Language and Cognition (2025), 17, e57, 1–22
doi:10.1017/langcog.2025.10014

ARTICLE

Temporal relationships between speech and hand gestures in the vicinity of potential turn boundaries in German and Swedish conversation

Margaret Zellers[1], Jan Gorisch[2] and David House[3]

[1]Institut für Skandinavistik, Frisistik, & Allgemeine Sprachwissenschaft, Kiel University, Kiel, Germany; [2]Department of Pragmatics, Leibniz-Institute for the German Language, Mannheim, Germany and [3]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
Corresponding author: Margaret Zellers; Email: mzellers@isfas.uni-kiel.de

(Received 15 November 2022; Revised 26 November 2024; Accepted 23 June 2025)

Abstract

Both gesture and talk are basic building blocks of face-to-face conversation. In this study, we address the temporal dynamics of hand gesture phases relative to places and types of turn transition. We annotated gesture features and measured temporal aspects of gesture related to speech in two languages, German and Swedish. We found variation in the temporal relationships of gesture types and alignment of gesture phases that relate to the management of turn-taking in conversation. Specifically, the frequency of different gesture phases accompanying the offset of speech differed depending on whether the same speaker held the floor or whether a new speaker took up a turn. In addition, we found that differences in temporal alignment of gesture phases can distinguish between the type of turn transition that is upcoming up to a second before the place of transition is reached. Our results emphasize the importance of the interaction of the verbal and the gestural modality to maintain the smooth flow of conversation.

Keywords: conversation; co-speech gesture; German; gesture phases; hand gestures; potential turn boundary; Swedish; temporal gesture alignment; turn transitions

1. Introduction

Turn-taking in conversation is managed in versatile ways, and even more so in multimodal settings. While conversational participants organize their turns easily, measuring turn-taking cues by analyzing recordings has proven complex, leading many researchers to study single phenomena and mostly independently of other phenomena. For example, phonetic, and especially prosodic, features have been studied extensively in terms of their ability to predict points of speaker change. Gestures, such as head nods, have also been studied extensively in the vicinity of places of potential turn transitions.

In attempts to combine spoken with gestural features, the temporal characteristics of gestures, such as the apices of gesture strokes, have been compared with the temporal characteristics of speech, such as the stressed syllable of a lexical affiliate of the corresponding gesture (for example, a spoken lexical item with which a gesture shares lexical content). These studies have often found timing differences suggesting that gestures come slightly before their lexical affiliate (Bergmann et al., 2011; Ferré, 2010; ter Bekke et al., 2020). Such findings on speech gesture synchrony often address implications, such as whether gestures facilitate lexical access, decrease processing time or the like. So far, however, it has not been investigated in depth what consequences the overall activity of gesturing or not gesturing has on the management of turn-taking or how the uptake or non-uptake of a turn depends on the timely dynamics of that gesturing activity.

In general, there seem to be two possibilities: on one hand, gesture strokes may mark an end-point of an utterance, thus yielding the turn. On the other hand, gesture activity may signal that an utterance is still in progress, hindering an uptake from the interlocutor. In the current study, we build on prior work calling for multimodal analyses of talk – for example, work by Mondada (2019) positioning multimodality in the context of social interaction, embodiment and multisensority from a conversation analytic (CA) and therefore qualitative perspective – and extend it by taking a quantitative approach toward annotating real data with traditional features and exploring them in an innovative way. Regarding annotations, we take the well-known gesture phases, that is, preparation, hold, stroke and retraction following Kendon (2004) and Kita et al. (1997), and basic turn transition types, such as keeping the turn, receiving a backchannel or yielding the turn. The simple but innovative analysis approach presented in this study takes a perspective focused around the offset of speech, that is, where a turn (potentially) comes to a syntactic and/or semantic end, and looks into the current speaker's gestures in the vicinity of this point in time. Our study therefore contributes to the body of research that investigates the resources that interactional participants employ at potential turn transition places for managing their turns.

In our quantitative approach, we only refer to hand gestures and disregard other gesture types such as head gestures or eye gaze. Although our annotations included whether we interpreted the gesture as referential or not, we do not differentiate between these referentiality types in the current work, cf. Loehr (2004), who does not distinguish between referential and non-referential gestures and Shattuck-Hufnagel and Ren (2018) whose results span across referentiality.

1.1. Gesture analysis

A gesture is a movement of some part of the body that accompanies speech and has communicative value for listeners (Kendon, 1994); gesturing can be distinguished from movement for its own sake, or movement which involves object manipulation (Novack et al., 2015). However, gestures are defined based on inferences about their communicative intent, not external criteria such as form (Bavelas, 1994). It has been shown that naïve observers are able to interpret gestural movements reliably (Goldin-Meadow & Sandhofer, 1999). Thus, any analysis of gesture makes the fundamental assumption that the body movement in question is intended to be communicative.

Early analyses of manual gestures focused on gestures that formed specific shapes, or referred to specific locations, sizes, objects, metaphors or ideas (c.f. Kendon, 1980;

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

McNeill, 1992). Since this time, however, a variety of classification systems have arisen, allowing the study of gestures across various parameters. The most popular of these systems were introduced by Kendon (1980) with the categories sign language, pantomime, emblems and gesticulation. McNeill arranged these bodily movements on a continuum with gesticulation being defined as co-speech gestures. To further classify the co-speech gestures, McNeill (2006) introduced four categories roughly related with gesture semantics or function: *iconic*, *metaphoric*, *deictic* and *beat* gestures. More recent behavioral and cognitive evidence indicates that there is not a clear divide between, e.g., *beat* gestures and *metaphoric* gestures (Casasanto, 2008, 2009); rather, gestures may be classified in more than one way simultaneously. McNeill had already raised the observation that a strict classification is not really realistic and that a dimensional description would better characterize the ways in which gestures are implemented. Thus, even if a single or specific function is attributed to a gesture, this attribution should not be considered as a unique or exclusive function but rather simply one aspect of the gesture in question.

1.2. Coordination of speech and gesture

A growing body of evidence supports the argument that linguistic research should treat speech and gesture as a unified system (cf. e.g., Kendon, 2004; McNeill, 2005). Wagner et al. (2014) provide an in-depth review of relationships between speech and gesture that have been reported in the literature. Their review raises the question of whether the auditory modality, that is, spoken language, and visual modality, that is, gesture, are used in parallel or as complements.

In speech production, a strong effect appears to arise in the context of prosodic features. Rhythmic or *beat* gestures have been demonstrated to appear in consistent temporal alignment with prosodic prominences in spoken language in adults as well as children (Ambrazaitis & House, 2017b; Esposito et al., 2007; Esteve-Gibert & Prieto, 2013; Florit-Pons et al., 2020; Knight, 2009; Krahmer & Swerts, 2007; Leonard & Cummins, 2011). Specifically, gesture apices tend to align with stressed syllables (e.g., Loehr, 2004; Rochet-Capellan et al., 2008) or intonation peaks (e.g., Esteve-Gibert & Prieto, 2013; Nobe, 1996; Pouw & Dixon, 2019).

Visual and auditory information have been found to be automatically integrated in the course of speech perception, and the combination of these different input streams influences speech intelligibility (Kelly et al., 2010; McGurk & MacDonald, 1976). Viewing speech-accompanying gesture has been demonstrated to lead to increased activity in the auditory cortex (Hubbard et al., 2009), as well as in brain areas involved with semantic processing (Dick et al., 2009).

Specific constellations of different prosodic and gestural prominence cues may also have different communicative effects than the individual cues alone (Ambrazaitis & House, 2017a, 2017b; Prieto et al., 2015). A manual McGurk effect has even been reported, where gestural beats were used to overwrite intonation cues for differentiating lexical stress, e.g., OBject versus obJECT (Bosker & Peeters, 2021). Guellaï et al. (2014) report that listeners can identify congruencies between even unintelligible speech and gesture and use gesture for disambiguation in cases when information in the speech signal is ambiguous or conflicting. It is thus clear that the temporal placement of speech-accompanying gesture can and does play a crucial role for speech understanding.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

1.3. The management of turn-taking in conversation

Conversation tends to proceed with a minimum of problematic (that is, disruptive) overlaps or silent gaps (Sacks et al., 1974), and the amount of silent time between conversational turns appears to have a stable mean of around 200 ms across a variety of typologically different languages, including sign languages (Buanzur et al., 2018; de Vos et al., 2015; Heldner & Edlund, 2010; Stivers et al., 2009). Many linguistic features, phonetic/prosodic and otherwise, play a role in signaling turn transition, including syntactic/semantic completion (e.g., Auer, 1996; de Ruiter et al., 2006; Schaffer, 1983), intonational features (e.g., Bögels & Torreira, 2015; Caspers, 2003; Local et al., 1986; Peters, 2006; Selting, 1996) and phonation quality/spectral characteristics (e.g., Kane et al., 2014; Ogden, 2001).

Studies using larger corpora (e.g., Gravano & Hirschberg, 2009, 2011; Hjalmarsson, 2011; Koiso et al., 1998) tend to find a hierarchy of various features correlated with speaker transition or floor hold, including lexico-syntactic as well as phonetic features; however, syntactic/semantic completion is not always a definitive cue to finality. Different types of conversational actions or turns have different degrees of 'projectability', or predictability as to their future direction, such as when a speaker tells a story which requires multiple conversational turns to complete (Auer, 2005).

Like linguistic cues, gestural cues have been shown to be relevant for the management of turn-taking in conversation. Schegloff (1984) points out that it is mostly current speakers who gesture, although gestures may be used by a current hearer to indicate the desire to take the floor, as has been found for a variety of languages (Li, 2014; Mondada & Oloff, 2011; Streeck & Hartge, 1992). Similarly, gestures may be used at turn ends to hold the floor during a pause or to invite a response from an interlocutor (Kendon, 1995; Mondada, 2007; Stivers & Rossano, 2010). Sikveland and Ogden (2012) demonstrate how hand gesturing across a turn end can help achieve the complex function of identifying and resolving a problem of understanding. Some gestural cues appear to parallel roles of prosodic structure; thus, Quek et al. (2002) find that hand gestures are temporally correlated with prosodic phrase boundaries, possibly contributing to the segmentation of speech into phrases. In addition, Chui (2005) and Graziano and Gullberg (2018) report that gesturing is linked with ongoing speech. From a turn-taking perspective, the end of a turn constitutes a break in continuity of speech, which might allow the absence of gesturing to be a turn-yielding cue, too. Similarly, Barkhuysen et al. (2008) report that speakers tend to look away from their interlocutor phrase-medially and to look back at them phrase-finally. These studies serve as evidence that gestures may provide information about the completeness of a spoken turn.

1.4. Aims of the current study

It is clear from the literature discussed above that close temporal relationships exist between spoken language and gesture and that coordinated speech and gesture are relevant for the management of turn-taking in conversation. At the same time, the literature reported above suffers to some degree from a lack of methodological unity, with the results of qualitative and quantitative studies not always brought into harmony with one another. The question of whether and how possible variation in temporal relationships between speech and gesture contributes to the management of turn-taking remains open.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

Thus, in the current study, we investigate the extent to which malleability in temporal relationships between speech and gesture is used for conversation management and how the use of such features may differ across languages.

Our specific research question is how and to what extent does the temporal relationship between speech and gesture contribute to the management of turn-taking in conversation? We operationalize the temporal relationship between speech and gesture as the temporal relationship between different phases of manual gestures produced by the speaker of the turn that is (potentially) coming to an end and the offset of speech at a location where turn transition may become relevant. We hypothesize that, at locations in conversation in which a current speaker reaches a point of possible completion but wishes to hold the floor, extra effort is needed, which will result in different temporal relationships between speech and gesture (cf. Kendrick et al., 2023; Schegloff, 1984).

We further address our research question in the context of two related languages with different prosodic structures, German and Swedish. While both are Germanic languages and thus have some substantial structural similarity, they differ in their intonational structure. German is an intonation language, where pitch movements are used exclusively for pragmatic purposes. In Swedish, however, pitch movements are part of the lexical specification of words, with words carrying one of two lexical pitch accents. The differences in the phonological systems have already been shown to be relevant for prosodic signaling of turn transition intentions (Rossi et al., 2022; Zellers et al., 2019a). Since gestural features are closely linked with prosodic features (cf. Section 1.2), it is thus possible that these prosodic differences could lead to differences in gesture use even in two relatively closely related languages.

2. Method

We adopt a quantitative, corpus-based approach that involves the annotation and analysis of video recordings. The data are spontaneous conversations from preexisting corpora in German and Swedish that have already been transcribed orthographically. In this section, we give more details on the selected recordings, the annotations we added for the purposes of this study, and an outlook on the statistical analyses we employed.

2.1. Data

The data used in the current study are drawn from two corpora of conversational speech. The Swedish data come from the Spontal corpus (Edlund et al., 2010), a corpus of two-party conversations collected in Stockholm, Sweden. Spontal comprises audio, video and motion-capture data, although only the video and audio data are used in the current study. The German data are taken from FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch, Research and Teaching Corpus of Spoken German) (Schmidt, 2014), a collection of speech taken from a variety of natural settings, comprising audio and video data.

An important goal of the current research was to use existing data rather than to collect new data, since so much data are already available. To do this, it was necessary to make a selection of the data that was maximally similar, while taking into account the fundamental differences between these two speech databases. The materials in the

Spontal corpus were more constrained in their form: all interactions involved two-party conversations, with participants sitting face to face and with no fixed topic of conversation in the portions of the data used. The topics of the conversations were quite varied although they generally fell into the categories of daily or common activities, such as hobbies, working out at the gym, buying a drill, moving into a new apartment, working as a translator, building a closet and travelling. Two of the speakers (09-22B and 09-35B) participated in more than one conversation, as indicated in Table 1.

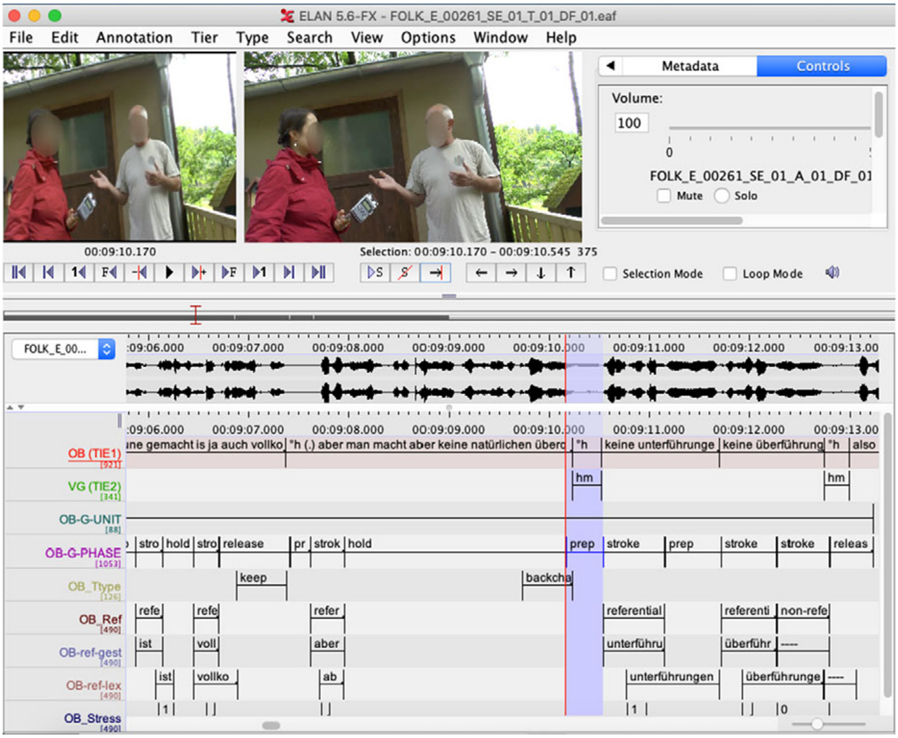
The selection constraints of two-party conversations with participants sitting face to face were also adopted while searching FOLK for appropriate data for a comparison. Three relevant conversations in FOLK were identified. In two of the conversations, two speakers interact in the context of a mock job interview (a third party is present but does not contribute to the conversation once the mock interview has begun; the excerpts we analyzed began after this point). In the third conversation, an expert in birds of prey is interviewed in an informal setting. Although these conversational settings may be more formally structured than in the Spontal data, observation of the data indicates that turn-taking proceeds similarly to in the fully spontaneous conversations in Spontal. Furthermore, we do not anticipate that the differences in topic or formality would have a large impact on the temporal coordination of speech and gesture, since this is likely to rest on cognitive processes rather than on the specific content of a conversation.

For each language, we used a total of 55 minutes of data. In Spontal, the 55 minutes comprise 5 minutes each from 9 conversations, and 10 minutes, in two separate chunks, from a tenth conversation (09-35; see Table 1). In FOLK, the 55 minutes comprise 17–20 minutes each from the three conversations. Due to the constraints of the available data and annotations, it was not possible to use data from an equivalent

Table 1. Metadata for the FOLK and Spontal files

Speaker ID	Recording ID	Sex
FOLK_S_00408	FOLK_E_00173, FOLK_E_00174	m
FOLK_S_00410	FOLK_E_00173	m
FOLK_S_00412	FOLK_E_00174	m
FOLK_S_00683	FOLK_E_00261	m
FOLK_S_00684	FOLK_E_00261	f
09–06A	spontal–09–06	f
09–06B	spontal–09–06	m
09–18A	spontal–09–18	m
09–18B	spontal–09–18	m
09–20A	spontal–09–20	m
09–20B	spontal–09–20	m
09–22A	spontal–09–22	m
09–22B	spontal–09–22, spontal–09–24, spontal–09–25	m
09–24A	spontal–09–24	f
09–25B	spontal–09–25	f
09–28A	spontal–09–28	m
09–28B	spontal–09–28	m
09–31A	spontal–09–31	m
09–31B	spontal–09–31	f
09–35A	spontal–09–35	f
09–35B	spontal–09–35, spontal–09–36	f
09–36A	spontal–09–36	m

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press



OB (TIE1) [921]: ıne gemacht is ja auch vollko | °h (.) aber man macht aber keine natürlichen überg | °h | keine unterführunge | keine überführung | °h | also

VG (TIE2) [341]: hm | hm

OB-G-UNIT [88]

OB-G-PHASE [1053]: stro | hold | stro | release | pr | strok | hold | prep | stroke | prep | stroke | stroke | releas

OB_Ttype [126]: keep | backcha

OB_Ref [490]: refe | refe | refer | referential | referenti | non-refe

OB-ref-gest [490]: ist | voll | aber | unterführu | überführ | ----

OB-ref-lex [490]: ist | vollko | ab | unterführungen | überführunge | ----

OB_Stress [490]: 1 | 1 | 0

Figure 1. Screenshot of the annotation environment in ELAN (data from FOLK).

amount of speakers while maintaining a similar amount of data as measured in minutes; we prioritized having a similar quantity of data per language so as to have a comparable number of completion points (see Section 2.2).

2.2. Annotations

Gesture and turn annotations were carried out using ELAN (Max Planck Institute, 2018), cf. Figure 1. Spoken features were annotated in Praat (Boersma & Weenink, 2021).

Gesture annotation was carried out using the video signal only (that is, with the audio muted) and proceeded one conversational participant at a time. The first step of the annotation process was to identify gesture phrases, that is, stretches of time when one or both of a participant's hands moved. In a second step, we segmented the gesture phases (*preparation, stroke, hold, retraction*, cf. Kendon (2004); Koiso et al. (1998)). In the analysis below, where we relate these gesture phases with syntactic/semantic features, we also labeled areas with no hand gesture as *none.* Strictly speaking, this is not a gesture phase *per se*, but it is implemented so that measurement points with gesture can also be compared to those without gesture. The boundaries of the gesture phases were refined by moving frame-by-frame through the video in ELAN; if a boundary was ambiguous between two frames, the earlier frame was chosen as the boundary.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

In a separate annotation phase in Praat, using the audio data only, we labeled locations where a speaker's turn was potentially complete and the possibility of speaker change thus became relevant (c.f. TRPs, Sacks et al. (1974); SYNCOMPS, Local and Walker (2012); Potential Turn Boundaries (completion points), Zellers (2017)). Since only locations in the conversation that were clearly syntactically or semantically complete in context were included in this classification, we adopt the term *completion points* for these locations, rather than, for example, TRPs, which are defined by constellations of features, not only syntactic/semantic completion. We excluded locations where the incoming speaker's turn or backchannel began in overlap with the end of the current speaker's turn, since these early incomings might represent an 'incorrect' prediction about the current speaker's turn-taking intentions. The completion points were then classified based on the sequential structure of the possible transition as one of the following: holding the floor (with or without a verbal backchannel from the other speaker), releasing the floor (either with or without an explicit question form) or ambiguous cases.

- **Floor hold without verbal backchannel (Keep)** Following the boundary location, the current speaker takes the next full turn, thus keeping the conversational floor; the interlocutor does not produce any kind of verbalization.
- **Floor hold with verbal backchannel (Backchannel)** After the completion point, the interlocutor produced a verbal backchannel but no other speech. Evidence from Truong et al. (2011) and Ferré and Renaudier (2017) indicates that verbal backchannels and gestural backchannels are positioned differently in conversation, with gestural backchannels tending to arise in overlap with ongoing speech, while verbal backchannels tend to be placed in silent gaps. Thus, verbal backchannels may also be produced in response to different speaker behavior than gestural backchannels. Furthermore, in our data, it was not always possible to identify whether the speaker in question was able to see a potential visual backchannel produced by a listener. To be as consistent as possible, we thus include only locations with a verbal backchannel in the current study.
- **Change** After the completion point, the interlocutor takes the next full turn.
- **Question** The current turn ended in a syntactically marked interrogative form (with, e.g., subject–verb inversion or a wh-word), and the next turn was taken up by the interlocutor. The role of the question label was to help distinguish cases with a clear invitation for a next speaker from speaker change cases where the lexical content does not specifically invite a contribution from the next speaker. Due to their rarity, questions are not included in the turn-taking analyses below.
- **Ambiguous** This label was used when no clear decision could be made, e.g., when the interlocutor laughed or produced unintelligible vocalisations, or when both speakers overlapped, e.g., talking collaboratively until the end of the turn. Ambiguous turns are also excluded from the turn-taking analyses.

2.3. Feature extraction and quantitative analysis

Using scripts, we extracted completion points and the ongoing gesture phase at the completion point, as well as over stretches of time beginning at 3 seconds before the completion point and ending at 3 seconds after the completion point.

Each analysis reported below has different requirements for the statistical analysis; thus the individual statistical analyses are reported in the Results. All statistical tests

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

were calculated using R version 4.2.1 (R Core Team, 2022); $\alpha = .05$. Figures were generated using `ggplot2` (Wickham, 2016). The extracted data and R code are available at https://osf.io/efs4c/?view_only=16af14465e314724aa44ba709f051860.

3. Results

3.1. *Temporal alignment of gestures with turn ends*

Parts of this analysis follow a similar procedure to that used by Zellers et al. (2019b); however, the current dataset is much larger than was used in that study, and some of the annotations were refined since the time of the previous analysis.

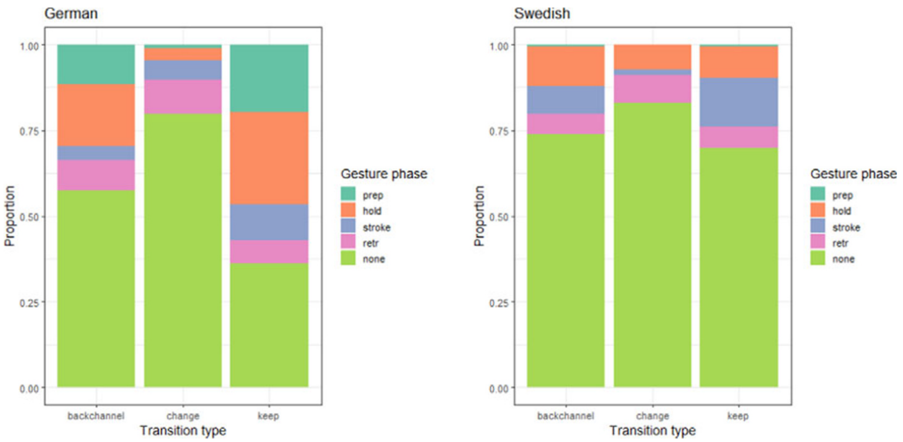
3.1.1. German

In the German data, we identified 451 completion points with the transition type *Backchannel*, *Change* or *Keep*. Of these, 223 (49.4%) had an ongoing gesture by the current speaker at the time of the offset of speech.

Figure 2a shows proportionally the gesture phase that was ongoing at the time of the offset of speech according to the type of turn transition in the German data; raw counts are given in Table 2. A χ^2 test shows that the difference in distribution of the gesture phases is different at different types of turn transition ($\chi^2(8) = 67.2, p < .05$). Specifically, significant residual values indicate that *preparations* and *holds* are more likely in *Keeps* and less likely in *Changes*. *Changes* are more likely to have no gesture (that is, *none*) and *Keeps* are less likely to have *none*. *Cramér's V = 0.546*, indicating a large effect size.

3.1.2. Swedish

In the Swedish data, we identified 511 completion points with the transition type *Backchannel*, *Change* or *Keep*. Of these, 125 (24.5%) had an ongoing gesture by the current speaker at the time of the offset of speech.



(a) German

(b) Swedish

Figure 2. Ongoing gesture phase at time of speech offset, German data left, Swedish data right. The y-axis shows the proportion of gesture phases at each transition type, rather than raw counts.

Table 2. Gesture phases according to transition types observed at the time of speech offset in completion points in German and Swedish

	German				Swedish			
	Backchannel	Change	Keep	total	Backchannel	Change	Keep	total
Preparation	14	1	47	62	1	0	1	2
Hold	22	3	65	90	23	10	15	48
Stroke	5	5	25	35	16	2	24	42
Retraction	11	9	16	36	12	11	10	33
None	70	71	87	228	152	114	120	386
Total	122	89	240	451	204	137	170	511

Figure 2b shows proportionally the gesture phase that was ongoing at the time of the offset of speech according to the type of turn transition in the Swedish data; raw counts are given in Table 2. As in German, a χ^2 test shows that the difference in distribution of the gesture phases is different at different types of turn transition ($\chi^2(8) = 19.86$, $p < .05$). Specifically, significant residual values show that *strokes* are more likely to arise in *Keeps* and less likely to arise in *Changes. Cramér's V = 0.282*, indicating a medium effect size.

3.1.3. *Cross-linguistic comparison*

In both languages, it was more frequent in general for turns to end without gesture than with gesture. Specifically, the likelihood that no gesture will be ongoing at the time of speech offset is highest in *Changes*, followed by *Backchannels* and *Keeps.* The pattern of the *retraction* phase is similar. The inverse can be seen for the *preparation* and the *hold* phases, which are more frequent at *Keeps* and *Backchannels* than at *Changes.*

There are almost no *preparations* at all at *Changes.* In both German and Swedish, there are more *stroke* phases in *Keeps* than in *Backchannels* or *Changes.*

In sum, we can say about the distribution of gesture phases at the offset of speech of the current speaker that the gesture phases which move into or take place within the gesture space (that is, *preparations*, *holds* and *strokes*) tend to correlate with the same speaker keeping the floor, while gesture phases which move out of the gesture space (*retraction*, or *none*) tend to correlate with a change in speakership.

3.2. *Timing of gestural activity around completion points*

The analysis in Section 3.1 provides a snapshot of what happens at the offset of speech at a potential turn boundary, that is, at a specific point in time. However, it is clear that the distribution of gesture phases must evolve over time. Thus we are left with the open question of how the distribution develops toward – and also away from – the single point of time where speech stops.

We therefore take the distributions of gesture phases as shown in Figure 2 and treat them as if they were spectral slices. Taking a slice every tenth of a second, starting from 3 seconds before the offset of speech to 3 seconds after, and arranging them horizontally according to gesture phases, we obtain distributions of gesture phases over time, which can also be divided for each transition type separately. The result of this analysis/transformation is shown in Figure 3. We chose 3 seconds following Pöppel (2009), who suggests this duration as the window of cognitive 'presence'.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

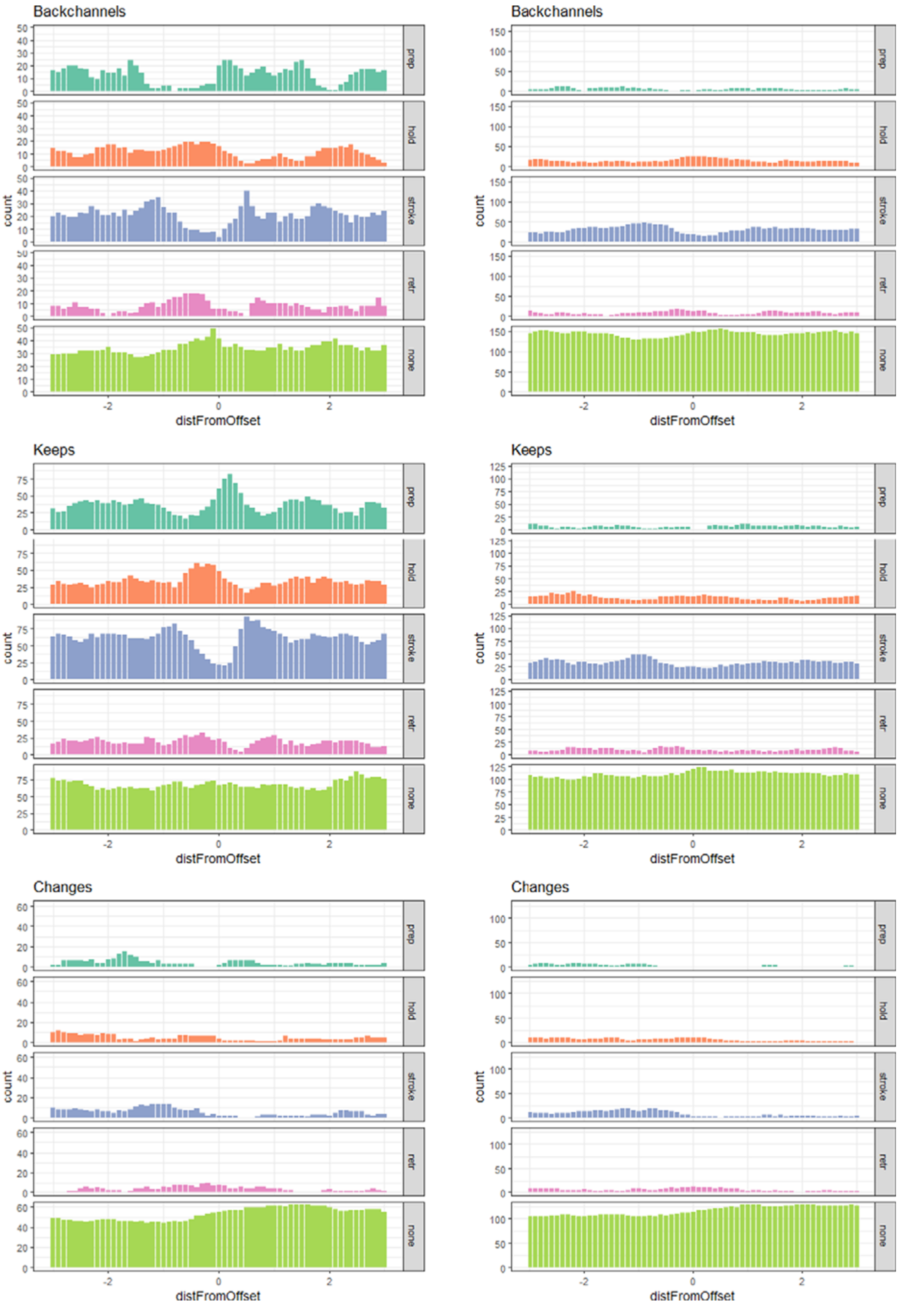


Figure 3. Distribution of gesture phases over time according to transition types; German data left, Swedish data right. The zero point is at the speech offset at a potential turn boundary (PTB). On the y-axis we count which gesture phase the current speaker is currently in. All counts across all phases sum up to the number of completion points in the data with the specific transition label. The counts in the slice at time point zero, accords with the numbers in Table 2.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

For each type of gesture phase, we conducted a mixed logistic regression with the criterion variable the frequency of the gesture phase and the fixed factors time, language, transition type and the interaction time:transitionType; the speaker was also included as a random factor (cf. function `glmer` in the R package `lme4`, Bartoń (2022)). The results for the model predicting the presence of gesture strokes are given in Figure 4; the model achieved an $R^2m = 0.091$ without the random factor of speaker and $R^2c = 0.283$ with the random factor. Expanded model results for all gesture phases are shown in Table 3, while the results of models for gesture phases other than

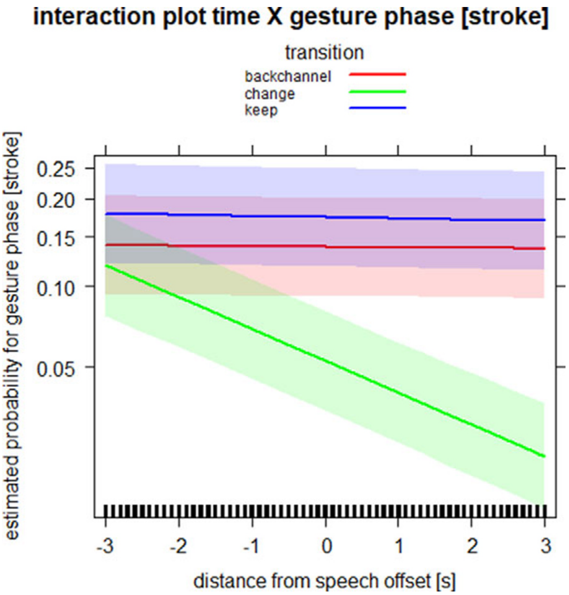


Figure 4. Model estimates for strokes. The x-axis shows the time offset from the completion point. The y-axis shows the estimated probability of strokes on a logarithmic scale. The random factor (speaker) is considered in the plot. We used the R package `effects` (Fox et al., 2022) in plotting the estimates.

Table 3. Upper part: statistical evaluation for each logistic model: R^2m = explained variation without random factor (speaker) and R^2c corrected for the random factor. Lower part: p-values (log probabilities) for each factor and the interaction with time. The reference group (Intercept) is backchannels in German. Bold text shows predictors that achieve statistical significance.

	prep	hold	stroke	retr	none
R^2m	0.138	0.078	0.091	0.017	0.153
R^2c	0.272	0.434	0.283	0.136	0.439
[Intercept]	**<0.0001**	**0.004**	**0.0006**	**<0.0001**	0.9
distFromOffset	**0.003**	**0.02**	0.7	0.1	**0.005**
Transitionchange	**<0.0001**	**<0.0001**	**<0.0001**	**0.0009**	**<0.0001**
Transitionkeep	**0.0001**	**<0.0001**	**<0.0001**	0.96	**<0.0001**
langSW	**0.002**	0.2	0.4	0.3	0.07
distFromOffset: transitionchange	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
distFromOffset: transitionkeep	**0.02**	0.6	0.7	**0.005**	0.3

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

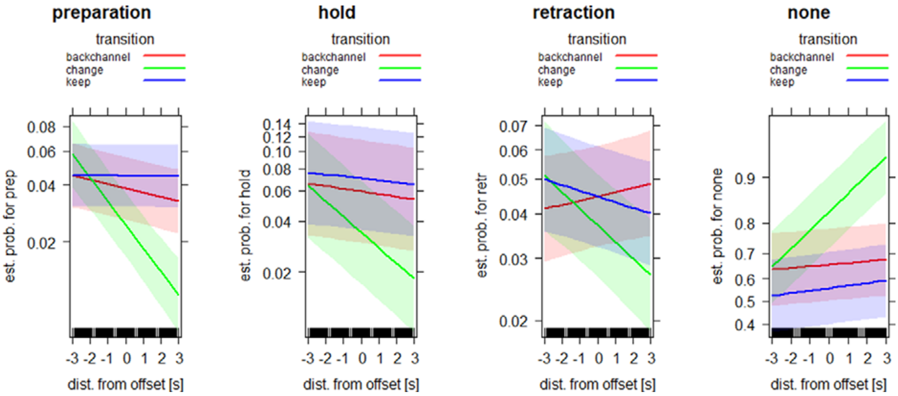


Figure 5. Estimates for the models for gesture phases *preparations*, *holds*, *retractions*, and *none.*

strokes are summarized in Figure 5. We calculated the R^2 values using the function `r.squaredGLMM` (Nakagawa & Schielzeth, 2013) in the R package `MuMIn` (Bates et al., 2022).

As the results in Section 3.1 have already shown, for *strokes*, there was a significant main effect of transition type on the frequency of *strokes*: Overall, participants produce more gesture *strokes* around *Keeps* than around *Backchannels* and more *strokes* around *Backchannels* than around *Changes.* The analysis here further shows a significant interaction between transition type and time: In *Changes*, *strokes* become progressively less frequent as the speech offset approaches and passes, while in *Keeps* and *Backchannels*, gesture *strokes* are equally probable preceding and following the end of the current turn. No significant effects were found for language in this model.

Although this finding is valid, it could be considered trivial, since it is already well-known that manual gestures are mostly performed by current speakers. However, a logistic regression attempts to fit a linear model, while, as the distributions shown in Figure 3 suggest, there might be more details in the evolution of gesture phases over time, which the logistic regression is unable to model. Therefore, to look deeper into the gesture dynamics, we calculated binomial tests for each point in time (that is, every tenth of a second), from which we obtained the probability of *stroke* activity at each time point as well as 95% confidence intervals for these probabilities. The resulting plot for gesture *strokes* is shown in Figure 6. Since previously no effect was found for language, both languages are modeled together.

While the results from the overall logistic regression did not show an effect of time for *Keeps* and *Backchannels*, Figure 6 shows that (i) *stroke* activity already differs as early as 3 seconds before the offset of speech between *Backchannels* (fewer *strokes*) and *Keeps* (more *strokes*). This difference however disappears between ca. −2.2 seconds before the completion point up to ca. 0.2 seconds after the completion point, where both in *Backchannels* and in *Keeps*, *stroke* activity first increases, reaching a peak at around 1 second preceding the speech offset, then decreases, reaching a valley at the speech offset and then starts to increase again. From ca. 0.4 seconds to 0.8 seconds following the speech offset, there is higher *stroke* activity at *Keeps* than at *Backchannels.*

This may relate to the *preparation* phases, shown in Figure 7. Interesting stretches of time here go from −0.5 seconds to +0.5 seconds and from ca. +1.8 to +2.2 seconds,

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

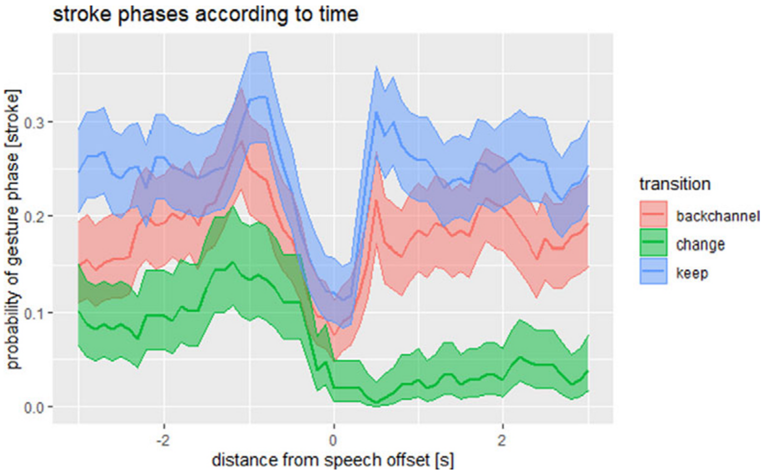


Figure 6. Probability of *strokes* over time (zero = offset of speech) according to transition types. Stretches where the confidence intervals do not overlap can be considered as significantly different.

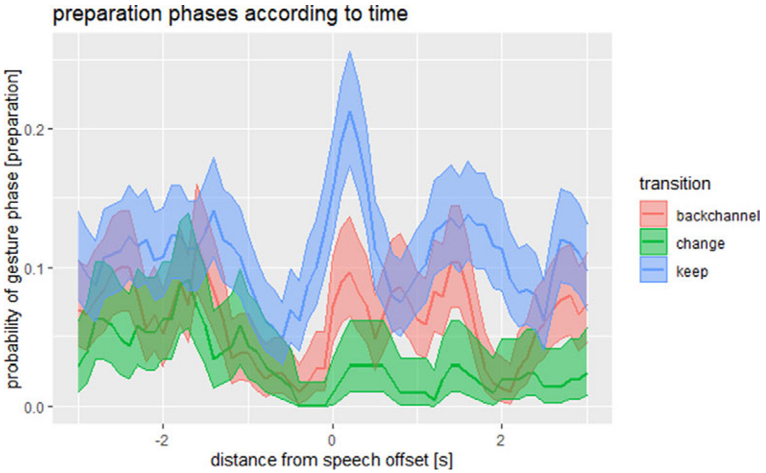


Figure 7. Probability of *preparations* over time (zero = offset of speech) according to transition types.

where the probability of gesture *preparations* at *Keeps* is higher than for *Backchannels*. The peak at the completion point could mean that at *Keeps*, the speaker already prepares upcoming *strokes* even before the completion point. The preparations at *Backchannels* seem to come slightly later.

The distribution of *hold* phases over time, shown in Figure 8, shows that *Backchannels* and *Keeps* do not differ significantly in terms of the presence of *holds*, but more gesture *holds* appear about 0.7 seconds before the completion point in *Keeps* than at *Changes*, and this higher probability of *holds* is retained throughout the remaining time.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

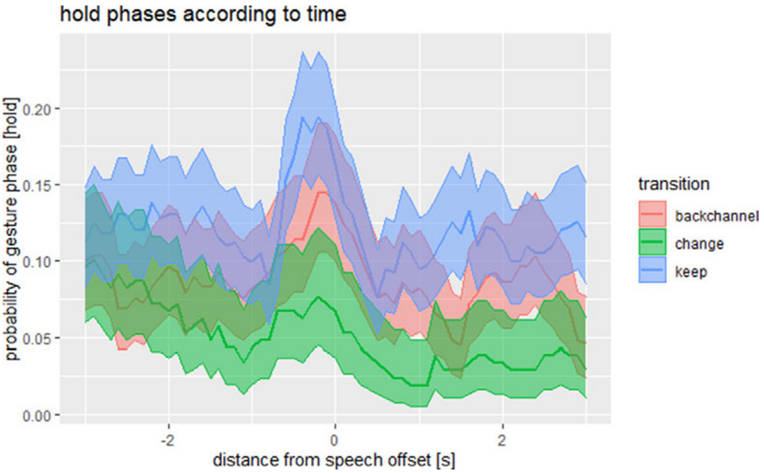


Figure 8. Probability of *holds* over time (0 = offset of speech) according to transition types.

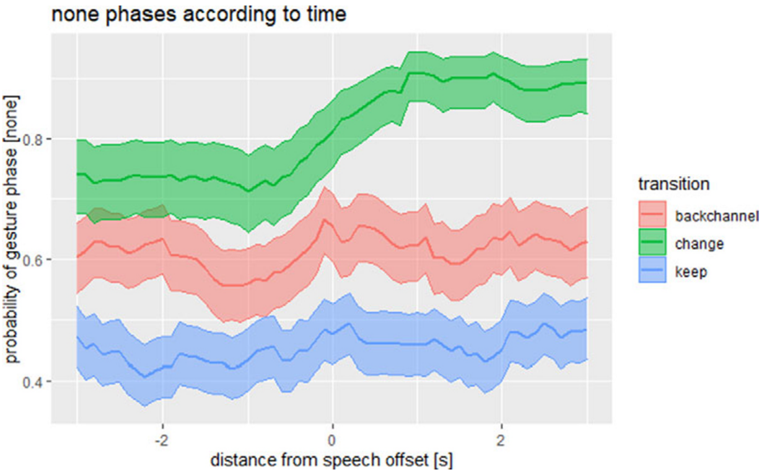


Figure 9. Probability of *none* over time (zero = offset of speech) according to transition types.

The distributions of *retractions* have overlapping confidence intervals throughout the entire time period investigated and are therefore not shown and discussed further.

The distributions of the last condition, *none*, are shown in Figure 9. The pattern for *Changes* seems to mirror that of the *strokes* (cf. Figure 6) with a higher probability of *none* phase after the completion point and the increase in likelihood of there being no gesture beginning about 0.5 second before the offset of speech.

In terms of the frequency of turn transitions arising without gesture (*none*), all three transition types are significantly different throughout the 6 seconds surrounding the completion point. The *none* category is highest at *Changes*, lowest for *Keeps*, with *Backchannels* in the middle.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

4. Discussion

With the current study, we took a first step to investigate precise temporal dynamics in the realm of turn-taking and gesture, where they have not yet been investigated quantitatively. We addressed the temporal relationships between manual gestures and the semantic and pragmatic content of conversational speech on two scales: the distribution of gestures in relation to turn-taking, as well as the overall distribution and alignment of hand gestures in the vicinity of potential turn boundaries.

We hypothesized that locations where a current speaker wishes to hold the floor (that is, *Keeps* and *Backchannels*) would demonstrate different temporal relationships between speech and gesture than cases in which the current speaker is ready to release the floor (that is, *Changes).*

We indeed found differences in gesturing behavior between *Keeps* and *Changes*, and to some extent between *Keeps* and *Backchannels.* The *Backchannel* locations may consist of a mixture of locations in which the current speaker wishes to keep the turn and locations where the current speaker would have been willing to allow an interlocutor to take up a turn but was 'refused' by the use of a backchannel (cf. Taboada, 2006; Yngve, 1970). Thus, the finding that gesture behavior in *Keeps* and *Backchannels* is similar but not identical is not unexpected. Gestural activity was more frequent and contained more strokes leading up to Keeps compared to Changes. Thus, the evidence from our study supports the interpretation suggested by Sacks et al. (1974), that participants must do more active work to keep the floor than to release it.

Kita et al. (1997) termed *preparation*, *stroke*, *partial retraction* and *retraction* as 'active phases' and distinguished them that way from gesture *holds* (p.34). Our results, however, indicate that the pattern of the *retraction* phase behaves in a way similar to the pattern of no gesture. So if we think in terms of movement effort of a gesture, we could rather classify *preparation*, *stroke* and *hold* as gesture phases that contribute to gesture activity and classify *retraction* and *none* as not contributing to an active gestural movement, that is, gesture passivity. In this sense, our results could mean that a high degree of gesture activity is more likely to lead to a *Keep*, while a reduction of gesture activity, that is, passivity, is more likely to invite a *Backchannel* or lead to a *Change* in speakership. Overall, gesture passivity (demonstrated by a gesture retraction or no gesture) might thus be an indication for interlocutors that at the upcoming turn end, some kind of contribution is expected.

From a multimodal perspective, gestures may be equally informative for participants of face-to-face conversations and complement other turn-taking cues, such as pitch, or even overwrite them. As Truong et al. (2011) have shown, compared to speech activity and mutual gaze, pitch was not a relevant factor in explaining the presence or absence of a backchannel signal (vocal, visual or bimodal). Our own previous work has also suggested that when gestural cues to turn-taking are available, pitch variation may be employed to a lesser extent (Zellers et al., 2019a).

Our investigation also included a cross-linguistic comparison. Findings on one language may not count for another language, but as we were analysing basic gestural properties (gesture phases) and not their emblematic use (Kendon, 1980), the influence of the language may be rather low, indicating a more general pattern of gesture dynamics and turn taking.

Although we did not find systematic or significant language differences, there appear to be differences in size of effects between German and Swedish (see Figure 3 and results for Cramér's V in Sections 3.1.1 and 3.1.2). These differences may arise

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

from other aspects of the data. First, our Swedish sample contained overall less gesturing than the German sample. Second, the conversational activities in the corpora differed, with the German conversations being in general more task-oriented than the Swedish ones. Thus, the current results are probably better understood as supporting the argument that gesture implementation in the vicinity of completion points is a universal communicative strategy, rather than one strongly mediated by linguistic structure.

4.1. Temporal features of hand gestures at turn ends

The analyses reported in Sections 3.1 and 3.2 provide evidence that hand gesturing overall is structured in a way that supports the structuring and management of turn-taking in conversation. This complements and expands upon findings by, for example, Kendrick et al. (2023), who report that *preparations* and *strokes* at TCU ends are associated with floor-holding by the current speaker, as well as by Kendon (1995) and Mondada (2007), who report functions of specific gesture shapes in terms of their contribution to the activity of holding or releasing the floor. By looking at all hand gestures, regardless of their form, we find that the *stroke* phase in general, that is, the obligatory and 'meaningful' portion of the gesture, also becomes rarer as a completion point approaches. At the time of speech offset, *strokes* are extremely rare in both German and Swedish, and this is further modulated by the type of turn transition that is taking place: ongoing *strokes* at the time of speech offset are more frequent in *Keeps*, where the same speaker intends to continue speaking, than at *Changes*, where a new speaker will take over. *Changes* are also the type of turn transition least likely to have any kind of ongoing gesture at the time of speech offset. Thus, for example, Schegloff's (1984) claim that hand gesturing is a current-speaker activity is supported by our quantitative analysis, and we can expand upon this by arguing that hand gesturing is also an activity that can indicate intentions about future speakership.

Expanding our view outward from the single time point of the offset of speech to look at the larger picture approaching and following a completion point, we see differences in gesturing behavior even at a substantial distance from this time point. Even in the few seconds preceding a completion point, gesturing is less frequent preceding *Changes* than preceding *Keeps* or *Backchannels.* In *Backchannels* and *Keeps*, *holds* remain similarly frequent, or may even increase, approaching the speech offset, and dip in frequency shortly afterward. In all types of transitions, the frequency of *strokes* is also at its peak about 1 second before the completion point. This suggests that stroking could be a useful early visual cue to an upcoming boundary, alerting an interlocutor to search for other cues about the current speaker's turn-taking intentions.

4.2. Limitations

The current study adopts the offset of speech as a reference point, thus taking the perspective of the current speaker. Different results might have arisen if our reference point was the onset of speech following our completion points, taking a more recipient-oriented perspective, as did Truong et al. (2011), who time-locked at the start of (verbal, visual or bimodal) backchannels and investigated the prior interlocutor's presence or absence of speech and mutual gaze. Our study also restricts gesture annotation to the four gesture phases following Kendon (2004). An addition to the *stroke* phase could be the annotation of the gesture apex, that is, the place of maximum effort (peak velocity, peak acceleration, or peak deceleration), cf. Pouw and Dixon (2019). A focus on the most prominent part of a gesture stroke might result in different (potentially sharper) distribution contours.

Another limitation of our study is our focus on the recipient's verbal behavior while not taking into account the visual resources of feedback such as head nods, facial expressions, gaze and so forth Including such annotations in future studies would better reflect the multimodal richness of face-to-face interaction. Similarly, it is beyond the scope of the current study to account for additional signaling on the part of the first speaker regarding his or her turn-taking intentions, either by lexical means or by variation in, e.g., pitch or duration; annotations of the turn-final pitch contours exist, and their relationship to gestural behavior will be explored in future research.

We were also limited by the available data. While many corpora of conversational speech exist, it is challenging to identify corpora that are sufficiently similar in terms of their structure. The interactional settings in particular are different for most corpora. In addition, our study used data from a different number of speakers in each language, meaning that for the χ^2 tests, which could not incorporate random factors, the differing amount of individual variability in the two languages could have influenced the statistical results. Once comparable data are available across languages, e.g., the parallel corpus of the PECII project (Kornfeld et al., 2023) with constant interactional settings, our study could be repeated, also taking the other shortcomings into account.

5. Conclusions

Annotating and carrying out quantitative analyses of conversational data could be interpreted as ignoring or overgeneralizing important complexity arising in conversational interaction; however, our larger-scale quantitative analysis has identified larger-scale temporal patterns arising across languages and across conversational settings. While interacting participants can still make sense of very context-specific cue organizations, we find evidence supporting the hypothesis that conversational participants systematically vary their gestural behavior in the approach to and at turn boundaries and that the temporal placement of different gesture phases (strokes versus holds versus other phases) shows a tendency to pattern similarly depending on the sequential structure of the conversational turn. We found differences only in the degree to which these patterns arose between German and Swedish, suggesting that these temporal patterns are either universal or that linguistic or cultural differences must be much larger to identify differences in timing behavior.

Future research will bring another prosodic parameter, the pitch contour at turn ends, into the equation. It will also expand the scope of the investigation beyond Germanic languages and beyond European cultures. These parameters will help us to refine our assessment of the universality of gesturing behavior at turn boundaries as well as its interaction with the linguistic system.

Data availability statement. Legal restrictions prohibit sharing of the raw video and audio dataset due to GDPR privacy restrictions and demands for anonymity of participants. The extracted data and R code for these analyses are publicly available at https://osf.io/efs4c/?view_only=16af14465e314724aa44ba709f051860.

Acknowledgments. We are grateful to Simon Alexanderson, Jonas Beskow and Jens Edlund for assistance with Spontal and to Caroline Kleen for supplementary annotation work. We also want to thank Sandra Hansen and Sascha Wolfer for valuable feedback on the statistical analyses and the anonymous reviewers for their comments and suggestions for improvement on previous versions of this paper. We gratefully acknowledge support from CLARIN SPEECH, the CLARIN Knowledge Centre for Speech Analysis.

Funding statement. This work was supported by the German Research Foundation (DFG; Aufbau Internationaler Kooperationen GO 3063/1–1, PE 2879/1–1, ZE 1178/1–1), the Swedish Research Council (VR-2017-02140) and the Riksbankens Jubileumsfond (P12–0634:1).

Competing interests. The authors declare none.

References

Ambrazaitis, G., & House, D. (2017a). Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization? In *Proceedings of 14th International Conference on Auditory-Visual Speech Processing*. Stockholm, Sweden.

Ambrazaitis, G., & House, D. (2017b). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, 95, 110–113. https://doi.org/10.1016/j.specom.2017.08.008.

Auer, P. (1996). On the prosody and syntax of turn-continuations. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 57–100). Cambridge: Cambridge University Press.

Auer, P. (2005). Projection in interaction and projection in grammar. *Text-Interdisciplinary Journal for the Study of Discourse*, 25(1), 7–36.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *Journal of the Acoustical Society of America*, 123(1), 354–365.

Bartoń, K. (2022). *lme4: Linear mixed-effects models using 'Eigen' and S4*. R package version 1.47.1. https://CRAN.R-project.org/package=lme4

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2022). *MuMIn: Multi-model inference*. R package version 1.1–31. https://CRAN.R-project.org/package=MuMIn

Bavelas, J. B. (1994). Gestures as part of speech: Methodological implications. *Research on Language and Social Interaction*, 27, 201–221.

Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*.

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. http://www.praat.org/.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52, 46–57.

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B*, 288(1943), 20202419. https://doi.org/10.1098/rspb.2020.2419.

Buanzur, T., Zellers, M., Namyalo, S., & Witzlack-Makarevich, A. (2018). A first investigation of the timing of turn-taking in Ruuli. In *Proceedings of Interspeech 2018*, Hyderabad, India (pp. 621–625).

Casasanto, D. (2008). Conceptual affiliates of metaphorical gestures. In *International Conference on Language, Communication, & Cognition*. Brighton, UK.

Casasanto, D. (2009). When is a linguistic metaphor a conceptual metaphor? *New Directions in Cognitive Linguistics*, 24, 127–146.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, 31, 251–276.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37(6), 871–887. https://doi.org/10.1016/j.pragma.2004.10.010.

de Ruiter, J., Mitterer, H., & Enfield, N. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3), 515–535.

de Vos, C., Torreira, F., & Levinson, S. C. (2015). Turn-timing in signed conversations: Coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*, 6, 268.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., & Small, S. L. (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping*, 30(11), 3509–3526.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proceedings of LREC 2010*, Valetta, Malta.

Esposito, A., Esposito, D., Refice, M., Savino, M., & Shattuck-Hufnagel, S. (2007). A preliminary investigation of the relationship between gestures and prosody in Italian. In A. Esposito, M. Bratanić, E. Keller, & M. Marinaro (Eds.), *Fundamentals of verbal and nonverbal communication and the biometric issue* (pp. 65–74). Amsterdam: IOS Press.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research*, 56(3), 850–864.

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In *Proceedings of LREC 2010* (pp. 86–91). Valetta, Malta.

Ferré, G., & Renaudier, S. (2017). Unimodal and bimodal backchannels in conversational English. *Proceedings of SEMDIAL*, 2017, 27–37.

Florit-Pons, J., Vilà-Giménez, I., Rohrer, P., & Prieto, P. (2020). The development and temporal integration of co-speech gesture in narrative speech: A longitudinal study. In: *Proceedings of GESPIN 2020*. Stockholm, Sweden.

Fox, J., Weisberg, S., Price, B., Friendly, M., & Hong, J. (2022). effects: Effect Displays for Linear, Generalized Linear, and Other Models. R package version 4.2–2. https://CRAN.R-project.org/package=effects

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, 2, 67–74.

Gravano, A., & Hirschberg, J. (2009). Turn-yielding cues in task-oriented dialogue. In P. Healey, R. Pieraccini, D. Byron, S. Young, & M. Purver (Eds.), *Proceedings of the SIGDIAL 2009 Conference* (pp. 253–261). London: Association for Computational Linguistics.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25, 601–634.

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology*, 9, 879.

Guellaï, B., Langus, A., & Nespor, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology*, 5, 700.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversation. *Journal of Phonetics*, 38, 555–568.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53, 23–35.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30(3), 1028–1037.

Kane, J., Yanushevskaya, I., de Looze, C., Vaughan, B., & Ní Chasaide, A. (2014). Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions. In *Proceedings of 15th Interspeech* (pp. 333–337). Singapore.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating speech and iconic gestures in a Stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience*, 22(4), 683–694.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Kay (Ed.), *The role of nonverbal communication* (pp. 207–227). Berlin/The Hague: De Gruyter Mouton.

Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction*, 27, 175–200.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in southern Italian conversation. *Journal of Pragmatics*, 23(3), 247–279. https://doi.org/10.1016/0378-2166(94)00069-V.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B*, 378(1875), 20210473. https://doi.org/10.1098/rstb.2021.0473.

Kita, S., van Gijn, I., & van der Hulst, H. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *International Gesture Workshop, Bielefeld, Germany, September 1997. Lecture Notes in Artificial Intelligence 1371* (pp. 23–35). Berlin: Springer.

Knight, D. (2009). *A multi-modal corpus approach to the analysis of backchanneling behaviour*. PhD Thesis, University of Nottingham.

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.

Kornfeld, L., Küttner, U.-A., & Zinken, J. (2023). Ein Korpus für die vergleichende Interaktionsforschung. Das Parallel European Corpus of Informal Interaction (PECII). In A. Deppermann, C. Fandrych, M. Kupietz, & T. Schmidt (Eds.), *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial* (pp. 103–127). Berlin/Boston: de Gruyter.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471.

Li, X. (2014). *Multimodality, interaction and turn-taking in Mandarin conversation* (Vol. 3). Amsterdam: John Benjamins Publishing Company.

Local, J., & Walker, G. (2012). How phonetic features project more talk. *Journal of the International Phonetic Association*, 42, 255–280.

Local, J. K., Kelly, J., & Wells, W. H. G. (1986). Towards a phonology for conversation: Turn-taking in Tyneside English. *Journal of Linguistics*, 22, 411–437.

Loehr, D. P. (2004), *Gesture and intonation*. PhD thesis, Georgetown University.

Max Planck Institute for Psycholinguistics (2018). ELAN *(version 5.2)*. Nijmegen, Netherlands. https://tla.mpi.nl/tools/tla-tools/elan/.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.

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

McNeill, D. (2005). *Gesture and thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2006). Gesture and communication. In K. Brown & A. Anderson (Eds.), *The Encyclopedia of language and linguistics* (2nd ed., pp. 58–66). Psycholinguistics Series. Amsterdam and Boston: Elsevier.

Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2), 194–225.

Mondada, L. (2019). Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics*, 145, 47–62.

Mondada, L., & Oloff, F. (2011). Gestures in overlap. The situated establishment of speakership. In Stam, G. & Ishino, M. (Eds.) *Integrating gestures. The interdisciplinary nature of gesture* (pp. 321–338). Amsterdam: John Benjamins.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133–142.

Nobe, S. (1996), *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production*. PhD thesis, University of Chicago.

Novack, M. A., Wakefield, E. M., & Goldin-Meadow, S. (2015). What makes a movement a gesture? *Cognition*, 146, 339–348.

Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1), 139–152. https://doi.org/10.1017/S0025100301001116.

Peters, B. (2006). *Form und Funktion prosodischer Grenzen im Gespräch – Ein phonetischer Beitrag zur Gesprächsforschung*. Saarbrücken: Südwestdeutscher Verlag für Hochschulschriften.

Pöppel, E. (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1525), 1887–1896.

Pouw, W., & Dixon, J. A. (2019). Quantifying gesture-speech synchrony. In A. Grimminger (Ed.), *6th gesture and speech in interaction conference – GESPIN 6* (pp. 75–80). Universitaetsbibliothek Paderborn.

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49(1), 41–54.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., & Ansari, R. (2002). Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9(3), 171–193. https://doi.org/10.1145/568513.568514.

https://doi.org/10.1017/langcog.2025.10014 Published online by Cambridge University Press

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, 51, 1507–1521.

Rossi, M., Feindt, K., & Zellers, M. (2022). Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish. In *Proceedings of Speech Prosody 2022* (pp. 185–189). Lisbon, Portugal.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, 50(4), 696–735.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, 11, 243–257.

Schegloff, E. A. (1984). On some gesture's relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 266–296). Cambridge: Cambridge University Press.

Schmidt, T. (2014). The research and teaching corpus of spoken German – FOLK. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)* (pp. 383–387). Reykjavik, Iceland: European Language Resources Association (ELRA).

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6, 357–388.

Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, 9, 1514.

Sikveland, R., & Ogden, R. (2012). Holding gestures across turns: Moments to generate shared understanding. *Gesture*, 12(2), 166–199.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26), 10587–10592.

Stivers, T., & Rossano, F. (2010). Mobilizing response. *Research on Language and Social Interaction*, 43(1), 3–31. https://doi.org/10.1080/08351810903471258.

Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer & A. di Luzio (Eds.), *The contextualization of language* (pp. 135–158). Amsterdam: Benjamins B.V.

Taboada, M. (2006). Spontaneous and non-spontaneous turn-taking. *Pragmatics*, 16(2–3), 329–360.

ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. *PsyArXiv*. https://doi.org/10.31234/osf.io/b5zq7

Truong, K. P., Poppe, R., de Kok, I., & Heylen, D. (2011). A multimodal analysis of vocal and visual backchannels in spontaneous dialogues. In *Proceedings of 12th Interspeech* (pp. 2973–2976). Florence, Italy.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232.

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

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting Chicago Linguistic Society, April 16–18, 1970* (pp. 567–578). Chicago: Chicago Linguistic Society.

Zellers, M. (2017). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. *Language and Speech*, 60(3), 454–478.

Zellers, M., Gorisch, J., House, D., & Peters, B. (2019a). Hand gestures and pitch contours and their distribution at possible speaker change locations: A first investigation. In *Proceedings of GeSpIn 2019* (pp. 93–98). Paderborn, Germany.

Zellers, M., Gorisch, J., House, D., & Peters, B. (2019b). Timing properties of hand gestures and their lexical counterparts at turn transition places. *Proceedings of FONETIK*, 2019, 119–124.

Cite this article: Zellers, M., Gorisch, J., & House, D. (2025). Temporal relationships between speech and hand gestures in the vicinity of potential turn boundaries in German and Swedish conversation, *Language and Cognition*, 17, e57, 1–22. https://doi.org/10.1017/langcog.2025.10014