**University of Cologne**
Faculty of Arts and Humanities


Master's thesis

# Backchannels in spontaneous and task-oriented speech: Prosody and lexical form

submitted for the degree Master of Arts

by
**Eduardo Möking**

born in Hamburg, Germany
emoekin1@uni-koeln.de

Two-subject M.A. Linguistics and Phonetics, and Philosophy
(specialization: Phonetics)
Supervisor: Prof. Dr. Martine Grice

Cologne, November 25, 2024

# Table of Contents

Abstract

This study investigates the use of backchannels in spontaneous and task-oriented conversations, focusing on their rate, duration, lexical form, and intonation, as well as exploring contour clustering as a way of revealing the intonational dynamicity of backchannels across conversational conditions. Backchannels, such as "mmhm" or "ja," serve crucial roles in maintaining conversational flow and providing feedback to manage mutual understanding between speakers. The analysis draws on data from dyadic conversations in German, comparing spontaneous discussions with task-oriented interactions using the Tangram game in a setting which allowed for participants to have eye contact. Results reveal differences in backchannel use across conversational conditions regarding all aspects of backchannel communication analyzed: spontaneous speech elicits more frequent backchannels, a higher rate of context-specific signals and a higher proportion of intonational falls. Task-oriented speech is marked by a lower backchannel rate, a more frequent use of non-lexical backchannels such as ‘mmhm’, and more intonational rises regardless of BC type. In addition, explorative results of contour clustering revealed type-specific prosodic dynamicity across conversations, such as a predominance of late-rising contours in ‘mm’ and of early-rising contours in ‘mmhm’, as well as a trend toward dynamic contours being used more often in task-oriented speech. These findings contribute to a deeper understanding of how backchannel intonation is shaped by conversational conditions. This study extends previous research while highlighting the need for further exploration of the prosodic dynamics of backchannels in connection with pragmatic functions and in light of speaker-specific behavior.

# 1 Introduction

Backchannels are short listener responses like ‘mmhm’, ‘yes’ and ‘okay’ that signal understanding and acknowledgement to the current speaker. Studies have highlighted their importance for communicative success, particularly by showing how they contribute to establishing mutual understanding and managing common ground (Dideriksen et al. 2023, Fusaroli et al. 2017). They have been found to fulfill relevant functions in the management of turns between speakers in a conversation (Drummond & Hopper 1993; Goodwin 1986; Hara et al. 2018; Gravano & Hirschberg 2011; Jefferson 1984; Levinson & Torreria 2015; Schegloff 1982), as well as affiliative and social functions (Bavelas et al. 2000; Cutrone 2005, 2011; Gardner 2001), and studies also suggest that they actively shape and even facilitate the exchange of information in conversations (Bangerter & Clark, 2003; Bangerter et al., 2004; Kuhlen & Brennan, 2010; Tolins et al., 2017; Tolins & Fox Tree, 2016; Tolins & Tree, 2014).

Despite the widely-attested significance of backchannels for communicative success, some aspects of backchanneling behavior remain underexplored, not least because studies differ regarding their research objectives and operationalization of backchannels, often complicating the comparability of results. To bridge the gap resulting from methodological and terminological differences, the current thesis aims to present a broad overview of the research on backchannels and its challenges before providing a multi-dimensional analysis of the rate, duration, lexical choice and

intonation of backchannels, as well as how each of these factors is related to different conversational conditions. For this study, data from 14 dyads was recorded in two spontaneous conversations and one task-oriented interaction. Importantly, all pairs were able to have eye contact in all three conversations in order to guarantee comparable conditions.

While previous studies have examined how backchannel frequency varies across different types of conversations (Dideriksen et al., 2023), and how prosodic features shift according to conversational condition, particularly in task-oriented dialogues (Janz, 2022; Spaniol et al. 2024), less attention has been paid to the dynamicity of intonation contours, partly due to methodological constraints. Therefore, the present study explores intonation through the use of a contour clustering application, developed by Kaland (2021), to allow for a more detailed analysis of prosodic variation, both within backchannel types and across different conversations.

# 2 A brief history of backchannels

## 2.1 Backchannels and mutual understanding

Research into backchannels now spans multiple decades, focusing first and foremost on the role they play in dialogue. This role is based on the basic notion of conversation as a collaborative process, in which interlocutors seemingly effortlessly coordinate who speaks when, establish coherence and manage shared knowledge (Schegloff 2006). For conversation to be successful, it is believed that a speaker needs to be assured that her message is not only received by the interlocutor, but also understood by the interlocutor, i.e. the speaker needs to know what the interlocutor knows. This principle of the need for mutual knowledge and its constant coordination has been referred to as *common ground*. According to this idea, interlocutors "try to establish that what has been said has been understood", thereby *grounding* what has been said, i.e. making it part of their common ground (Clark & Brennan 1991: 223). In order for common ground to be maintained, interlocutors rely on the constant exchange of cues signaling understanding or the lack thereof. These cues can take the form of either negative or positive evidence of understanding: Negative evidence (or feedback) are mechanisms such as repair requests, which signal that a message has not

been fully received or understood and that there is need for clarification. Thus, negative feedback indicates that "mutual understanding is potentially compromised and needs to be reestablished or indeed repaired" (Dideriksen et al. 2023: 876). Backchannels, on the other hand, are one example of positive feedback, signaling, instead, that a message has been received and understood. Such backchannel signals may serve the basic function of acknowledging the speaker's turn and thereby contribute to grounding what has been said, that is, establishing and maintaining the common ground (Clark & Brennan 1991).

## 2.2 Backchannels in turn-taking

In addition to serving as a device for establishing and maintaining mutual understanding, backchannels have also been described as playing a role in making communication work in another crucial way: The term 'backchannel' (communication) was originally introduced by Yngve in the context of turn-taking (1970), and indicates the existence of a "back channel", separate from the "front channel" occupied by the primary speaker, through which the listener communicates by sending feedback signals that are not interpreted as interruptions of the main speaker's turn. There have been numerous studies on the system of turn-taking across different fields of research, primarily driven by the question of how interlocutors in a conversation coordinate the smooth transition of turns without any prior planning as to who speaks when, while mostly avoiding speech overlaps and long gaps. While being a phenomenon that is widely taken for granted in everyday conversations, smooth turn transitions involve complex cognitive processes: Considering that gaps between turns have mean durations of around 200 ms, while speech production latencies require between 600 and 1500 ms (Levinson & Torreira 2015), this implies that listeners must be able to simultaneously process the interlocutor's speech and plan their own turn. The rapid transition of turns in natural conversations has therefore led to theories about how this system of turn-taking works and what cues speakers must be able to pick up in order for it to work as it does. Sacks et al. (1974) formulated one of the first theories on the matter, proposing that turn-taking is primarily governed by a set of rules applying at the end of turns, which constitute "turn relevance places" (TRP). The authors note that turn-taking is organized to fulfill two primary conversational goals, namely that only one party talks at a time, and that gaps and

overlaps between turns are minimized when a change of speaker takes place. In order for gaps and overlaps to be minimized, two groups of turn-allocational techniques are followed: 1) the current speaker selects the next speaker (e.g. by means of gaze or by defaulting to the other person; Levinson & Torreira 2015: 11), or 2) if the current speaker does not select the next speaker, the next speaker may self-select. These techniques are integrated in a set of rules to ensure a smooth and organized transition of turns, stipulating how and when the transition takes place after the next speaker is either selected or self-selects (Sacks 2004).

Turns in this system could vary in length and are composed of "turn-constructional units", including sentential, clausal, phrasal, and lexical constructions. Since turns can be long or short, this raises the question of how the listener is able to identify at what point the speaker has concluded or is about to conclude a turn, so as to avoid an interruption or gap. Addressing this issue, Sacks et al. (1974) point to the importance of intonation serving as a cue for turn completion. Looking primarily at such, and other, cues in the context of turn-taking, Duncan (1972) proposed a signals-based approach, in which turn transitions are organized by turn-yielding and turn-maintaining cues, including a variety of different prosodic, gestural and lexical/syntactic cues. According to this model, signals such as a shift in the intonation contour, the end of a hand gesture, a drop in loudness, or the completion of a grammatical clause, among several others, (Duncan & Fiske 1979) serve as signals to the listener that he or she may take the turn. Though the signals-based approach is considered outdated, as it implied that the main responsibility in turn management lay with the main speaker sending signals to the listener rather than it being a collaborative, reciprocal process (Levinson & Torreira 2015), its strength consisted in pointing to important links between turn transitions and prosody as well as visual cues such as gestures and gaze. Since then, a number of studies have looked into the interrelation of turn-taking, prosody and backchannels (Gravano & Hirschberg 2011; Hara et al. 2018; Jefferson 1984; Jurafsky et al. 1998; Koiso et al. 1998; Savino 2014; Sbranna et al. 2022; Schegloff 1982), or investigated turn-taking and backchannels from a multimodal perspective (Bertrand et al. 2007; Harrigan 1979; Neiberg & Gustafson 2011; Oertel et al. 2012; Spaniol et al. 2024).

Backchannels play a subtle yet important role within turn management: Taking a turn and communicating in the back channel are seen as two different paths a listener may take (Yngve 1970), with a backchannel signal thereby serving as a turn-yielding move. Similarly, Duncan and Fiske (1979) note that backchannels are not speaking turns or attempts to take the floor, but rather provide the speaker with information on how the auditor is following and reacting to the speaker's message. Accordingly, the notion of backchannels is that in addition to signaling attention and understanding (Fries 1952; Kendon 1967; Duncan & Fiske 1977), they provide structure to the discourse by giving the speaker a go-ahead sign and conveying the listener's unwillingness to take the floor. Indeed, Schegloff (1982: 78) viewed the latter as the primary property of backchannels, pointing out that they "at best *claim* attention and /or understanding, rather than *showing* it or evidencing it". The author argued that it was unclear why backchannels would even be needed to claim or show attention and understanding, particularly if other manifestations of attention, such as continued gaze direction at the speaker, were present. More important, on the other hand, was their use in exhibiting on the part of the listener the understanding that an extended unit of talk was underway, and, in doing so, granting the speaker the possibility to continue with and complete the turn. Backchannels used in this way can therefore be termed "continuers", as Schegloff (1982: 81) proposes:

> "'Uh huh's, etc. as continuers do not merely claim an understanding without displaying anything of the understanding they claim. The production of talk in a possible turn position which is nothing other than 'uh huh' claims not only 'I understand the state of the talk', but embodies the understanding that extended talk by another is going on by declining to produce a fuller turn in that position."

This understanding of backchannels as continuers underlines the collaborative nature of turn management, with backchannels on the one hand signaling understanding of the state of the speaker's turn, and on the other hand passing over the possibility of the listener to take the floor when it would have been possible to do so. Due to the continuers' function of acknowledging the primary speaker's continued turn and signaling the listener's momentary passivity, Jefferson (1984) referred to such utterances as acknowledgement tokens marking 'passive recipiency'. More importantly, Jefferson noted that not all backchannels acted as continuers. Instead, some backchannels occurred before a turn transition, i.e. as turn-claiming signals on

the part of the listener; an observation that had been made also by Duncan and Fiske (1979). Backchannels used with the intention of taking the floor and shifting from recipiency to speakership were referred to by Jefferson (1984) as marking 'incipient speakership'. It has to be pointed out that the idea of backchannels functioning as floor-claiming signals is debatable. In a sense it contradicts the original notion of '*back* channel' communication as inherently reflecting passive recipiency, and therefore backchannels as turn-yielding signals. From a broader perspective, however, Jefferson's approach can be viewed as an attempt to disentangle the term 'backchannel', which up to that point had been used to denote a quite extensive list of tokens and utterances. Indeed, the differentiation of passive recipiency and incipient speakership revealed that instances of the former category were often realized as 'mm-hm', while items such as 'yeah' were often used to signal incipient speakership (Jefferson 1984), an observation that was also confirmed by Drummond and Hopper (1993). Regarding the question of the categorization of tokens of incipient speakership as backchannels, it can be said that whether or not it is reasonable to do so depends on the aspect of backchannel communication under investigation. For example, the studies by Jefferson (1984) and Drummond and Hopper (1993) showed that not all types of short utterances formerly subsumed under the term 'backchannel' are used with the same turn-taking function in relation to passive recipiency and incipient speakership. This in turn led to studies investigating how backchannels can be differentiated prosodically to indicate the intention to either let the primary speaker continue or claim the turn, finding that tokens marking passive recipiency are often produced with a rising intonation, while those marking incipient speakership tend to be realized with flat or slightly falling intonation (Savino 2010; Sbranna et al. 2022).

## 2.3 Backchannels as active contributions

In sum, there is substantive evidence for the extent to which backchannels as a conversational device are relevant for establishing common ground and coordinating turn-taking. Still, as the term "*passive* recipiency" might suggest, backchannels were for a long time regarded as secondary phenomena before studies began revealing their significance in the active shaping of conversations. While early research did recognize the importance of backchannels in providing a speaker with the means for "monitoring the quality of communication" (Yngve 1970: 568) and understood discourse as a joint

activity (Clark & Brennan 1991; Goodwin 1997; Sacks et al. 1974), the actual effect of listener contributions on communication were rather assumed than explicitly tested. In a study investigating the effect of two kinds of listener responses (backchannels and responses such as gestural and exclamatory reactions) in asymmetrical conversations (a storytelling scenario in which one participant has the role of a speaker and the other of a listener), Bavelas et al. (2000) found that listeners acted as co-narrators through their responses and that narrators told their stories less well when listeners were distracted and produced less responses. In an experiment looking at feedback from a more general perspective, including multimodal cues such as eye gaze and head nods, etc., Clark and Krych (2004) reported that speakers monitor addressees for understanding and *alter* their utterances if necessary, while addressees collaborate by displaying their understanding in the process. Speakers were found to be sensitive to feedback signals as a way of providing and monitoring positive evidence of mutual understanding, whereas a lack of such signals had detrimental effects on communication. Tollins and Fox Tree (2014) found that storytellers reacted in distinct ways to different types of backchannels: *context-generic* backchannels, such as 'yeah' and 'uh huh', which respond to the need to signal understanding and continued attention, thereby serving as grounding displays, and *context-specific* backchannels, such as 'oh wow', which are understood as responses to and commentaries on the content of the preceding utterance. Other studies had made similar distinctions, referring to generic and specific backchannels as *continuers* and *assessments* (Goodwin 1986) or *alignment* and *affiliation* respectively (Stivers 2008). In line with these studies, Tollins and Fox Tree (2014) argue that backchannels are not merely reactive phenomena but actively shape conversations, with speakers continuing narrating by providing discourse-new events after generic backchannels, while taking specific backchannels as cues for confirming previously presented information and thus elaborating on the preceding turn. Similar results, indicating that different kinds and forms of backchannels influence the ways in which speaker narration unfolds, were reported in several other studies (Bangerter & Clark 2003; Kuhlen & Brennan 2010; Tollins & Fox Tree 2016; Tollins et al. 2017). It can be argued that this function of backchannels at least partly explains the observations made by Dideriksen et al. (2023), who found a positive relation between backchannels and performance in Map Task conversations. The authors suggested that this was due the asymmetrical access to information in this type of task, requiring participants to share information more

precisely in order to perform it more successfully. In contrast, less backchannels were produced in a different task with equal access to information, where no relation between backchannel use and performance was found (Dideriksen et al. 2023: 882). Such evidence supports the notion of backchannels contributing to communication in more intricate and proactive ways than by regulating turn-taking and mutual knowledge alone.

Further evidence in support of this observation stems from research on the use of backchannels in cross-cultural contexts, which has shed light on the potentially detrimental effects on communication if the use of backchannels deviates from language-specific norms. Given that backchannels facilitate floor transfer processes, differences in turn-taking systems across languages may affect the smoothness of conversations in cross-linguistic contexts. Differences were found by Berry (1994) in the turn-taking styles of speakers of Spanish and North-American English. Backchannels were more frequent and longer among Spanish speakers, resulting in longer stretches of overlapping speech, compared to English. Interviews with the participants after interactions with speakers of the other culture indicated that the potential for cross-cultural misunderstanding was greater in those areas where the turn-taking styles of Spanish and English differed, especially regarding the quantity of overlapping speech and backchannel behaviors. In another study, backchannels were found to be potentially misleading and cause miscommunication in inter-cultural conversations between Canadian and Chinese speakers, while the opposite was observed when participants were paired with speakers with the same linguistic background as them (Li 2006). The prosodic realization of backchannels was also found to affect cross-linguistic communication. Prosodic patterns of backchannels reportedly differed in task-oriented conversations of speakers of Vietnamese and German, with backchannels in Standard Vietnamese being produced consistently with falling or level pitch contours, while in German they are produced with predominantly rising contours (Ha, Ebner & Grice 2016). The authors suggest that Vietnamese speakers may interpret rising pitch as impolite, whereas for German speakers the same is likely to be the case with falling or flat pitch. In a similar study, confirming the intonation patterns of Vietnamese and German backchannel productions, Wehrle and Grice (2019) found that Vietnamese learners of German, as opposed to native speakers of German, did not prosodically distinguish backchannels from filled pauses, which

may lead to negative character attributions and misunderstandings on behalf of native German listeners. Another study (Cutrone 2005), investigating social rather than purely communicative aspects of backchannel use, found several differences in the use of backchannels in dyadic conversations between Japanese and British participants, as well as evidence for the hypothesis that backchannel conventions that differ between cultures contribute to negative perceptions and stereotyping.

The studies summarized above provide substantial evidence for the idea that the role of backchannels in discourse supersedes the more basic functions of helping to establish common ground and structuring turn-taking. By producing backchannels, listeners actively shape the direction of the dialogue and contribute to the success of the communication in certain task-related contexts, but they also shape the rapport with the interlocutor and the impression they make on them.

## 2.4 Systemizing the term 'backchannel'

One of the major difficulties in trying to obtain a clear and comprehensive overview of the research that has been carried out on backchannels is that the term 'backchannel' has been used in different ways by various authors. One of the reasons for this might be that backchannels have been a source of interest for researchers from various fields of research, from sociology and psychology, to pragmatics, phonetics and second-language acquisition, with researchers in each discipline operationalizing the term in different ways and applying methods of analysis suitable to the requirements of their respective research questions.

In the following I will attempt to systematize the term 'backchannel' based on the literature reviewed so far, in order to provide more terminological clarity for the subsequent section dealing with the main features of backchannels under investigation in this analysis.

Studies on backchannels diverge along the lines of two major questions regarding the use of the term: 1) what counts as a backchannel? And 2) what are the functions of backchannels? These questions are partly interrelated, meaning that how one of them is answered has an influence on how the other one will be answered. If, for instance, negative feedback such as repair requests are regarded as a form of backchannel communication (Duncan 1974), then its analysis will have to focus on

other features (in terms of types of utterances and turns) than if repair requests were considered a different conversational device, with BCs restricted to positive feedback (Dideriksen et al. 2023). Similarly, if backchannels are defined as fulfilling the function of passive recipiency (or continuer) only, and this function is distinguished from affiliative functions (Bavelas et al. 2000; Truong & Heylen 2010), then this will have implications for the types of backchannels that will be encountered (e.g. 'oh wow' as an affiliative signal, 'mmhm' as a typical continuer). Consequently, if backchannels are defined as involving a larger variety of functions, such as passive recipiency, incipient speakership, agreement, assessment, etc. (Mereu et al. 2024), this will result in a larger variety of types and forms being labeled as backchannels, as well as, likely, a greater complexity in their prosodic realization.

Naturally, the definition used will depend on the individual aim of the analysis and one cannot expect complete uniformity with respect to terminology. Nevertheless, I will suggest the following terminological structure when referring to backchannels: Backchannels are those feedback signals uttered from the *back* channel, that is, they are not turns on their own and do not claim the turn (in line with most previous literature). Their functions can be categorized into at least three broader domains (I do not claim this list to be exhaustive): First, adopting the differentiation proposed in previous literature (Goodwin 1986; Bavelas et al. 2000; Tolins & Fox Tree 2014), backchannels can occur with the function of conveying understanding and attention in a 'generic' (or context-generic) way, which includes the common case of *continuers*, or, second, in a 'specific' (context-specific) way, comprising affiliative signals such as assessment, agreement and acceptance. The third function is related to turn-taking, where the term *acknowledgement token* has been proposed (Jefferson 1984; Drummond & Hopper 1993). These tokens can be used with the two sub-functions of signaling passive recipiency (no claim to the turn) or incipient speakership (claim to the turn).

In line with Dideriksen et al. (2023), backchannels will, in the following, be regarded as distinct from other conversational devices such as repairs, which constitute turns in their own right, as they actively solicit a response from the interlocutor, unlike backchannels.

# 3 Dimensions of backchannels: Prosody, rate, type

Multiple aspects of backchannels have been investigated, including their forms, discourse functions and prosodic features, as well as how these features may be interrelated and/or shaped by the conversational condition. The main findings for each of the features relevant for the present analysis will be summarized below. Due to the overall majority of studies that analyze backchannel functions linking these functions to prosody, the aspect of function will not be discussed individually but instead be addressed in the section on prosody. Functional aspects related to backchannel type will be discussed in the corresponding section.

## 3.1 Prosody

The prosodic characteristics of backchannels have been analyzed in a large variety of different languages (Beňuš 2016 for Slovak; Beňuš, Gravano & Hirschberg 2007 for English; Caspers et al. 2000 for Dutch; Edlund, Heldner & Pelcé 2009 for Swedish; Ha, Ebner & Grice 2016 for Vietnamese and German; Heldner, Edlund & Hirschberg 2010 for American English; Jurafsky et al. 1998 for American English; Keevallik 2003 for Estonian; Savino 2014 for Italian; Sbranna et al. 2022 for Italian and German; Zellers 2021 for Ruruuli/Lunyala). A majority of studies report that backchannels are (predominantly) produced with a rising intonation contour (Beňuš 2016; Beňuš, Gravano & Hirschberg 2007; Caspers et al. 2000; Edlund, Heldner & Pelcé 2009; Ha, Ebner & Grice 2016; Heldner, Edlund & Hirschberg 2010; Keevallik 2003), while a few studies report mostly falling (or level) backchannel intonation (Gardner 2001; Jurafsky et al. 1998; Müller 1996; Pipek 2007; Zellers 2021). However, some crucial differences between the studies' methodologies put the results into perspective and point to important underlying correlations: Firstly, all studies reporting backchannels to be realized with rising final pitch have used Map Tasks to elicit data, while the studies reporting falling intonation contours mostly analyzed data from spontaneous (free) conversations. Secondly, of those studies that found backchannels to possess rising intonation, most attribute those rises to backchannels acting as continuers, categorically separating them from other functions such as agreement, assessment or affiliative functions (Beňuš 2016; Beňuš, Gravano & Hirschberg 2007; Keevallik 2003). Consequently, Jurafsky et al. (1998), observing mostly falling intonation in backchannels produced in spontaneous dialogues, note that

continuers constitute an exception to the general pattern, being realized with rising pitch. This is a clear indication that backchannel prosody is on the one hand affected by backchannel function (continuer vs assessment/agreement) and by the conversational condition (task-oriented vs spontaneous speech) on the other.

Indeed, studies taking backchannel functions into account when analyzing prosodic patterns report that pitch contours are linked to individual functions: Savino (2010) and Sbranna et al. (2022) found an overall tendency for acknowledgement tokens marking passive recipiency to be produced with rising intonation and tokens marking incipient speakership to be realized with falling or level pitch contours. Moreover, Sbranna et al. (2022) also reported a stronger relation between backchannel types and intonation, which has been confirmed by Janz (2022) and Spaniol et al. (2024). Ha, Ebner and Grice (2016) found that backchannels with a continuer function were generally produced with rising contours, while backchannels with a closing confirmation function were realized with falling intonation. And Janz (2022), besides reporting that continuers exhibit more intonation rises than other backchannel functions, found an effect of conversational condition, in that more backchannels with rising intonation were produced in the Map Task condition, while spontaneous speech exhibited mostly backchannels with falling intonation. These results suggest a complex interplay between backchannel intonation, function and conversational condition.

Some authors describe backchannels as being produced with the intention of being inconspicuous, so that they are not interpreted by the primary speaker as an interruption or a claim to the turn (Gardner 2001; Heldner, Edlund & Hirschberg 2010; Müller 1996; Zellers 2021). In general, speakers have been shown to match the prosodic features used by previous speakers, particularly in contexts where they align with the interlocutor, i.e. show understanding or empathy (Gorisch 2012; Reed 2006). Backchannels tend to be produced more quietly than the primary speaker's speech (Gardner 2001; Müller 1996), however they have also been reported to match the interlocutor's preceding utterance in terms of pitch and pitch movement (Heldner, Eldund & Hirschberg 2010). It has been suggested that by making the utterance more similar to the interlocutor's speech, a backchannel is rendered unobtrusive (ibid.). Nevertheless, in addition to fulfilling supportive functions, backchannels can also prosodically mark dis-alignment (Gorisch 2012) and bring the interlocutor's turn to an

end (Stivers 2004). Analyses of the German (multi-unit) backchannel ‘ja ja’ have shown that the same lexical item can convey a variety of different meanings, with different consequences for the conversation, depending on the prosodic form: ‘ja ja’ with a pitch peak on the first syllable indicates “I already got it, so stop”, while ‘ja ja’ with a pitch peak on the second syllable conveys to the speaker that there might be a problem in the sense of “hold on, you didn’t get it” (Golato & Fagyal 2008). Barth-Weingarten (2011) reports further prosodically-marked interactional functions of ‘ja ja’, including (re)claiming epistemic priority and agreeing/ acknowledging with reservation.

As the studies summarized above suggest, backchannel functions are inherently related to their prosodic characteristics. This becomes clear from the fact that a listener has to be able to capture the meaning of the utterance on the basis of tokens that, on their own, carry little to no semantic content. Whether ‘ja’ signals the wish for the primary speaker to continue or the intention of the secondary speaker to take the turn will depend on the prosodic features of the utterance, for the lack of other informative characteristics. As previous studies have reported, continuers will most likely bear a rising intonation contour (Beňuš, Gravano & Hirschberg 2007; Caspers et al. 2000; Edlund, Heldner & Pelcé 2009; Keevalik 2003).

### 3.1.1 Duration

Backchannel features other than pitch have been less frequently studied. Nevertheless, some studies report on BC duration and intensity and their potential effects: As for temporal aspects, Young and Lee (2004) report a mean BC duration of 0.39 seconds (s) for American English, also noting that final sonorant lengthening was a common feature of BCs in Korean. Peters and Wong (2015) report a similar mean duration, noting, however, that the context in which backchannels occur (as part of a string of BCs or as standalone BCs) has an effect on their duration, as the median durations of ‘yeah’ and ‘mmhm’ were longer in initial position of a sequence than in final position. Mereu et al. (2024), focusing in particular on the functional and prosodic differences between single and multiple-unit backchannels (MUB), report a mean duration of 0.34 s for single BCs compared to 0.7 s for MUBs. The authors also note that the MUBs with the shortest durations tended to be used with the function of signaling incipient speakership, while MUBs conveying assessment were usually

longer. Non-lexical BCs used as reaction tokens (displaying an affective stance towards the previous turn) were found to have relatively longer durations compared to BCs conveying other functions (Zellers 2021). On the other hand, continuer backchannels were found to be significantly longer than tokens signaling acknowledgement or agreement in Slovak, which was not the case in Standard American English (Beňuš 2016). In monosyllabic BCs, perceived surprise and interest were found to be correlated with longer duration and higher average F0 (Neiberg et al. 2013).

### 3.1.2 Intensity

As far as intensity is concerned, studies have reported mixed results: backchannels tend to be produced quietly to minimize disruption (Zellers 2021), since they are considered 'listener behavior' and not intended to claim a turn. Many non-lexical BCs were reported to be produced perceptually very quiet (Ward 2004). And BCs exhibited low and dropping intensity slopes in a study comparing the prosodic realization of different conversational stances, such as general agreement, rapport-building agreement, reluctance, disagreement and strongly-expressive stances (Freeman 2019). However, the study defined backchannels as "minimal stances" and distinguished them from the other conversational stances, which in principle could be conveyed by backchannels as well. Therefore, it can be assumed that their results refer to instances of generic backchannels only. Agreement tokens in Slovak were found to be produced with higher intensity than continuer backchannels tokens, while the opposite had been observed for Standard American English (Beňuš 2016). Other studies found that the use of intensity can in fact vary in backchannel productions, leading to different effects. For instance, "supportive" listener responses tend to be produced relatively loudly in New Zealand English (Stubbe 1998). And higher intensity was found to be a salient cue to attentiveness in both bisyllabic and monosyllabic BCs (Oertel et al. 2016). Overall, however, results seem to suggest a relationship between BC function and intensity, in that backchannels with a generic function appear to be produced with less saliency in terms of loudness than backchannels supporting other discourse functions.

## 3.2 Rate

Studies reporting on the rate of backchannels show mixed results when it comes to the effect of conversational condition. Fusaroli et al. (2017) and Janz (2022) report a higher frequency of backchannels in task-oriented speech compared to free conversations. Contrary to this pattern, Dideriksen et al. (2023) and Spaniol et al. (2024) found higher backchannel rates in spontaneous dyadic dialogues. In addition to comparing spontaneous and task-oriented speech, Dideriksen et al. (2023) compared two different task conditions, using both a Map Task and an adapted version of the Alien Game (Tylén 2020), a joint decision task. The authors found a higher backchannel rate in the Map Task conversations than in the Alien Game. Explaining this discrepancy, they suggest that the use of backchannels could be more frequent in the Map Task condition due to as asymmetrical sharing of information, with one of the participants performing the role of a *director* (and talking more) and the other participant taking the role of a *matcher* (talking less but providing more verbal feedback). Compared to the Alien Game, where the information sharing is more equal, more backchannels were indeed found to be produced in the Map Task. In spontaneous conversations, the authors note, backchannels might be used as a device to manage shared attention and thus play a different role than in the task-based settings. In sum, Dideriksen et al. (2023) report a higher rate of backchannels in the Map Task than in the Alien Game – which the authors attribute to the different contextual demands in the asymmetric director-matcher context – but an overall higher BC rate in the spontaneous condition compared to the task-based ones. This is not the case in Janz (2022), who found a higher overall rate in the Map Task condition compared to spontaneous speech, in line with Fusaroli et al. (2017). Spaniol et al. (2024) analyzed the interplay between gaze behavior and lexical and prosodic aspects of backchannels in spontaneous and task-oriented speech, although not using a Map Task but instead a different (Tangram) game, similar in structure to the Alien Game. They found a higher BC rate in spontaneous speech than in the task condition (in line with Dideriksen et al. 2023).

Given that these studies come to contradictory results regarding the rate of backchannels produced by speakers in spontaneous and task-oriented speech – with Dideriksen et al. (2023) and Spaniol et al. (2024) reporting a higher BC rate in spontaneous speech compared to task-oriented speech, while Fusaroli et al. (2017) and Janz (2022) observed the opposite – more research is needed to shed light on

backchanneling behavior across different conversational conditions, as well as how task-related contextual requirements might affect the frequency of backchannel utterances across different kinds of tasks.

## 3.3 Type

As was the case for prosodic- and rate-related analyses, studies on backchannels performing analyses on speakers' choices of types follow different methodological approaches and come to different conclusions.

When studies refer to backchannels, they usually name non-lexical examples such as 'uh-huh', 'mmhm', 'mm' and 'yeah' (Bangerter & Clark 2003; Goodwin 1986; Jefferson 1984; Mereu et al. 2024; Poppe et al. 2011; Schegloff 1982; Truong & Heylen 2010, Ward 2004, 2006), and lexical examples such as 'yes', 'okay' and 'all right' (Bangerter & Clark 2003; Beňuš, Gravano, & Hirschberg, 2007; Janz 2022; Mereu et al. 2024; Sbranna et al. 2022, Spaniol et al. 2024; Wehrle 2023). In addition to these more common types, also subsumed under the term 'generic', studies report on the occurrence of tokens such as 'oh wow' and 'ah' (Goodwin 1986; Janz 2022; Tollins & Fox Tree 2014), which have been referred to as 'specific' backchannels, due to their more context-dependent nature of providing a reaction to the content of the interlocutor's utterance. However, most studies differ with regard to which and how many types are analyzed, which conversational functions they are linked to (see section 2.1), and come to different conclusions about what the most preferred types are in the language under investigation:

In an analysis of backchannels in Slovak, the word 'no' (equivalent to *yes*) was found to be the most frequent type, followed by 'mmhm' (Beňuš 2016). Looking at Standard American English, Beňuš, Gravano, and Hirschberg (2007) performed a similar analysis, reporting the non-lexical BC types 'mmhm' and 'uh-huh' to be the most frequent categories, followed by 'okay', while 'yes' was the least produced type. It should be noted that backchannels were strictly defined as continuers and distinguished from tokens signaling acknowledgement and/or agreement, which might explain the marginal use of 'yes'. Another study looking at the type choice of backchannels in American English (following a similar methodology of defining backchannels as continuers only) found that 'uh-huh' was the most common type, followed by 'yeah' (Jurafsky et al. 1998). Unlike in the previous study, however, 'okay'

accounted for only 1% of continuer-backchannels. Pipek (2007) found the non-lexical category to be the most frequent one in American English, followed, however, by 'yes' as the second most frequent, and 'yeah' as the third most frequent category. The author did not restrict the analysis of backchannels to continuers alone, including other functions such as agreement and assessment. This could potentially explain the higher frequency of 'yes' tokens among the analyzed backchannels compared to the studies summarized above.

In an analysis of German and Italian (Sbranna et al. 2022), 'ja' (yes) was found to be the most common choice by German speakers (43%) and 'okay' by Italian speakers (41%), while the non-lexical category accounted for only 25% and 22% of backchannels respectively. Unlike the studies summarized above, which analyzed data from spontaneous conversations, this study used data elicited with the Map Task. In a study looking at backchannel productions by German speakers in two spontaneous and one task-oriented conversation, Janz (2022) found that, overall, 'ja' and 'mmhm' were the most common BC types in all three conversations. However, 'mmhm' was the most preferred type in the task-oriented condition (Map Task) and in one of the spontaneous conversations, while 'ja' was the most frequent type in the other spontaneous conversation. Results from the Map Task setting, therefore, contradict the findings by Sbranna et al. (2022), who found 'ja' to be the most prevalent type, as well as an overall wider margin between this and the non-lexical category. In addition, Janz (2022) reports that the more generic types 'ja', 'okay', 'mmhm' and 'genau' (right) in her analysis were more prevalent in the Map Task condition, while more specific types occurred in the spontaneous conversations, suggesting that the conversational condition has an effect not only on the BC rate but also on the variety of types uttered by the speakers. Choosing a similar approach, Spaniol et al. (2024) looked at backchannels in task-oriented and spontaneous speech, however using a Tangram task instead of a Map Task, which had been shown to affect the speakers' backchanneling behavior differently (Dideriksen et al. 2023). Their results showed 'mmhm' and 'ja' as the most frequent BC types, in line with Janz (2022) and Sbranna et al. (2022). However, while the two categories were almost equally frequent in the first spontaneous conversation and the task condition, a clear tendency toward 'ja' was observed in the second spontaneous conversation. More evidence and a more in-depth analysis are needed to shed light on how the conversational condition influences the speakers' BC type choices.

### 3.3.1 Type and function

A number of studies link different or individual types of backchannels to specific discourse and conversational functions (Bangerter & Clark 2003; Kjellmer 2009; Mereu et al. 2024; Sbranna et al. 2022; Tartory et al. 2024; Ward 2004; Wong & Peters 2007). One of the earliest studies linking individual types to functions was Jefferson (1984), who found that acknowledgment tokens such as 'mmhm' were used as continuers, whereas 'yeah' is produced when a speaker intends to take the turn. Analyzing the most frequent backchannel types in German and Italian and their functional use as markers of passive recipiency and incipient speakership, Sbranna et al. (2022) confirm that the non-lexical type 'mmhm' is used predominantly when a speaker does not intend to take the turn. When signaling the intention to take the turn, German speakers preferred 'ja' (*yes*) and 'okay', while Italian speakers opted for 'okay' in the vast majority of cases (72%). In a study on the pragmatic functions of the prosodic features in non-lexical utterances only, Ward (2004) found that disyllabic variants of non-lexical items (e.g. 'uh-huh', 'mm-hm') were more often used to signal the intention to keep a listening role (continuer function), compared to the monosyllabic versions of the same tokens (e.g. 'uh', 'mm'). In addition, he reported that when non-lexical types such as 'mm' involved more thought, they were produced with longer durations, whereas their shorter counterparts appeared to be more appropriate for lighter topics. Bangerter and Clark (2003), referring to backchannels as *project markers*, analyzed how backchannels are used to navigate projects, reporting that different types are specialized for marking different transitions within a conversation: 'm-hm', 'uh-huh' and 'yeah' are used primarily as "horizontal markers", allowing a current speaker to continue with the action they are performing, while tokens such as 'okay' and 'all right' act as markers of "vertical transitions", such as digressions, that is, for entering into and exiting from subprojects.

Looking primarily into the use of multi-unit backchannels (MUBs), Mereu et al. (2024) report that most single-unit BCs are used as continuers, while MUBs in a majority of cases convey multiple functions simultaneously, usually involving the function of agreement. Moreover, the authors report that MUBs occurred relatively frequently in their data set, accounting for 29% of BC signals. Analyzing different varieties of English, Wong and Peters (2007) reported that single-unit BCs mostly support the speaker holding the floor, while an increased complexity in terms of BC clusters was associated with a shift in importance from supporting the current speaker

to the content of the speech itself. Also, complex BC tokens accounted for 22.6% of backchannels in Australian English, compared to 36.9% in New Zealand English.

### 3.3.2 Type and prosody

A few studies investigated the relation between BC type and prosodic form in German, finding a type-to-prosody mapping that appears to be consistent regardless of function for certain BC types (Janz 2022; Sbranna et al. 2022): the non-lexical type 'mmhm' displays rising intonation, while the lexical type 'genau' (*exactly/right*) is realized with falling intonation irrespective of their use as markers of passive recipiency or incipient speakership. Janz (2022) and Spaniol et al. (2024) analyzed the effect of conversational condition on the prosodic forms of different backchannel types, confirming, on the one hand, the type-to-prosody mapping reported by Sbranna et al. (2022), but also showing that, overall, more instances of backchannels with rising intonation were produced in the task-oriented condition.

### 3.3.3 Type summary

In sum, backchannel types have been shown be related to various conversational functions. Some studies have reported that certain types are linked to specific conversational functions, such as continuers and incipient speakership. Other studies suggest that the choice of backchannels is related to the navigation of topics in a conversation by the speakers, with some types being preferred for ending a topic and moving on to the next and other types being involved in staying with and elaborating on a topic. Moreover, backchannel types have been found to fulfill different functions depending on their form, with syllabification and higher complexity leading to a different functional use. Studies on the prosodic realization of backchannels suggest type-specific intonation patterns ('mmhm' and 'genau') as well as context-dependent intonation contours ('ja' and 'okay'). When it comes to the conversational condition, more specific backchannels were found to be produced in spontaneous speech than in task-oriented speech, while more rising intonation contours and greater pitch excursion were observed in backchannels uttered in task-oriented speech.

These results indicate a complex interrelation between BC type, function, intonation and conversational condition and suggest that an analysis of formal

characteristics of backchannels has to take all of these factors into account in order to reach a better understanding.

# 4 Aims and methodological considerations

The studies summarized above have investigated a wide range of different aspects of backchannels, such as their formal, temporal, functional and intonational properties, as well as how the use of backchannels changes according to the conversational requirements (spontaneous vs task-oriented dialogues). A comparison of these studies' results suggests a strong interrelation between all of these properties. Since some reports indicate, for instance, that the rate of backchannels is influenced by whether the conversation is spontaneous or task-oriented (Dideriksen et al. 2023; Fusaroli et al. 2017; Janz 2022), it is important to bear in mind this effect when looking at the results of studies that have analyzed backchannels elicited in one condition alone. Similarly, it can be argued that backchannel intonation is inherently related to the discourse functions that are ascribed to backchannels: Defining them as continuers only will likely lead to more rising intonation being observed, while backchannels with other functions, such as agreement and assessment, are realized with falling intonation (Ha, Ebner & Grice 2016). Moreover, the conversational condition might also have an influence on backchannel functions, as spontaneous conversations yield more affiliative signals, while task-oriented speech elicits more continuers, due to the different contextual requirements (Dideriksen et al. 2023). Thus, the choice of conversational condition could ultimately skew results for BC intonation towards more rises or falls.

Against the background of these considerations, previous studies should be interpreted with attention to these crucial cross relations. Indeed, methodological differences with regard to the classification of backchannels, data elicitation and analyzed features make it either difficult to compare results, or even lead to contradictory conclusions in some cases. In other cases, even similar approaches produce diverging outcomes: Dideriksen et al. (2023) report higher backchannel frequency in spontaneous speech compared to task-oriented speech, while Fusaroli et al. (2017) report *lower* backchannel frequency in spontaneous speech, despite overall

comparable study designs. What makes matters more complicated is that most studies on backchannels focus either on prosody *or* on the conversational condition (task-oriented vs spontaneous speech), with only a few taking both into account (Janz 2022; Spaniol et al. 2024).

Since the purpose of this study will be to offer a multidimensional study of backchannels' rate, duration, type and intonation in two different conversational conditions (spontaneous and task-oriented speech), special attention will be given to the above-mentioned studies (Dideriksen et al. 2023; Fusaroli et al. 2017; Janz 2022; Spaniol et al. 2024). In order to provide an outline for the motivation of the present study, some crucial limitations of these previous studies will be discussed: Participants in the study by Janz (2022) were not able to see each other in the Map Task and in one of the two spontaneous conversations, as they were separated by a screen making their communication audio only. In addition, the participants were friends (and flat mates). In Dideriksen et al. (2023) and Spaniol et al. (2024) participants were able to see each other in all conversational conditions and were unfamiliar with each other. Based on previous literature, the audiovisual condition can be assumed to have an impact on the subjects' backchanneling behavior, as eye contact, which is only possible in audiovisual, has been reported to elicit verbal and non-verbal BCs (Neiberg & Gustafson 2011). Although this finding could not be confirmed by Spaniol et al. (2024), their analysis of gaze behavior and backchanneling showed a slight trend toward more rising intonation in the absence of mutual gaze and more level intonation during eye contact. Thus, performing an analysis of data elicited under similar conditions as in Dideriksen et al. (2023) and Spaniol et al. (2024) might provide more reliable grounds for comparison.

In the context of previous work on backchannels, open questions remain regarding the relation between BC rate, duration, type, intonation and conversational condition in a setting in which participants are able to have eye contact. This study will therefore aim at providing an in-depth multi-dimensional analysis of these aspects of backchanneling behavior and how they are interrelated, building on and extending previous research on backchannels and creating the foundation for further conversational analyses to investigate the intricate interplay between BC functions and prosody. To do so, this analysis will look at the intonation of backchannels from both

categorical and continuous perspectives following an approach used in previous studies (Janz 2022; Sbranna et al. 2022; Wehrle 2023), which concentrates on the difference between two pitch points taken from each BC to determine the pitch excursion and categorize the intonation as rising, level or falling. Since this intonational analysis inevitably reduces intonation contours to a linear slope, concealing intonational nuances, the present thesis will explore a novel way of visualizing and examining the intonation contours of backchannels using contour clustering (Kaland 2021). This will allow for more detailed representations of BC intonation contours, as well the subtle changes they might undergo when adapted to conversational condition.

# 5 Method

## 5.1 Participants and data collection

Data was analyzed from 14 dyads (28 native speakers of German; 13 female, 15 male), who were matched for age and gender, with the exception of one mixed-gender dyad (22). Each dyad performed a total of 30 minutes of conversation divided into 3 sections of 10 minutes each (two spontaneous conversations and one task-oriented section) and was afterwards asked to answer several questionnaires. Before the recordings, an assistant familiarized the participants with the recording set-up and provided instructions for the following conversation/task. Recording started as soon as the assistant left the room and ended when the assistant re-entered the room after 10 minutes. The first part (*Introduction*) consisted in a spontaneous conversation in which the two interlocutors were asked to get to know one another and given the opportunity to ask questions. After 10 minutes, the assistant entered the room and provided instructions for the second part, the *Tangram* task (Fig. 1). This is a joint-decision task in which the participants have to describe figures presented to them on a sheet of paper and come to a joint conclusion about whether or not the figures they are each seeing match. Speaker A is given a sheet containing four figures with an arrow pointing at one of them, while Speaker B is presented with only one figure. The aim is for Speaker B to describe her figure as precisely as possible, so as to allow Speaker A to evaluate whether it matches the figure indicated by the arrow on her sheet. Once the pair have

reached a joint decision, they show each other their respective sheets to check whether or not they had gotten in right. In the next turn, the roles are switched, so that the speaker who had previously received the sheet with one figure gets the one with four figures, vice versa. The game was played until the 10 minutes were over and the assistant entered the room to provide instructions for the next and final phase. The third conversation (*Discussion*) was another spontaneous conversation, however, the participants were instructed to discuss the Tangram game and talk about whether or not they liked it, whether they had developed a strategy and whether they believed their interlocutor had developed a strategy.

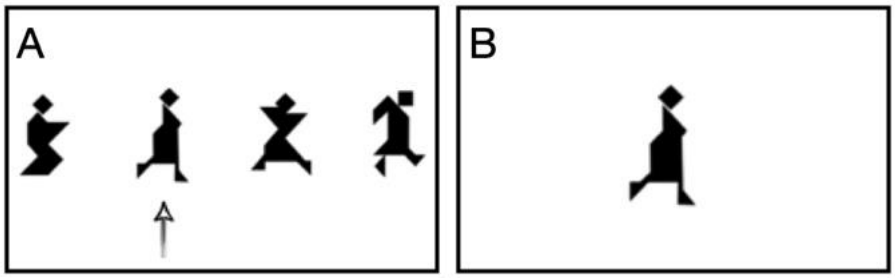


**Fig. 1** Example of Tangram task sheets for each speaker. Sheet A indicates the figure that is to be matched with the one on sheet B.

## 5.2 Annotation

In line with previous studies (Janz 2022; Möking 2021; Sbranna et al. 2022; Wehrle 2023), backchannels were defined as those feedback signals uttered by the interlocutor not currently holding the turn and intended to signal understanding, acknowledgment and/or agreement with the speaker's utterance. Defined in this way, backchannels do not constitute a turn of their own, i.e. an interruption of the primary speaker's turn, and are not followed by a turn of the speaker producing them. Therefore, only those tokens were labeled as backchannels that were not directly adjacent to speech of the speaker uttering the backchannel. A minimum silence interval of 400 ms between the backchannel and a turn from the same speaker was chosen as a threshold. Due to backchannels not being considered a turn and not requesting a turn transition, they have been referred to as markers of 'passive recipiency' (PR), as opposed to feedback signals conveying understanding and acknowledgement but simultaneously initiating a turn, which have been termed markers of "incipient speakership", IS, (Drummond & Hopper 1993). Only feedback signals marking passive recipiency were considered backchannels in this analysis. It should be noted

that ‘passivity’ in this sense is not to be confused with a passive or disengaged stance towards the conversation, but rather refers to the role of a listener that does not intend to take the floor. Backchannels were further distinguished from other very short utterances of similar form that are induced, for instance, by polar questions, whereby the primary speaker passes the floor to the secondary speaker by asking a question, which requires a response.

Backchannels were categorized as non-lexical types ‘mmhm’ and ‘mm’, as well as lexical types ‘ja’ (*yes*), ‘okay’ and ‘genau’ (*exactly/right*). It should be noted that, despite their segmental and perceptual similarity, ‘mmhm’ and ‘mm’ were treated as distinct BC types since previous analyses had shown that they are realized with noticeable prosodic and functional differences (Malisz et al. 2012, Ward 2004). Backchannels were furthermore categorized as *multi-unit backchannels* (MUB) if they were either reduplicated (‘ja ja’, ‘okay okay okay’, etc.), with less than 200ms of silence in between each item, or combined (‘ja okay’, ‘mmhm genau’, etc.). This category is based on but not identical to the definition of MUBs used by Mereu et al. (2024). In the current study, in order to be categorized as MUBs, the tokens had to be reduplicated or combined versions of any of the five main types defined above. Due to the commonness of these forms and their lack of specificity regarding the conveyed meaning, they will also be referred to as *(context-)generic* forms. Any tokens of a different form, such as ‘cool’, ‘ah’, ‘gut’ (*good*), or combined with any such forms (e.g. ‘*ah* okay’) were categorized as ‘other’. As these signals refer to the interlocutor’s utterance in a specific way, that is, conveying a reaction to the content of the speech, they will also be referred to as *(context-)specific* forms. The proposed categorization resulted in the seven BC categories ‘mmhm’, ‘mm’, ‘ja’, ‘okay’, ‘genau’, ‘MUB’, and ‘other’.

The recorded conversations were then annotated in Praat (Boersma & Weenink 2024) with all backchannels being transcribed orthographically on a ‘Token’ tier and then labelled in accordance with one of the seven main categories listed above on a ‘Type’ tier. All figures were created using the programing language R and the software RStudio (R Core Team, 2022; RStudio Team, 2022).

## 5.3 Intonation analysis

Using a Praat script, all backchannels and their annotated labels were extracted from the full audio recording of the conversations for further processing. Each of the tokens' pitch contours was then manually corrected and smoothed using *Mausmooth* (Cangemi 2015), the purpose of which was to correct octave jumps in the F0 trajectory, resulting from creaky voice portions or falsely detected pitch points in unvoiced fricatives. Following the approach used in Janz 2022, Sbranna et al. 2022 and Wehrle et al. 2023, pitch values were sampled at 10% and 90% of token duration. If due to missing pitch information (e.g. as a consequence of voiceless material), no values could be extracted at 10% and/or 90% of token duration, the point of pitch extraction was moved by 10 percentage points to 20% and/or 80% respectively. The furthest possible extraction points for comparison were 40% and 70% of token duration. If still no pitch values were available at either of these points, the pitch information of that token was declared *NA* and it was excluded from the prosodic analysis. In order to determine contour categories, the distance between the two sampled pitch values was calculated in semitones. Positive values of 1 semitone and above were defined as intonational rises, while negative values of –1 semitone and below were defined as falls. Values in between –1/+1 semitones indicated level contours. This method of tracking f0 movement for intonation analyses will be referred to in the following as the method of *linear interpolation*.

While being a simple and efficient method that provides a useful overview of general intonation patterns, this approach of quantifying the pitch slopes has one noticeable limitation: Taking only two points of a trajectory essentially results in a linear representation of intonation contours, which inevitably ignores any dynamicity. Complex pitch movements occurring in between the two measurement points taken, such as fall-rise or rise-fall patterns, would therefore not be detected. Given that the meaning of backchannels is conveyed and perceived to a large extent through intonation (Ha, Ebner & Grice 2016; Ward 2004; Wehrle & Grice 2019), and that contour shapes have been shown to differ in relation to different types and discourse functions (Beňuš, Gravano & Hirschberg 2007; Beňuš 2016; Edlund, Heldner & Pelcé 2009; Freeman 2019), any analysis of how BC intonation is modulated according to conversational conditions needs to be able to pick up on the subtle but potentially

meaningful characteristics that pitch contours might present, beyond whether they are simply rising or falling.

To address this requirement, this thesis explores a contour clustering (CC) approach using an application developed by Kaland (2021) to determine, visualize and analyze contour types with greater attention to detail. This approach offers several advantages, including the minimal need for manual annotation prior to analysis and the applicability to spontaneous speech. Having originally been developed for field-data analyses in the initial stages of prosodic research, where prior descriptions of prosody may be lacking, the CC analysis tool can be expected to be suitable for the explorative nature of this BC intonation analysis.

Prior to the cluster analysis, two subsets of the full data set were created, one for the non-lexical BC types 'mmhm' and 'mm', and one for the lexical type 'ja', in order for the analysis to be performed individually for each of the two subsets. This was done to make the clusters as homogenous as possible in terms of their segments in order to facilitate the comparison of different contours. The analysis was restricted to these BC types due to the relative low frequency of use of the remaining BC types ('okay', 'genau' and 'MUB'). The decision to exclude these types from the intonation analysis was thus motivated by the fact that more data was available for the types 'mmhm', 'mm' and 'ja', allowing for a more reliable and informative analysis. Apart from the intonation analysis, all other parts of the analysis (on rate, duration and type choice) were carried out with the full data set.

The data used in the CC analysis consisted of the audio files (in .wav format) of the individual backchannel tokens and the corresponding annotations (in the form of TextGrid files), containing only their orthographic transcriptions. First, a number of parameters were set to determine the way the f0 measurements needed for the analysis are taken: The lower and upper boundaries of f0 calculation were set to the default minimum of 75 Hz (*f0 floor*) and default maximum of 500 Hz (*f0 ceiling*). Next, the *time-step* setting, which refers to the frame duration for each calculated f0 measurement point, was set to the default of 10 ms. This means that for a token duration of 200 ms (corresponding to the median duration of the BCs analyzed) 20 f0 measures with a window length of 10 ms each were taken. Of these 20 tracked f0 points, the *number of f0 measurement points,* used to represent the contour, was set to 10 measurement points. A smaller number of points for the representation of the

contour was chosen due to the fact that BCs tend to be very short, such that 10 measurements points are enough to capture the essential f0 movements. Choosing too many measurement points relative to the unit of analysis could result in insignificant f0 measures being given too much importance (Kaland 2021), thus mischaracterizing the contours. In addition, the *f0 fit* setting, indicating the minimum probability for an f0 measurement to be accurate, was raised from the default of 0.52 to 0.6, making the algorithm stricter and thereby guaranteeing more accurate f0 candidates at the expense of the quantity of candidates.

In addition to the f0 measurements, the contours were speaker normalized in order to account for differences in the speakers' individual f0 levels and ranges. The method of standardization used in the CC application also preserves register differences within each speakers' range. Furthermore, the backchannels' duration values were taken into account in the cluster analysis, to allow for the contours to be differentiated not only on the basis of their f0 trajectories, but also based on their durations. Clustering was performed with 'complete' linkage and 'Euclidean' distances between the time-series f0 and duration measures.

More detailed information on the analysis parameters can be found in Kaland (2021) and in the application's manual[1]. The analysis was performed using the 2024-08 version of the Contour Clustering application.

## 5.4 Data

A total of 3.196 backchannels were collected and used in the broader analysis. For the intonation analysis, BCs of the 'other' category (450 items) were excluded, as well as further 74 items due to insufficient availability of voiced material. This resulted in a total of 2.672 BCs used in the intonation analysis based on the 10/90 method of f0 tracking, including 328 'mmhm', 394 'mm', 259 'okay', 1.417 'ja', 98 'genau', and 176 'MUB' tokens. Pitch information of 76% of these items were extracted at the ideal 10% and 90% of token duration. The data that had been processed with the linear interpolation method was used in the analysis of the pitch excursion of BC types (continuous intonation analysis) and to determine the distribution of rise, level and falling contours (categorical intonation analysis) across BC types and conversional

[1] The contour clustering application and its manual, as well as other resources, can be retrieved via: https://constantijnkaland.github.io/contourclustering/#download

conditions. Parts of the continuous intonation analysis and all of the categorical intonation analysis were performed on the types ‘mmhm’, ‘mm’ and ‘ja’ only.

The CC analysis, too, was performed exclusively on the BC types ‘mmhm’, ‘mm’ and ‘ja’. From 1.476 ‘ja’ tokens, 588 had been marked for error removal by the CC application, as the extraction of f0 measures in accordance with the settings laid out above had not been possible for these items. In the subset of non-lexical BCs, 249 items of the original 730 were removed. Therefore, after time-series f0 measures were taken for all backchannels in these subsets, 1.396 items were available for the contour clustering (481 ‘mmhm’ and ‘mm’, and 888 ‘ja’ tokens). The relatively high number of excluded items (around 30% of tokens from either category) could be explained by the time-series f0 measurement settings chosen, particularly the number of measurement points in combination with the raised f0-fit value. With a lower number of measurement points and a reduced f0-fit, leading to a less strict algorithm, the proportion of items fit for the analysis could have been increased. However, since f0 contours are not manually inspected for the CC analysis, with the f0 tracking being a fully automated process, the decision was made to prioritize stricter settings and therefore more reliable contours at the cost of a reduced number of available data points.

The results of the BC intonation analysis, using both the linear interpolation method and the CC approach, are reported in the subsequent chapter in the section on intonation (5.4–5.6). A comparison of the two approaches together with an evaluation of their limitations and advantages is provided in the discussion (6.4.2).

# 6 Results

## 6.1 Backchannel rate

The total rate of backchannels per minute of conversation across all dyads and tasks was 7.33 BCs/min. Taking the conversational condition into account, it was found that backchannels were produced at a higher rate in the spontaneous conditions compared to the task-oriented dialogue (Fig. 2): In the first spontaneous conversation (Introduction), speakers produced a mean of 9.00 BCs/min, compared to 4.5 BCs/min

during the Tangram game and 8.48 BCs/min in the second spontaneous conversation (Discussion).

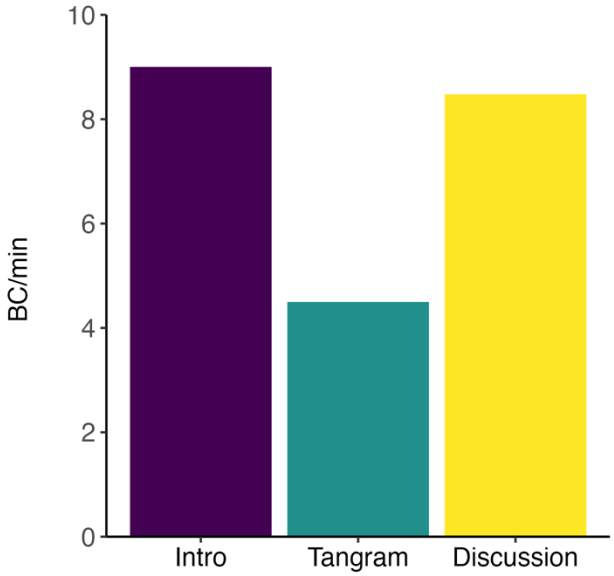


**Fig. 2** Backchannel rate, calculated as BCs per minute of dialogue, by conversational condition.

BC rates by dyad cover a wide range, from an overall mean of 4.23 BCs/min, uttered by dyad 06, to a mean of 12.58 BCs/min produced by dyad 04, suggesting that backchanneling behavior in terms of the quantity of produced feedback signals can vary to a large extent. A look at by-dyad results across conversational conditions (Fig. 3) nonetheless reveals that, despite the difference in rates, all dyads conform the same general pattern of producing more backchannels in spontaneous conversations than in task-oriented speech.

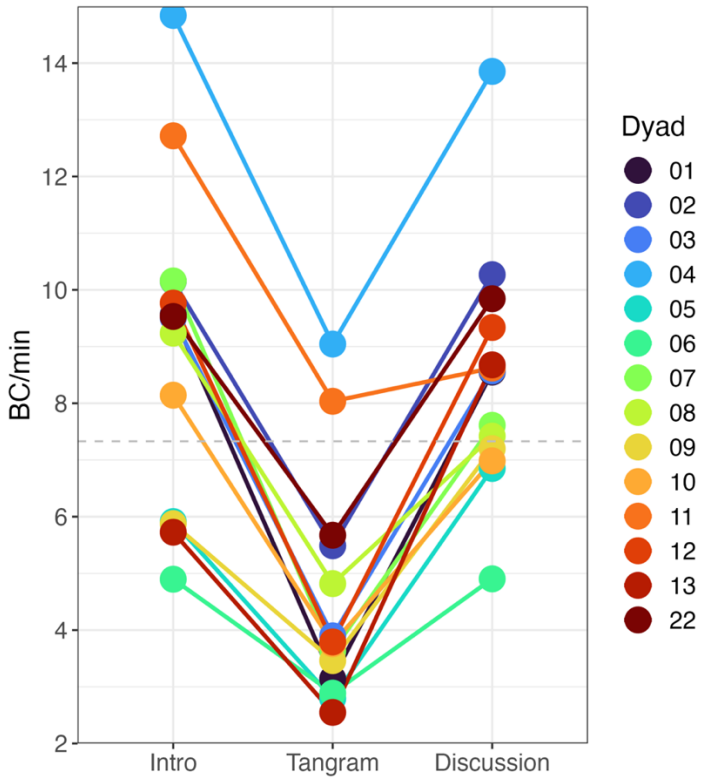


**Fig. 3** BC rate by dyad across the three conversational conditions Introduction, Tangram and Discussion. The horizontal dashed line indicates the overall mean of 7.33 BCs/min.

## 6.2 Backchannel duration

The mean backchannel duration across all speakers, conditions and BC types was 318 ms (SD = 210 ms). By conversational condition, BC duration was longest in the Tangram condition (331 ms; SD = 251) and shortest in the Discussion part (301 ms; SD = 184). The mean BC duration during the first spontaneous conversation (Introduction) was only marginally shorter (M = 328 ms, SD = 210) than during the Tangram game. However, these results should be interpreted with caution, since looking at type-related differences in BC duration appeared to be more informative (Fig. 4) than generalizing over BC types. Hence, the differences in the overall BC duration across tasks are likely an artifact of certain BC types with longer durations being used more often in the task condition, rather than backchannels in general being produced longer.

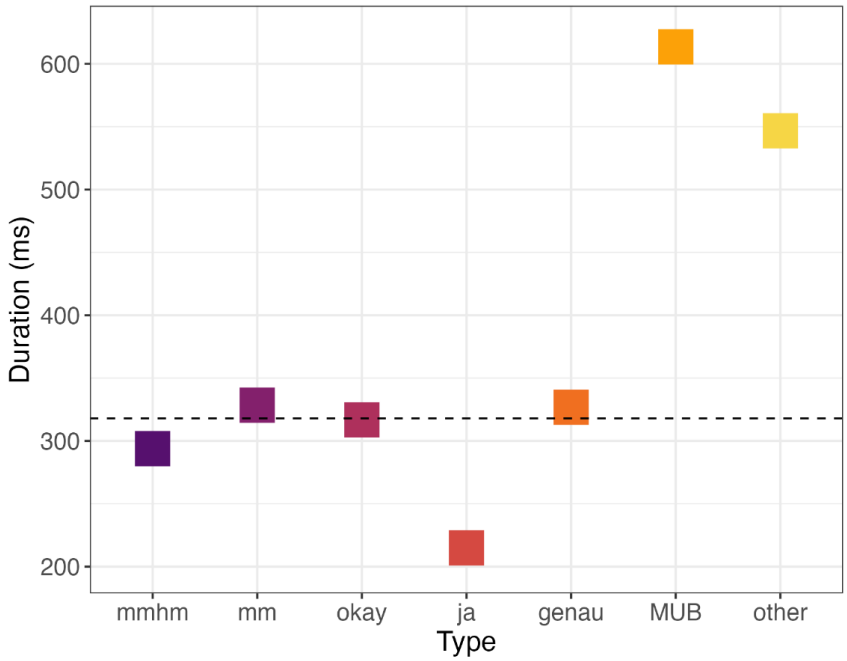


**Fig. 4** Mean backchannel durations according to BC types. The dashed horizontal line indicates the overall mean duration of 318 ms.

The shortest BC type in the analyzed dataset was found to be 'ja'. Items of this category had a mean duration of 215 ms (SD = 86). Conversely, the category of multi-unit backchannels (MUB) was shown to be produced with the longest overall mean duration of 604 ms (SD = 277 ms). The category of 'other' backchannels showed the second-longest mean duration (M = 554 ms, SD = 344). This category is composed context-specific backchannels and includes single- as well as multi-unit types. Multi-unit types account for around half of the types in the category, which might explain the longer mean duration compared to the generic single-unit BC types, with mean durations of 294 ms ('mmhm'), 329ms ('mm'), 317 ms ('okay'), and 329 ms ('genau'). Noticeably, the non-lexical type 'mm' showed a longer mean duration than its non-lexical counterpart 'mmhm'. In terms of duration, it is therefore more similar to the group of disyllabic backchannels ('okay', and 'genau') and distinct from the monosyllabic type 'ja'. Taking context-specific BCs from the 'other' category into account, 'mm' is in line with the non-lexical monosyllabic type 'ah', which showed a mean duration of 353 ms (SD = 256).

Due to the large discrepancy in duration between the generic single-unit BCs on the one hand and MUBs as well as 'other' types on the other hand, when duration was calculated individually for the single-unit BCs 'mmhm', 'mm', 'ja', 'okay' and 'genau', the resulting mean duration of 257 ms (SD = 99) was naturally shorter than the mean duration of 318 ms (SD = 210 ms) across all types. It should be noted that

the single-unit items from the 'other' category, i.e. context-specific BC types, showed longer mean durations than the generic single-unit backchannels: Taken together, the mean duration of *specific* single-unit BCs is 428 ms (SD = 320), and thus considerably longer (66.3%) than the mean duration of the *generic* single-unit types. Table 1 provides a list of the 10 most frequent specific types and their individual mean durations.

Context-specific *multi*-unit BCs are also longer in duration compared to the generic MUBs, however by a smaller margin: the mean duration of specific multi-unit BCs was 688 ms (SD = 317) and therefore 13.9% longer than generic MUBs with a mean duration of 604 ms (SD = 277).

| Type | Mean duration (ms) | Count |
|---|---|---|
| **ah** | 353 | 40 |
| **ach so** | 610 | 24 |
| **cool** | 345 | 19 |
| **stimmt** | 465 | 19 |
| **natürlich** | 607 | 18 |
| **das stimmt** | 518 | 16 |
| **gut** | 286 | 13 |
| **krass** | 393 | 11 |
| **nice** | 417 | 11 |
| **voll** | 328 | 9 |

**Table 1** List of the 10 most frequent *context-specific* single-unit BC types from the 'other' category, including their mean durations (middle column) and the number of occurrences (*count* column) across all conversations.

In sum, 'ja' is the shortest generic BC type in the analyzed dataset, thus sticking out from the remaining single-unit BC types, which show overall similar durations to one another. Non-lexical type 'mm' is more similar to disyllabic BC types in terms of duration than to 'ja'. As it was to be expected, MUBs showed the longest mean durations, followed by BCs in the 'other' category, which includes single- and multi-unit versions of context-specific backchannels. Single-unit *specific* backchannels (Table 1) have shown to be produced with overall longer durations than single-unit *generic* backchannels (Fig. 4).

### 6.2.1 Duration across conversations

Looking at the duration of BC types across conversational conditions (Fig. 5) showed that the single-unit BC types were produced with relatively stable durations across conditions. Among the single-unit types, the largest cross-condition difference in duration was observed for ‘genau’, whose mean duration was 53 ms (15.3%) shorter in the Tangram task (M = 290 ms, SD = 47) compared to the Discussion (M = 342 ms, SD = 82). Utterances of the non-lexical type ‘mmhm’ were slightly longer (by 36 ms or 13.3%) in the Tangram condition (M = 310 ms, SD = 95) compared to the Introduction (M = 274 ms, SD = 74). Overall, the mean duration of the generic single-unit BCs is higher in the task-oriented condition, albeit by a slight margin: Introduction (M = 259 ms, SD = 98), Tangram (M = 277 ms, SD = 109), Discussion (M = 244 ms, SD = 93). The difference in the mean BC duration across types between task-oriented and spontaneous speech is therefore 25 ms, which is equivalent to a 10.0% increase in the Tangram condition.

Cross-condition changes in duration were larger for MUBs. The largest difference was found between the first spontaneous and the task-oriented conversation, were the mean duration of MUBs was 129 ms longer than in the Introduction. In total, the mean duration of MUBs was 115 ms longer in the Tangram task (M = 696 ms, SD = 278) compared to the spontaneous conditions taken together (M = 581 ms), which is a 19.8% durational increase in the task-oriented speech condition.

Despite MUBs showing a trend toward longer durations in the task-oriented condition compared to spontaneous speech, individual BC types in that category show the opposite pattern of shorter mean durations in task-oriented speech than in the spontaneous conditions: For instance, ‘ja genau’ (Tangram: 479 ms < Introduction: 537 ms, Discussion: 524 ms); and ‘ja okay’ (Tangram: 456 ms < Introduction: 529 ms, Discussion: 492 ms) showed a reversed trend, with 9.8% shorter (‘ja genau’) and 11.9% shorter (‘ja okay’) mean durations in task-oriented speech compared to spontaneous speech.

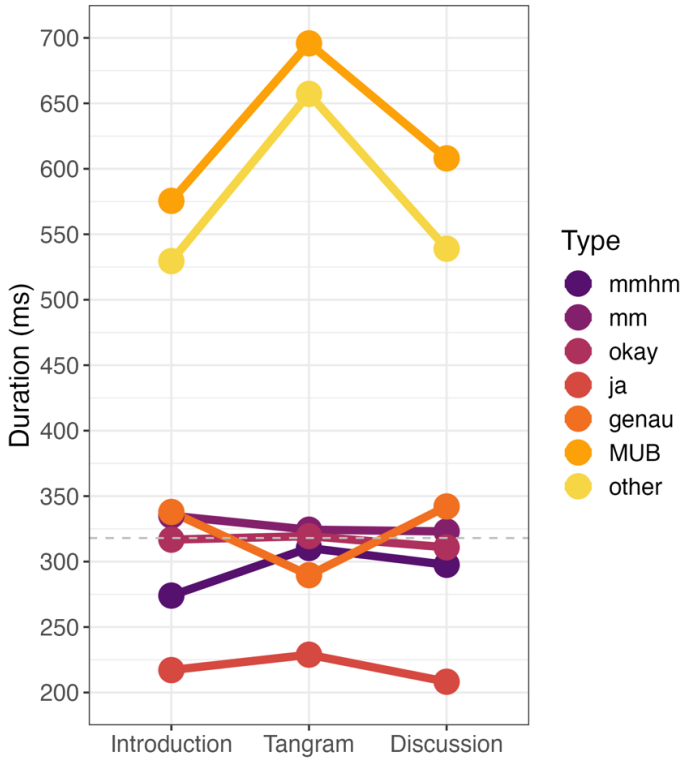


**Fig. 5** Mean BC duration (in ms) by BC type across conversational conditions.

The category of context-specific backchannels ('other'), which includes both single- and multi-unit specific types, conformed to the pattern of longer durations in task-oriented speech. The mean duration of backchannels in this category was 137 ms longer in the task-oriented condition (M = 677 ms, SD = 618) compared to the spontaneous speech conditions taken together (M = 540), which is equivalent to an increase of 25.3%. Interestingly, for single- and multi-unit types within the 'other' category a stronger increase in durations in task-oriented speech was observed for single-unit types than for multi-unit ones, which was not the case for the generic BC types, where MUBs showed a stronger increase in durations than single-unit items (Fig. 5b). In comparison, single-unit backchannels from the context-*specific* category showed a stronger increase in the mean duration in the task condition: Introduction (M = 407 ms, SD = 219), Tangram (M = 558 ms, SD = 826), Discussion (M = 425 ms, SD = 175). Their mean duration in task-oriented speech was thus 34.0% longer in relation to spontaneous speech (for generic single-unit BCs the difference was 10.0%). Multi-unit specific types showed a similar pattern in terms of longer durations in the Tangram condition: Introduction (M = 667 ms, SD = 324), Tangram (M = 774 ms, SD = 383), Discussion (M = 693 ms, SD = 278). However, for these types the mean duration in

task-oriented speech was 13.9% longer compared to spontaneous speech (while for generic MUBs it was 19.8%).

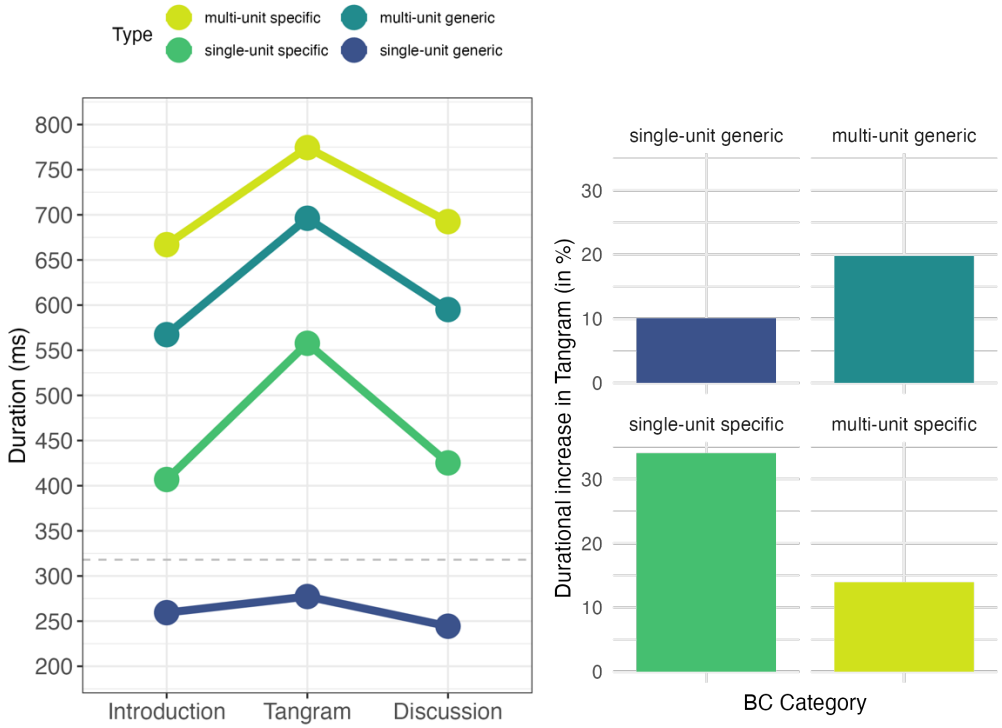


**Fig. 5b** On the left: Mean duration across conditions by single- and multi-unit BC categories. On the right: Durational increase in Tangram compared to spontaneous speech (in %) by BC category.

### 6.2.2 Summary: Duration

Overall, results for BC duration across conversational conditions showed small differences for the group of generic single-unit BCs, with a 10% longer mean duration in task-oriented compared to spontaneous speech. The conversational condition appeared to have a greater effect on the duration of generic BCs when they form complex units (MUBs), as their mean duration was 19.8% longer in task-oriented than in spontaneous speech. The 'other' category, which includes single- and multi-unit versions of context-specific backchannels showed the longest durations overall in terms of absolute temporal values. However, specific single-unit BCs showed greater cross-condition changes than multi-unit items in this category, with a 34% longer mean duration in the Tangram condition for single-unit BCs and a 13.9% longer mean duration for multi-unit BCs. The opposite pattern was the case for generic BCs, where MUBs showed a greater increase in duration than single-unit items in the task condition. Therefore, the Tangram task appears to affect generic and specific BCs differently as far as their duration is concerned.

Aggregated results should be interpreted with caution, as individual BC types within each category (single-unit and multi-unit) have shown patterns that oppose the group patterns. Furthermore, the lower overall number of specific BCs ('other') compared to the main generic ones should be taken into consideration.

## 6.3 Backchannel type

Among the five main BC types under investigation as well as the broader categories MUB and 'other', the most preferred BC type across all speakers and conditions was "ja", which made up for 46.2% of all uttered backchannels. Non-lexical BCs were the second most preferred category, making up for 22.6% in total (10.3% "mmhm", 12.4% "mm"). Instances of "okay" accounted for 8.3%, while the least frequent single-unit generic type was "genau", representing 3.1% of all backchannel utterances. MUBs and context-specific ('other') backchannels accounted for 5.5% (176 utterances) and 14.1% (444 utterances) respectively.

Within the category of multi-unit backchannels, there were 38 individual types, i.e. combinations and/or repetitions of the five main generic forms summarized above, of which 23 (60.5%) were unique forms that occurred only once. The most frequent MUB was 'ja ja', accounting for 37.5% of all MUBs. This is at least three times as much as the second-most frequent type 'ja genau' (12.5%) and the third-most frequent type 'ja okay' (11.4%). Table 2 shows a list of the 15 most frequent MUB types. Most items in this category are either repetitions of single generic types ('ja ja', 'okay okay', etc.) or combinations of two different generic types, often involving repetitions of one of them ('ja ja genau', 'ja okay okay'). Combinations of more than two different types were rare, occurring only twice in this dataset: 'ja genau okay' and 'ja ja okay mm'. The longest concatenations of single types included strings of 4 to 7 'ja's and strings of 4 to 8 'okay's, all of which were produced by the same speaker (speaker 08 of dyad 04). Notably, 89.2% of all MUB utterances included the type 'ja'.

The category of 'other' BCs, which comprises specific types of single- and multi-unit form, has shown a wider range of individual types and a more even distribution, with less distinct preferences for specific types across all speakers and conditions. In total, 444 specific BCs were uttered, including 226 single-unit and 218 multi-unit items. Among those utterances, 150 individual types were found, of which 101 (67.3%) occurred only once, with the remaining 49 types having been uttered at

least twice in the analyzed dataset. A list of the 15 most frequent types from this category are listed in table 2 (right half of the table). The most frequent type was the single-unit BC 'ah', which had been produced 40 times and accounted for 8.9% of all 'other' BCs. The distribution of *specific* types is more even compared to the *generic* types, where there is a clearer preference for one individual type, with 'ja' accounting for almost half of all BCs, and 'ja ja' being the most preferred by a wide margin in the group of MUBs.

Overall, the group of multi-unit specific BCs is composed of a larger variety of lexical forms (118) than the category of multi-unit generic BCs (38). A full table of all MUB and 'other' types can be found in the Appendix.

| **MUB types** | | | **'Other' types (context-specific)** | | |
|---|---|---|---|---|---|
| **Type** | **Percentage %** | **Count** | **Type** | **Percentage %** | **Count** |
| **ja ja** | 37.5 | 66 | **ah** | 8.9 | 40 |
| **ja genau** | 12.5 | 22 | **ach so;<br>ah okay** | 5.4 | 24 |
| **ja okay** | 11.4 | 20 | **cool;<br>stimmt** | 4.3 | 19 |
| **ja ja ja** | 7.4 | 13 | **natürlich** | 4.1 | 18 |
| **ja ja ja ja** | 3.4 | 6 | **das stimmt** | 3.6 | 16 |
| **ja ja genau** | 2.8 | 5 | **gut** | 2.9 | 13 |
| **mm ja;<br>mmhm ja;<br>ok ok ok ok ok** | 1.7 | 3 | **krass;<br>nice** | 2.5 | 11 |
| **genau ja;<br>ja genau ja;<br>ok ok ok ok;<br>okay ja;<br>okay okay;<br>okay okay okay** | 1.1 | 2 | **ah ja** | 2.3 | 10 |
| | | | **ja ja klar;<br>voll** | 2.0 | 9 |
| | | | **ach cool;<br>ja stimmt** | 1.8 | 8 |

**Table 2** List of the 15 most frequent MUB types (left side) and *context-specific* types from the 'other' category (right side), including their percentages and the number of occurrences (*count* column) across all speakers and conversations.

### 6.3.1 Backchannel type across conversations

Looking at the proportionate use of types across the three conversational phases (Fig. 6) revealed the following patterns: "ja" accounted for a slightly lower rate in the Tangram condition (37.3%) than in the Introduction (42.3%) and was used the most in the Discussion (55.3%). At the same time, non-lexical backchannels ('mmhm' and

'mm') were used overall more frequently in the task-based condition (30.7%) compared to the spontaneous speech conditions (Introduction: 22.3%; Discussion 19%). *Within* the category of non-lexical BCs, however, there is a preference for "mm" to be uttered more frequently in spontaneous speech (Introduction: 13%, Discussion 13.1%) than during the Tangram task (9.8%), while "mmhm", on the other hand, is produced more than twice as often in task-oriented speech (20.9%) compared to spontaneous speech (Introduction: 9.3%, Discussion: 5.9%).

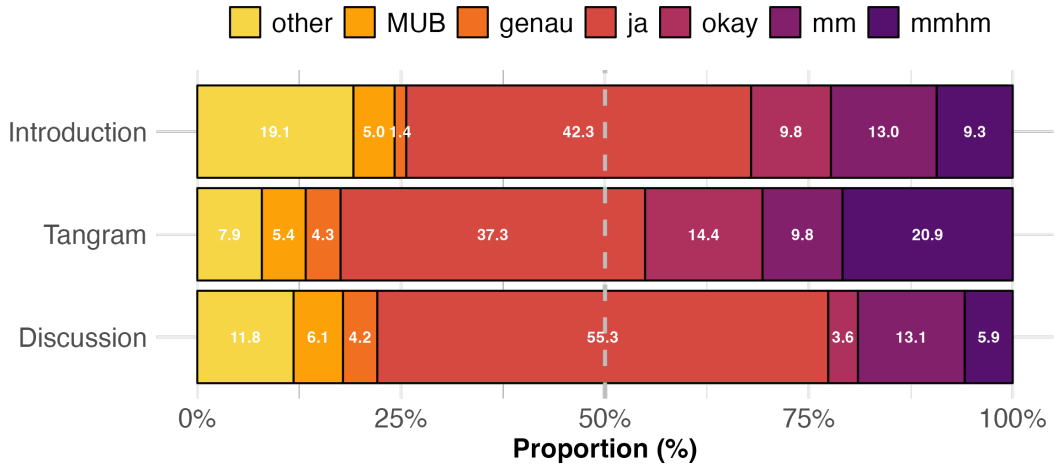


**Fig. 6** Percentage of BC types (color-coded) produced across the three conversational conditions.

The lexical BC "okay" is produced most frequently in task-oriented speech (Tangram: 14.4%), followed by the first spontaneous conversation (Introduction: 9.8%), and the least in the second spontaneous condition (Discussion: 3.6%). The least frequent generic BC 'genau' accounted for 1.4% in the Introduction compared to 4.3% in the Tangram condition and 4.2% in the Discussion.

In sum, context-specific BCs as well as 'ja' and 'mm' were produced more often in spontaneous speech than in the task interactions, while the opposite is the case for 'mmhm'.

### 6.3.2 MUB types

In contrast to the types described above, MUBs were produced at a relatively low but stable rate across all conversational conditions: (Introduction: 5%, Tangram; 5.4%, Discussion: 6.1%). The most frequent type within the category, 'ja ja', was used almost exclusively in spontaneous speech (93.9%, 62 utterances) and only a few times

in task-oriented speech (6.1%, 4 utterances). For the overall second most frequent type ‘ja genau’, on the other hand, no such preference was found across the three conversations (Introduction: 31.8%, Tangram: 31.8%, Discussion 36.4%). However, with only 22 utterances in total, cross-condition results for ‘ja genau’ should be interpreted with caution.

### 6.3.3 Context-specific backchannel types

BCs in the “other” category accounted for 7.9% in the Tangram task, which is less than in the spontaneous conditions (Introduction: 19.1%, Discussion: 11.8%). The non-lexical specific type ‘ah’, which had been shown to be the most frequent single-unit BC across conditions, has been used mostly in spontaneous speech, where 87.5% of all ‘ah’ utterances occurred. Within the two spontaneous conversations it was produced more often in the Introduction (60%, 24 utterances) than in the Discussion (27.5%, 11 utterances). In task-oriented speech, the least instances of ‘ah’ were produced (12.5%, 5 utterances). However, the proportionate use of ‘ah’ in relation to other specific types remained stable across all three conversations. The multi-unit item ‘ah okay’ was preferred in spontaneous speech (Introduction 75%, Discussion 20.8%; 23 utterances in total), with only one recorded instance during the Tangram task. The same distribution was observed for ‘ach so’ (Introduction 75%, Discussion 20.8%, Tangram 4.2%), which together with ‘ah okay’ was the second most frequently used context-specific BC type (see Table 2).

Other overall less frequently used specific BC types were also not distributed equally across conditions: items such as “cool”, “ach krass/krass” and “nice”, which express an evaluation of and a stance towards the primary speaker’s utterance were used almost exclusively in spontaneous speech, with the only exception being one utterance of “nice” in the Tangram condition. In task-oriented speech, on the other hand, more neutral and less colloquial items were preferred, such as ‘gut’ (good), which was the most frequent type in this condition. This item was also used in a variety of other less frequent combined forms, e.g. ‘ah gut’, ‘okay gut’ and ‘sehr gut’ (very good).

In sum, a smaller variety of specific BC types was observed in task-oriented compared to spontaneous speech (Fig. 6b), with various types occurring predominantly or exclusively in spontaneous speech. However, a low overall number of specific BCs in the Tangram condition (49 items, compared to 250 in the

Introduction and 145 in the Discussion) makes it difficult to reliably analyze individual types across conditions and pinpoint their distributions. Regardless, the results indicate on the one hand a trend towards a less varied use of specific BCs in task interactions, and on the other hand a trend towards different types of specific BCs being used in each of the spontaneous conversations.

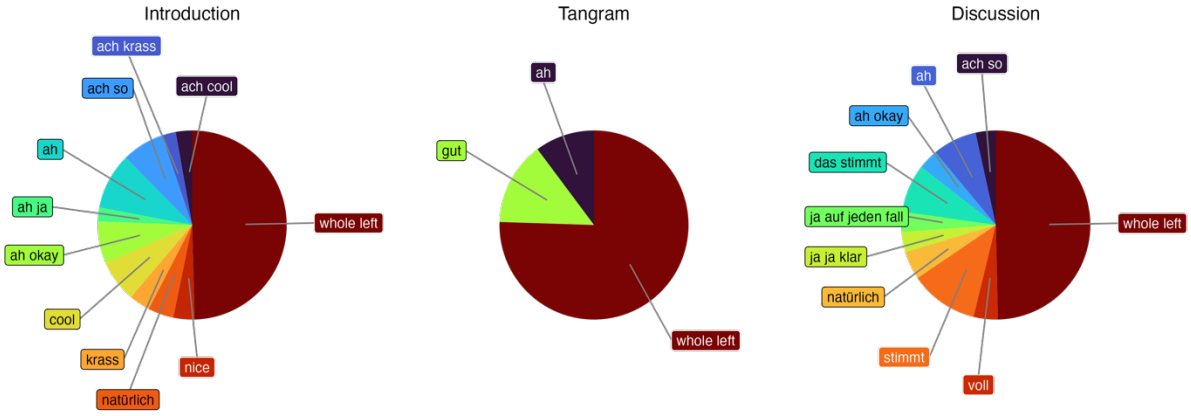


**Fig. 6b** Pie plots showing the proportionate use of context-specific BC types across the three conditions. "Whole left" refers to remaining types that occurred less than five times in each condition.

### 6.3.4 Backchannel type – by dyad

Results for dyad-specific type choice (Fig. 7) suggest that individual variability is an important factor to be taken into consideration when analyzing backchanneling behavior. Dyad 01 produced the highest proportion of "other" BCs in the *Introduction* (36.5%), frequently using types such as "ah", "krass" and "chillig". This dyad also produced the lowest percentage of non-lexical BCs in the Tangram condition (6.25% "mmhm", no "mm"), opposing the overall mean of higher rates of non-lexical BCs in task-oriented speech and instead opting for lexical types (56.3% "ja", 21.9% "okay"). Dyad 06, on the other hand, consistently produced the highest proportions of non-lexical BCs throughout all three conversations (Introduction: 54.4%, Tangram: 58.6%, Discussion: 65.2%). Their pattern stands out in particular in the *Discussion*, where their most preferred type was "mm" (63%), while for the majority of dyads "ja" was the most frequent BC type in the final spontaneous conversation. While otherwise conforming to the mean type choice pattern, dyad 04 produced higher-than-average rates of MUBs in all three conditions: Introduction (10.5%), Tangram (15.6%), Discussion (21.1%). One particularly noticeable idiosyncratic behavior from this dyad

was one of the two speakers (speaker 8) producing the longest reduplications, mostly of types "ja" and "okay", with concatenations of up to seven instances of "ja" and six instances of "okay". Of the 10 longest BCs produced in the analyzed dataset, 8 were produced by this speaker.

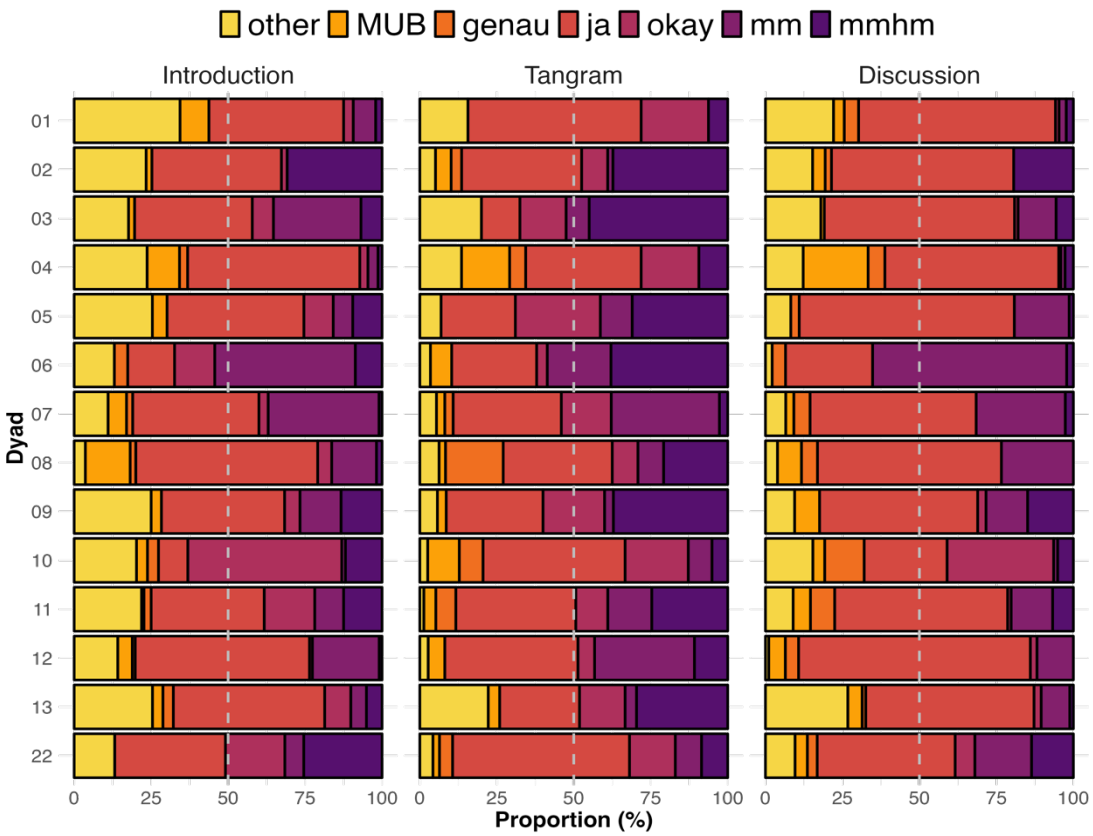


**Fig. 7** Percentage of BC types (color-coded) produced by dyad across the three conversational conditions.

## 6.4 Intonation – Continuous measurements

Aggregated continuous results for pitch excursion in all generic backchannels (Fig. 8), i.e. excluding the category 'other', show that for all conversational conditions there is a mode around flat or slightly falling pitch movement, approximately at the –1 ST mark. Only the Tangram condition additionally features more rising values with greater pitch excursion than the other two conditions, indicated by a higher mean pitch excursion of M = 1.57 ST (SD = 5.22), compared to –0.46 (SD = 3.05) in the Introduction and –0.83 ST (SD = 2.54) in the Discussion.

As stated above, a semitone difference of ± 1 between the two pitch points sampled from each token is defined as a level intonation in this analysis. Even though the means show a trend toward rising intonation in the Tangram task and flat or slightly

negative pitch excursion in the spontaneous discussion, there is still a substantial number of BCs with the opposite pitch excursion pattern in either condition. Therefore, the mean values should be interpreted as describing a rough trend.

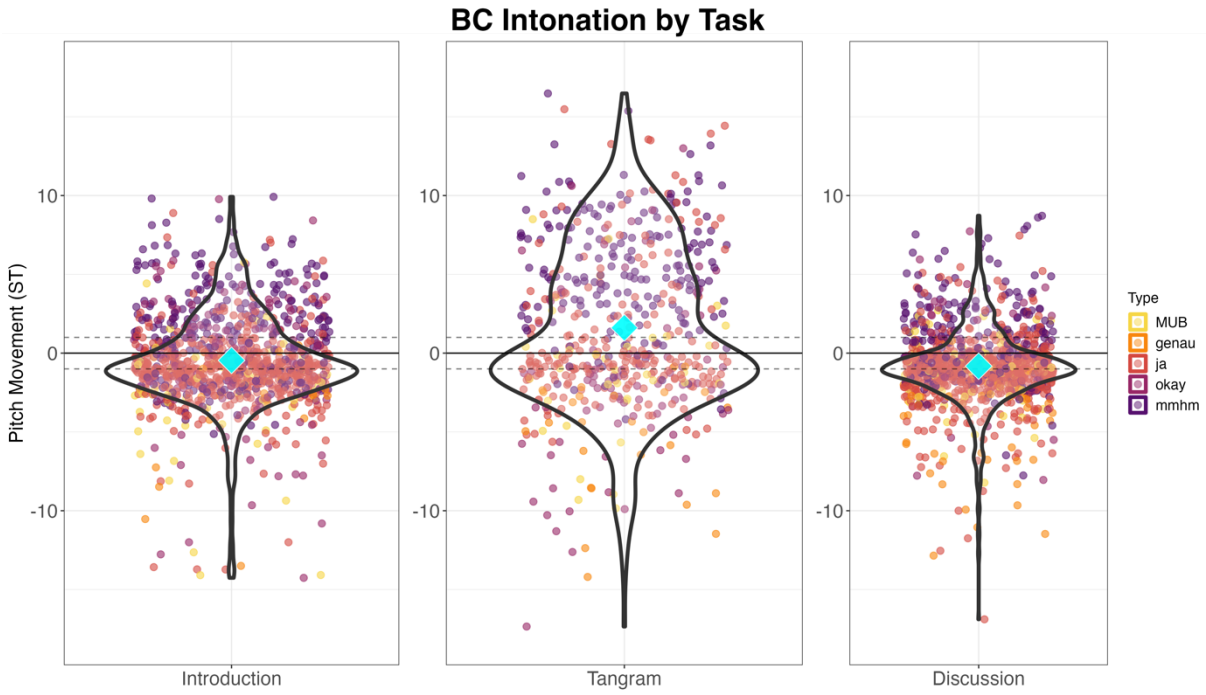


**Fig. 8** Violin plots showing the pitch excursion in semitones (ST, on the y-axis) of backchannels across the three conversations. Backchannel types are color-coded. Mean pitch excursion is indicated by the cyan diamond. The area of pitch movement of ± 1 semitone, defined as level intonation, is indicated by the dashed lines.

Due to the relatively low number of "genau", "okay" and "MUB" tokens found in this dataset (see Fig. 6), results of the prosodic analysis going forward will be focused on the three most frequent generic BC types, for which more data points suitable for a prosodic analysis were available: In this dataset, the most frequently produced BC type was "ja", with a total number of 1.414 tokens being available for the prosodic analysis. The second most frequently produced category was the category of non-lexical backchannels, comprising the BC types "mmhm" and "mm", for which a total of 722 items were available for prosodic analysis (328 "mmhm" tokens, 394 "mm" tokens).

Specific BC types ('other') will be excluded from the prosodic analysis, as well, due to the limited availability of data for individual types within the category.

### 6.4.1 Type-specific intonation – continuous

Results for type-specific pitch excursion across conversational conditions (Fig. 9) provide a more differentiated picture, as the non-lexical type "mmhm" shows higher mean values overall compared to the lexical type "ja", suggesting that there is a general type-related tendency for "mmhm" to be produced with rising intonation and for "ja" to be realized with flat or falling intonation. Moreover, there is a contrast between the two non-lexical types, as "mmhm" tends to be produced with more and stronger rising contours (indicated by a higher mean pitch movement) overall than "mm", which shows more negative pitch movement, i.e. intonational falls, in all three conditions.

Finally, there is a general pattern of greater positive pitch excursion in the task condition compared to spontaneous speech regardless of BC type. This is indicated by the higher mean values in that condition for each type, which reflect more items being produced with stronger intonational rises during the Tangram task.

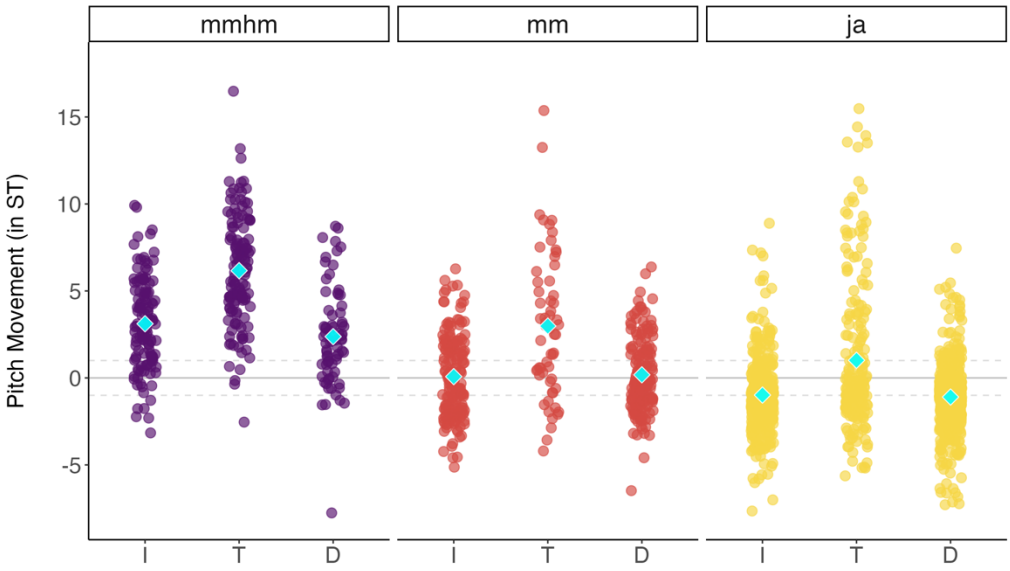


**Fig. 9** Pitch movement by BC type across the three conversations (I = Introduction, T = Tangram, D = Discussion). Pitch movement, measured in semitones (ST), shown on the y-axis. BC types and conditions shown on the x-axis.

## 6.5 Intonation – Categorical measurements

Categorical results for prosodic realization show the proportions of rising, level and falling intonation by BC type across conditions (Fig. 10). Results show that the non-lexical type "mmhm" is produced with the highest rate of rising intonation in

spontaneous speech among the BC types under comparison (Introduction: 78%, Discussion: 68.5%). Its non-lexical counterpart "mm" shows an even distribution of intonation rises and falls in the spontaneous condition: In the *Introduction* it is realized with rising intonation in 36.8% of the cases, and with falling intonation 38.6% of the time. In the Discussion, 33.3% of "mm" utterances are produced with rising intonation, compared to 32.7% with falling intonation. In the latter condition, level intonation accounts for 34%. The lexical type "ja", by contrast, is realized predominantly with falling intonation in spontaneous speech (Introduction: 53.7%, Discussion: 54.8%). Rising intonation accounts for 8.9% in the *Introduction* and 7.5% in the *Discussion*.

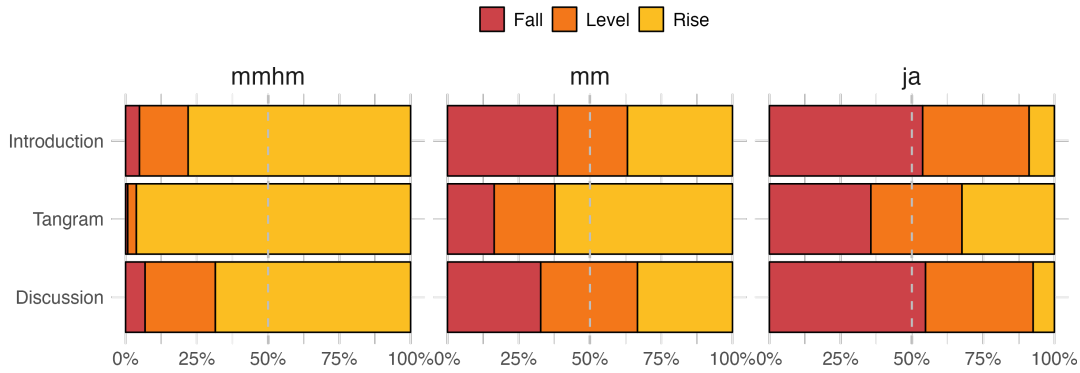


**Fig. 10** Percentage of contour categories *fall*, *level* and *rise* by BC type across conversational conditions.

In all three types, the pattern of greater pitch excursion in task-oriented speech (seen in Fig. 9) is reflected here in a greater proportion of rising intonation in the Tangram task compared to spontaneous speech: "mmhm" shows a 96.2% rate of intonation rises, "mm" is realized with rising intonation in 62.3% of cases, and "ja" in 32.3% of cases. This suggests a clear and stable pattern of both greater pitch excursion and more items produced with rising intonation in the task-oriented Tangram condition for all three examined BC types.

## 6.6 Intonation – Contour clustering analysis

The following section provides results for the CC analysis using the corresponding application developed by Kaland (2023). The analysis was performed separately for the non-lexical types ('mmhm' and 'mm') and the lexical type ('ja') to maximize the similarity of the segmental material and therefore facilitate the differentiation of the clusters based on the contour shapes alone.

To find the ideal number of clusters for the representation of the contours, a method of evaluation, proposed by Kaland and Ellison (2023), was applied that seeks to determine the minimal description length (MDL) for a dataset, assuming the optimal representation of the set. Prior to running the evaluations, some contours were removed from each subset that were identified as outliers, indicated by a higher standard deviation in the clusters they were grouped into. This meant that for the 'ja' subset the final analysis was carried out with 859 contours (initially 888), and with 425 contours (initially 481) in the non-lexical subset. This was done to further ensure that each cluster reliably represents the contours in it, keeping the standard deviation as low as possible.

In order to capture potential durational differences by BC type across conversational conditions, which had been reported to be functionally relevant in previous literature and were visible in the results on duration reported above, *duration* was factored into the contour clustering analysis. Contours will therefore be clustered not only on the basis of pitch movement patterns, but also durational characteristics.

### 6.6.1 Cluster evaluation – MDL

For the MDL cluster evaluation, the number of clusters ranged from 2 to 10, while the bending factor, referring to the degree of dependency between adjacent f0 measurement points (Kaland & Grice 2024), was set to the recommended application default of 4. The evaluation curves for each data subset are presented in figures 11 and 12. The lowest point in each curve represents the ideal number of clusters according to the MDL method. Thus, the optimal cluster number was 2, both for 'ja' (Fig. 11) and 'mmhm' and 'mm' (Fig. 12). The visualized f0 contours generated in the cluster analysis under the assumption of two ideal clusters are presented below (Fig. 13 for 'ja'; Fig. 14 for 'mmhm' and 'mm').

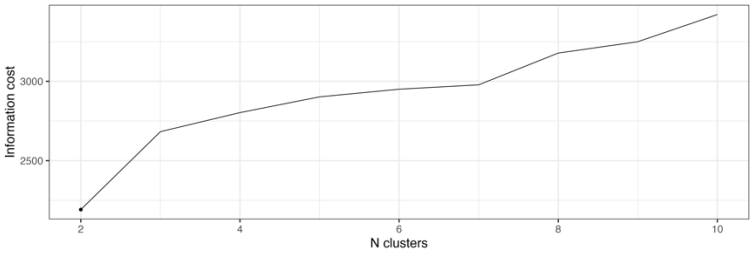


**Fig. 11** Evaluation of the optimal cluster number for the 'ja' subset.

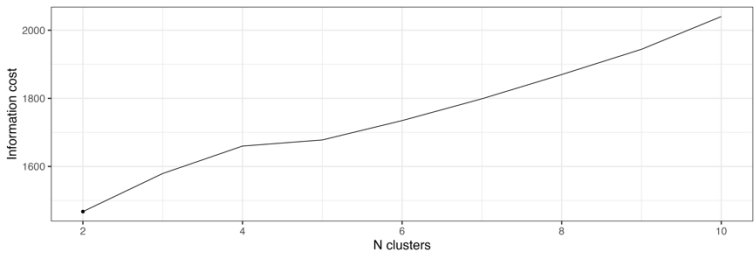


**Fig. 12** Evaluation of the optimal cluster number for the non-lexical subset.

Assuming two clusters for the subset of 'ja' BCs, the cluster analysis shows one falling contour and one rising contour with a shallow fall and a late rise (Fig. 13). The first cluster has a mean duration of 210 ms, while the second cluster, with a more complex shape, has a mean duration of 440 ms. Notably, the vast majority of contours (839) fall into the first cluster, while only 20 contours belong to the second cluster.

The number of f0 measurement points taken from each token in the time-series f0 measurement, which had been set to 10 points for both subsets, is given on the x-axis. The y-axis shows a speaker-standardized f0 scale, in which 0 marks the speakers' mean f0.

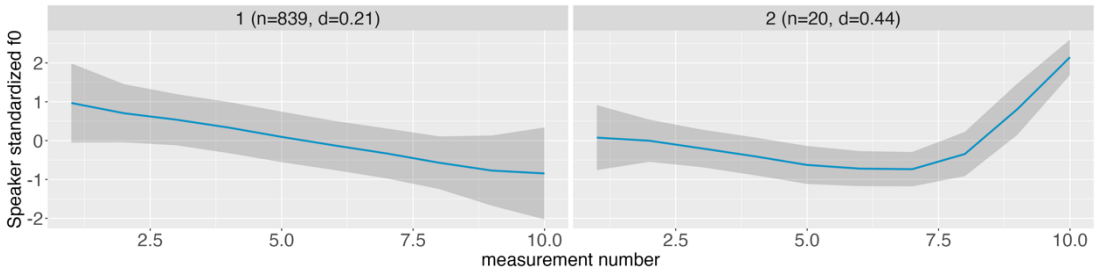


**Fig. 13** Speaker-standardized f0 contours in two clusters. 'ja' subset.

For the non-lexical subset, including the BC types 'mmhm' and 'mm', the cluster analysis shows one falling mean contour with a late shallow rise and one rising contour (Fig. 14). Here, both contours have a similar mean duration of 310 ms (1) and 320 ms (2). The distribution of contours across the two clusters is more even compared to the clusters in the 'ja' subset, with 288 contours in (1) and 137 in (2).

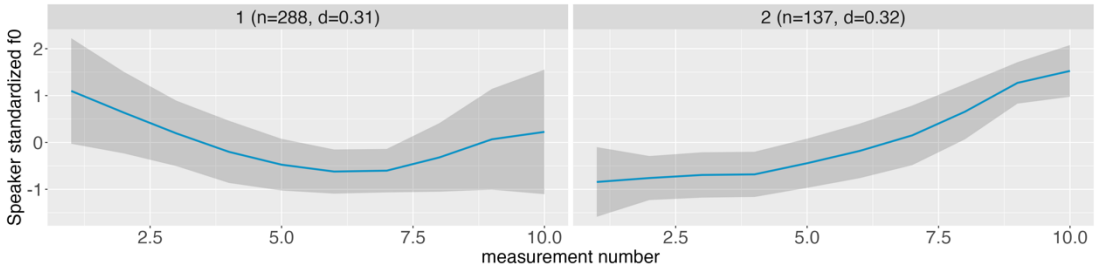


**Fig. 14** Speaker-standardized f0 contours in two clusters. Non-lexical ('mmhm', 'mm') subset.

The clusters shown represent the contours with a minimal degree of complexity and variation. In principle, they should capture the most salient, coarse-grained differences between the contours of each subset. However, they do not match the proportions of intonational rises and falls reported under section 6.5, especially in the case of 'ja'. Based on the results of the linear interpolation method, a higher number of rising intonation contours can be expected for this type than the 20 instances suggested by the above contour clusters (Fig. 13). Therefore, there is reasonable ground to assume a higher number of clusters to be more fitting to capture these intonational patterns.

### 6.6.2 Cluster evaluation – W/B variance

In order to allow for more variability to be captured, the number of assumed clusters was raised to a higher number expected to be able to pick up on more fine-grained differences such as the contours' steepness and curvature, as well as the possibility of contours with similar shapes to be produced with different durations. To avoid the setting of the cluster number to be made at random, another cluster evaluation method (W/B cluster variance) was applied, which aims to find the crossover point of within and between cluster variance. *Between* variance refers to the degree of variance between the clusters for each of the number of clusters evaluated, while *within* variance indicates the variance within each cluster for each number of clusters assumed. The lower the variance, the more similar the clusters are to one another (within) a given cluster. The less variation is allowed within a cluster, the greater the variance will be *between* the clusters, as they will have to be distributed among a greater number of narrowly defined clusters. Therefore, the greater the number of clusters assumed, the more neatly the contours in each cluster will fit the mean contour in each cluster, as they will be more similar to one another, and the

greater the differences will be between the clusters. The risk in choosing a higher number of clusters is that the clusters may be overfitted and as a consequence disproportionately represent differences in the contours that could be of minor significance. Therefore, the number of clusters assumed should allow for as much variance as necessary between the clusters, while keeping the number of clusters as low as possible. The W/B method of cluster evaluation provides useful indications for the appropriate range of variance.

For the W/B method, the number of clusters for the evaluation was set from 2 to 10. Figures 15 and 16 show the W/B evaluation results for the two subsets.

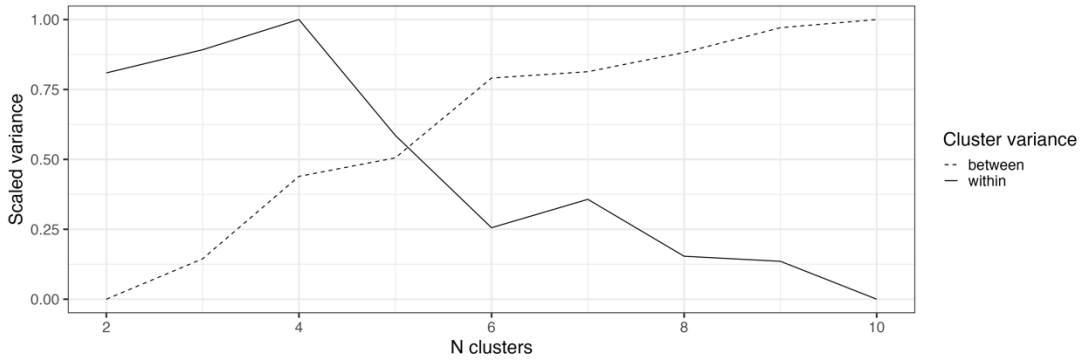


**Fig. 15** Results of the W/B Cluster Variance evaluation for the 'ja' subset. Degree of variance shown on the y-axis, number of assumed clusters shown on the x-axis.

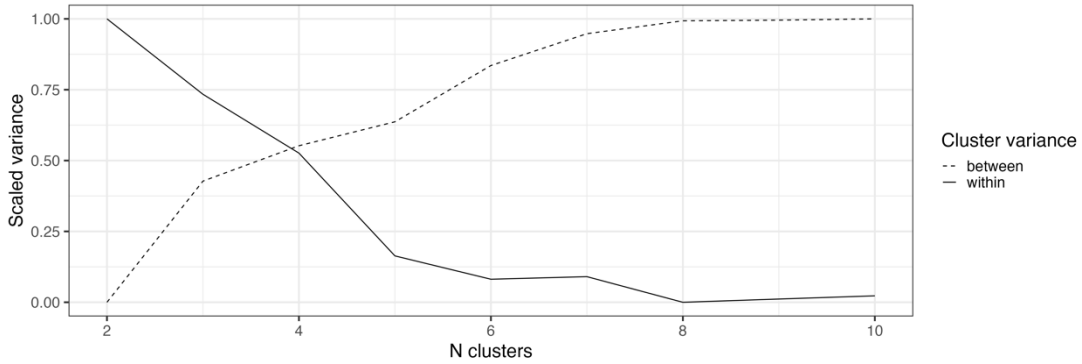


**Fig. 16** Results of the W/B Cluster Variance evaluation for the non-lexical subset

Based on the results of the W/B cluster variance evaluation, the number of assumed clusters was set to 6 for the 'ja' subset and to 5 clusters for the subset of 'mmhm' and 'mm' BCs. This number is located after the crossover point of within and between cluster variance, allowing for more variance between clusters and keeping variance within clusters as low as necessary without overestimating the number of clusters.

### 6.6.3 Contour clusters – 'ja'

The following results (Fig. 17) show the contour clusters for the lexical BC type 'ja', assuming an ideal number of 6 clusters. The resulting contours have been categorized as follows: (1) fall long, (2) rise, (3) fall short, (4) rise-fall, (5) fall-rise long, (6) fall-rise short.

The most frequent contours belong to clusters (3) and (1), with 441 and 231 corresponding contours respectively. Both clusters represent falling contours of overall similar shapes, differentiated by duration. Cluster (3) has a shorter mean duration of 160 ms, compared to 270 ms in cluster (1). The contours in cluster (4) are characterized by an initial rise followed by a fall, with a mean duration of 260 ms. This cluster is the least frequent in the group of clusters showing a final fall in intonation, consisting of only 36 contours.

The group of clusters with rising intonation contours comprises clusters (2), (5) and (6). Cluster (2) shows a rising mean f0 with a subtly curved contour and a mean duration of 250 ms. Clusters (5) and (6) have similar contours, characterized by an initial fall and a final rise. Their shapes differ in terms of the steepness of the fall and rise respectively, with cluster (5) showing a shallow fall initiated with a lower f0 than cluster (6). Moreover, the rise in cluster (5) begins later and is steeper. In the following, contour shapes with a late rise, as in clusters (5) and (6), will be referred to as *complex* rises or contours to distinguish them from contours with an early rise and no (clear) change in f0 direction, as in cluster (2).

Clusters (5) and (6) differ with regard to their durations and in terms of their frequencies: Cluster (5) has a mean duration of 440 ms, which is more than twice the overall mean duration of 'ja' BCs (214.8 ms). Contours in cluster (6) have a mean duration of 290 ms. The contours belonging to this cluster are more frequent, with 80 observations showing this pattern, compared to 20 in cluster (5).

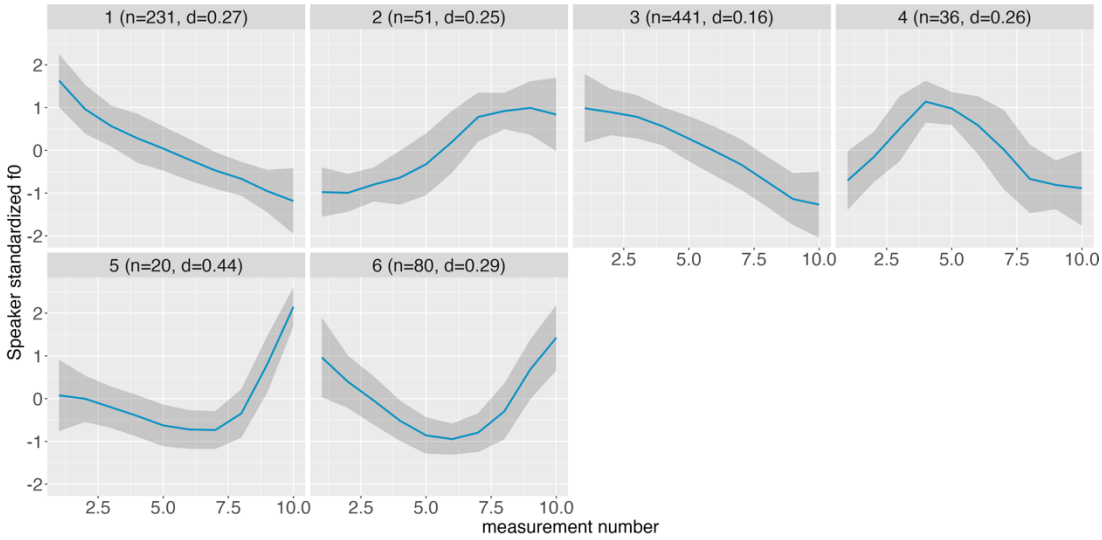


**Fig. 17** Speaker-standardized f0 contours for the lexical category 'ja' in six clusters. Measurement number (10 per contour) represents time-normalized scale; n = number of observations in cluster; d = average duration.

The distributions of the contours across the different conversational conditions are shown in Fig. 18. The most frequent cluster across conditions is cluster (3), containing short falling contours (Introduction: 48.9%, Tangram: 55.8%, Discussion: 52.3%). These contours are realized with a similar proportion in all three conversations. Cluster (1), containing *long* falling contours, is the second most frequent cluster overall, but its contours are used at a proportionately lower rate in the Tangram condition (17.9%) than in spontaneous speech (Introduction: 29.9%, Discussion: 26.4%). Cluster (4), with a mean fall-rise contour, is represented at a low but stable rate in the Introduction (2.3%) and Tangram (2.1%) condition, and a slightly higher rate in the Discussion (6.2%). However, the overall number of contours in this cluster is relatively low (36).

In the group of clusters showing contours with final rises (color-coded with different shades of yellow), only cluster (5) accounts for a low but steady rate across conditions (Introduction: 2.6%, Tangram: 2.1%, Discussion: 2.2%). The rising contours from cluster (2) are found at a higher rate in task-oriented speech (8.4%) than in spontaneous speech (Introduction: 5.7%, Discussion: 5.5%). Similarly, the short fall-rising contours from cluster (6) are found more in the Tangram conversations (13.7%) compared to spontaneous speech (Introduction: 10.3%, Discussion: 7.4%).

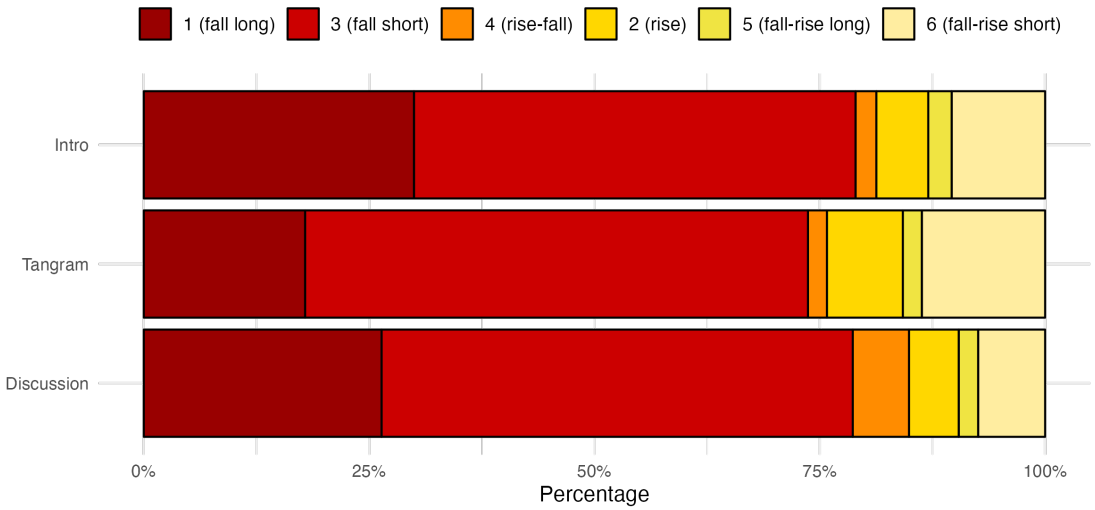


**Fig. 18** Percentage of contour types in each of the conversational conditions for BC type "ja". Falling contours in shades of red and orange, rising contours in shades of yellow.

Overall, contour clusters with falling contours account for the majority of contours found in the 'ja' subset, while rising contours make up a quarter or less proportionately. Across conditions, the composition of clusters was found to change in favor of more rising contours in the Tangram task, particularly with clusters (2) and (6) being found more in that condition. At the same time, 'ja' appears to be realized with less long falling contours (1) in task-oriented speech compared to spontaneous speech. These results suggest that conversational conditions have an influence on intonation contours not only in terms of the quantity of falling or rising contours, but also on the preference for particular shapes within the group of falling and rising contours.

### 6.6.4 Contour clusters – 'mmhm' and 'mm'

The following results (Fig. 19) show the contour clusters for the non-lexical subset, including the BC types 'mmhm' and 'mm'. In accordance with the W/B cluster variance evaluation (Fig. 16), the ideal number of clusters in this subset is 5. The resulting contours have been categorized as follows: (1) fall short, (2) fall long, (3) rise short, (4) fall-rise short, (5) fall-rise long. Note that cluster (4) shows only a subtle and shallow initial downward movement. For the sake of simplicity, it is labeled as a 'fall-rise', but will be treated and discussed as a complex contour, together with cluster (5), as will be explained further down below.

Clusters (1) and (2) contain falling contours, differentiated by their mean durations of 230 ms (1) and 340 ms (2) respectively, as well as a slightly more curved

shape in cluster (2), starting with a higher f0 and a sharper fall. The remaining clusters (3), (4) and (5) show contours with overall rising intonation. Cluster (3) is the shortest of this group, with a mean duration of 270 ms. It is characterized also by the least complex f0 movement, as it shows an early rise that stretches across most of the contour. Clusters (4) and (5) show an initial fall, which is steeper in (4), followed by a sharp final rise. In cluster (4), which has a shorter mean duration of 330 ms, the falling intonation in the first half of the contour is more pronounced, while the rise begins earlier (around measurement point 5). This cluster contains the highest number of contours (115) in this subset. The mean contour of cluster (5) indicates a shallow initial fall and a late steep rise. This contour has the longest mean duration (460 ms) in this subset, which is above the mean duration of both non-lexical BC types 'mmhm' (293.8 ms) and 'mm' (328.5ms). Contour shapes of 'mmhm' and 'mm' with a late rise, as in clusters (4) and (5), will in the following be referred to as *complex* rises or contours. Despite cluster (5) lacking an initial f0 fall similar to that of cluster (4), for the sake of this analysis, greater emphasis in the categorization of these contours will be put on the position and steepness of the rise, which is believed to present a relevant distinction between clusters (4) and (5) and the early-rising cluster (3).

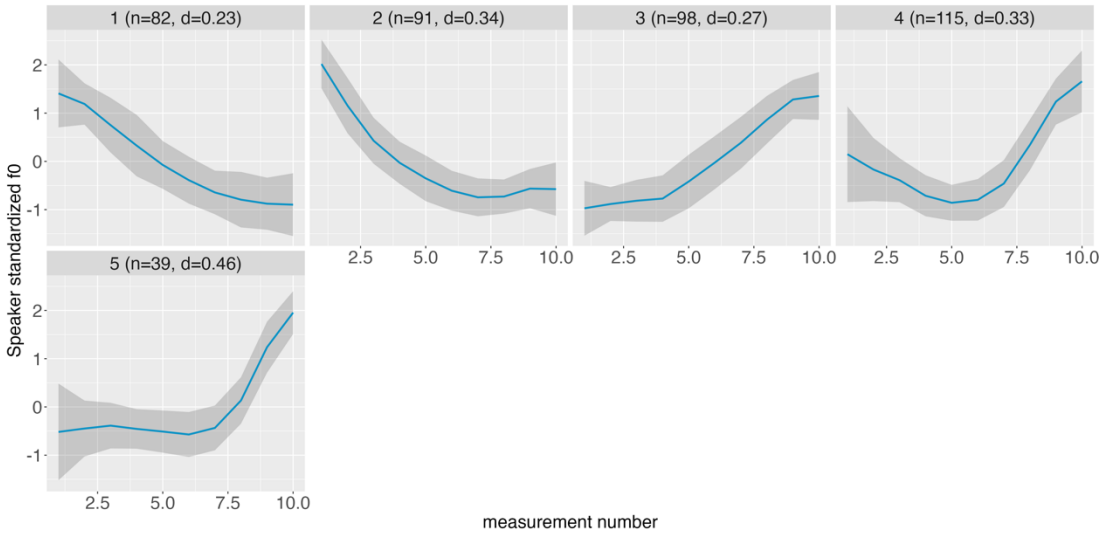


**Fig. 19** Speaker-standardized f0 contours for non-lexical types 'mmhm' and 'mm' in five clusters. Measurement number (10 per contour) represents time-normalized scale; n = number of observations in cluster; d = average duration.

The distribution of contour clusters for 'mmhm' and 'mm' across the different conversational conditions are shown in Fig. 20. According to the contour clustering analysis, the non-lexical type 'mmhm' was realized with rising intonation mostly, regardless of the conversational condition. Results suggest a cross-condition

preference for the short rising contour (3) as the most frequent contour type overall: Introduction 48.6%, Tangram 50%, Discussion 39.5%. The short fall-rising contour of cluster (4) has been the second most frequent contour shape for this BC type across conditions: Introduction 25.7%, Tangram 26.5%, Discussion 34.9%.

Across conditions, results show a shift towards a higher overall percentage of rising contours in the Tangram condition, which appears to be driven primarily by an increased proportion of the long fall-rising contours of cluster (5): Introduction 7.1%, Tangram **14.7**%, Discussion 4.7%. Clusters (3) and (4), on the other hand, show similar proportions in the Tangram condition compared to the Introduction.

The non-lexical type 'mm' shows a higher overall proportion of falling contours, primarily in the spontaneous speech conditions, where the falling contours from clusters (1) and (2) account for over 50% of contours. In the Introduction, the short falling contours (1) make up 31%, while the long falls (2) account for 24.8%. In the Discussion, the distribution shifts in favor of the long falls (38.3%), while the short falling contours account for a smaller proportion of 19.2%. Within the group of rising contours results suggest a preference for the short fall-rising contour (4): 22.5% in the Introduction, 25.8% in the Discussion. The short rising contour of cluster (3) accounted for a low but stable rate of 8.5% and 10.8% in the Introduction and Discussion respectively. The long fall-rising pattern shows a reducing trend across conditions: Introduction 13.2%, Tangram 10.3%, Discussion 5.8%.

Looking at the Tangram condition in comparison to the spontaneous contexts, the falling contours are reduced by more than half, with the short falling contours (1) accounting for 17.2% and the long falls (2) for 6.9%. The large increase in rising contours is primarily driven by the short fall-rise pattern in cluster (4) accounting for 44.8% in task-oriented speech, and thus approximately twice as much as in the spontaneous conditions. This contour type was the most commonly used one in the Tangram condition. The short rising contours in cluster (3) accounted for a proportion of 20.7%, being the second most frequent contour type in this condition.

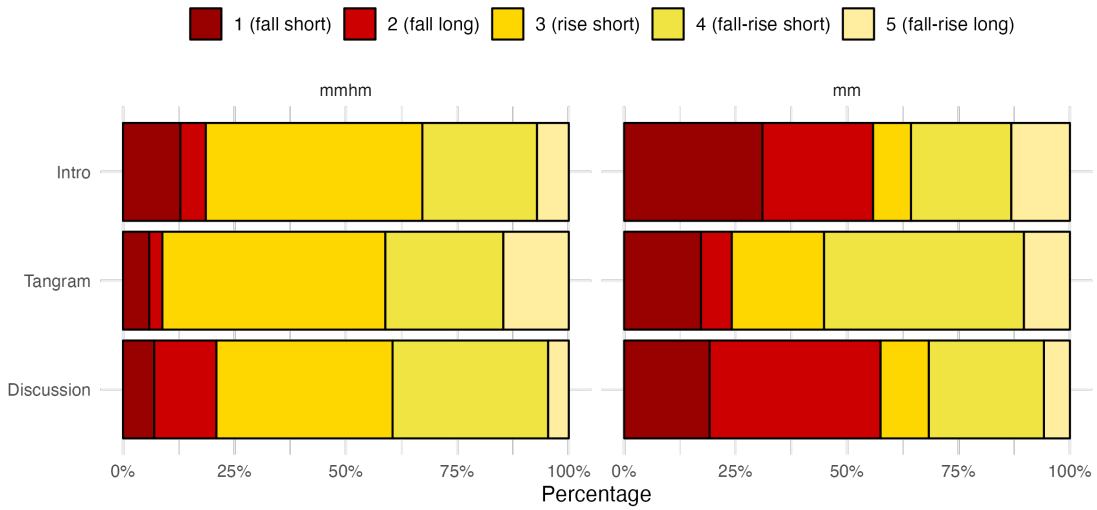


**Fig. 20** Percentage of contour types in each of the conversational conditions for non-lexical BC types "mmhm" and "mm".

### 6.6.5 Summary: Contour clustering

The results of the Contour Clustering analysis suggest that the backchannel types 'ja', 'mmhm' and 'mm' are realized with various different intonation contours, varying not only with regard to their slopes, but also in terms of the steepness and position of rises and falls, their pitch range and their durations. The cluster visualizations provide an informative and advantageous addition to the categorical intonation results reported in section (5.5). While the overall pattern of more rising intonation in the Tangram condition (see Fig. 10) matches the proportions of rising contours that resulted from the cluster analysis, the Contour Clustering approach complements the results from the linear interpolation technique with a perspective on the types of contours that are present in each category of rises and falls, as well as on their distributions across conditions. For instance, 'ja' backchannels have been shown to be produced with mostly falling intonation, while Contour Clustering results (Fig. 17) reveal a preference for short falling contours, which is clearer in the Tangram condition than in the two spontaneous conversations. At the same time, a majority of the rising 'ja' BCs are realized with a fall-rising pattern, seen in clusters (5) and (6). The shorter and overall straighter rising contour in cluster (2) is less frequent across conditions relative to the fall-rising patterns.

For the non-lexical types 'mmhm' and 'mm', the Contour Clustering (Fig. 19) has, again, confirmed the categorical results of a higher proportion of rising contours in the Tangram condition (as in Fig. 10). But in addition, it suggests that 'mmhm' is realized primarily with an early rising shape (3) and to a lesser extent with the complex

fall-rise shape of cluster (4). The increased proportion of rises in the task condition is mainly due to an increase in cluster (5), while otherwise showing relatively stable contour shape proportions. On the other hand, ‘mm’ showed greater overall variation in terms of its intonational realization, as well as the clearest cross-conditional difference in global intonation patterns. Within its rising realizations, ‘mm’ is predominantly produced with complex contours, showing either an early fall and a late rise or an initial flat movement and a late rise. Moreover, taking duration into account, the relatively longer contours (2), (4) and (5) are in total realized more frequently with the type ‘mm’, which might explain the higher mean duration observed for this type compared to ‘mmhm’ (see Fig. 4).

# 7 Discussion

## 7.1 Rate

The overall backchannel rate across dyads was higher in the spontaneous conversations (Introduction: 9.00 BCs/min; Discussion: 8.48 BCs/min) than in the task-oriented speech condition (Tangram: 4.5 BCs/min). These results confirm the observations made by Dideriksen et al. (2023), who had reported a higher BC rate in spontaneous speech compared to two different task-based conversations, although the two studies are not directly comparable. The authors measured the BC rate in terms of the proportion of utterances identified as backchannels (in %), whereas in this study BC rate has been operationalized as the number of BCs uttered per minute of dialogue. While it is not known whether the BC rate reported by Dideriksen et al. (2023) matches the rate reported here in terms of the frequency of BCs per minute, the relation between the BC rate in spontaneous and task-oriented speech could be tentatively compared with the relation observed in the present study: Dideriksen et al. (2023) identified 33.52% of utterances as backchannels in the spontaneous conversations, against 14.93% in the Alien game (similar in structure to the Tangram task), which means that the BC rate in spontaneous speech was about twice as high as in task-oriented speech. In this study, a similar relation was observed, as the rate of BCs per minute was found to be about twice as high in the Tangram task compared to the free conversations (Fig. 2). These results therefore contradict the observations of Fusaroli et al. (2017) and Janz

(2022), who had reported higher BC rates in task-oriented speech. However, the different contextual requirements resulting from different tasks, such as the Tangram game or the Map Task, should be taken into consideration. Furthermore, in the study by Janz (2002), which used the Map Task, the participants were not able to see each other in the task condition, which might have further increased the reliance on verbal feedback and explain the higher rate compared to the spontaneous speech condition.

To better understand the impact of collaborative tasks on the rate of backchanneling, in light of the different contextual requirements they pose, Dideriksen et al. (2023) analyzed two different task-based conversations under comparable conditions, reporting that the BC rate was 10.7 percentage points higher in Map Task conversations than during the Alien game. Similarly, the BC rate reported in the current study, using the Tangram task, was lower (4.5 BCs/min) than the rate reported by Sbranna et al. (2022) for native speakers of German (6.12 BCs/min) using Map Tasks and an overall similar methodology.

While studies on backchanneling behavior should generally be compared with caution, mainly due to sometimes widely different operationalizations of the concept of backchannels, the current study provides further evidence for the observations by Dideriksen et al. (2023), according to which, on the one hand, different tasks lead to different backchanneling rates, and, on the other hand, more backchannels are used in spontaneous speech compared to task-oriented interactions.

Interestingly, all dyads have shown to conform to the pattern of producing considerably less backchannels during the Tangram game than in the two spontaneous conversations (Fig. 3). It is noticeable, however, that while there is no exception to this pattern, there is a large variation in dyad-specific BC rates, with dyad 14 producing almost 15 BCs per minute in the first spontaneous conversations, compared to approximately 5 BCs per minute uttered by dyad 06 in the same context. Furthermore, these large discrepancies mean that the rate of BCs produced by dyad 14 in the Tangram task is higher than the BC rates produced by several dyads (05, 06, 09, 10, 13) in spontaneous speech, despite the group average indicating a higher BC rate in spontaneous than in task-oriented speech. This suggests that group averages should be interpreted with caution and that considerable attention should be given to dyad- and speaker-specific behavior.

In sum, the results for backchannel rate suggest that the conversational condition has an effect on the number of backchannels that listeners produce. Data from all 14

dyads recorded and analyzed in this study indicates that more backchannels are used in spontaneous conversations than in task-oriented speech, hinting at backchannels fulfilling a different conversational function in spontaneous speech. Importantly, these observations were made in a setting in which participants were able to have eye contact.

## 7.2 Duration

The analysis of BC duration has shown that the mean duration of single-unit generic backchannels (‘mmhm’, ‘mm’, ‘okay’, ‘ja’, ‘genau’) was 257 ms, while MUBs, which are multi-unit versions of generic BC types, were found to have a mean duration of 604 ms. Both the single- and multi-unit BCs analyzed here were thus slightly shorter than in Mereu et al. (2024), who had investigated BCs in Italian and found a mean duration of around 340 ms for single and 700 ms for multi-unit BCs. In an analysis of backchannels in American English, Young and Lee (2004) reported a mean duration of 390 ms. These results suggest an overall shorter mean duration of BCs in German. However, since individual BC types have different mean durations (Fig. 4), the mean across BC types is likely to be influenced by the preference for a particular BC type in each language. For instance, the BC type ‘ja’, which has the shortest mean duration of all types (215 ms), was found to be the most frequent BC type in this analysis. This means that the mean duration across all types will be skewed towards the mean duration of this type and away from less frequent types such as ‘mm’. ‘okay’ and ‘genau’ with mean durations of above 300 ms. Since type-choice preferences have been found to be language specific, this has to be taken into account when examining mean BC durations across types.

In addition to type preferences, BC functions had been found to have an influence on BC duration in previous studies (Beňuš 2016; Neiberg et al. 2013; Zellers 2021). In this study, backchannels have been defined broadly, as encompassing all feedback signals uttered in the *back channel* that do not constitute or claim a turn. Other studies have restricted their definitions of backchannels to continuers, excluding agreement and assessment functions, among others. This will inevitably impact results on BC duration, making direct comparisons difficult.

Results on the duration of context-specific BCs (categorized as ‘other’) have shown that single-unit specific types were produced with a mean duration of 428 ms,

while multi-unit specific BCs showed a mean duration of 688 ms. Both categories of context-specific BCs were therefore found to have longer mean durations than single- and multi-unit context-*generic* BCs. Given that backchannels used as reaction tokens, conveying an affective stance towards the speaker's utterance (Zellers 2021), as well as monosyllabic BCs conveying surprise and interest (Neiberg et al. 2013), have shown relatively longer durations compared to BCs with other functions, it can be hypothesized that specific types, such as 'ah', 'cool', 'krass' (colloquial: sick/sweet/wicked) and 'nice' fulfill such reactive functions and therefore contribute to the longer durations by the group of context-specific types. A closer inspection of the category of context-specific BCs has shown that many types found in this category, based on their lexical form and semantic content alone, can be assumed to convey affiliative and social functions, such as surprise, interest and empathy. An overview of the most frequent context-specific types can be found in Table 1 and 2.

### 7.2.1 Duration across conditions

The conversational condition seemed to have no effect on the generic single-unit BCs. Their individual mean durations, as well their aggregate mean duration, were found to be relatively stable across conditions. The effect of conversational condition appeared to be stronger for generic MUBs and specific single BCs (Fig. 5 and 5.1), which showed longer overall durations in the Tangram condition. The largest durational difference between spontaneous and task-oriented speech was found for context-specific single BCs, which showed a 34% longer mean duration in the task context. As for the reason for this observation, either of the following explanations seems plausible: Firstly, the low overall quantity of data points, particularly in the Tangram condition, were there was a total of only 50 context-specific BCs, might have resulted in outliers skewing the mean duration in a particular direction without any actual systematic effect being the reason for it. On the other hand, since previous literature suggests surprise and interest to correlate with relatively longer durations, particularly in non-lexical monosyllabic types, it could be hypothesized that the slightly more asymmetric information structure in the Tangram task evokes more emphatic expressions of surprise, especially in the sense of "*ah* (now I get it)". At the same time, it may be less likely for speakers to utter such feedback signals in the same way in spontaneous speech, due to different contextual factors compared to the task condition. Overall, context-specific BCs were found to be produced more often in

spontaneous speech. However, contextual demands in the structurally different Introduction, which resembles small talk, or Discussion, where speakers share their thoughts about the Tangram task, may lead to feedback signals of the same type being produced with a different function than in task-oriented speech, reflected in different prosodic forms, including duration.

One type that has frequently been found to convey surprise in reaction to an unexpected piece of information was 'ah', which was coincidentally the most frequent context-specific BC in the analyzed dataset. Its mean duration in the Tangram condition was in fact longer (M = 422 ms) than in spontaneous speech (Introduction: M = 363 ms; Discussion: M = 298 ms). As mentioned above, however, the relatively low rate of tokens in this condition should be taken into consideration. Moreover, due to different types being preferred across the different conversations, which will be discussed in detail in the subsequent chapter, not all BC types were available from each condition for a direct, type-specific comparison of their durations. For instance, the evaluative reaction token 'cool' was uttered 35 times in total in the Introduction, including 18 times as a single BC, and 17 times as a multi-unit BC ('cool wow', 'oh cool' and 'ja cool'), whereas in the Discussion there was only one single-unit utterance of 'cool', and none in the Tangram task (only one multi-unit BC: 'ah cool ja'). In other cases, such as the one of 'gut' (*good*), several utterances were available from each conversation, however, the longest mean duration was found in the Discussion, contradicting the overall tendency of longer durations in task-oriented speech. Regarding the interpretation of this trend, it has to be taken into consideration that different types of single-unit specific BCs are being compared across conversations. Since BC types have different intrinsic durations (see Fig. 4), cross conditional durational differences might be a consequence of varying type-choice patterns, primarily for the categories of MUBs and 'other' BCs, which each include a wide variety of different types. Nevertheless, if the duration of these categories of BCs is longer in task-oriented speech due to other types being chosen, this remains an interesting finding in itself that warrants further investigation.

### 7.2.2 Summary: Duration

The analysis of backchannel duration has shown slightly shorter average durations for generic single-unit as well as multi-unit BCs compared to previous studies. A look at individual BC types suggests that the average duration of

backchannels as a whole might be influenced by type-related preferences: In this analysis, ‘ja’ was found to be produced with the shortest mean duration, while the non-lexical type ‘mm’ showed the longest mean duration among the generic single BCs.

Context-specific single BCs were found to have a longer mean duration than context-generic single BCs, which might be attributed to this kind of backchannels fulfilling a different function than generic BCs, such as the social and affiliative function of reacting to the content of the speaker’s utterance. Reactions in the form of expressions of interest and surprise had been found to correlate with longer durations in previous studies.

Cross-conversational differences revealed a minor change towards longer durations in task-oriented speech in the case of ‘ja’ and ‘mmhm’, towards shorter durations for ‘genau’, and mixed patterns for ‘mm’ and ‘okay’. Therefore, the group of generic single BCs indicate no clear cross-condition tendency. A stronger trend toward longer durations in the Tangram condition was observed for (generic) MUBs and specific single-unit BCs, with the largest increase in duration having been observed for the latter category. However, this category is composed of a large variety of different lexical and non-lexical types, many of which were not used consistently across the three conversations, making a systematic comparison of specific types in different conversational conditions difficult. In order to determine the validity of this observation, more data on specific BCs is needed, as well as an analysis that takes the pragmatic context into account to determine potential function-related effects on BC duration.

## 7.3 Type

The analysis of BC type choice has shown that ‘ja’ was the most frequent backchannel across all conversations and dyads, accounting for 46.3% of all BCs, followed by the two non-lexical types ‘mmhm’ and ‘mm’, which together accounted for 22.7% of all BCs uttered. This is in line with previous studies on BCs in German showing the same two categories being the most prevalent (Janz 2022; Sbranna et al. 2022). Context-specific backchannels accounted for a total of 14.1%. The least frequent generic single BCs were ‘okay’ (8.3%) and ‘genau’ (3.1%), while MUBs were used 5.5% of the time.

Looking at cross-conversational differences, two tendencies stand out: Firstly, the two non-lexical 'mmhm' and 'mm' types taken together showed a higher rate in the Tangram task than in spontaneous speech. Individually, however, they show opposing patterns, with 'mmhm' being produced more in task-oriented than in spontaneous speech, and the opposite being the case for 'mm' (see section 5.3.1).

The non-lexical type 'mmhm' can be considered a prototypical backchannel that is commonly used as a *continuer*, or marker of passive recipiency (Bangerter & Clark 2003; Beňuš 2016; Drummond & Hopper 1993; Jefferson 1984; Ward 2004). The higher use of this type of BC in task-oriented speech may be attributed to task-related contextual requirements. Tasks such as the Map Task or the Tangram game are characterized by more or less clearly defined roles. This is mainly the case in the Map Task, where one speaker (the information giver) receives a map with a path that has to be described to the interlocutor (the information follower). In the Tangram game, the roles are not fixed, as the speakers switch roles after each picture. And since neither of the speakers knows in advance whether or not their pictures match, the access to information is more symmetrical than in the Map Task. Nevertheless, in each turn one speaker is tasked with describing the target picture, which means that the interlocutor takes a listening role, knowing that a longer stretch of speech is coming from the speaker. It can be presumed that situations like these generate more continuers, compared to spontaneous speech, where the exchange of information, and therefore the change of turns, is more instantaneous.

The higher proportion of 'ja' in spontaneous speech further suggests that the function of backchannels in this context differs from the function of feedback signals in task-oriented speech, shifting away from the more basic function of maintaining and developing mutual understanding, and moving towards social and affiliative functions of displaying agreement and alignment. Interestingly, the highest proportion of 'ja' backchannels is found in the Discussion (55.3%). As opposed to the Introduction, where the speakers had to introduce themselves in a small talk-like scenario and establish where potential commonalities lied to keep the conversation going, the Discussion provided them with the opportunity of sharing their thoughts and feelings about the task they had just performed together. This context might have invited more signals of agreement in the form of 'ja' than the topically more open Introduction. Apart from showing the highest proportion of 'ja', the Discussion also showed the lowest rates of 'mmhm' (5.9%) and 'okay' (3.6%). These results indicate that, while

the strongest differences in the backchanneling behavior can be seen between spontaneous and task-oriented speech due to greater structural differences, free conversations should not be assumed to guarantee homogenous outcomes.

Another indication of the structural and contextual difference of the spontaneous conversations can be found in the dyad-specific results for type choice (Fig. 7), which hint at a more aligned pattern across dyads in the Discussion and a higher degree of variability in the Introduction: With the exception of dyad 06 and 10, all dyads displayed a dominant preference for 'ja' in the Discussion, while no such distinct preference is visible in the Introduction. This suggests that backchanneling behavior differs not only between task-oriented and spontaneous speech, but that there is also potential for systematic differences, albeit to a lesser degree, between different kinds of spontaneous conversations, depending on thematic factors (e.g. small talk vs shared experience).

### 7.3.1 MUB types

MUBs were produced at a low rate of around 5% in all three conversations (see section 5.3.2 for detailed results). This is a substantially lower rate than reported by other studies (Mereu et al. 2024; Wong and Peters 2007), which analyzed spontaneous speech. Mereu et al. (2024) included backchannels with the function of signaling incipient speakership (IS), i.e. turn-claiming signals, which had been excluded in the present study. Since the authors report MUBs to often convey multiple functions simultaneously, including IS, it can be expected that the rate of complex backchannels would be higher if feedback signals with this function had been included. Moreover, in the current study, 'MUBs' had been defined as multi-unit versions of generic backchannels, while complex backchannels involving specific types (e.g. '*ah* okay') were labeled as 'other'. Counting both generic and specific multi-unit BCs, the overall rate of complex BCs would be higher than 5%, considering that about half of all context-specific BCs were complex. Nevertheless, even in this case the rate of MUBs found in this dataset would still be considerably lower than the rates reported by Mereu et al. (2024) and Wong and Peters (2007). More data on multi-unit backchannels in German is needed to confirm and elaborate on the observation made in the current study regarding the frequency of complex backchannels.

As far as MUB types are concerned, the most frequent type overall was found to be 'ja ja'. Interestingly, this type occurred almost exclusively in spontaneous speech,

accounting for almost half of all MUBs in the Introduction and Discussion, while 'ja genau' was instead the most preferred type in task-oriented speech. The preference for 'ja ja' to be used in spontaneous speech might be explained by its pragmatic meaning. Analyzing the functional difference of this BC type and its monosyllabic counterpart 'ja', Golato and Fagyal (2008) note that 'ja ja' is especially used in contexts where the prior speaker says something that is "obvious and/or known" by the secondary speaker. Given the different contextual requirements of spontaneous and task-oriented speech, it can be assumed to be less likely for the speaker to utter something that is considered obvious by the listener in a scenario such as the Tangram game, in which the interlocutors do not have equal access to relevant information. The information the speakers share with each other in this task scenario is mostly new and relevant. Thus, backchannels with the meaning of "I already got it, so stop" can be expected to be uttered less often, which is reflected in the low rate of 'ja ja' in the Tangram condition (6.1%). As for the pragmatic meaning and cross-conditional use of other MUBs found in this dataset, further (conversational) analysis is needed to reveal potential type-function relations.

### 7.3.2 Context-specific backchannel types

Context-specific backchannels, summarized in the category of 'other' BCs, were found to be overall more frequent in spontaneous than in task-oriented speech. This is in line with Janz (2022), who reported a higher proportion of similarly defined specific BCs in spontaneous speech compared to Map Task dialogues. The analysis of context-specific types showed that a wide variety of types were used, with no clear preference for any particular type (see sections 5.3 and 5.3.3 for detailed results). Context-specific backchannels provide feedback to the content of speech, rather than merely signaling continued attention and serving as a go-ahead sign. Accordingly, based on their form and/or lexical meaning, specific types found in this dataset fulfilled functions such as conveying surprise and interest (e.g. 'ah', 'oh', and 'wow'), providing an evaluative or emotional reaction (e.g. 'cool', 'nice', and 'krass'), and expressing acceptance towards or alignment with the preceding utterance (e.g. 'natürlich', '(das) stimmt', 'voll', and 'auf jeden Fall' – roughly translating to *of course*, *(that's) correct*, *totally* and *definitely*).

Interestingly, the conversational condition has shown to affect the use of specific backchannels not only in terms of rate, with fewer specific backchannels being uttered

during the Tangram task, but also regarding the choice of types. A greater variety of types were used in spontaneous compared to task-oriented speech (see Fig. 6.1). And different type preferences were not only observed between spontaneous and task-oriented speech, but notably also between the two spontaneous conversations: Most evaluative backchannels were found in the Introduction, while expressions of acceptance and alignment were uttered more frequently in the Discussion. This suggests again that, as the topic of the spontaneous conversation changes, the backchanneling behavior is adapted accordingly. In the Introduction, the participants get to know each other by talking about themselves (including topics such as work, hobbies and vacations, etc.), which invites evaluative and emotional reactions as well backchannels signaling surprise and interest. In the Discussion, they talk primarily about the Tangram game, exchanging thoughts about their strategies and whether they believed it was enjoyable, among other things. Results suggest that the specific backchannels uttered in this condition signaled acceptance and alignment more often than in the previous two conversations, indicating that the speakers were in agreement regarding their impressions of the game. During the Tangram task, on the other hand, context-specific backchannels were not only fewer, but also more neutral: evaluative and emotional signals such as ‘cool’/’ach cool’, ‘krass’/’ach krass’ and ‘nice’ were either not used at all or very rarely.

### 7.3.3 Summary: Type

Results from the analysis of backchannel types suggest that the conversational condition has an impact on the choice of backchannel types. Among the main generic BCs, ‘ja’ was the most frequent type in all three conversations, but it showed a higher rate in spontaneous speech compared to the task condition. The non-lexical type ‘mmhm’ was used primarily in task-oriented speech, while ‘mm’ showed the reverse pattern of being produced more in spontaneous speech and less during the Tangram task.

Multi-unit backchannels (MUB) showed low and steady proportions across conditions with no observable conversational effect, however, the most frequently used MUB type ‘ja ja’ was found almost exclusively in spontaneous speech.

Regarding the different use of context-generic and context-specific backchannels, results suggest a more frequent use of specific backchannels, which convey a reaction to the content of the previous utterance, in spontaneous speech. The

overall variety of backchannels is reduced in task-oriented speech in favor of a stronger preference for generic types. Moreover, specific BC types, due to their context-sensitive nature, appear to be adapted according to the context of spontaneous speech, as more signals conveying evaluative or emotional reactions tended to be produced in the Introduction, while more backchannels expressing acceptance and alignment were uttered in the Discussion.

## 7.4 Intonation

The analysis of intonation, measured in terms of pitch movement, across all generic BC types has shown a trend toward greater pitch movement and more intonational rises in task-oriented speech (Fig. 8) compared to both spontaneous speech conditions. Results for the two spontaneous speech conditions suggest slightly negative pitch movement, indicating more items with flat or falling intonation. Due to the low number of items of 'genau' and 'okay', as well as the large variety of types in the MUB category, complicating comparability, the type-specific analysis of intonation was carried out with those BC types for which the most data was available: 'ja', 'mmhm' and 'mm'. All three types showed greater pitch excursion in the Tangram condition, indicating that these backchannels tended to be produced with more pronounced intonational rises in that condition.

Results of the categorical analysis (see section 5.5), showing the proportions of rising, level and falling intonation for each BC type, confirmed that more rising contours were produced in task-oriented speech compared to the spontaneous conversations. Moreover, the results further affirm type-specific intonational differences, showing that 'mmhm' is produced mostly with rising intonation, while 'ja' shows mostly flat or falling contours, and 'mm' being positioned approximately in between. Irrespectively, the pattern of more rising contours in the Tangram condition appears to hold true for all three types. This confirms results by Janz (2022) and Spaniol et al. (2024), who had reported more rising backchannel intonation and greater pitch excursion in task-oriented speech, as well as results by Sbranna et al. (2022) on the general trend of backchannel intonation being type-specific.

Overall, the results of the continuous and categorical intonation analyses suggest that the intonational realization of backchannels is affected by the conversational condition. The task context appears to evoke more backchannels with rising intonation,

regardless of the backchannel type. Moreover, continuous measurements indicate that the rises in the task condition tend to be produced with greater pitch excursion.

Previous studies suggested that the contextual requirements of a conversation shape the use of backchannels. In a task-oriented context, the primary function of backchannels may be to establish common ground, with the listener signaling that he understood the message and that the speaker may continue. And since in such a scenario the speaker relies on the listener to have received and understood the message, the feedback will have to be distinct and clear. Studies have reported continuer backchannels, which fulfill this function, to be produced with rising intonation. And results on BC intonation from the present study show that more backchannels were produced with rising intonation in the Tangram condition compared to spontaneous speech. Therefore, an explanation for the increased proportion of rising backchannels could be that the collaborative task scenario led the participants to produce more continuer backchannels to successfully and smoothly navigate the conversation.

In addition to finding a greater proportion of intonational rises in task-oriented speech, it was observed that rises tended to be produced with larger pitch movements in this condition (Fig. 9), indicating that the task interactions evoked not only more but also *different* rises. One reason for this could be that the success and smoothness of the task-oriented conversations relies to a greater extent on clearly identifiable prosodically-encoded feedback than it is the case in spontaneous speech, where backchannels might primarily serve social and affiliative functions. Research into the cognitive and linguistic significance of pitch rises has demonstrated the importance of rises for the orientation of attention and therefore for successful communication more generally (Lialiou et al. 2024). This could explain why (in some cases) more distinct rises are produced during the Tangram game, where the speaker requires more informative feedback from the listener, while less rises are produced in spontaneous speech, where backchannels contribute to communicative success in functionally a different way.

It should be noted that dyad-specific behavior is an important factor to be taken into consideration. While the majority of dyads approximate the group pattern regarding the proportions of rising, level and falling intonation for the individual BC types, a few dyads stand out by showing distinct intonational preferences. Dyad 01

was the only dyad to produce ‘ja’ exclusively with falling intonation in the Tangram condition. Overall, this dyad produced only two rising instances of ‘ja’ out of 97 utterances of this type. Since ‘ja’ was the only BC type produced consistently in all conversations by this dyad, there is not enough data to check the intonation patterns of other types across conditions. While otherwise conforming to the group-level averages, Dyad 03 realized the non-lexical type ‘mm’ almost exclusively with rising intonation in spontaneous speech, opposing the group pattern of predominantly falling intonation in that condition. Dyad 06 produced no rising instances of ‘ja’, ‘mmhm’ and ‘mm’ in the Introduction. In the Tangram task, only a single utterance of ‘mm’ was realized with rising intonation. This dyad showed the highest rates of the non-lexical type in spontaneous speech. Dyad 13 realized almost all utterances of ‘ja’ in the Tangram condition with rising intonation and particularly large pitch excursion.

Interestingly, a majority of dyads whose intonational patterns were found to deviate from the group-average had already shown noticeably deviant type-choice patterns. This suggests that particular idiosyncratic behavior with regard to one of the analyzed dimensions of backchanneling behavior might be reflected in particular preferences in others as well. In general, the results of this study point to the relevance of individual variability in the analysis of backchannels.

### 7.4.1 Contour clusters

The explorative analysis of contour clusters, carried out with an application developed by Kaland (2021), has shed light on a few intonational nuances regarding the contour shapes and durations of backchannels in spontaneous and task-oriented speech. The analysis was performed on the BC types ‘ja’, ‘mmhm’ and ‘mm’, for which more data was available in the analyzed dataset, allowing for a more reliable intonation analysis.

It was found that, overall, a variety of contour types were used with each of the analyzed BC types. The lexical type ‘ja’ was realized in the vast majority of cases with two types of falling contours, differing primarily in terms of duration, with the shorter fall accounting for the largest proportion. In the task-oriented condition, the proportion of rising contours was higher. This increase was partly due to one fall-rising contour being produced more in this conversational condition.

The non-lexical types ‘mmhm’ and ‘mm’ showed different patterns, not only from ‘ja’, but also from one another: ‘mmhm’ was realized with rising contours almost

exclusively, while ‘mm’ was realized with mostly rises only in task-oriented speech. For both types there was an increase in the proportion of rising contours in the Tangram game. They differed, however, in terms of the preferred rising contour type they were produced with, as ‘mmhm’ showed a slight tendency towards an early rising contour, compared to the clear preference for a particular late-rising shape in ‘mm’.

These results further suggest that ‘mmhm’ and ‘mm’, despite their relative segmental similarity, should be treated as two categorically different BC types. Evidence from the type choice analysis (section 5.3) suggested that ‘mmhm’ is used more in task-oriented than in spontaneous speech, while the opposite is the case for ‘mm’. Taken together with the intonation analysis revealing different intonation patterns and contour shapes for each of them, this indicates that they potentially serve different conversational functions. This would be in line with Ward (2004), who observed that ‘mm-hm’ was used more often as a continuer than ‘mm’. Moreover, results from the current analysis indicate that ‘mm’ is the more adaptable type, showing a much stronger change in its intonational patterns across conditions.

Apart from confirming the trend toward more rising intonation in task-oriented speech, the results of the Contour Clustering show that the increased rate of rises is either partly or mainly reflected in more complex contour shapes in the Tangram condition, differing mainly in terms of the position and steepness of the rise. This could be seen as an effect of the different requirements of the task compared to spontaneous speech. Ward (2004) suggested that utterances of ‘mm’ involving more thought were correlated with longer durations, whereas shorter productions appeared to be more appropriate for lighter topics. Based on this observation, it could be hypothesized that backchannels involving more thought occurred more often in the more demanding task-oriented context, and that in addition to longer durations, more complex contours are evoked in this context.

Future studies should analyze the intonation contours of the BC types discussed here with attention to the particular contexts in which they are produced to shed light on the pragmatic functions they are each related to. More work should also be devoted to the analysis of speaker-specific (in addition to dyad-specific) behavior and how it might affect mutual understanding and social aspects of conversations.

### 7.4.2 Linear interpolation and contour clustering: A comparison

For this study, two different methods of analyzing and visualizing pitch contours were used: linear interpolation and contour clustering. In the former approach, two pitch points are taken from each backchannel token and the distance between them is calculated in semitones (ST) to determine whether the intonation rises, falls or is level, as well the range of the pitch movement. Taking two pitch points results in a linear description of the pitch trajectory, which inevitably ignores any dynamicity in between. In order to gain a better understanding of the potentially meaningful intonational characteristics of backchannels that would not be captured with linear interpolation, this study explored contour clustering (Kaland 2021) in the context of analyzing and visualizing backchannel intonation.

The linear interpolation method provides a simple but informative overview of the global pitch trajectory. It could be argued that reducing complexity to a certain degree is beneficial, as it reveals some of the more fundamental patterns, such as a basic tendency towards rises or falls for each of the analyzed backchannels, as well as a crucial condition-related effect of more rises in task-oriented speech. Complementing these results, the pitch excursion measures showed that this effect was not only a quantitative one – suggesting a higher quantity of rises alone –, but also a qualitative one, as rises were produced with greater pitch excursion in the Tangram condition. However, the linear interpolation method does not provide insights into the contour shapes apart from suggesting more and greater rises in one condition. The contour clustering method was useful in compensating this deficiency. Results from this method revealed that, despite their relative shortness, backchannels show a great deal of intonational dynamicity and complexity with regard to various parameters. In addition, categorizing the clusters allowed for a visualization of the proportionate use of each contour type across the different conversations, similar to the categorical results of the linear interpolation method, but with added information about the contour shapes (see figures 18 and 20). Interestingly, this showed that both methods provided matching results in terms of the proportions of rises and falls for each BC type and the trend of a greater proportion of rises in task-oriented speech.

Both methods presented drawbacks nonetheless: The linear interpolation of pitch contours conceals the dynamicity that backchannel intonation has been shown to have. Complex pitch movement such as falls and rises happening in between the sampled

pitch points, as well as information about their shapes, is ignored. As a result, rises and falls are involuntarily treated as uniform. Moreover, the category of level intonation, assuming a straight pitch trajectory, potentially contains instances of fall-rising or rise-falling intonation that are mischaracterized as flat. Results from the contour clustering seem to conform with the proportion of intonational rises found based on the linear interpolation method, which speaks to the reliability of both methods regarding this category. At the same time, the clustering approach revealed contours that showed the potential for being falsely identified as 'level' if linear interpolation was applied. This mainly concerns clusters (4) and (6) of 'ja', and cluster (4) of 'mmhm' and 'mm'. This suggests that the category of 'level' intonation should be interpreted with caution.

The contour clustering method captured the dynamicity that is present within signals as short as backchannels. It provided further understanding of backchannel intonation and how it is adapted according to the conversational condition, revealing that rises were realized in a variety of ways and that some shapes were used more than others. Despite the advantages of this approach, some limitations have to be pointed out: A considerable share of data points had been excluded due to errors in the sampling of pitch points. This problem could potentially have been resolved by making the algorithm less strict and/or lowering the number of measurement points. However, this would have come at the expense of obtaining less reliable f0 contours. While this may not present a problem if the units of analysis are longer phrases, there is arguably less room for error with utterances as short as backchannels, which sometimes present very limited amount of pitch information. The margin of potential mischaracterization might therefore be higher.

The contour clusters change considerably depending on the number of clusters deemed optimal. Therefore, the reliability and captured nuances of the represented contours hinge on finding the optimal cluster number through cluster evaluation methods (discussed in sections 6.6.1 and 6.6.2). It was found that the W/B cluster evaluation approach was useful in approximating a number of clusters suitable for the purpose of this explorative analysis. The MDL evaluation, on the other hand, had suggested a different cluster number as optimal. But the resulting contours conflicted with prior information (obtained from the linear interpolation method) about the proportion of rises and falls that would have to be expected for the analyzed backchannels. Thus, it was found that having a certain degree of prior knowledge of

the intonational characteristics of the object of analysis was at the very least helpful for choosing the appropriate cluster number.

In sum, the linear interpolation method is a simple and efficient way of capturing and visualizing some more global, but no less informative, intonational characteristics, while it lacks the resolution required to represent intonational nuances that proved relevant. The contour clustering method shed light on some of these nuances of backchannel intonation, including aspects of pitch and duration. In doing so, it advanced the notion of how backchannel intonation is at the same time type-specific and adapted to the conversational condition. It was found, however, that in order to determine the optimal number of clusters that would allow for an appropriate degree of complexity to be represented, results from the intonational analysis using the linear interpolation method were useful for making an informed decision. Therefore, a case can be made for the complementary value of the linear interpolation method, despite contour clustering having overall turned out as the more suitable method for more in-depth backchannel intonation analyses moving forward.

### 7.4.3 Summary: Intonation

Results of the intonation analysis have, overall, confirmed the type-specificity of backchannel intonation, as well as a trend towards more rising intonation and greater pitch excursion, regardless of BC type, in task-oriented speech. This suggests that backchannel intonation is affected by the conversational condition, potentially leading to backchannels fulfilling different functions, which is reflected in their intonation shifting towards the perceptually more relevant rises. The contour clustering analysis provided further evidence that the conversational condition has an effect on backchannel intonation, as all three BC types under investigation were found to be produced with a greater proportion of rising contours in task-oriented speech. Furthermore, the proportion of late-rising contours appeared to be increased in the BC productions during task interactions. The most noticeable change in intonational patterns across conversations was found for the non-lexical type ‘mm’, which appeared to be the most context-sensitive type. It was also the type for which the most dyad-specific patterns were found. The remaining types were produced with predominantly rising (‘mmhm’) or falling (‘ja’) intonation regardless of the conversational condition.

The results suggest, on the one hand, that the different backchannel types are realized not only with a preference for rises or falls, but also with a preference for specific contour shapes, and, on the other hand, that more complex rising contours are produced during task-oriented speech as a potential result of a function-related shift in the use of backchannels in this condition. In the task scenario, the speaker might require more distinct feedback signals to ensure a smooth and goal-directed flow of information, while anything other than a clear go-ahead signal may be interpreted by the speaker as a sign of a problem having occurred, therefore leading to potential disfluencies.

# 8 Conclusion

This thesis has aimed at providing a multi-dimensional analysis of the rate, duration type choice and intonation of backchannels and how each of these features respond to different conversational conditions in which speakers are able to have eye contact. In addition, a contour clustering method was used as a way of exploring backchannel intonation. For the experiment, 28 speakers grouped into 14 dyads were invited to perform a series of three conversations, including two spontaneous conversations and one task-oriented conversation. In the task condition, participants played the Tangram game, a joint-decision task in which they had to describe figures presented to them and come to a joint conclusion about whether or not their respective figures match. The participants' backchannel utterances were then annotated, labeled and analyzed.

The analysis of backchannel rate has shown that backchannels were used less frequently in task-oriented compared to spontaneous speech. All dyads conformed to this general pattern despite large dyad-specific differences regarding backchannel rate. In line with previous literature, this observation suggests that backchannels play a different role in task-oriented speech.

The mean duration of generic single- and multi-unit backchannels was slightly shorter than the durations reported in previous studies. Context-specific backchannels were found to be longer in duration than context-generic backchannels, which could be an effect of the former fulfilling different functions, such as providing expressions of interest and surprise, which had been found to correlate with larger durations in

previous studies. Durational differences across conditions were greater for context-specific single as well as generic multi-unit backchannels. This is, however, potentially a result of type-choice preferences, as different types of backchannels were used across conditions within these two categories, complicating a direct comparison. More data is needed on these backchannels categories to confirm this trend.

The conversational condition appeared to have an impact on the choice of backchannel types. While 'ja' was found to be the most frequent backchannel type in all three conversations, it was used less often in the task condition than in spontaneous speech. Conversely, the non-lexical type 'mmhm' was used primarily in task-oriented speech and less in spontaneous conversations. This trend can be attributed to its primary use as a continuer backchannel, which is likely a more common backchannel function in task-oriented speech, due the different contextual requirements compared to spontaneous speech. Another indication for this is the more frequent use of context-specific backchannels in spontaneous speech than in the task-oriented condition. The overall variety of backchannel types was shown to be reduced in the task setting in favor of a larger proportion of generic types. Backchannels which, on the basis of their form and lexical content, signaled social and affiliative functions occurred more often in spontaneous speech. This suggests that task-oriented speech evokes more backchannels with the more basic functions of managing and maintain common ground and mutual understanding. However, differences were not only found between spontaneous and task-oriented speech, but also between the two spontaneous conditions, as more signals conveying evaluative and emotional reactions were used in the Introduction, while more backchannels expressing acceptance and alignment were uttered in the Discussion, suggesting that the types of backchannels used vary across spontaneous conversations depending on the topic of conversation.

The analysis of intonation has shown that backchannel types have intrinsic intonation patterns, as 'mmhm' is realized almost exclusively with rising intonation, and 'ja' mostly with falling intonation, regardless of the conversational condition. In addition to type-specific intonation patterns, however, a trend toward greater pitch movement and more intonational rises was found in task-oriented speech. All three types that underwent intonational analysis were realized with more intonational falls in spontaneous speech and more rises in the task condition. The explorative analysis of backchannel intonation using contour clustering (Kaland 2021) has revealed further type-specific intonational patterns by capturing nuances in the backchannels'

intonation contours, as well as a tendency towards more complex, i.e. early- and late-rising, contour shapes in task-oriented speech, most notably in the case of 'mm', which showed an overall preference for late rises. This complements the observation that more rises are produced in the task setting with a more detailed grasp of how particular melodic and durational features of backchannels are adapted across conversational conditions. Crucially, these results shed light on type-specific intonational differences, showing that the trend of more rises in task interactions is realized differently in each BC type: 'mm', which was produced with more variable intonation regarding rises and falls, showed a clear predominance of complex rises in the task condition, whereas 'ja' and 'mmhm', which are predominantly falling and rising respectively, showed a subtle increase in complex rises. Overall, these results provide further evidence for the notion that intonational rises play an important role in speech, not only in phrases, but also in backchannels, and in particular in task-oriented speech, where clear and distinct rises may support the continuer function of backchannels, contributing to a smooth conversational flow. The contour clustering approach in particular has shed light on some of these potentially meaningful intricacies in backchannel intonation, underscoring its viability and usefulness for future backchannel intonation analyses.

## 8.1 Limitations and future outlook

Overall, the results of this analysis indicate that the conversational condition has an impact on all features of backchannels analyzed. The results also shed light on the intricate interplay between the rate, duration, type and intonation of backchannels. This study has implications for future analyses of backchannels, as it suggests, on the one hand, that backchannel use is sensitive to contextual factors, and, on the other, that none of the dimensions of backchannels analyzed should be looked at individually without considering the influence of other dimensions.

One of the principal limitations of this study has been the partial exclusion of the backchannels' pragmatic functions from the analysis. The backchannels analyzed here were not distinguished on the basis of their functions, such as continuer, agreement, and assessment, among other functions, as some previous studies have done. It was believed that, while it would have been insightful to do so, including a functional analysis would have exceeded the scope of this study. Previous studies reported a close relation of backchannel types, intonation and the conversational condition to pragmatic

functions. Therefore, it can be expected that the results of the present study may serve as a basis for future backchannel analyses focusing on pragmatic functions. Especially the results of the intonational analysis could provide a productive foundation for future studies to examine the intonational features linked to particular pragmatic functions.

The contour clustering approach has shown to be a promising tool, capable of capturing potentially meaningful details in the backchannels' intonation contours. Future studies could look further into the intonation-function relation of backchannel utterances, as well as extending research into other backchannel types and categories. Due to limited data being available for the generic backchannel type 'okay', contour clustering was not performed on it, despite it being used frequently by some dyads. Moreover, future studies could look into context-specific backchannels, focusing especially on the intonational features that might contribute to them being perceived as displaying more or less interest, surprise and sympathy, among other things.

Finally, this study has been of explorative and descriptive nature. Although some of the observations made indicate patterns clear enough to suggest statistical significance, a definitive test is still pending. Nevertheless, one of the most important takeaways of this analysis should be that individual variability is a key factor to be taken into consideration in the analysis of backchannel communication. While group-level results serve as a major source of information about general patterns, much can and should be learned from idiosyncratic behavior.

# Acknowledgements

I would like to thank Martine Grice and Simon Wehrle for their support, guidance and feedback throughout the process of conceptualizing and writing this thesis, Malin Spaniol for allowing me to be part of her project, sharing her audio recordings and always providing supportive advice, Constantijn Kaland for his help with technical and conceptual aspects of the contour clustering analysis, and last but not least Simona Sbranna, despite not being directly involved in this project, for helping me become a more resilient and confident person.

## References


Anderson, A. H., Bader, M., Bard, E.G., Boyle, E.H., Doherty, G.M., Garrod, S.C., (...) & Weinert, R. (1991). The HCRC map task corpus. *Language and speech, 34*(4), 351-366.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive science, 27*(2), 195-225.

Barth-Weingarten, D. (2011). Response tokens in interaction: prosody, phonetics and a visual aspect of German "jaja". *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion, 12*, 301-370.

Bavelas, J. B., Coates, L., & Johnson, T. (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

Beňuš, S., Gravano, A., & Hirschberg, J. B. (2007). The prosody of backchannels in American English. *Proceedings of ICPhS XVI,* 2007, Saarbrücken, Germany.

Beňuš, Š. (2016). The prosody of backchannels in Slovak. In *Proceedings of 8th International Conference on Speech Prosody*, 75-79.

Berry, A. (1994). Spanish and American Turn-taking Styles: A Comparative Study. In L. F. Bouton (Ed.), *Pragmatics and Language Learning*, (Vol.5, pp. 180-190). University of Illinois, Urbana- Champaign: Division of English as an International Language.

Berry, A. (2003). Are you listening? (Backchannel behaviors). Teaching Pragmatics. *US Department of State, Office of English Language Programs, Washington, DC.*

Bertrand, R., Ferré, G., Blache, P., Espesser, R., & Rauzy, S. (2007). Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*, 1-5.

Boersma, P., & Weenink, D. (2024). *Praat: doing phonetics by computer* [Computer program]. Version 6.3.09, retrieved 02 March 2024 from http://www.praat.org/

Cangemi, F. (2015). *mausmooth*. [computer program]. Retrievable online at http://phonetik.phil-fak.uni-koeln.de/fcangemi.html.

Caspers, J., Yuan, B., Huang, T., & Tang, X. (2000). Melodic characteristics of backchannels in Dutch map task dialogues. In *INTERSPEECH*, 611-614.

Clancy, P. M., Thompson, S. A., Suzuki, R., & Tao, H. (1996). The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of pragmatics, 26*(3), 355-387

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition, 13*, 127–149. https://doi.org/10 .1037/10096-006

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of memory and language, 50*(1), 62-81.

Clark, H. H. (2009). Context and common ground. In M. L. Jacob (Ed.), *Concise encyclopedia of pragmatics*, 116–119. Elsevier.

Cutrone, P. (2005). A case study examining backchannels in conversations between Japanese–British dyads. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication, 24*(3), 237–274. https://doi.org/10.1515/mult.2005.24.3.237

Cutrone, P. (2011). Politeness and face theory: Implications for the back-channel style of Japanese L1/L2 Speakers. *Language Studies Working Papers, 3*, 51–57.

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2023). Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General, 152*(3), 864.
Drummond, K., & Hopper, R. (1993). Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on language and Social Interaction, 26*(2), 157-177.
Duncan, S. (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, S., & Fiske, D.W. (1977). *Face-to-Face Interaction: Research, Methods, and Theory* (1st ed.). Routledge. https://doi.org/10.4324/9781315660998
Duncan, S., & Fiske, D. W. (1979). Dynamic patterning in conversation: Language, paralinguistic sounds, intonation, facial expressions, and gestures combine to form the detailed structure and strategy of face-to-face interactions. *American Scientist, 67*(1), 90-98.
Edlund, J., Heldner, M., & Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In *Nordic Prosody-Proceedings of the Xth Conference*, 57-68. Frankfurt am Main.
Freeman, V. (2019). Prosodic features of stances in conversation. *Laboratory Phonology, 10*(1).
Fries, C. C. (1952). *The structure of English*. New York: Harcourt, Brace.
Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In *the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, 2055–2060.
Gardner, R. (1997). The Conversation Object Mm: A Weak and Variable Acknowledging Token. *Research on Language and Social Interaction, 30*(2), 131–156.
Gardner, R. (2001). *When listeners talk*. Benjamins. https://doi.org/10 .1075/pbns.92
Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology*. New York: Irvington Publishers.
Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies, 9*(2–3), 205–217. https://doi.org/10.1007/BF00148127
Golato, A., & Fagyal, Z. (2008). Comparing single and double sayings of the German response token ja and the role of prosody: A conversation analytic perspective. *Research on language and social interaction, 41*(3), 241-270.
Gorisch, J. P. (2012). Matching across turns in talk-in-interaction: The role of prosody and gesture (Doctoral dissertation, University of Sheffield).
Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language, 25*(3), 601-634.
Ha, K.P., Ebner, S. & Grice, M. (2016). Speech prosody and possible misunderstandings in intercultural talk - A study of listener behaviour in Vietnamese and German dialogues. *Proc. Speech Prosody 8*, Boston, 801-805.
Hara, K., Inoue, K., Takanashi, K., & Kawahara, T. (2018). Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener, 162*, 364.
Harrigan, J. A. (1979). Relationship between the auditors' nonverbal behavior and turn-taking in social conversation. PhD thesis, University of Cincinnati.

Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. In *Interspeech 2010*. https://doi.org/https://doi.org/10.7916/D8WS92R4

Janz, A. (2022). Navigating Common Ground Using Feedback in Conversation - A Phonetic Analysis, MA thesis, University of Cologne, Cologne, Germany.

Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens “yeah”; and “mm hm.” *Papers in Linguistics, 17*(2), 197–216.

Jurafsky, D., Shriberg, E., Fox, B., Curl T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. *Proc. ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, 114-120.

Kaland, C. (2021). Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association*. doi:10.1017/S0025100321000049 [pdf]

Kaland, C. & Ellison, T. M. (2023). Evaluating cluster analysis on f0 contours: An information theoretic approach on three languages. In Radek Skarnitzl & Jan Volín (eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, 3448–3452. Prague: Guarant International.

Kaland, C. & Grice, M. (2024). Exploring and explaining variation in phrase-final f0 movements in spontaneous Papuan Malay. *Phonetica, 81*(3). doi:10.1515/phon-2023-0031

Keevallik, L. (2003). Terminally rising pitch contours of response tokens in Estonian. *Crossroads of Language, Interaction, and Culture, 5*, 49-65.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica 26*. 22–63.

Kjellmer, G. (2009). Where do we backchannel?: On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics, 14*(1), 81-112.

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

Kuhlen, A. K., & Brennan, S. E. (2010). Anticipating distracted addressees: How speakers’ expectations and addressees’ feedback influence storytelling. *Discourse Processes, 47*(7), 567–587. https://doi.org/10.1080/01638530903441339

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology, 6*, 731.

Li, H.Z. (2006). Backchannel responses as misleading feedback in intercultural discourse. *Journal of Intercultural Communication Research, 35*(2), 99-116.

Lialiou, M., Grice, M., Röhr, C. T., & Schumacher, P. B. (2024). Auditory processing of intonational rises and falls in German: rises are special in attention orienting. *Journal of cognitive neuroscience, 36*(6), 1099-1122.

Malisz, Z., Wodarczak, M., Buschmeier, H., Kopp, S., & Wagner, P. (2012). Prosodic Characteristics of Feedback Expressions in Distracted and Non-distracted Listeners. In *Proceedings of The Listening Talker. An Interdisciplinary Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions,* 36-39.

Mereu, D., Cangemi, F., & Grice, M. (2024). Backchannels are not always very short utterances. The case of Italian Multi-Unit Backchannels. *Journal of Pragmatics, 228*, 1-16.

Müller, F. E. (1996). Affiliating and disaffiliating with continuers: prosodic aspects of recipiency. *Prosody in conversation: Interactional studies, 12*, 131.

Neiberg, D., Salvi, G., & Gustafson, J. (2013). Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication, 55*(3), 451-469.
Neiberg, D., & Gustafson, J. (2011, August). Predicting Speaker Changes and Listener Responses with and without Eye-Contact. In *INTERSPEECH*, 1565-1568.
Oertel, C., Włodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). Gaze patterns in turn-taking. In *13th annual conference of the International Speech Communication Association*, Portland, USA, 2246–2249.
Oertel, C., Gustafson, J., & Black, A. W. (2016). Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Feedback Utterances. In *INTERSPEECH*, 2915-2919.
Peters, P., & Wong, D. (2015). Turn management and backchannels. *Corpus Pragmatics*, 408.
Pipek, V. (2007). On backchannels in English conversation (Doctoral dissertation, Masarykova univerzita, Pedagogická fakulta).
Poppe, R., Truong, K. P., & Heylen, D. (2011). Backchannels: Quantity, type and timing matters. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011*, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11 (pp. 228-239). Springer Berlin Heidelberg.
R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Reed, B. S. (2006). *Prosodic orientation in English conversation*. Springer.
RStudio Team. (2022). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/
Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50*, 696–735.
Sacks, H. (2004). An initial characterization of the organization of speaker turn-taking in conversation. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 35–42). John Benjamins Publishing Company.
Savino, M. (2010). Intonational strategies for backchanneling in Italian Map Task dialogues. In *Third ISCA workshop on experimental linguistics*.
Savino, M. (2014). The intonation of backchannel tokens in Italian collaborative dialogues. In: Vetulani, Zygmunt, Mariani, Joseph (Eds.), *Human Language Technology Challenges for Computer Science and Linguistics*. LTC 2011. Lecture Notes in Computer Science. Springer, Cham, 17-28.
Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing Discourse: Text and Talk, 71*, 93.
Schegloff, E. A. (2006). Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In S. C. L. N. J. Enfield (Ed.), *Roots of human sociality: Culture, cognition and interaction* (pp. 70–96). Berg.
Sbranna, S., Moeking, E., Wehrle, E., Grice, M. (2022). Backchannelling across languages: rate, lexical choice and intonation in L1 Italian, L1 German and L2 German. In: *11th International Conference on Speech Prosody*, 2022. https://doi.org/10.21437/SpeechProsody.2022-149.
Sbranna, S., Wehrle, S. & Grice, M. (2024). A multi-dimensional analysis of backchannels in L1 German, L1 Italian and L2 German. Language, Interaction, and Acquisition. *PsyArXiv*. DOI: 10.31234/osf.io/am248.

Spaniol, M., Wehrle, S., Janz, A., Vogeley, K., & Grice, M. (2024). The influence of conversational context on lexical and prosodic aspects of backchannels and gaze behaviour. *Proceedings of Speech Prosody 2024*, Leiden, The Netherlands.

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on language and social interaction, 41*(1), 31-57.

Stubbe, M. (1998). Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics, 29*(3), 257-289.

Tartory, R., Al-khawaldeh, S., Azieb, S., & Al Saideen, B. (2024). Backchannel forms and functions in context and culture: The use of backchannels in Arab media discourse. *Discourse Studies*. https://doi.org/10.1177/14614456241236904

Tolins, J., & Tree, J. E. F. (2014). Addressee backchannels steer narrative development. *Journal of Pragmatics, 70*, 152-164.

Tolins, J., & Fox Tree, J. E. (2016). Overhearers use addressee backchannels in dialog comprehension. *Cognitive Science, 40*(6), 1412–1434. https://doi.org/10.1111/cogs.12278.

Tolins, J., Namiranian, N., Akhtar, N., & Fox Tree, J. E. (2017). The role of addressee backchannels and conversational grounding in vicarious word learning in four-year-olds. *First Language, 37*(6), 648–671. https://doi.org/10.1177/0142723717727407.

Truong, K. P., & Heylen, D. K. J. (2010). Disambiguating the functions of conversational sounds with prosody: the case of 'yeah'. In *Proceedings of Interspeech 2010*, 2554-2557.

Tylén, K., Fusaroli, R., Smith, P., & Arnoldi, J. (2020). The social route to abstraction: Interaction and diversity enhance rule-formation and transfer in a categorization task. *Psyarxiv*. https://doi.org/10.31234/osf.io/qs253.

Ward, N. (2004). Pragmatic functions of prosodic features in non- lexical utterances, In *Proceedings of Speech Prosody*, 325-328.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition, 14*(1), 129-182.

Wehrle, S., & Grice, M. (2019). Function and Prosodic Form of Backchannels in L1 and L2 German. *Poster at Hanyang International Symposium on Phonetics and Cognitive Sciences of Language 2019*, Seoul, South Korea.

Wehrle, S. (2023). *Conversation and intonation in autism: A multi-dimensional analysis*. Berlin: Language Science Press. DOI: 10.5281/zenodo.10069004

Wong, D., & Peters, P. (2007). A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification. *International Journal of Corpus Linguistics, 12*(4), 479-510.

Yngve, V. (1970). *On getting a word in edgewise*. In Chicago Linguistics Society, 6th meeting (pp. 567–577).

Young, R. F., & Lee, J. (2004). Identifying units in interaction: Reactive tokens in Korean and English conversations. *Journal of Sociolinguistics, 8*(3), 380-407.

Zellers, M. (2021). An overview of forms, functions, and configurations of backchannels in Ruruuli/Lunyala. *Journal of Pragmatics, 175*, 38-52.

# Appendix

This appendix contains further information, tables and plots that were excluded from the main text for reasons of space and stringency.

| All ‚other' types | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ah | ach so | ah okay | cool | stimmt | natürlich | das stimmt | gut | krass | nice |
| ah ja | ja ja klar | voll | ach cool | ja stimmt | ja das stimmt | ach krass | ja auf jeden Fall | klar | schön |
| perfekt | ah ja okay | ja cool | stimmt ja | ach | ah cool | chillig | ja gut | richtig | yes |
| ach geil | ach nice | ach schön | ach so okay | ach witzig | ah chillig | ah gut | ah krass | ah nice | eben |
| ja ja das stimmt | ja klar | ja richtig | ja voll | lustig | oh schön | ok krass | absolut genau | ach du schande | ach echt |
| ach ja stimmt | ach krass ok ok ok | ach lustig | ach so crazy ok | ach so ja | ach so ja ja | ach toll | ah ach krass ja | ah cool ja | ah ja genau |
| ah ja ja | ah ja ja ja | ah ja krass | ah ok ok ok ok ja | ah okay ja | ah okay ja gut | ah ok jetzt versteh ich | ah ok ok ja ja | ah ok spannend | ah verstehe ja |
| ah wie cool | alles klar | bin bei dir | boah | boah krass | cool wow | das ist krass | das stimmt ja | doch auf jeden Fall | das glaube ich |
| fantastisch | geil | glaub‘ ich auch | gut ja | gut ja ja | hab‘ ich auch | ich auch | is‘ so | ist anstrengend ja | ja chillig |
| ja gut ja | ja gut ne | ja ich auch ja | ja ist echt so | ja ist gut | ja ja ja das stimmt | ja ja ja gut | ja ja ja richtig genau ja | ja ja klar das macht sinn | ja ja voll voll voll voll |
| ja kenn‘ ich ja | ja könnte hinkommen | ja krass | ja macht sinn | ja nice | ja ok gut gut gut gut | ja richtig genau | ja safe | ja schön | ja spannend |
| ja stark | ja voll voll | ja witzig | kann sein | kenn‘ ich | korrekt | mm ok ich verstehe | mmhm stimmt | na ja richtig das stimmt | naja |
| oh | oh cool | oh gott | oh krass | oh nice mmhm | oh schön ja | oh wie cool | ok gut | ok gut gut gut | ok gut gut gut gut gut |
| ok perfekt | ok perfekt perfekt perfekt | ok ich verstehe | okay perfekt | richtig genau | richtig ja | richtig richtig genau | sehr gut | sehr gut okay | super |
| sweet | uh | verstehe | voll gut | wahrschein-lich ja | what | witzig | | | |

**Table A1**: List of all ‘other’ (context-specific) BC types produced across dyads and conversations. Ordered alphabetically and from most to least frequent. All types from ‘absolut genau’ to ‘witzig’ occurred only once in the dataset.

| All MUB types | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ja ja | ja genau | ja okay | ja ja ja | ja ja ja ja | ja ja genau | ok ok ok ok ok | mmhm ja | mm ja | genau ja |
| okay okay | okay ja | ok ok ok ok | okay okay okay | ja genau ja | ja ja ja ja ja ja | ja ja klar | genau ja ja | ja genau genau | ok ok ok ok ok ok ok ok |
| ok ok ok ok ok ok ok | ok ok ok ok ok ok | ja genau okay | ok ok ok | ok ok | ok ja ja ja | mmhm okay | mmhm mmhm genau | mmhm mmhm | ja ja genau genau |
| mm ok ok ok ok | ja ja ja ja ja | ja okay okay | ja ja ja ja ja ja | ja mm | ja ja okay mm | ja ja okay | genau genau genau | | |

**Table A2**: List of all multi-unit backchannel (MUB) types produced across dyads and conversations. Ordered from most frequent ('ja ja') to least frequent ('genau genau genau').

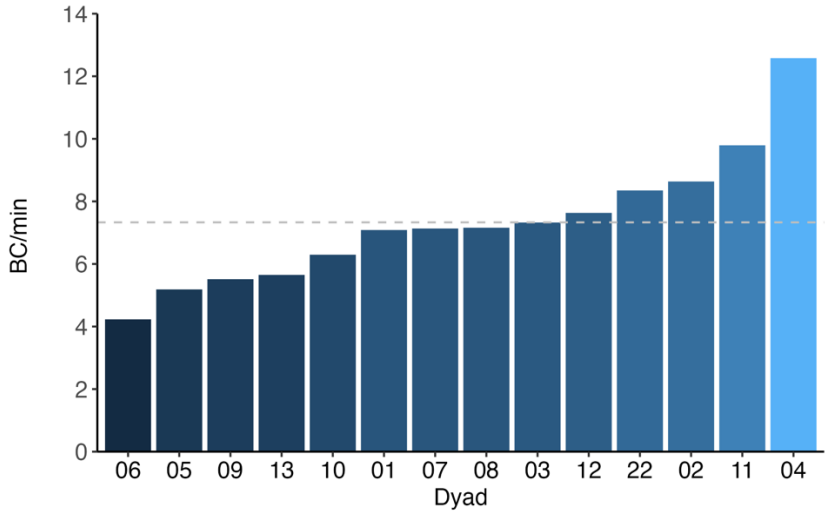


**Fig. A1**: Total rate of BCs per minute by each dyad. Dashed horizontal lines indicates the mean of 7.33 BCs/min

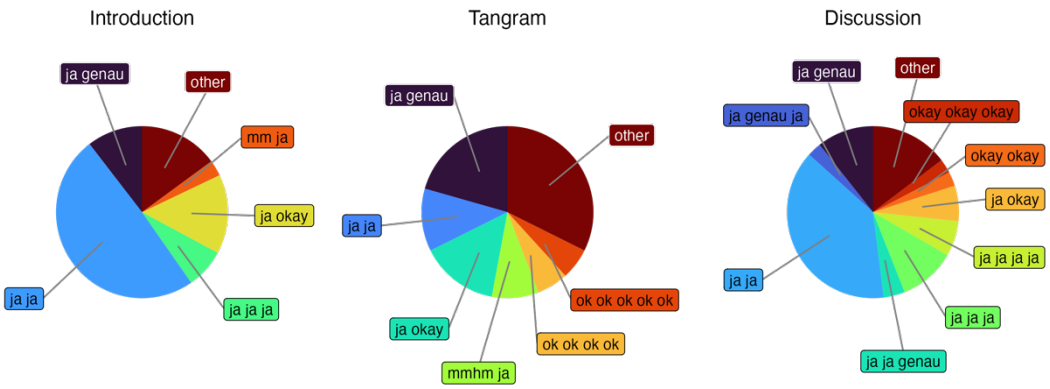


**Fig. A2**: Pie plots showing the proportionate use of MUB types across the three conditions. "other" refers to remaining types that occurred less than twice in each condition.


**Fig. A3**: Percentage of contour categories *rise*, *level* and *fall* within the BC types 'mmhm', 'mm' and 'ja' by dyad across conversational conditions.

**Abschlussarbeit – Philosophische Fakultät**

**Vorlage Eidesstattliche Versicherung:**

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche unter Angabe der Quelle kenntlich gemacht. Diese Arbeit ist in gleicher oder ähnlicher Form im Rahmen einer anderen Prüfung noch nicht vorgelegt.

Köln, 25.11.2024 Unterschrift: