 17[17] and Figure 18 show PR BCs and nods produced by a listener respectively as they occur within 0.5 s before, and 1.5 s after a high (H-%) or low (L-%) boundary tone produced by a speaker. This analysis is motivated by previous findings which suggest that syntactically complete phrases (which are often reached at the end of intonation phrases) are used as BC opportunity points (cf. Duncan (1972) (AmE), Koiso et al. (1998) (Japanese)). Further, both low and high sentence final pitch has previously been described as a means to elicit feedback in other languages (cf. Gravano & Hirschberg, 2011 (AmE); Ward & Tsukahara, 2000 (AmE & Japanese)).

Boundary tones are treated as intervals here to gather associated nods and BCs. A negative offset was applied, since notable amounts of feedback were produced shortly before the annotated boundary tone is reached. While the boundary tone itself may not have been produced when the feedback signal starts, the pitch trajectory of the intonation phrase has been indicated by this point. The (arbitrary) 2 s interval used to associate the end of the speaker's intonation phrase and a following feedback signal establishes the successful elicitation of feedback through the boundary tone produced by the listener.

[17]Due to lack of data in dyad 5, only one speaker is represented per condition.

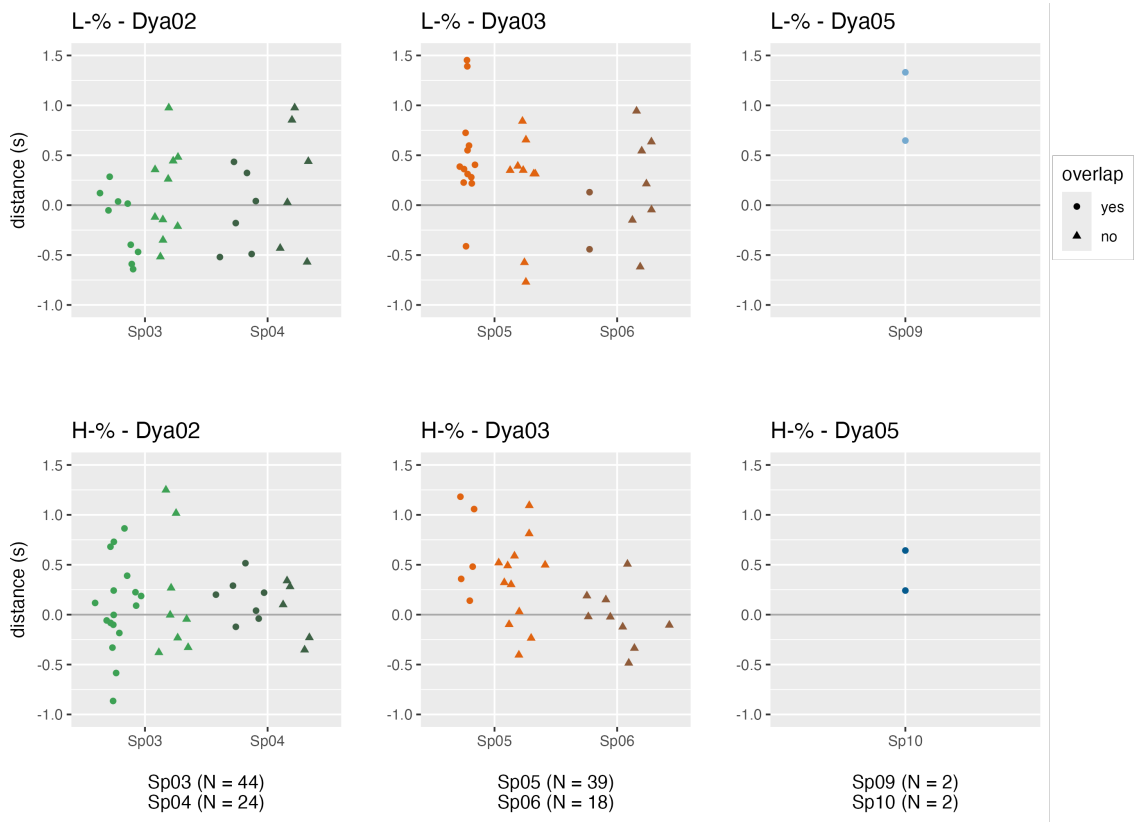

Fig. 17: Distance between speaker boundary tone and listener BC, distinguishing between (non-)overlapping annotations

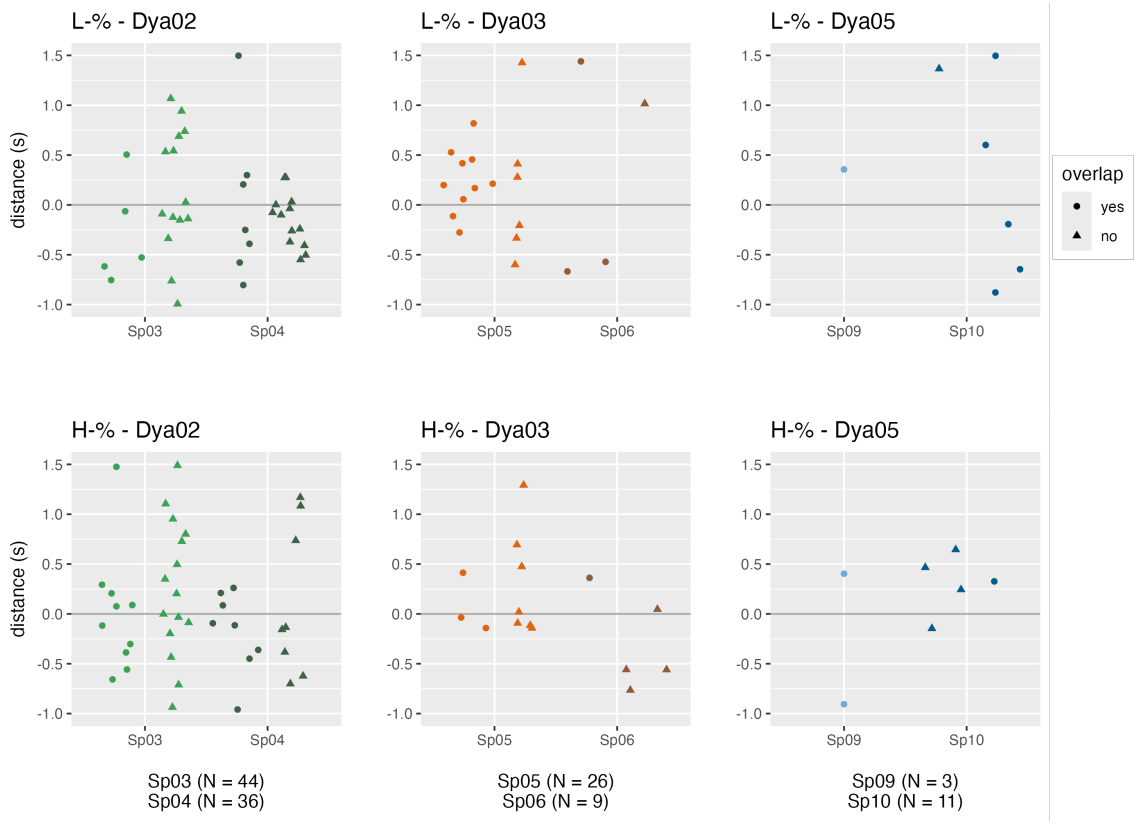

Fig. 18: Distance between speaker boundary tone and listener nod, distinguishing between overlapping (circles) and non-overlapping (triangles) annotations

Note that some boundary tones elicit more than one feedback signal from the listener, both within- and across modalities. This is the case, e.g. when multiple BCs are uttered intersecting with the boundary tone interval. Nod and BC Feedback signals that overlap across modalities are distinguished in Fig.s 17 and 18.

The amount of total feedback signals that are elicited by a boundary tone varies starkly by speaker (ca. 5-75 %, both values for BCs). Overall, compared to the total feedback signals by modality on average more BCs (50 %) follow boundary tones than nods (42 %) do.

There is no apparent difference in feedback behaviour produced by a listener following high and low boundary tones produced by a speaker. Regarding BCs, this contrasts with previous findings by Ward and Tsukahara (2000, p. 1185) who found that after a speaker produced low pitch for 110 ms, listeners produced more vocal BCs in AmE and Japanese, and Gravano and Hirschberg (2011), who describe high pitch as a BC inviting cue in AmE.

Comparing feedback modality, individual preferences appear to be defining. Speakers 4 and 10 produce more nods, while speakers 5 and 6 produce more BCs in proximity to speaker boundary tones. Speakers 3 and 9 produces equal amounts in both modalities. Descriptions of speakers 9 and 10 need to be treated with caution due to limited associated feedback. Across all speakers a dynamic related to the amount of feedback is apparent however: the more feedback is produced, the lower the offset to the boundary tone, i.e. the quicker feedback follows the speaker's speech.

5.5 Summary of findings

In the data presented, preference for feedback modality appears to be dyad-specific. Two dyads produce similar amounts of nods and PR BCs, while one dyad produces more nods than BCs. The mean duration of nods (1.14 s) is about four times as long as PR BCs (299 ms). The amount of overlaps varies drastically between speakers without a common tendency related to both BC and nod count.

On average, almost two thirds of all nods recorded are produced as pro-speech gestures, i.e. without overlapping with a BC. Per speaker this number varies between 50-80 % of nods. The mean overlap size of BCs with nods varies by speaker between ca. 20-40 %. This indicates, in combination with the difference in total feedback produced by listeners, a notable degree of individual variability in terms of multimodal feedback strategy.

The central finding is that when feedback overlaps, nods predominantly fully encompass overlapping PR BCs (*nod-BC-nod*). IS BCs and nods overlap less frequently than PR BCs, and are more likely to start later, in *nod-BC* type overlaps. There

does not appear to be any connection between lexical type, pitch, or overlap type with regard to the duration of overlapping PR BCs and nods.

The evaluation of the use of a window to associate feedback across modalities has shown that it needs to be handled with caution. When adding a window, the number of overlaps increase linearly per speaker, without indication of a plateau before more than 100 % of annotations overlap.

The small data set is strongly characterised by what appears to be individual variability. Within dyads speakers appear to behave more similarly compared to their interlocutor in terms of nod-, BC-, and overlap counts than across dyads. Dyad 5 appears to have a different dynamic compared to the other dyads, evident in their production of longer nod gestures, fewer BCs, produced with predominantly falling intonation contours, and notably less feedback after speakers conclude intonation phrases (independent of the boundary tone produced).

6 Considerations for generalisability

The findings presented in 5 should be interpreted with caution, since the interaction between participants was externally motivated and recorded in a controlled environment and is based on a small subsample of just three dyads. To elaborate on these limitations, the following section explores the validity of lab speech 6.1, ecological validity 6.2, as well as the problem of sampling as it pertains to generalisation of findings 6.3. Section 6.4 summarises limitations of the data in terms of the previously established considerations.

6.1 Validity of Lab Speech

The debate around the validity of "lab speech" in linguistics, specifically phonetics, questions the entrenched reliance on speech data gathered in laboratory settings. Lab speech is (via the term) generally negatively connotated, because it implies that data is "unnatural, overly clear, over planned, monotonous, lacking of rich prosody, and devoid of communicative functions, interactions and emotions" (Xu, 2010, p. 329).

Thus, recorded speech data can be categorised as one of two extremes: "lab speech" – which is highly controlled and *natural(istic)*-, or *spontaneous* speech – which is more natural. Lab speech is often used for theory-driven-, confirmatory research because it allows researchers to focus on specific phenomena of interest, because of the control exerted in elicitation. In stereotypical "lab speech", participants are asked to repetitively carry out short tasks in an unfamiliar context. Conversely, naturalistic speech is more suited to data-driven-, exploratory research because it

is less focussed. Naturalistic speech is recorded in speakers' normal environment, recording their minimally influenced natural behaviour.

Xu (2010) defends lab speech by arguing that concerns about lacking naturalness are unfounded. Xu argues that the control exerted in laboratory settings is an advantage, because it affords researchers the ability to correlate manipulated variables. Wagner et al. (2015) argue that the term "natural" should not be used due to lack of terminological clarity. Instead, Wagner et al. call for a gradient scale to capture the quality of data elicited and more attention to the effect of elicitation methods on the quality of data collected. Feedback behaviour for example differs depending on the experimental task.[18]

Specific behaviours, such as e.g. head movements, require specific equipment to be recorded, often at the cost of impeding participants' natural behaviour. But technology available for the recording of participant behaviour has improved dramatically in recent years. While Dittmann & Llewellyn (1968) and Hadar et al. (1985) had to rely on obtrusive custom rigs for the recording of head gestures, digital video recording in conjunction with computer vision can be used to gather similar data, while restricting the participants' behaviour to a lesser degree. In the end lab speech is not a problem per se, as long as authors, and by extension their readers, are aware of possible effects of the elicitation methods on data and associated findings.

6.2 Ecological validity

The concept of Ecological validity (EV) – now widely used as a buzzword with reference to the validity of experimental scientific inquiry – was first described by Brunswik (1956). Ecological validity conventionally assesses "whether or not one can generali[s]e from observed behavio[u]r in the laboratory to natural behavio[u]r in the world" (Schmuckler, 2001, p. 419). Schmuckler describes three central dimensions to assess EV: setting, stimuli and response. The concept itself was, however, originally not stipulated as "ecological validity" (this refers to another, earlier theory by Brunswik), but rather "experimental design". Holleman et al. (2020) summarises representative design as an approach in which researchers design and conduct their experiments by making sure that participants, situation, task, and stimuli are representative of a sample population.

The term and concept of ecological validity and its use are widely critiqued. Hammond (1998)[19], a student of Brunswik, decries the skewed (ab)use of the concept by contemporary scholars. Holleman et al. advocate for an explicit disclosure of what

[18] Spontaneous interaction in Danish and German yielded more BC (Dideriksen et al., 2019; Spaniol et al., 2024) compared to task oriented interaction (Fusaroli et al., 2017; Janz, 2022).

[19] As acknowledged by Hammond, the writing made available online had been rejected by the American Psychologist. While it is technically grey literature, it offers valuable insight on the matter from a Brunswikian's perspective.

an author understands to be ecological validity. Similar to lab speech, ecological validity is best viewed as a continuous-, rather than categorical measurement.

6.3 WEIRD samples

Globally, academic research focusses heavily on western, educated, industrialised, rich, democratic – in short *WEIRD* – cultures and communities. As Henrich et al. (2010) report in their landmark article, a majority of experimental research in psychology and related fields (such as linguistics) relies on WEIRD participants (or subjects in Henrich et al.'s terms), some of the least representative of all of humanity. Participants are frequently undergraduate students, since they are easily recruited, constricting generalisability even more. But the problem is not the sampling of easily accessible populations per se, rather the implicit assumptions researchers make about the generalisability of their findings to wider populations based on the sampled populations.

A related problem, is the bias of research towards the study of English-speaking participants. Blasi et al. (2022) warns about an overreliance on languages such as English (or German realistically) possibly impeding progress in cognitive science. Focussing heavily on a particular language (or -family) entrenches assumptions about features that are then thought be universal. Linking back to the production of multimodal feedback signals, studies like Jakobson's (1972) show that there is diversity across languages and cultures that ought to be studied and embraced.

Most of the experimental research consulted for background reading focussed almost exclusively on languages that are spoken by WEIRD speakers. The research consulted analysed the following spoken Indo-European languages: English (UK, US, CA – 12, 25, 3 articles), German (18), Dutch (5), Danish (7), Swedish (5), Norwegian (1), Finnish (1), Bulgarian (1), Italian (7), French (5), Catalan (1), Polish (1) and a minority of Non-Indo-European languages: Japanese (6), Russian (2), Mandarin Chinese (2), and Turkish (1). In addition, single articles investigating German, Dutch, Russian, and Turkish sign language were sampled.

Research focussing on the most heavily researched languages predominantly fails to overtly state which language it is describing. This underscores the privileged status of these languages with regard to funding and ongoing research. When the population sampled is not sufficiently described, implicit claims for generalisability of findings are easily inferred by readers. In researching this thesis, it was a typical occurrence that the limited, non-standardised demographics given about sampled populations had to be consulted just to infer the language under analysis.

From a readers' perspective, an explicit indication of the language (variety) re-

searched in the title, abstract or keywords would be very useful.[20] By doing this it could be easier to find research on a specific language through indexing and also remind readers of the fact that the results presented pertain to a specific language.

6.4 Limitations of the dataset

Following the questionnaire provided in Naumann et al. (2022) as part of MARC (a tool to measure ecological validity), the experimental setup can be described as a mix of naturalistic and controlled aspects. The task in the *Discussion* segment is partially naturalistic – participants are asked to interact casually (which is directly representative of the research interest), but are required to wear eye-tracking glasses while being recorded with cameras and microphones. While some impact of the recording situation is unavoidable – as per Labov's observer's paradox (1972) – the eye-tracking glasses look unremarkable when worn Similarly, the testing site is partially naturalistic, as recordings took place in a seminar room at the university. It should be noted that the recordings had to be gathered in a controlled environment, due to some of the equipment. The sample is one of convenience, consisting of primarily WEIRD undergraduate students. The subset analysed here is small and consists exclusively of female participants. It would be very easy to generalise to participants of all genders, but as previous research showed this would risk being an overgeneralisation.[21]

The ecological validity of the recordings is an improvement over previous research. Modern tools used to record data are less intrusive and impede participants less. Nonetheless, careful generalisations are key, especially from the limited subsample analysed here. This is doubly important when considering the strong apparent individual variation. WEIRD research agendas remain a systemic problem. Still, the greater dataset has been used for a variety of analyses due to its multimodal nature, pointing to what might be described as economic validity.

7 Future directions

Future work with the dataset should expand annotations to more dyads to confirm observations made with the limited data available at the time of writing. Broadening the scope, listener nod behaviour during the *Tangram* segment of the experiment could be annotated. The task-oriented segment is particularly interesting because previous analyses has shown differences in BC behaviour (Spaniol, Forthcoming).

[20] See De Stefani (2021) for a near perfect example of this.

[21] Helweg-Larsen et al. (2004) showed (with a sample of US American undergraduate students) that female listeners nod more than male listeners in dyadic peer interactions. Similar results were found for Japanese speakers (Maynard, 1987, p. 602).

Comparison to the matched ASD dyads could prove interesting when investigating (head-)gestural feedback behaviour in contrast to vocal BCs.

The eye-tracking glasses worn by participants during the experiment contained an accelerometer which recorded relative spatial movement of the glasses at high temporal resolution. The data contains among others measurements along the pitch axis (the axis that is used while nodding), capturing nod movements. This data could, in conjunction with manually annotated nod gestures, be exploited to extract amplitude and cycles of nods. Additionally, employing the existing annotations, a machine learning classifier could be trained to automatically extract nod gestures, amplitude, number of cycles, and possibly other head gestures. Expanding the manual annotations through categorisation of the communicative function of nods could be an interesting avenue for the comparison of feedback across modalities.

Finally, the study, particularly with respects to the modalities recorded and particularly the spontaneous *Discussion* segment, should be replicated using a different sample and more importantly in other languages.

References

Aburumman, Nadine, Gillies, Marco, Ward, Jamie A., & Hamilton, Antonia F. de C. (2022). Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *International Journal of Human-Computer Studies*, *164*. https://doi.org/10.1016/j.ijhcs.2022.102819

Agirrezabal, Manex, Paggio, Patrizia, Navarretta, Costanza, & Jongejan, Bart. (2023). *Multimodal detection and classification of head movements in face-to-face conversations: Exploring models, features and their interaction.* 533376. https://doi.org/10.17617/2.3527200

Akamine, Sho, Dingemanse, Mark, & Özyürek, Aslı. (Forthcoming). *Validating dynamic time warping as a measure of gesture form similarity.*

Allwood, Jens, & Cerrato, Loredana. (2003). A study of gestural feedback expressions. In P. Paggio, K. Jokinen, & A. Jönsso (Eds.), *First nordic symposium on multimodal communication* (pp. 7–22).

Allwood, Jens, Cerrato, Loredana, Jokinen, Kristiina, Navarretta, Costanza, & Paggio, Patrizia. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3-4), 273–287. https://doi.org/10.1007/s10579-007-9061-5

Andonova, Elena, & Taylor, Holly A. (2012). Nodding in dis/agreement: A tale of two cultures. *Cognitive Processing*, *13*(S1), 79–82. https://doi.org/10.1007/s10339-012-0472-x

Baltrusaitis, Tadas, Zadeh, Amir, Lim, Yao Chong, & Morency, Louis-Philippe. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. https://doi.org/10.1109/FG.2018.00019

Bauer, Anastasia, Kuder, Anna, Schulder, Marc, & Schepens, Job. (2024). Phonetic differences between affirmative and feedback head nods in German Sign Language (DGS): A pose estimation study. *PLOS ONE*, *19*(5). https://doi.org/10.1371/journal.pone.0304040

Bavelas, Janet Beavin, Coates, Linda, & Johnson, Trudy. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*(6), 941–952. https://doi.org/10.1037/0022-3514.79.6.941

Bavelas, Janet Beavin, Coates, Linda, & Johnson, Trudy. (2002). Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication*, *52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Bavelas, Janet Beavin, & Gerwing, Jennifer. (2011). The Listener as Addressee in Face-to-Face Dialogue. *International Journal of Listening*, *25*(3), 178–198. https://doi.org/10.1080/10904018.2010.508675

Bavelas, Janet Beavin, Gerwing, Jennifer, Sutton, Chantelle, & Prevost, Danielle. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520. https://doi.org/10.1016/j.jml.2007.02.004

Berry, Ann. (1994). Spanish and American Turn-taking Styles: A Comparative Study. In Lawrence F. Bouton & Yamuna Kachru (Eds.), *Pragmatics and Language Learning* (Vol. 5, pp. 180–190). Division of English as an International Language Intensive English Institute University of Illinois at Urbana-Champaign.

Birdwhistell, Ray L. (1970). *Kinesics and context: Essays on Body Motion Communication*. University of Pennsylvania Press.

Blasi, Damián E., Henrich, Joseph, Adamou, Evangelia, Kemmerer, David, & Majid, Asifa. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Blomsma, Peter, Vaitonyté, Julija, Skantze, Gabriel, & Swerts, Marc. (2024). Backchannel behavior is idiosyncratic. *Language and Cognition*, *16*(4), 1158–1181. https://doi.org/10.1017/langcog.2024.1

Boersma, Paul, & Weenink, David. (2024). *Praat: Doing phonetics by computer*. https://www.praat.org/

Brunswik, Egon. (1956). *Perception and the representative design of psychological experiments*. University of California Press. https://doi.org/10.1525/9780520350519

Butterworth, Brian, & Beattie, Geoffrey. (1978). Gesture and Silence as Indicators of Planning in Speech. In Robin N. Campbell & Philip T. Smith (Eds.), *Recent Advances in the Psychology of Language* (pp. 347–360). Springer US. http://link.springer.com/10.1007/978-1-4684-2532-1_19

Caldognetto, Emanuela, Poggi, Isabella, Cosi, Piero, Cavicchio, Federica, & Merola, G. (2004). *Multimodal score: An ANVIL™ based annotation scheme for multimodal audio-video analysis*.

Cangemi, Francesco. (2015). *Mausmooth*. http://phonetik.phil-fak.uni-koeln.de/fcangemi.html

Cao, Zhe, Hidalgo, Gines, Simon, Tomas, Wei, Shih-En, & Sheikh, Yaser. (2018). *OpenPose: Realtime multi-person 2D pose estimation using part affinity fields*. https://doi.org/10.48550/ARXIV.1812.08008

Castello, Erik, & Gesuato, Sara. (2019). Holding up one's end of the conversation in spoken English: Lexical backchannels in L2 examination discourse. *International Journal of Learner Corpus Research*, *5*(2), 231–252. https://doi.org/10.1075/ijlcr.17020.cas

Cathcart, Nicola, Carletta, Jean, & Klein, Ewan. (2003). A shallow model of backchannel continuers in spoken dialogue. *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - EACL '03*, *1*, 51. https://doi.org/10.3115/1067807.1067816

Cerrato, Loredana. (2007). *Investigating communicative feedback phenomena across languages and modalities* [PhD thesis]. Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.

Cheong, Jin Hyun, Jolly, Eshin, Xie, Tiankang, Byrne, Sophie, Kenney, Matthew, & Chang, Luke J. (2023). Py-Feat: Python Facial Expression Analysis Toolbox. *Affective Science*, *4*(4), 781–796. https://doi.org/10.1007/s42761-023-00191-4

Chui, Kawai. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, *37*(6), 871–887. https://doi.org/10.1016/j.pragma.2004.10.016

Clark, Herbert H., & Fox Tree, Jean E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, Herbert H., & Krych, Meredyth A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81. https://doi.org/10.1016/j.jml.2003.08.004

Cowley, Stephen J. (1998). Of timing, turn-taking, and conversations. *Journal of Psycholinguistic Research*, *27*(5), 541–571. https://doi.org/10.1023/A:1024948912805

Darwin, Charles. (1899). *The Expression of the Emotions in Man and Animals*. D. Appleton; Company.

De Ruiter, J. P., Mitterer, Holger, & Enfield, Nick J. (2006). Projecting the end of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, *82*(3), 515–535. https://www.jstor.org/stable/4490203

De Stefani, Elwys. (2021). Embodied Responses to Questions-in-Progress: Silent Nods as Affirmative Answers. *Discourse Processes*, *58*(4), 353–371. https://doi.org/10.1080/0163853X.2020.1836916

Dideriksen, Christina, Fusaroli, Riccardo, Dingemanse, Mark, & Christiansen, Morten H. (2019). Contextualizing Conversational Strategies: Backchannel, Repair and Linguistic Alignment in Spontaneous and Task-Oriented Conversations. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 261–267). Cognitive Science Society. https://cognitivesciencesociety.org/wp-content/uploads/2019/07/cogsci19_proceedings-8July2019-compressed.pdf

Dittmann, Allen T., & Llewellyn, Lynn G. (1968). Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social*

Psychology, *9*(1), 79–84. https://doi.org/10.1037/h0025722

Drummond, Kent, & Hopper, Robert. (1993). Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency. *Research on Language & Social Interaction*, *26*(2), 157–177. https://doi.org/10.1207/s15327973rlsi2602_3

Duncan, Starkey Jr. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292. https://doi.org/10.1037/h0033031

Duncan, Starkey Jr., & Fiske, Donald W. (1985). *Interaction structure and strategy*. Cambridge University Press; Editions de la maison des sciences de l'homme.

Edelsky, Carole. (1981). Who's got the floor? *Language in Society*, *10*(3), 383–421. https://doi.org/10.1017/s004740450000885x

Ekman, Paul, & Friesen, Wallace V. (1977). *Manual for the Facial Action Coding System*. Consulting Psychologists Press.

Esselink, Lyke D., Oomen, Marloes, & Roelofsen, Floris. (2024). Evaluating Inter-Annotator Agreement for Non-Manual Markers in Sign Languages. *Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages*, 66–76. https://aclanthology.org/2024.signlang-1.7/

Ferré, Gaëlle. (2010). Timing relationships between speech and co-verbal gestures in spontaneous french. *Language Resources and Evaluation Conference (LREC). Workshop on Multimodal Corpora*. https://hal.science/hal-00485797

Ferré, Gaëlle, & Renaudier, Suzanne. (2017). Unimodal and Bimodal Backchannels in Conversational English. *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, 20–30. https://doi.org/10.21437/SemDial.2017-3

Finstad, Mikael. (2021). *LosslessCut*. https://www.mifi.no/losslesscut/

Fusaroli, Riccardo, Tylén, Kristian, Garly, Katrine, Steensig, Jakob, Christiansen, Morten H., & Dingemanse, Mark. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2055–2060). Cognitive Science Society. https://pure.mpg.de/rest/items/item_2449737_7/component/file_2464033/content

Gardner, Rod. (2001). *When Listeners Talk: Response tokens and listener stance*. John Benjamins Publishing Company. http://www.jbe-platform.com/content/books/9789027297426

Goffman, Erving. (1976). Replies and Responses. *Language in Society*, *5*(3), 257–313. https://www.jstor.org/stable/4166887

Gratch, Jonathan, Okhmatovskaia, Anna, Lamothe, Francois, Marsella, Stacy, Morales, Mathieu, Van Der Werf, R. J., & Morency, Louis-Philippe. (2006).

Intelligent virtual agents. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, & Patrick Olivier (Eds.), *Virtual rapport* (Vol. 4133, pp. 14–27). Springer Berlin Heidelberg. http://link.springer.com/10.1007/11821830_2

Gravano, Agustín, & Hirschberg, Julia. (2009). *Backchannel-inviting cues in task-oriented dialogue*. https://doi.org/10.7916/D86W9KDP

Gravano, Agustín, & Hirschberg, Julia. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601–634. https://doi.org/10.1016/j.csl.2010.10.003

Grice, Martine, Baumann, Stefan, & Benzmüller, Ralf. (2005). German Intonation in Autosegmental-Metrical Phonology. In Sun-Ah Jun (Ed.), *Prosodic Typology* (1st ed., pp. 55–83). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0003

Gurion, Tom, Healey, Patrick, & Hough, Julian. (2020, July). Comparing models of speakers' and listeners' head nods. *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. http://semdial.org/anthology/Z20-Gurion_semdial_0013.pdf

Hadar, U., Steiner, T. J., & Clifford Rose, F. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, *9*(4), 214–228. https://doi.org/10.1007/BF00986881

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. Clifford. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, *2*(1-2), 35–46. https://doi.org/10.1016/0167-9457(83)90004-0

Hammond, Kenneth R. (1998). *Ecological validity: Then and now*. https://brunswiksociety.org/wp-content/uploads/2022/06/essay2.pdf

Heldner, Mattias, Hjalmarsson, Anna, & Edlund, Jens. (2013). Backchannel relevance spaces. In Eva Liina Asu-Garcia & Pärtel Lippus (Eds.), *Nordic prosody; proceedings of the XIth conference* (pp. 137–146). Peter Lang. https://doi.org/10.3726/978-3-653-03047-1

Helweg-Larsen, Marie, Cunningham, Stephanie J., Carrico, Amanda, & Pergram, Alison M. (2004). To Nod or Not to Nod: An Observational Study of Nonverbal Communication and Status in Female and Male College Students. *Psychology of Women Quarterly*, *28*(4), 358–361. https://doi.org/10.1111/j.1471-6402.2004.00152.x

Henrich, Joseph, Heine, Steven J., & Norenzayan, Ara. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hjalmarsson, Anna. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, *53*(1), 23–35. https://doi.org/10.1016/j.specom.2010.08.003

Holleman, Gijs A., Hooge, Ignace T. C., Kemner, Chantal, & Hessels, Roy S. (2020). The "Real-World Approach" and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*, 721. https://doi.org/10.3389/fpsyg.2020.00721

Hömke, Paul, Levinson, Stephen C., Emmendorfer, Alexandra K., & Holler, Judith. (2025). Eyebrow movements as signals of communicative problems in human face-to-face interaction. *Royal Society Open Science*, *12*(3), 241632. https://doi.org/10.1098/rsos.241632

Invisible. (n.d.). https://docs.pupil-labs.com/invisible/

Ishi, Carlos Toshinori, Ishiguro, Hiroshi, & Hagita, Norihiro. (2014). Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, *57*, 233–243. https://doi.org/10.1016/j.specom.2013.06.008

Jakobson, Roman. (1972). Motor signs for 'Yes' and 'No'. *Language in Society*, *1*(1), 91–96. https://doi.org/10.1017/s0047404500006564

Janz, Alicia. (2022). *Navigating common ground using feedback in conversation - A phonetic analysis* [Master's thesis, University of Cologne]. https://kups.ub.uni-koeln.de/72407

Jefferson, Gail. (1984). Notes on a systematic deployment of the acknowledgement tokens "yeah" and "mm hm". *Paper in Linguistics*, *17*(2), 197–216. https://doi.org/10.1080/08351818409389201

Karpiński, Maciej, Jarmołowicz-Nowikow, Ewa, & Malisz, Zofia. (2009). Aspects of gestural and prosodic structure of multimodal utterances in polish task-oriented dialogues. *Speech and Language Technology*, *11*, 113–122.

Kassambara, Alboukadel. (2025). *Ggpubr: 'ggplot2' based publication ready plots*. https://rpkgs.datanovia.com/ggpubr/

Kendon, Adam. (1972). Some Relationships Between Body Motion and Speech. In Aron Wolfe Siegman & Benjamin Pope (Eds.), *Studies in Dyadic Communication* (pp. 177–210). Elsevier. https://doi.org/10.1016/B978-0-08-015867-9.50013-7

Kendon, Adam. (1980). Gesticulation and Speech: Two Aspects of the Process of Utterance. In *The Relationship of Verbal and Nonverbal Communication* (pp. 207–228). De Gruyter Mouton. https://doi.org/10.1515/9783110813098.207

Kendon, Adam. (1988). How gestures can become like words. In Fernando Poyatos (Ed.), *Cross-cultural perspectives in nonverbal communication* (pp. 131–141). C. J. Hogrefe.

Kendon, Adam. (2002). Some uses of the head shake. *Gesture*, *2*(2), 147–182. https://doi.org/10.1075/gest.2.2.03ken

Kendrick, Kobin H. (2017). Using Conversation Analysis in the Lab. *Research on Language and Social Interaction*, *50*(1), 1–11. https://doi.org/10.1080/08351813.

2017.1267911

Kita, Sotaro, & Ide, Sachiko. (2007). Nodding, aizuchi, and final particles in Japanese conversation: How conversation reflects the ideology of communication and social relationships. *Journal of Pragmatics*, *39*(7), 1242–1254. https://doi.org/10.1016/j.pragma.2007.02.009

Knight, Dawn. (2009). *A multi-modal corpus approach to the analysis of backchanneling behavior* [PhD thesis, The University of Nottingham]. https://eprints.nottingham.ac.uk/10786/

Koiso, Hanae, Horiuchi, Yasuo, Tutiya, Syun, Ichikawa, Akira, & Den, Yasuharu. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, *41*(3-4), 295–321. https://doi.org/10.1177/002383099804100404

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

Kügler, Frank, Baumann, Stefan, & Röhr, Christine. (2022). Deutsche intonation, modellierung und annotation (DIMA) richtlinien zur prosodischen annotation des deutschen. In Cordula Schwarze & Sven Grawunder (Eds.), *Transkription und annotation gesprochener sprache und multimodaler interaktion: Konzepte, probleme, lösungen* (pp. 23–54). Narr.

Labov, William. (1972). The Study of Language in its Social Context. In J. B. Pride & Janet Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 180–202). Penguin Books.

Levinson, Stephen C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Li, Han Z. (2006). Backchannel Responses as Misleading Feedback in Intercultural Discourse. *Journal of Intercultural Communication Research*, *35*(2), 99–116. https://doi.org/10.1080/17475750600909253

Lorenzen, Janne, Baumann, Stefan, Pelageina, Nadia, & Grice, Martine. (2025). Signalling information status in face-to-face dialogue. *6th Phonetics and Phonology in Europe*. https://agenda.uib.es/120122/section/53648/6th-phonetics-and-phonology-in-europe-pape-2025.html

Lüdecke, Daniel, Patil, Indrajeet, Ben-Shachar, Mattan, Wiernik, Brenton, Waggoner, Philip, & Makowski, Dominique. (2021). See: An r package for visualizing statistical models. *Journal of Open Source Software*, *6*(64), 3393. https://doi.org/10.21105/joss.03393

Lugaresi, Camillo, Tang, Jiuqiang, Nash, Hadon, McClanahan, Chris, Uboweja, Esha, Hays, Michael, Zhang, Fan, Chang, Chuo-Ling, Yong, Ming Guang, Lee,

Juhyun, Chang, Wan-Teh, Hua, Wei, Georg, Manfred, & Grundmann, Matthias. (2019). *MediaPipe: A framework for building perception pipelines*. https://doi.org/10.48550/arXiv.1906.08172

Malisz, Zofia, Włodarczak, Marcin, Buschmeier, Hendrik, Skubisz, Joanna, Kopp, Stefan, & Wagner, Petra. (2016). The ALICO corpus: analysing the active listener. *Language Resources and Evaluation*, *50*(2), 411–442. https://doi.org/10.1007/s10579-016-9355-6

Maynard, Senko K. (1987). Interactional functions of a nonverbal sign Head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, *11*(5), 589–606. https://doi.org/10.1016/0378-2166(87)90181-0

McClave, Evelyn Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878. https://doi.org/10.1016/S0378-2166(99)00079-X

McNeill, David. (1992). *Hand and mind: what gestures reveal about thought*. The University of Chicago Press.

Mereu, Daniela, Cangemi, Francesco, & Grice, Martine. (2024). Backchannels are not always very short utterances. The case of Italian Multi-Unit Backchannels. *Journal of Pragmatics*, *228*, 1–16. https://doi.org/10.1016/j.pragma.2024.05.003

Möking, Eduardo. (2025). *Backchannels in spontaneous and task-oriented speech: Prosody and lexical form* [Master's thesis, University of Cologne]. https://kups.ub.uni-koeln.de/78814

Naumann, Sandra, Byrne, Michelle L., De La Fuente, Alethia, Harrewijn, Anita, Nugiel, Tehila, Rosen, Maya, Van Atteveldt, Nienke, & Matusz, Pawel J. (2022). Assessing the degree of ecological validity of your study: Introducing the multidimensional assessment of research in context (MARC) tool. *Mind, Brain, and Education*, *16*(3), 228–238. https://doi.org/10.1111/mbe.12318

O'Connell, Daniel C., & Kowal, Sabine. (2004). The History of Research on the Filled Pause as Evidence of The Written Language Bias in Linguistics (Linell, 1982). *Journal of Psycholinguistic Research*, *33*(6), 459–474. https://doi.org/10.1007/s10936-004-2666-6

Ogden, Richard. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association*, *43*(3), 299–320. https://doi.org/10.1017/S0025100313000224

Oomen, Marloes, Esselink, Lyke D., de Ronde, Tobias, & Roelofsen, Floris. (2023). *First steps towards a procedure for annotating non-manual markers in sign language*. https://doi.org/10.21942/UVA.25563453.V1

Otsuka, Kazuhiro, & Tsumori, Masahiro. (2020). Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, *8*, 217169–217195. https://doi.org/10.1109/ACCESS.

2020.3041672

Paggio, Patrizia, & Navarretta, Costanza. (2013). Head movements, facial expressions and feedback in conversations: empirical evidence from Danish multimodal data. *Journal on Multimodal User Interfaces*, *7*(1-2), 29–37. https://doi.org/10.1007/s12193-012-0105-9

Pedersen, Thomas Lin. (2025a). *Ggforce: Accelerating 'ggplot2'*. https://ggforce.data-imaginist.com

Pedersen, Thomas Lin. (2025b). *Patchwork: The composer of plots*. https://patchwork.data-imaginist.com

Poggi, Isabella, D'Errico, Francesco, & Vincze, Laura. (2010). Types of Nods. The Polysemy of a social signal. *7th International Conference on Language Resources and Evaluation*.

Poppe, Ronald, Truong, Khiet P., & Heylen, Dirk. (2011). Backchannels: Quantity, type and timing matters. In Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, & Kristinn R. Thórisson (Eds.), *Intelligent virtual agents* (Vol. 6895, pp. 228–239). Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-23974-8_25

Pöppel, Ernst. (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1887–1896. https://doi.org/10.1098/rstb.2009.0015

Pouw, Wim, & Akamine, Sho. (2025). *Using media-pipe for full-body tracking, masking, blurring, and movement tracing*. https://github.com/WimPouw/envisionBOX_modulesWP/tree/main/Mediapipe_Optional_Masking

Pouw, Wim, Yung, Bosco, Shaikh, Sharjeel Ahmed, Trujillo, James, Rueda-Toicen, Antonio, De Melo, Gerard, & Owoyele, Babajide. (2025). *EnvisionHGdetector: A computational framework for co-speech gesture detection, kinematic analysis, and interactive visualization*. https://doi.org/10.31234/osf.io/psg5f_v1

Python 3.11.3. (2023). https://www.python.org

R: A language and environment for statistical computing. (2024). R Foundation for Statistical Computing. https://www.R-project.org/

Rohrer, Patrick Louis, Tütüncübasi, Ulya, Vilà-Giménez, Ingrid, Florit-Pons, Júlia, Esteve-Gibert, Núria Esteve, Ren-Mitchell, A., Shattuck-Hufnagel, Stefanie, & Prieto, Pilar. (2023). *The MultiModal MultiDimensional (M3D) labeling system - version 2*. https://doi.org/10.17605/OSF.IO/ANKDX

RStudio: Integrated development environment for r. (2024). Posit Software, PBC. http://www.posit.co/

Sacks, Harvey, Schegloff, Emanuel A., & Jefferson, Gail. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696. https://doi.org/10.2307/412243

Sbranna, Simona, Möking, Eduardo, Wehrle, Simon, & Grice, Martine. (2022). Backchannelling across Languages: Rate, Lexical Choice and Intonation in L1 Italian, L1 German and L2 German. *Speech Prosody 2022*, 734–738. https://doi.org/10.21437/SpeechProsody.2022-149

Schegloff, Emanuel A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen (Ed.), *Analyzing discourse: text and talk* (pp. 71–93). Georgetown Univ. Pr.

Schlenker, Philippe. (2019). Gestural semantics: Replicating the typology of linguistic inferences with pro- and post-speech gestures. *Natural Language & Linguistic Theory*, *37*(2), 735–784. https://doi.org/10.1007/s11049-018-9414-3

Schmuckler, Mark A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, *2*(4), 419–436. https://doi.org/10.1207/S15327078IN0204_02

Selkirk, Elizabeth O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press.

Shattuck-Hufnagel, Stefanie, & Ren, Ada. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, *9*, 1514. https://doi.org/10.3389/fpsyg.2018.01514

Spaniol, Malin. (Forthcoming). *Face to face: Exploring human social interaction through gaze, speech and neurodiversity* [PhD thesis]. University of Cologne.

Spaniol, Malin, Janz, Alicia, Wehrle, Simon, Vogeley, Kai, & Grice, Martine. (2023). Multimodal signalling: The interplay of oral and visual feedback in conversation. In Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences – ICPhS 2023*. International Phonetic Association. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/49.pdf

Spaniol, Malin, Wehrle, Simon, Janz, Alicia, Vogeley, Kai, & Grice, Martine. (2024). The influence of conversational context on lexical and prosodic aspects of backchannels and gaze behaviour. *Speech Prosody 2024*, 607–611. https://doi.org/10.21437/SpeechProsody.2024-123

Stivers, Tanya. (2008). Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation. *Research on Language & Social Interaction*, *41*(1), 31–57. https://doi.org/10.1080/08351810701691123

Stivers, Tanya, Enfield, N. J., Brown, Penelope, Englert, Christina, Hayashi, Makoto, Heinemann, Trine, Hoymann, Gertie, Rossano, Federico, De Ruiter, Jan Peter, Yoon, Kyung-Eun, & Levinson, Stephen C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Swerts, Marc, & Krahmer, Emiel. (2020). Visual prosody across cultures. In Carlos

Gussenhoven & Aoju Chen (Eds.), *The oxford handbook of language prosody* (pp. 476–485). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.013.45

The GIMP Development Team. (2023). *GNU image manipulation program (GIMP), version 2.10.34 community, free software (license GPLv3)*. https://gimp.org/

Tolins, Jackson, & Fox Tree, Jean E. (2014). Addressee backchannels steer narrative development. *Journal of Pragmatics*, *70*, 152–164. https://doi.org/10.1016/j.pragma.2014.06.006

Tonsen, Marc, Baumann, Chris Kay, & Dierkes, Kai. (n.d.). *A high-level description and performance evaluation of pupil invisible*. https://doi.org/10.48550/arXiv.2009.00508

Türk, Olcay, Lazarov, Stefan, Wang, Yu, Buschmeier, Hendrik, Grimminger, Angela, & Wagner, Petra. (2024). Predictability of Understanding in Explanatory Interactions Based on Multimodal Cues. *ICMI '24: Proceedings of the 26th International Conference on Multimodal Interaction*, 449–458. https://doi.org/10.1145/3678957.3685741

Türk, Olcay, Lazarov, Stefan, Wang, Yu, Buschmeier, Hendrik, Grimminger, Angela, & Wagner, Petra. (2025). Acoustic/Kinematic Detection of False Positive Backchannels of Understanding in Explanations. *LingCologne 2025*.

Wagner, Petra, Malisz, Zofia, & Kopp, Stefan. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Wagner, Petra, Trouvain, Jürgen, & Zimmerer, Frank. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12. https://doi.org/10.1016/j.wocn.2014.11.001

Ward, Nigel, & Tsukahara, Wataru. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, *32*(8), 1177–1207. https://doi.org/10.1016/S0378-2166(99)00109-5

Wehrle, Simon. (2023). *Conversation and intonation in autism: a multi-dimensional analysis*. Language Science Press. http://langsci-press.org/catalog/book/404

Wickham, Hadley. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12). https://doi.org/10.18637/jss.v021.i12

Wickham, Hadley. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer International Publishing.

Wickham, Hadley, Averick, Mara, Bryan, Jennifer, Chang, Winston, McGowan, Lucy, François, Romain, Grolemund, Garrett, Hayes, Alex, Henry, Lionel, Hester, Jim, Kuhn, Max, Pedersen, Thomas, Miller, Evan, Bache, Stephan, Müller, Kirill, Ooms, Jeroen, Robinson, David, Seidel, Dana, Spinu, Vitalie, … Yutani,

Hiroaki. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, Hadley, François, Romain, Henry, Lionel, Müller, Kirill, & Vaughan, Davis. (2025). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org

Wilke, Claus O. (2025). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. https://wilkelab.org/cowplot/

Wittenburg, Peter, Brugman, Hennie, Russel, Albert, Klassmann, Alex, & Sloetjes, Han. (2023). *LREC'06*. http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

Włodarczak, Marcin, Buschmeier, Hendrik, Malisz, Zofia, Kopp, Stefan, & Wagner, Petra. (2012). Listener head gestures and verbal feedback expressions in a distraction task. *Interdisciplinary Workshop on Feedback Behaviors in Dialog*. www.isca-archive.org/fbid_2012/wodarczak12_fbid.pdf

Xu, Yi. (2010). In defense of lab speech. *Journal of Phonetics*, *38*(3), 329–336. https://doi.org/10.1016/j.wocn.2010.04.003

Yngve, Victor H. (1970). On getting a word in edgewise. *Papers from the Sixth Regional Meeting Chicago Linguistic Society*, *6*, 567–578.

Yung, Bosco. (2022). *Nodding pigeon*. https://github.com/bhky/nodding-pigeon