REGISTER
Aspects of
Language
in Situation

Turn-taking in video-mediated and co-present dialogues

A corpus-based study of German

Qiang Xia[a]
[a]Humboldt-Universität zu Berlin

Conversational turn-taking is a well-practiced activity in our daily life. Although the use of video-conferencing tools like Zoom has increased enormously in the past few years, it is still unclear whether interlocutors adapt their turn-taking behaviours to the situational change from co-presence to virtual communication. The exploratory study thus aims to compare several turn-taking behaviours (e. g. turns, backchannels, gaps and overlaps) in these two situations. Spontaneous dialogues in the Berlin Dialogue Corpus (Belz et al. 2021) were investigated. Twenty native German speakers were given 10 minutes in each situation to complete two spot-the-difference tasks in pairs. The results show that Zoom conversations have lower articulation rates than face-to-face conversations, and have less turns and backchannels, longer gaps and overlaps. Interestingly, more overlaps are found in face-to-face interactions. Speaker transition time in German conversations demonstrated bi-modal distribution in both situations, contrary to earlier studies on other languages. The temporal differences may be partially explained by transmission delay. More and longer overlaps are thought to be associated with a speech register employed in collaborative tasks and in communication within close relationships.

Keywords: Turn-taking, spontaneous speech, Zoom interaction, multimodality, corpus study

1 Introduction

Conversation is one of the most common and frequent activities in human communication (Mehl et al. 2007, Levinson & Torreira 2015). It is a well-practiced behaviour in which interlocutors cooperate and take turns speaking in a smooth way, usually without noticeable acoustic silence in-between or overlapping with each other (Sacks et al. 1974). Yet recent studies have proposed that the suggested

BerlinUP Journals
Qiang Xia. 2023. Turn-taking in video-mediated and co-present dialogues — A corpus-based study of German. *Register Aspects of Language in Situation (REALIS)*. 2(1), pp. 1–29, doi=https://doi.org/10.18452/27484
ISSN 2750-9370

no-gap-no-overlap pattern in turn-taking should not be understood in a strict sense. Empirically, it has been demonstrated that unnoticeable silences in conversation last about 200 milliseconds on average in English, and they differ from one language to another with the mean duration within the range of 0-250 ms (Wilson & Wilson 2005, Stivers et al. 2009, Heldner & Edlund 2010). At the same time, overlaps are not rare in a conversation. According to Levinson & Torreira (2015), overlapping speech accounts for 3.8 % of the total length of a conversation and 30.1 % of all turn transition cases.

Turn-taking behaviour seems to have been extensively researched from both theoretical and empirical perspectives. However, it is important to point out that our knowledge of this conversational behaviour is based primarily on face-to-face situations. Over the past few years, our use of video-mediated communication has increased significantly. Some recent studies indicate that the differences found in the online form may result from the unavoidable transmission delay (Boland et al. 2021) or the increased difficulty in sending and receiving (para)linguistic signals (Bailenson 2021) in online conversation, which are considered indispensable for turn-end anticipation (Duncan 1972, Kendon 1967, Gravano & Hirschberg 2011). But still little is known about how the turn-taking mechanism in online conversation differs from that in face-to-face situations.

Therefore, the aim of the current study is to compare turn-taking behaviours in these two situations with respect to the count of turns and backchannels, gap and overlap durations, as well as speakers' articulation rate. Twenty co-present and twenty Zoom German conversations in the Berlin Dialogue Corpus (BeDiaCo, Belz et al. 2021) were investigated.

2 Background

2.1 Terminology of turn-taking

Sacks, Schegloff & Jefferson were some of the researchers who first focused on the riddle of "speech exchange systems" (1974: 696). Ever since then, the mechanism of conversation has received more and more attention, leading to a number of studies related to the components of a conversation and their temporal sequences.

Although the topic of turn-taking has been intensively researched, an uncontroversial definition of *turn* has not yet been established. The general understanding of a *turn* is the speech held by a speaker before a speaker-change takes place. It is also figuratively described as a conversational *floor* momentarily taken by the speaking party (e. g. Sacks et al. 1974, Edelsky 1981). In practice, ten Bosch

et al. technically identified "stretches of one or more utterances that are not interrupted by another speaker" (2005: 82) as a turn. However, there is no further illustration in their work as to when a stretch of speech is to be considered as "interrupted". In the following text, *turn* is still used as a general concept to describe a speaker's speech before a speaker-change takes place in a conversation.

Another term that is frequently used but lacks a precise definition is backchannel. In contrast to the *mainchannel*, where the current speaker holds the floor, the listening party of the conversation usually gives some verbal and nonverbal feedback in the *backchannel*, for instance, head nodding or giving short messages like "uhm", to indicate that they are paying attention to the ongoing speech. Yngve (1970) first coined the term "backchannel" to describe these kinds of feedback messages. Despite the introduction of this term, many researchers still treat these brief verbal feedback as a minimal version of turns (e. g. Sacks et al. 1974, Heldner & Edlund 2010, Levinson & Torreira 2015).

Turns and backchannel responses have mostly been investigated separately in the turn-taking research field, though sometimes with the same focus, such as on their anticipatory mechanism (e. g. turn: Levinson & Torreira 2015, backchannel: Heinz 2003), realisation forms (e. g. turn: Sacks et al. 1974, backchannel: Dideriksen et al. 2019), multi-functionality (e. g. turn: Fusaroli & Tylén 2016, backchannel: Peters & Wong 2014). But they are still rarely discussed on the same page. Therefore, it is interesting to investigate whether and how turns and backchannels interact with each other.

Reading the temporal aspect of a conversation, the turn-taking relevant units, for example, turns and backchannels, can follow each other either smoothly, with acoustic silence or with overlaps. In previous studies, the identification of a silence and an overlap has mainly been based on acoustic signals and is thus less controversial than determining turns or backchannels. But there are some alternative terms for the same temporal interval in a conversation.

Sacks et al. divided acoustic silences in dialogue into three groups based on its surrounding context: Pauses are short silences within a turn, while gaps and lapses can only be found when speakership has changed, and gaps are shorter than lapses (1974: 715). In adherence to these terminologies, Heldner & Edlund (2010) added "between-speaker intervals" as a cover term for silences and overlaps between speech from different interlocutors. In addition, between-speaker intervals are also mentioned as *turn transition offsets* (Levinson & Torreira 2015, de Ruiter et al. 2006).

The concept of overlap is relatively unambiguous as well. In the literature, between-speaker silences and overlaps are often illustrated as two ends of a continuum (Stivers et al. 2009, Heldner & Edlund 2010, Levinson & Torreira 2015):

When measuring the temporal interval between the ending point of the first speaker and the starting point of the second speaker, gaps have positive values and overlaps negative values. Therefore, overlaps can also be referred to as *negative floor transfer offsets* (de Ruiter et al. 2006) as well as *interruption*, *simultaneous talk* or *double talk* (Schegloff 2000).

2.2 Turn-taking is time-sensitive

The mechanism that ensures a smooth conversation is complicated. Sacks et al. (1974) concluded that human conversation overwhelmingly follows the *one-speaker-at-a-time* pattern where transitions with no gap and no overlap are common. A speaker's turn consists of turn-constructional unit(s), usually abbreviated as TCU, which can take various shapes, from a single word to a sentence or even longer. When a TCU is about to end, it reaches a transition relevant place, where the speaker can either self-select or allocate the floor to the next speaker. Since neither order nor length of turns are fixed or specified in advance, a *more-than-one-speaker-at-a-time* scenario happens on occasion, but very briefly. Different techniques are applied to repair these turn-taking mistakes, such as stopping talking prematurely. Thereby, the one-speaker-at-a-time pattern is soon restored.

Among the factors involved in the allocation of the floor, time plays an important role. Wilson & Wilson (2005) demonstrated that the timing of turn-taking in ordinary conversation is highly precise. They studied the turn transition phenomena and proposed that listeners' readiness to speak counterphased with that of the speaker. The speech coordination is achieved by the intrinsic oscillators in the brains of interlocutors, which become entrained with each other based on speakers' syllable rates, approximately 200 milliseconds per syllable in informal English speech (2005: 962). Therefore, interlocutors are able to minimise the likelihood of simultaneous speech commencements and sustain the *one-speaker-at-a-time* dynamics.

Heldner & Edlund (2010) challenged the classic conversational model which indicates the overwhelming dominance of the *no-gap-no-overlap* pattern in turn-taking behaviour. They argued empirically that the real turn-transition time is rarely zero and is not as precise as suggested, but more distributed. Categorising the between-speaker intervals from -10 ms to 10 ms as *no-gap-no-overlap*, they found that only 0.4 % to 0.7 % of the turn transitions could be counted as *smooth* (2010: 562-563). Later, Heldner (2011) argued that 120 ms is the more reliable perception threshold for listeners to detect gaps and overlaps. Levinson & Torreira (2015: 13) criticised their extremely strict interval ranges, since 10 ms is

not realistic for human performance. If the range were set to 200 ms, the majority of their speech data would have been *no-gap-no-overlap*. As can be seen from the debate, the *one-speaker-at-a-time* pattern is rather a conceptualised model of conversation, which is not necessarily equivalent to literally zero millisecond gaps or overlaps, but depicts the most prominent characteristics of natural dialogues to a great extent. Brief gaps or overlaps within 200 ms often co-occur with turn transitions, at least for Dutch, Swedish and English (Heldner & Edlund 2010, Levinson & Torreira 2015).

Turn-taking behaviour does differ from one language to another. Stivers et al. (2009) found that a language has its own conversational style in terms of the appropriate timing of turn transitions. They observed that Danes and Laos prefer longer transition times, while quicker responses are desired in Japanese. On the other hand, some cross-linguistical phenomena can be found. All 10 languages investigated share a uni-modal turn-taking distribution with the most frequent offset time between 0-200 ms, and a mean offset time of about 250 ms averaged across languages. The tendency to avoid overlaps and silence between speech turns is thus considered to be cross-linguistically valid.

Quantitative data on the temporal aspects of German turn-taking behaviours is still relatively scarce. A distribution analysis of gaps and overlaps in German turn transition has only been reported in Weilhammer & Rabold (2003), which also investigated English and Japanese. They examined natural conversation between German native speakers in a scenario of planning a business trip and observed Gaussian distributions in the logarithmic domain of durations. The geometric mean of gap duration is 363 ms, and the geometric mean of overlap durations is 331 ms.

It is worth pointing out that the research on the temporal aspect of turn-taking mentioned here, such as Weilhammer & Rabold (2003), Heldner & Edlund (2010) and Levinson & Torreira (2015), has mainly taken those speech chunks which are automatically detected and segmented based on audio signals as "turns". That is to say, every audible interval separated by silences is identified as a speech turn, regardless of its syntactic, semantic or pragmatic function in the conversational context. According to the definition in Sacks et al. (1974), it is rather the *turn-constructional units* than the *turns*, in which syntactic and semantic completeness plays an important role, that are investigated.

Whichever category is used to classify turn and whichever unit is used as the object of the temporal investigation, the fact that speakers share a social norm requiring them to take a turn at the right moment does not change. The reason why the timing of turn-taking matters is believed to be associated with its pragmatic

and social functions. Replying too soon or too late are both considered inappropriate. Heritage (1984: 267-268) discussed different cases of accepting invitations from the aspect of conversation analysis and described an early acceptance with slight overlap as normal; however, a delayed acceptance or an early refusal may be interpreted as reluctant, rude and even hostile. In social communication, a delayed response, for example, a noticeable silence, is usually associated with refusal or interpreted as an indication of a dispreferred response (Riest et al. 2015, Robinson 2020). It is the social understanding that makes taking the turn at the right time particularly important.

2.3 Signals are important for turn-taking

Given the time-sensitivity of turn-taking, the anticipation of upcoming turn-endings is crucial. Researchers have used different approaches to explain the mechanism of turn-end anticipation.

Sacks et al. (1974) saw syntactic and prosodic completion as the most important projecting indicator for a turn-end. They believed that listeners use contextual and structural information of the current turn as a basis to project its end and prepare for their own turn. Several studies have examined how interlocutors rely largely on syntactic structures (Selting 1996, Auer 2005) and semantic information (Riest et al. 2015) to predict a possible turn-end.

Different from the *projection* approach, some researchers have demonstrated that interlocutors make use of a spectrum of signals to anticipate when the current speaker is about to complete the turn. The more available cues there are, the higher the probability is that a turn transition is going to take place (see Kendon 1967, Yngve 1970, Duncan 1972, Gravano & Hirschberg 2011).

Both linguistic and paralinguistic cues contribute to a correct anticipation of turn transition. Gravano & Hirschberg (2011) investigated a long list of prosodic and acoustic parameters and observed that, for example, falling or high-rising intonations and a lower pitch level at the end of interpausal units, signal the coming turn switching. Local & Walker (2012) put their focus on phonetic features and confirmed that reduced consonants and vowels, continuation of voicing and avoidance of durational lengthening at the transition relevant place contribute to projecting a turn-end. Koiso et al. (1998) emphasised the role of prosodic features (e. g. rising F0 patterns and decreasing sound volume) in signalling turn changes in Japanese.

Some paralinguistic cues make major contributions to turn-end anticipation as well. Kendon (1967) observed that speakers used gaze direction to convey multiple pieces of information: When speakers avert their gaze, they are concentrating

on utterance organisation, signalling their intention of holding the floor. As the end of the turn approaches, speakers are very likely to look towards the listener again. Holler & Kendrick (2015) investigated gaze movement in multi-person interactions. They observed that unaddressed participants will shift eye gaze to the next speaker right before the end of the current turn. A recent study involving eye-tracking method found similar results (Auer 2021): Gaze selection would frequently lead speakership to the gazed-at person in multi-party settings, regardless the usage of second person pronouns. Hence, the study claimed gaze as the most ubiquitous speaker-selection technique.

In addition, Rochet-Capellan & Fuchs (2014) argued that breathing may help dialogue partners to coordinate their turns. Inhalations are found before most turns, while smooth and interrupted transitions show different profiles of alignment to partner breathing. Body movements including noding and hand movement (Holler & Levinson 2019), smiles and facial expressions (Brunner 1979, Kaukomaa et al. 2013) have been shown to be functionally beneficial for turn-end anticipation.

2.4 Zoom-meetings are special

Most studies mentioned above treat face-to-face situations as the default conversational setting, where interlocutors are able to perceive both verbal and non-verbal cues directly. A video-conference however constitutes more than just a change of the conversational partner from a person to a screen. It differs from a co-present situation in many ways.

Firstly, cues that are easy to perceive in a face-to-face situation may be hardly noticeable or even totally absent in a video-conference. Bailenson (2021) pointed out that interlocutors have to work harder to send and receive cues. From the perspective of cue senders, they had to nod exaggeratedly for a few seconds to make sure that their agreement had been seen. They shifted their eye gaze consistently between the camera and the faces on the screen, but they failed to have *direct* eye contact due to the limited situation. Their unsuccessful attempt to make *direct* eye contact with their conversational partner increased however their cognitive load. This observation was also made in the study of Seuren et al. (2021), who demonstrated that the lack of direct eye contact was associated with unintended interruption and extensive silence between speech turns. From the perspective of cue receivers, Bailenson (2021) argued that the excessive amounts of close-up eye gaze might be stressful. Using the default speaker view in Zoom with a typical laptop configuration, the size of the speaker's face was about the size one would see when standing face-to-face only 50 cm away. Such a close

interpersonal distance was so intimate that it would normally only be reserved for families or loved ones. Hence, the close conversation might cause discomfort and social stress in a one-to-one video meeting. However, the only possibility to show respect or to catch cues related to turn-taking is for interlocutors to stare at the screen throughout the meeting, leading to the so-called "Zoom fatigue" (Bailenson 2021).

Secondly, latency in online conversation is an objective factor that cannot be ignored. Seuren et al. (2021) suggested that the observed latency during the interaction hampered the turn-taking behaviour. Silence occurred when speech was expected. As a result, speakers started at the same time and then stopped soon after, for the sake of the *one-speaker-at-a-time* pattern. Boland et al. (2021) observed simultaneous starts and mutual silence as well. They explained that latency disrupted the natural rhythm of turn-taking, which usually synchronises on the syllable rate (see Wilson & Wilson 2005). Egger-Lampl et al. (2010) investigated the impact of network delay on perceived speech quality and conversational interactivity. The more interactive a conversation is, the more sensitive interlocutors are to delays and the worse the perceived speech quality is. As the delay length increases up to 800 ms, the number of unintentional interruptions grows steadily.

Even without technical problems or inconsistent internet connections, it still takes roughly 30-70 milliseconds for audio transmission in a very ideal situation, as estimated in Boland et al. (2021). A precise measurement of the exact technical delay is not known to the public, unfortunately. Though brief, 30 to 70 ms is still long enough to disturb the oscillators entrained during the conversation. Jones (2019: 139) found that as long as the violations of fuzzy rhythms are within 15-20 % of an oscillator's period, the entrainment can still accommodate. In other words, for a speaker speaking with an averaged rate of 200 ms per syllable, it is not hard to catch the rhythm when the actual syllable rate varies from 160 to 240 ms, since 40 ms is 20 % of 200 ms, the maximum violation value based on the entrained oscillation period. Thus it is reasonable to assume that it is highly possible for an electronic transmission delay of 30-70 ms to disturb the oscillatory pattern that speakers have entrained during the conversation. To sum up, both temporal aspects and signal aspects can contribute to the differences between the video-mediated and co-present situations.

2.5 Research questions

Given the unavoidable transmission delay in online conversation and the increased time for communication, it is reasonable to assume that (H1) Zoom con-

versations have fewer turn transitions than in co-present situations. In a similar vein, (H2) there may be fewer backchannels in Zoom conversation, since speakers would probably confuse a backchannel with a turn, causing unnecessary misunderstandings. For example, when a backchannel is misunderstood as the beginning of a new turn, both parties might relinquish the floor without finishing their turn. In general, (H3) a positive correlation between the count of turns and backchannels is expected in both situations.

Due to these uncertainties in Zoom conversations, (H4) speakers might slow down their articulation rate, so that their conversational partner can better comprehend conveyed information.

In addition, face-to-face situation might have (H5) shorter turn transition gap durations and (H6) shorter overlap durations than Zoom situation where speakers' ability to anticipate the turn-end and the rhythmical entrainment might be hampered by the irregular transmission delay (Wilson & Wilson 2005, Boland et al. 2021). It is reported in Egger-Lampl et al. (2010) that transmission delays resulted in more unintended interruptions in a scenario demanding high interactivity. In collaborative tasks that require frequent information exchange, such as the Diapix tasks employed in the experiment, (H7) more overlaps are expected in Zoom situations where delays can hardly be avoided.

Apart from the hypotheses listed above, an open question remains to be answered as well. Whether speakers adapt their conversational behaviours during the task, will also be examined in the study.

3 Methods

3.1 Corpus and experiment

The subcorpus *Videocall* from the Berlin Dialogue Corpus Version 2 (BeDiaCo, Belz et al. 2021) was used in the current study.

It contains 104,000 word tokens from 40 dialogues between 20 German native speakers (mean age = 25.7, SD = 3.8, 10 females, 10 males) conversing in pairs. Due to the pandemic situation and the strict hygiene regulations in 2020, recordings could only be collected from people living together. All dyads knew each other well prior to the experiment; 6 were heterosexual couples, 2 were brothers, and 2 were female roommates. They were asked to participate in two Diapix tasks (Van Engen et al. 2010, Baker & Hazan 2011, Bullock & Sell 2022) in each of the following conditions: over Zoom and face-to-face. During the task, each participant received one of two nearly identical images (see Figure 1). They had about 10 minutes to find the 10-13 differences between the images, without seeing the

counterpart in their partner's hand. This setting constructs an optimal situation to elicit naturalistic and spontaneous conversation.

In the face-to-face situation, participants sat across from each other in the phonetics laboratory of the university. For Zoom conversations, participants were guided into two adjoining rooms (the phonetics laboratory and the adjacent office), connected via Zoom installed on two tablets. Zoom was utilised only for communication and not for recording. There are two reasons for using extra microphones in the experiment. Firstly, the tablets used in the experiment do not allow connections with additional microphones, and the built-in microphones on the tablets did not meet the precision requirements necessary for subsequent phonetic analysis. Secondly, conversations conducted over Zoom inherently introduce varying levels of latency for all involved parties, depending on real-time internet connectivity at that given moment. Considering these two facts, recording with external microphones was preferred. This methodology can put the perceived audio signals on an external timeline, rather than relying on a singular source whose latencies fluctuate consistently. Given these considerations, the speech of each participant was recorded separately by using different microphones in both situations. The collected speech data were stored and processed using the same computer. As the corpus is aimed at investigating spoken languages, video was not recorded.

Based on the post-experiment questionnaire, among the 20 participants, 13 reported using Zoom on a "daily" or "weekly" basis, the remaining seven either "monthly" or "never". Yet, 15 participants felt "comfortable" or "very comfortable" during the Zoom conversations, five "neither comfortable nor uncomfortable" (Belz et al. 2021).

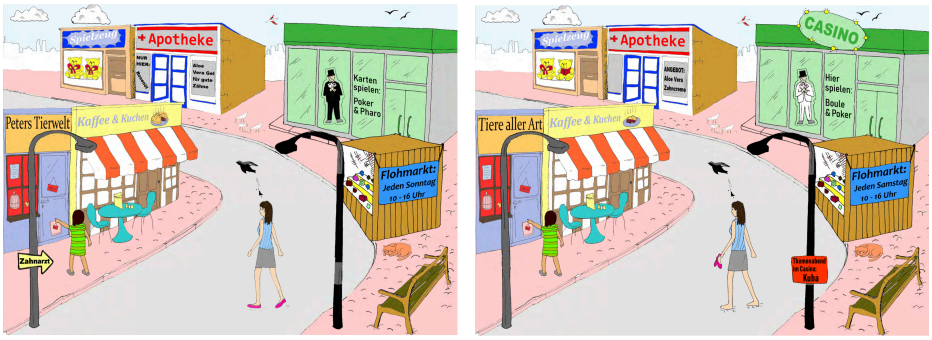

Figure 1: A picture pair from the Diapix task used in BeDiaCo v2 (Belz et al. 2021)

3.2 Corpus annotation

For the research questions posed in Section 2, an annotation scheme was developed (for the full documentation, see supplementary file 6). Several annotating systems used in prior work (Belz et al. 2021, Gravano & Hirschberg 2011, Heldner & Edlund 2010, ten Bosch et al. 2005) were referenced. Turns, backchannels and TCUs were manually annotated on two separate levels. The TCU-level was subordinate to the turn-level. On the turn-level, a speech chunk can either be a *turn* or a *backchannel* (shorter than 1 second). On the TCU-level, more values are possible: *polar question, w-question, answer, description* etc. Gaps and overlaps were automatically detected, based on the TCU level. Temporal aspects of turn-taking behaviours, such as gap and overlap durations, are discussed on the TCU level rather than the turn level. It is in accordance with previous studies that took automatically detected speech chunks as the basis to identify silence and overlap (Weilhammer & Rabold 2003, Heldner 2011, Levinson & Torreira 2015).

3.3 Data preparation

Given that the actual length of each conversation differs, the occurrences of turns and backchannels per minute were additionally counted by dividing the occurrences by the speech length in minutes.

In order to compare an individual's articulation rate in different situations, realised syllables were counted based on the manual transcription of speech, by using the R-package Sylly.de developed by Michalke (2018). Pauses within an utterance were subtracted before its duration was divided by the sum of syllables produced within it. Then, the articulation rate was calculated by diving the sum of syllables with the total speaking time of a speaker.

Some extremely long gaps have been removed. It is observed in the annotating practice that seconds of silence are common in the first and/or last minute of the dialogues, because participants are still trying to adjust themselves to the new conversational settings or are hesitant about when they should start or end the task. For this reason, outliers of gap durations (calculated by the upper quartile of the data plus 1.5 times of its interquartile range, Q3+1.5IQR) were not considered.

After the data were transformed into an Emu-Database (Winkelmann et al. 2017), the duration of annotation units could be obtained. As suggested in previous studies (Heldner & Edlund 2010, de Ruiter et al. 2006), overlaps were treated as negative floor transfer offsets, and calculated by 0 minus the measured duration to get the corresponding negative values. Similar to long gaps, long overlaps below the lower bound (Q1-1.5IQR) were omitted.

To investigate whether the specific task setting has any influences on the turn-taking behaviour, for example, speakers might slow down after having spotted obvious differences, causing longer gaps and fewer overlaps, the relationship between these phenomenon and task time span was studied. The time span of each dialogue was normalised by dividing the start time of a gap by the whole length of the dialogue. The normalised time point was then rounded to one decimal place. For each normalised time point, there is possibly more than one instance across the whole corpus. Gap durations whose starting points fall within the range were averaged. In a similar vain, overlap occurrences were summed up for each normalised time slot.

3.4 Statistical analysis

In order to examine whether the face-to-face and Zoom situation has influences on the speakers' turn-taking performance (e. g. longer gaps and more overlaps), separate mixed-effects linear regression models were fitted for the tested variables using the LME4 package (Bates et al. 2015) in R (R Core Team 2022).

4 Results

4.1 Turn and backchannel

In total, there are 2640 speech turns in the face-to-face and 2375 turns in the Zoom interactions. The count per minute amounts to 17.9 and 15.5 for the two situations, respectively. Both the absolute count of turns and the count per minute are normally distributed over conversations, as demonstrated by the results of Shapiro-Wilk test for face-to-face situation and Zoom situation in Table 1. Face-to-face and Zoom situations are abbreviated to f and z in the tables and plots throughout the study. Paired t-test was used to examine the differences between the two situations. The null hypothesis that the true difference in the number of turns in the two situations is equal to zero can be rejected at the 5 % significance level ($p = 0.039$), as well as the null hypothesis for the number of turns per minute ($p = 0.003$). Hypothesis 1 that face-to-face interactions have more turns than Zoom conversations is hence confirmed.

The counts of backchannels are summarised in Table 2. Only the counts in face-to-face conversations are normally distributed ($p = 0.62$). Hence, the Wilcoxon test for paired samples was employed to the backchannel data set. For both backchannel counting groups (count and count per minute), the differences between the two situations are not statistically significant. Hypothesis 2 that Zoom conversations have fewer backchannels can not be confirmed.

Table 1: Normality test and paired t-test on the count of turns

Situation	Count	Shapiro-Wilk test	Paired t-test
f	2640	W = 0.95 (p = 0.09)	t = 2.13 (p = **0.039**)
z	2375	W = 0.96 (p = 0.31)	
	Count per minute		
f	17.90	W = 0.95 (p = 0.16)	t = 3.19 (p = **0.003**)
z	15.49	W = 0.97 (p = 0.47)	

Table 2: Normality test and Wilcoxon test on the count of backchannels

Situation	Count	Shapiro-Wilk test	Wilcoxon test
f	1620	W = 0.97 (p = 0.62)	V = 472 (p = 0.07)
z	1423	W = 0.94 (p = **0.04**)	
	Count per minute		
f	11.53	W = 0.93 (p = **0.02**)	V = 533 (p = 0.10)
z	9.36	W = 0.97 (p = **0.02**)	

In addition, the counts of turns and backchannels per minute in the face-to-face situation show a positive correlation. The estimate of the Pearson correlation coefficient is 0.41 ($p < 0.05$); see Figure 2. However, a strong correlation between the two turn-taking relevant utterance types in the Zoom conversation could not be revealed ($p = 0.17$). The total number of turns and backchannels do not correlate either. Hypothesis 3 that turns and backchannels correlate positively is only valid for the co-present interactions.

4.2 Articulation rate

To examine Hypothesis 4, averaged articulation rates of speaker in each dialogue are shown in Figure 3. Speakers produced about 4.9 syllables every second in the face-to-face situation, faster than the 4.6 syllables in Zoom conversations. The difference was significant (paired t-test: $t = 2.3$, $df = 39$, $p < .01$). The mean difference in the articulation rates was 0.26 syllables (95 % CI: 0.17-0.35). Since

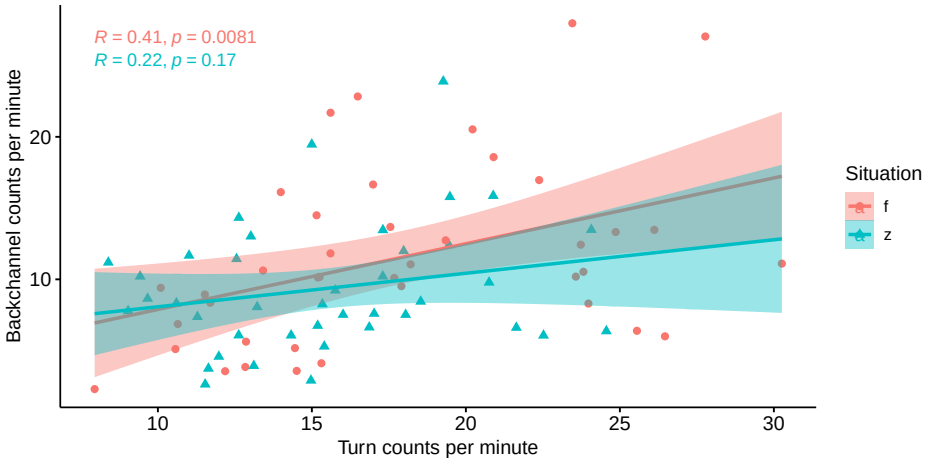

Figure 2: Pearson estimated correlation between the number of turns and backchannels per minute.

the articulation rates were calculated by dialogue, and each speaker had only two dialogues in each situation, the error bars may be strongly extended due to the scarcity of data.

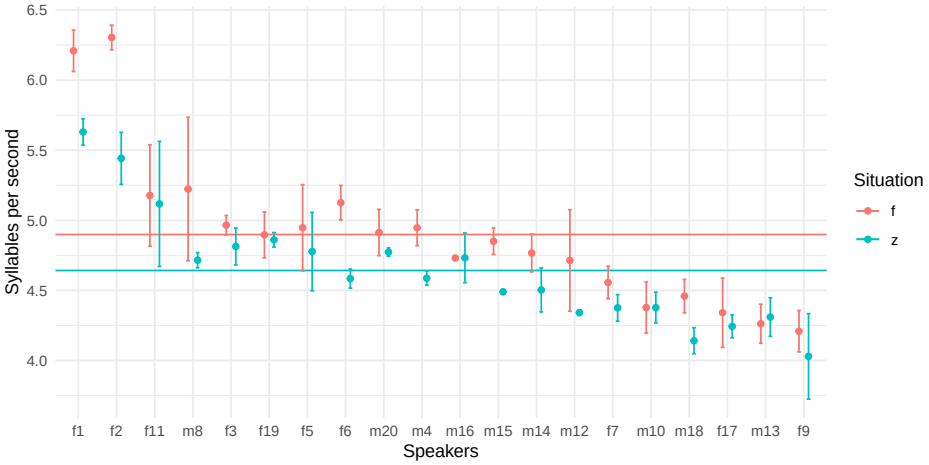

Figure 3: Point diagram with standard errors of individual differences in articulation rate. The horizontal lines show the mean value in the two situations.

4.3 Gap

Figure 4 shows the histogram of gap durations in the face-to-face and Zoom situations with the estimated distributions. Outliers beyond the range of [0, Q3+1.5IQR] have been excluded. 3088 of 3317 gaps in co-present conversation and 2722 of 2936 gaps in Zoom conversation were analysed (about 92.92 % of the entire data set); see Table 3.

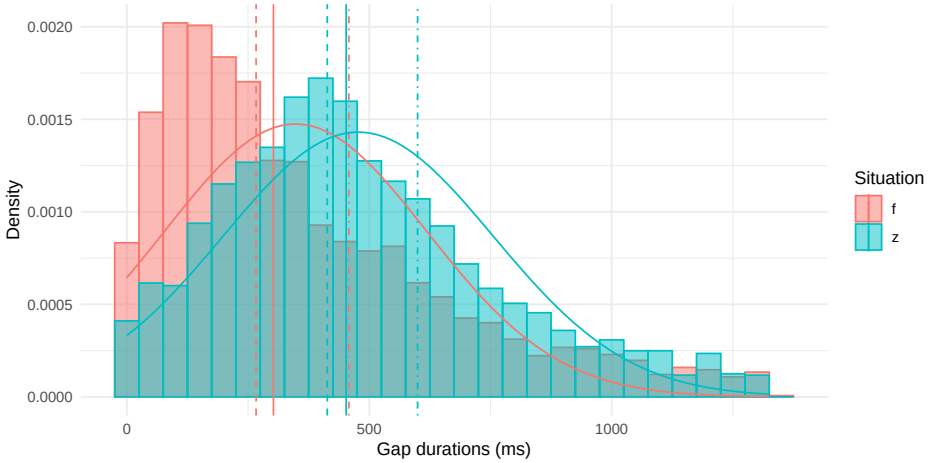

Figure 4: Histogram of gap durations in the face-to-face and Zoom situations with the estimated distributions. Bin size 50 ms. Dashed lines show the geometric means, solid lines the medians, and dot-dash lines the arithmetic means.

Hypothesis 5 that gaps in face-to-face conversations are generally shorter than those over Zoom is confirmed, as the peak of the estimate distribution, the mode, and the median value of the data (shown in red) are closer to zero. Previous studies argued that geometric means can estimate the data central tendency more realistically than arithmetic means (e. g. Heldner & Edlund 2010, de Ruiter et al. 2006). The geometric means (dashed lines) are therefore included which indeed offer more meaningful estimates than arithmetic means (dot-dash lines).

Apparently, 200 ms is not the most frequent gap duration in either of these situations. In face-to-face conversations, the most frequent gap duration (i. e. the mode) is 152.38 ms, shorter than the 200 ms threshold suggested for gap detection in previous studies (e. g. Walker & Trimboli 2010, Wilson & Wilson 2005). Still, 33.83 % of the gaps fall below this threshold. Over Zoom, the mode is much higher than 200 ms; see Table 3.

Table 3: Descriptive statistics of gap durations in the two situations (in ms) in the left panel. Counts and percentages of different gap duration thresholds in the right panel.

Situation	f	z	Threshold	f	z
Mean	348.04	477.08	< 10 ms	56	42
Geometric mean	231.60	365.76		1.69 %	1.43 %
Median	269.56	437.49	< 200 ms	1122	396
Mode	152.38	402.38		33.83 %	13.49 %
Standard deviation	270.52	278.80	< 250 ms	1389	561
Skewness	1.01	0.67		41.88 %	19.11 %
Kurtosis	3.34	3.09	< 500 ms	2282	1585
Without outliers	3088	2722		68.80 %	53.99 %
Total N	3317	2936	< 1000 ms	2994	2554
Percentage	93.10 %	92.71 %		90.26 %	86.99 %

Since speakers were asked to find all the differences between the pictures (see Section 3), the duration of gaps is assumed to increase because more time is needed for consideration and observation. Figure 5 presents how gap durations vary on a timeline. The median value of gap durations within a time slot is consistently shorter in face-to-face interactions.

A mixed effects model was fitted to determine whether the conversational situation and the position in a dialogue where a gap occurs had any effect on gap durations. The situation, the normalised time of gap and their interaction were added as fixed effects, while task and speaker as random intercepts. The gap durations in face-to-face dialogues were set as the dependent variable. The zoom situation significantly increases the gap durations, as shown in Table 4. The gaps tend to extend as time goes by. The interaction between situation and time slot did not improve the model.

4.4 Overlap

In the face-to-face situation, overlaps between dialogue components happened 2047 times, more than the 1880 times in Zoom conversations. After the normality assumption of overlaps per conversation was confirmed in both situations,

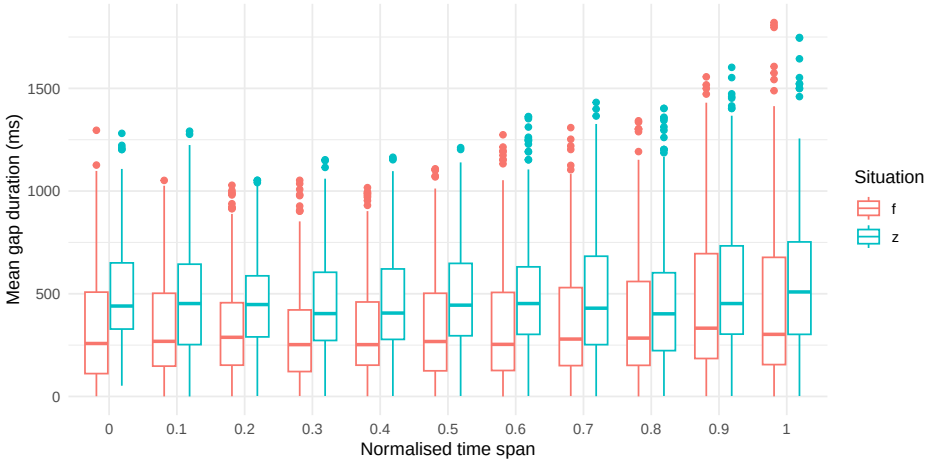

Figure 5: Mean gap durations change on a normalised time span.

Table 4: Fixed effects results (situation and time) of linear regressions for gap duration (intercept: face-to-face)

`gap duration ~ Situation * t.norm + (1|task) + (1|speaker)`					
	Estimate	Std. error	df	t value	Pr(>|t|)
(Intercept)	287.72	31.21	13.54	9.220	<0.001
Situationz	186.28	26.00	7256.24	7.166	**<0.001**
t.norm	268.12	31.25	7245.84	8.579	**<0.001**
Situationz:t.norm	-37.70	45.28	7252.74	-0.833	0.405
			R_c^2: 0.067		R_m^2: 0.038

the paired t-test was applied to the data to compare the group differences, see Table 5. Surprisingly, there are more overlaps between dialogue components in face-to-face conversations than over Zoom. But the difference is not significant. Hypothesis 7 can therefore not be accepted.

Regarding the temporal aspects, overlap durations are shorter in the face-to-face situation than over Zoom on average, supporting Hypothesis 6. Outliers falling below Q1-1.5IQR have been removed. 92.8 % of the overlap data is shown in Figure 6.

Table 5: Shapiro-Wilk test and paired t-test on the counts of overlaps in the face-to-face and Zoom situations

Situation	Count	Shapiro-Wilk test	paired t-test
f	2047	W = 0.97 (p = 0.75)	t = 1.24 (p = 0.23)
z	1880	W = 0.94 (p = 0.28)	

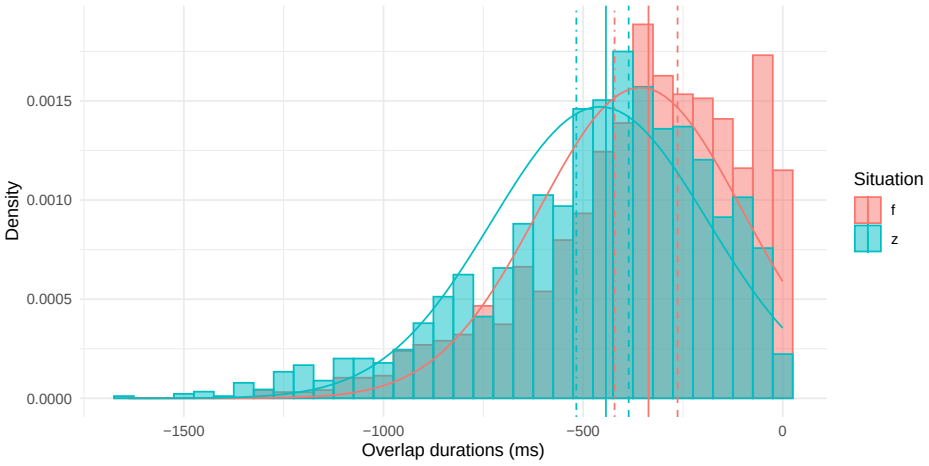

Figure 6: Histogram of overlap duration in face-to-face and Zoom situations with the estimated distributions. Bin size 50 ms. 92.8 % of the data included. Dashed lines show the geometric means, solid lines the medians, and dot-dash lines the arithmetic means.

Descriptive statistics of overlap durations are summarised in Table 6. Different thresholds from -10 ms to -1000 ms were applied to the data. Cases up to the respective thresholds and their proportions are listed in the right panel. Less than a quarter of the data can be represented by the 250 ms threshold in both situations.

Figure 7 shows that the occurrences of overlaps vary with time. In general, face-to-face interactions have more or comparable overlaps compared to Zoom conversations. In both situations, the first and the last slot have less overlaps than in the middle.

In order to test whether the conversational situation and the position where an overlap occurs in a conversation affects overlap occurrences, a mixed effect linear model of Poisson regression was fitted, which is more appropriate for count

Table 6: Descriptive statistics of overlap durations in the two situations (in ms) in the left panel. Frequencies and percentages of different overlap duration thresholds in the right panel.

Situation	f	z	Threshold	f	z
Mean	-356.19	-457.95	> -10 ms	14	5
Geom.mean	-238.27	-359.05		0.62 %	0.25 %
Median	-319.41	-416.35	> -200 ms	324	147
Mode	-247.62	-397.62		14.38 %	7.44 %
Standard deviation	254.56	271.35	> -250 ms	491	245
Skewness	-0.78	-0.66		21.84 %	12.39 %
Kurtosis	3.09	3.0	> -500 ms	1428	982
Without outliers	1935	1790		63.38 %	49.67 %
Total N	2047	1880	> -1000 ms	1956	1643
Percentage	94.53 %	95.21 %		86.82 %	83.11 %

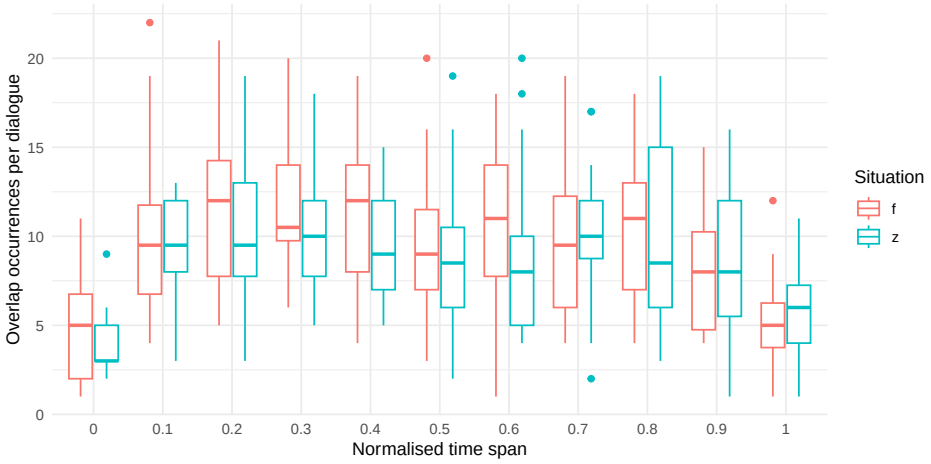

Figure 7: Overlap occurrences per dialogue on normalised time span

data such as overlap occurrences (Winter & Bürkner 2021). The overlap occurrences were seen as the dependent variable. The conversational situation, the normalised time point of overlaps and their interaction were set as fixed effects. The model was instructed to estimate by-speaker varying intercepts and by-speaker

varying slops. As a result, the factor Zoom situation did have a negative effect on the overlap occurrences comparing to co-present situation. The normalised time did not improve the model in a significant way; however, its interaction with Zoom situation did, see Table 7.

Table 7: Fixed effects results (situation and time) of linear regressions for overlap occurrences (intercept: face-to-face)

`overlap occurrences ~ Situation * t.norm + (1+t.norm|speaker)`				
	Estimate	Standard error	z value	Pr(>|z|)
(Intercept)	1.34442	0.07584	17.728	< 1.01
Situationz	-0.38138	0.07607	-5.013	**< 0.01**
t.norm	-0.17519	0.08968	-1.954	0.0508
Situationz:t.norm	0.51349	0.12501	4.108	**< 0.01**
		R_c^2: 0.230		R_m^2: 0.031

5 Discussion

5.1 Turns and backchannels in Zoom conversations

Based on our knowledge about turn-taking from previous studies, we may easily assume that conversations over Zoom would function the same way as talking face-to-face. However, the results show that there are multiple differences between these two situations. For example, the number of backchannels per minute in Zoom conversations does not distribute normally; Turns and backchannels do not correlate with each other positively as they do in co-present conversation. In terms of temporal aspects, Zoom conversation includes longer gaps and longer overlaps with lower articulation rate.

Due to commercial interests and legal limitations, statistics on electronic transmission delays over Zoom are not available to the public. The true proportion of latency in gap and overlap durations can therefore not be removed from the current calculations. Even if the assumed 30-70 ms delay (Boland et al. 2021) could be subtracted properly, Zoom conversation would still show a different pattern from that in a face-to-face dialogue, for instance, the means and medians of turn transition durations would still be longer over Zoom. Also the counts of turns

and backchannels do not correlate in Zoom situation as they do in a face-to-face scenario.

It can be assumed that the latency in Zoom hampers speakers' ability to anticipate an upcoming turn-end and to synchronise themselves with their partner's syllable rate (Boland et al. 2021, Wilson & Wilson 2005). Previous studies have argued that speakers have already started planning their turn before the current turn reaches the end (Levinson & Torreira 2015, Barthel et al. 2017). That is to say, speakers' cognitive load is increased near the end phase of a turn, as they not only need to comprehend and process the input information, but also have to consider simultaneously what to say, and more crucially, when to say it. Adding this to the irregular transmission delay, they may find it difficult to predict when their partner is going to finish or start a turn. Of course, the lack of signals can also cause uncertainty in Zoom communication, which correspondingly results in unexpected overlaps and interruptions in dialogue (Duncan 1972, ten Bosch et al. 2005, Sheng 2021). As a result, speakers spoke less and slower, as indicated by the lower turn and backchannel counts, and the lower articulation rate, trying to avoid interruption and misunderstanding.

5.2 Distribution of transition times

If the speech transition times are put onto one single scale with overlaps having the negative values and gaps having the positive ones (see Figure 8), we will have bi-modal distributions for face-to-face and Zoom situations. However, the floor transfer times are predominantly reported to have a uni-modal distribution in previous studies, where the most frequent duration falls within the range of 0-250 ms, depending on the researched language (e. g. Stivers et al. 2009, Heldner & Edlund 2010, Levinson & Torreira 2015).

Such quantitative research on the temporal aspects of German turn-taking behaviour is relatively scant. Weilhammer & Rabold (2003) is the only study which has reported the durational aspects of German conversation (apart from Japanese and American English). They also found Gaussian distributions in gap and overlap durations respectively, but in the logarithmic domain. If they re-transformed the data in milliseconds and put the two sequences onto one scale, they would have received a similar bi-modal distribution. It warrants further investigation on the question whether the bi-modal distribution of transfer durations is a German-specific phenomenon.

Another reason for the bi-modal distribution could be the specific task setting in the experiment. In the Diapix task, participants were asked to cooperate to spot all the differences between two very similar pictures. On the one hand, this

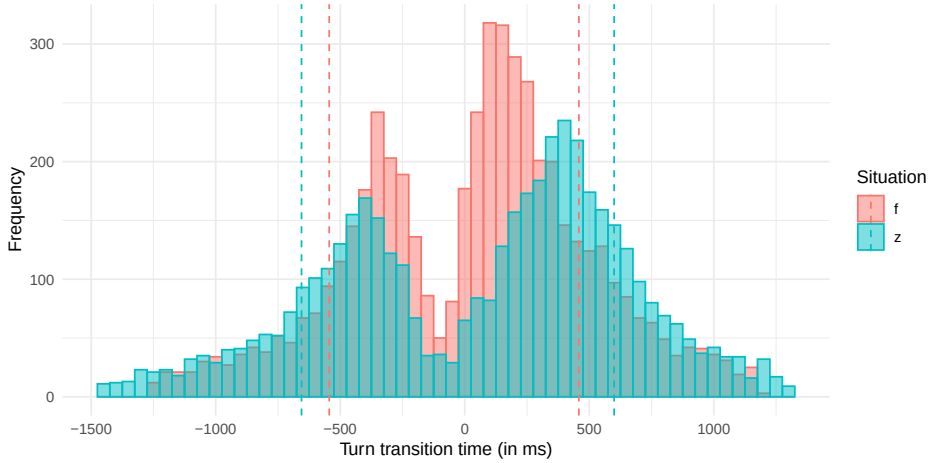

Figure 8: Histogram of turn transition time in the face-to-face and Zoom situations. Bin size 50 ms. Outliers excluded. Dashed lines show the geometric means.

allows easy elicitation of spontaneous conversations balanced between speakers; on the other hand, the cognitive load of the participants might increase due to the growing difficulty in finding more differences as time goes by. To minimise the potential impact of the irregularly lengthened intervals that occurred when the participants were searching for less-obvious differences, gaps and overlaps in the first parts of the conversations were extracted additionally. Nonetheless, the bi-modality of distribution did not change. The special bi-modal distribution seems to be irrelevant to the task settings.

One could assume that the observed longer gaps might be resulted from participants' gaze shifts between the task image and their partner. They needed to compare the picture in their hand and what their partner described throughout the experiment, simultaneously. The constant shift of eye gaze might have increased the reaction time slightly. Though such shifting is extremely quick, it may still contribute significantly to the rapid floor transition time. As a result, the gaps between dialogue components become longer and the modes of gaps shift to the right in the distribution plot. Considering the unavoidable latency in video-based conversation, the gaps are even more salient in Zoom interaction. Whether the extremely rapid gaze-shifting would indeed increases gap durations should be further investigated with other methods that enable the measurement of gaze shifting duration.

Even if gaze shifting could explain the shifting of gap durations to the right

in comparison to previous studies, it can hardly explain the shifting of overlap durations to the left, indicating longer overlapped speech.

The reason may be twofold: Firstly, the longer overlaps may be traced back to the collaborative situation required in the experiment task. Speakers may correspondingly use a different speech register from that investigated in previous studies on turn-taking behaviours. Edelsky (1981) put forward two organisational models for conversation: *singly produced floor* and *collaboratively constructed floor*. The first model is in accordance with the *no-gap-no-overlap* pattern specified by Sacks et al. (1974), where overlapping of any duration signals conversational malfunction, because the current speaker's right to have the exclusive floor is challenged. In contrast, in the second model, the collaboratively constructed floor is meant to be potentially open to all conversation participants, and overlapped speech is considered a sign of active engagement of the participants. Speakers share the floor with each other, so there is no need to compete to seize it (see Edelsky 1981). From this aspect, the longer overlap durations in the Diapix tasks can be interpreted not as malfunctioning turn-taking, but as a sign of high interactivity and engagement. Therefore, speakers did not have to relinquish their speech prematurely for the sake of the *one-at-a-time* pattern. Long overlaps are acceptable in collaborative conversations.

Secondly, the close relationship between speakers might have caused the longer overlaps. When communicating with friends and family members, we tend to use an informal register, in other words, a more relaxed style of speech. This style is distinguished from the style applied to communication with strangers (Redeker 1986, O'Leary & Gallois 1985). It is embodied in several behaviours, for example, more frequent overlaps and more turn switches (Coates 1994). The speakers from whom the conversations were collected in the BeDiaCo corpus had close relationships to their conversational partners in the dyad. They did not have to use the same register employed in conversations between people with different social status (see e. g. Seuren et al. 2021, Duran et al. 2023). Even though their speech overlaps, (at least) one party will continue till the turn is finished, without concerns about being "reluctant, rude or hostile" (Heritage 1984) or creating an aggressive impression. As a result, overlaps detected in the face-to-face situation are longer than those reported in previous studies.

For the longer transition time in Zoom conversations, the electronic transmission delay might have played an indispensable role. If speech signal is received with latency, reaction on the signal will also be delayed. Speakers can not relinquish their speech turn *on time* even if they would like to restore the *one-at-a-time* pattern. As a consequence, overlap continues a bit till it is perceived, causing longer durations.

6 Conclusion

In the current study, turn-taking behaviours in German conversation in face-to-face and Zoom situations were compared. For the analysis, the BeDiaCo V2 corpus was investigated, where task-oriented conversations between speakers who know each other were collected. It has been found that Zoom conversations have lower articulation rate, slightly fewer turns and backchannels, longer gaps, and more and longer overlaps than face-to-face interactions. Based on these results, we can conclude that Zoom conversation differs from a co-present dialogue in many ways. The rhythm of conversation and speakers' ability to anticipate turn-ends are probably disrupted by the omnipresent transmission delay and the absence of signals (e. g. gaze direction, breath rhythm). As indicated by the lack of correlation between turns and backchannels in Zoom conversation, speakers' conversational behaviours were less regular and less predictable than in face-to-face situations. The conversational pattern used in video-mediated communication invites more research.

In addition, the study offers an exploratory description of the temporal aspects of turn-taking in German conversation, which has only been scantly investigated in previous research. Speaker transition times are found to distribute bi-modally. Whether the bi-modal distribution is German-specific, as also observed in Weilhammer & Rabold (2003), or it is due to the specific speech register triggered by the task setting and by the close relationship between speakers, should be further examined with more data on German conversations.

Supplementary files

Full documentation of corpus annotating practice for turn-taking research can be found in https://box.hu-berlin.de/f/1dd95d4d174545f0bf42/.

Funding information

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

Acknowledgements

This paper is an excerpt from Qiang Xia's master thesis. The author would like to express gratitude to Christine Mooshammer for supervision and warm support

before and throughout the pandemic. A lot of thanks to Malte Belz for guidance in corpus annotation and valuable comments. Also thanks to Daniela Palleschi for advise on statistics.

Competing interests

The authors have no competing interests to declare.

References

Auer, Peter. 2005. Projection in interaction and projection in grammar. *Text - Interdisciplinary Journal for the Study of Discourse* 25(1). 7–36. DOI: 10.1515/text.2005.25.1.7.

Auer, Peter. 2021. Turn-allocation and gaze: A multimodal revision of the "current-speaker-selects-next" rule of the turn-taking system of conversation analysis. *Discourse Studies* 23(2). 117–140. DOI: 10.1177/1461445620966922.

Bailenson, Jeremy N. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior* 2(1). DOI: 10.1037/tmb0000030.

Baker, Rachel & Valerie Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3). 761–770. DOI: 10.3758/s13428-011-0075-y.

Barthel, Mathias, Antje S. Meyer & Stephen C. Levinson. 2017. Next speakers plan their turn early and speak after turn-final "go-signals". *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.00393.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: 10.18637/jss.v067.i01.

Belz, Malte, Alina Zöllner, Megumi Terada, Robert Lange, Lea-Sophie Adam & Bianca Sell. 2021. Dokumentation und Annotationsrichtlinien für das Korpus BeDiaCo (version 2). *Zenodo*. DOI: 10.5281/zenodo.4593351.

Boland, Julie E., Pedro Fonseca, Ilana Mermelstein & Myles Williamson. 2021. Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*. 1272–1282. DOI: 10.1037/xge0001150.

Brunner, Lawrence J. 1979. Smiles can be back channels. *Journal of Personality and Social Psychology* 37(5). 728–734. DOI: 10.1037/0022-3514.37.5.728.

Bullock, Oliveira Maggie & Bianca Sell. 2022. *PDF and PSD files of DiapixGEtv picture materials – German version adapted to elicit tense vowels*. DOI: 10.5281/ZENODO.6510724.

Coates, Jennifer. 1994. No gap, lots of overlap: Turn-taking patterns in the talk of women friends. In David Graddol, Janet Maybin & Barry Stierer (eds.), *Searching language and literacy in social context*, 177–192. Clevedon: Multilingual Matters. DOI: 10.1017/CBO9780511791147.011.

de Ruiter, Jan-Peter, Holger. Mitterer & N. J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82(3). 515–535. DOI: 10.1353/lan.2006.0130.

Dideriksen, Christina, Riccardo Fusaroli, Kristian Tylén, Mark Dingemanse & Morten H. Christiansen. 2019. *Contextualizing conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations*. Preprint. PsyArXiv. DOI: 10.31234/osf.io/fd8y9.

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

Duran, Daniel, Melanie Weirich & Stefanie Jannedy. 2023. Assessing register variation in local speech rate. In Radek Skarnitzl & Jan Volín (eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, 2314–2318. Prague: Guarant International.

Edelsky, Carole. 1981. Who's Got the Floor? *Language in Society* 10(3). 383–421.

Egger-Lampl, Sebastian, Raimund Schatz & Stefan Scherer. 2010. It takes two to tango - Assessing the impact of delay on conversational interactivity on perceived speech quality. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 1321–1324. DOI: 10.21437/Interspeech.2010-412.

Fusaroli, Riccardo & Kristian Tylén. 2016. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science* 40(1). 145–171. DOI: 10.1111/cogs.12251.

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

Heinz, Bettina. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 35(7). 1113–1142. DOI: 10.1016/S0378-2166(02)00190-X.

Heldner, Mattias. 2011. Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America* 130(1). 508–513. DOI: 10.1121/1.3598457.

Heldner, Mattias & Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38(4). 555–568. DOI: 10.1016/j.wocn.2010.08.002.

Heritage, John. 1984. *Garfinkel and ethnomethodology*. Cambridge: Polity Press.

Holler, Judith & Kobin H. Kendrick. 2015. Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.00098s.

Holler, Judith & Stephen C. Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences* 23(8). 639–652. DOI: 10.1016/j.tics.2019.05.006.

Jones, Mari Riess. 2019. Tuning in to slow events. In *Time will tell: A theory of dynamic attending*, 135–157. Oxford: Oxford University Press. DOI: 10.1093/oso/9780190618216.003.0007.

Kaukomaa, Timo, Anssi Peräkylä & Johanna Ruusuvuori. 2013. Turn-opening smiles: Facial expression constructing emotional transition in conversation. *Journal of Pragmatics* 55. 21–42. DOI: 10.1016/j.pragma.2013.05.006.

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

Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa & Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task Dialogs. *Language and Speech* 41(3-4). 295–321. DOI: 10.1177/002383099804100404.

Levinson, Stephen C. & Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.00731.

Local, John & Gareth Walker. 2012. How phonetic features project more talk. *Journal of the International Phonetic Association* 42(3). 255–280. DOI: 10.1017/S0025100312000187.

Mehl, Matthias R., Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher & James W. Pennebaker. 2007. Are women really more talkative than men? *Science (New York)* 317(5834). 82. DOI: 10.1126/science.1139940.

Michalke, Meik. 2018. *Sylly: Hyphenation and Syllable Counting for Text Analysis*. (Version 0.1-5). https://reaktanz.de/?c=hacking&s=sylly (5 October, 2023).

O'Leary, Maria J. & Cynthia Gallois. 1985. The last ten turns: Behavior and sequencing in friends' and strangers' conversational findings. *Journal of Nonverbal Behavior* 9(1). 8–27. DOI: 10.1007/BF00987556.

Peters, Pam & Deanna Wong. 2014. Turn management and backchannels. In Christoph Rühlemann & Karin Aijmer (eds.), *Corpus Pragmatics: A Handbook*, 408–429. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139057493.022.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Manual. R Foundation for Statistical Computing. Vienna, Austria.

Redeker, Gisela. 1986. *Language use in informal narratives: Effects of social distance and listener involvement*. United States – California: University of California, Berkeley. (Doctoral dissertation).

Riest, Carina, Annett B. Jorschick & Jan P. de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.00089.

Robinson, Jeffrey D. 2020. Revisiting preference organization in context: A qualitative and quantitative examination of responses to information seeking. *Research on Language and Social Interaction* 53(2). 197–222. DOI: 10.1080/08351813.2020.1739398.

Rochet-Capellan, Amélie & Susanne Fuchs. 2014. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369(1658). 20130399. DOI: 10.1098/rstb.2013.0399.

Sacks, Harvey, Emanuel A. Schegloff & Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4). 696–735. DOI: 10.1353/lan.1974.0010.

Schegloff, Emanuel A. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society* 29(1). 1–63. DOI: 10.1017/S0047404500001019.

Selting, Margret. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics* 6(3). 371–388. DOI: 10.1075/prag.6.3.06sel.

Seuren, Lucas M., Joseph Wherton, Trisha Greenhalgh & Sara E. Shaw. 2021. Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics* 172. 63–78. DOI: 10.1016/j.pragma.2020.11.005.

Sheng, Yu. 2021. How the function of video-conference software, Zoom, interfere with turn-taking in the online classroom? In *2021 International Conference on Education, Information Management and Service Science (EIMSS)*, 209–215. IEEE Computer Society. DOI: 10.1109/EIMSS53851.2021.00052.

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

ten Bosch, Louis, Nelleke Oostdijk & Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*. In Honour of Louis Pols 47(1). 80–86. DOI: 10.1016/j.specom.2005.05.009.

Van Engen, Kristin J., Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim & Ann R. Bradlow. 2010. The wildcat corpus of native-and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53(4). 510–540. DOI: 10.1177/0023830910372495.

Walker, Michael B. & Carmelina Trimboli. 2010. Smooth transitions in conversational interactions. *The Journal of Social Psychology*. DOI: 10.1080/00224545.1982.9713444.

Weilhammer, Karl & Susen Rabold. 2003. Durational Aspects in Turn Taking. In M. J. Solé, D. Recasens & J. Romero (eds.), *15th International Congress of Phonetic Sciences (ICPhS)*, 2145–2148. Barcelona: Causal Productions Pty Ltd.

Wilson, Margaret & Thomas P. Wilson. 2005. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review* 12(6). 957–968. DOI: 10.3758/BF03206432.

Winkelmann, Raphael, Jonathan Harrington & Klaus Jänsch. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45. 392–410. DOI: 10.1016/j.csl.2017.01.002.

Winter, Bodo & Paul-Christian Bürkner. 2021. Poisson regression for linguists: a tutorial introduction to modelling count data with brms. *Language and Linguistics Compass* 15(11). e12439. DOI: https://doi.org/10.1111/lnc3.12439. https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12439.

Yngve, Victor. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, 567–578. Chicago, IL, USA: Chicago Linguistic Society.