Language Sciences 108 (2025) 101704

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Rhythm as an integration principle for modeling speech-action intersemiosis in classroom interaction: a social semiotic perspective

Xiaoqin Wu

College of International Studies, Southwest University, No.2 Tiansheng Road, Beibei District, 400715, Chongqing, China

A R T I C L E I N F O

Article history:
Received 27 January 2024
Received in revised form 6 December 2024
Accepted 13 December 2024
Available online xxx

Keywords:
Intersemiosis
Rhythm
Video studies
Classroom interaction
Multimodal transcription
Systemic-functional linguistics

A B S T R A C T

A continuing challenge for scholars working with multimodal educational research is to devise theoretical and methodological tools that can effectively navigate the complexity and emergent meaning when different semiotic resources interact. This paper demonstrates how rhythm, as an integration principle, coordinates the interaction of speech and embodied action in classroom settings at multi-scalar temporalities. Transcription designs are also devised to capture and visualize the patterns of multimodal rhythmic interaction. Drawing on a social semiotic theorization of rhythm, the paper conducts nuanced multimodal analyses of video data documenting teacher-student embodied interaction. The paper first reports four types of multimodal rhythmic patterns in classroom interaction, showcasing how rhythms coordinate across participants and semiotic resources. It then demonstrates how the tempo of the speech rhythmically structures the embodied actions at different time scales, resulting in multimodal synchronies that are semantically motivated. Finally, the paper reveals that the multiple actions in a pedagogic practice, while themselves rhythmical, may not always be rhythmically integrated with speech. The paper contributes to existing studies of speech-action interplay by developing theoretical and methodological tools to capture and visualize their interactions. Observations developed in this paper can also potentially inform pedagogic practices that involve the co-deployment of speech and embodied action.

© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Multimodal classroom interaction, intersemiosis and rhythm

The advent of digital technology prompted an upsurge of video studies in language and communication research (e.g. Baldry and Thibault, 2006; Deppermann, 2013; Deppermann et al., 2010; Fitzgerald et al., 2013; Flewitt, 2006; Goodwin, 2000; Hannula et al., 2022; Heath et al., 2010; Jacobs et al., 1999; Mondada, 2011, 2012, 2016, 2018; Norris, 2009). This video boom further occasioned the emergence and prevalence of multimodality studies informed by Systemic-Functional Linguistics (hereafter SFL) (e.g. Bateman et al., 2017; Bezemer and Mavers, 2011; Cowan, 2014) that examined 'the use of several semiotic modes in the design of a semiotic product or event together with the particular way in which these modes are combined' (Kress and Van Leeuwen, 2001: 20). The role of images, sound, animation, movement and other embodied

E-mail address: xiaoqinwu415@gmail.com.

https://doi.org/10.1016/j.langsci.2024.101704
0388-0001/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

semiotic resources were increasingly highlighted in the communication landscape. The social and cultural reshaping of the communication landscape spawned a large body of SFL-informed multimodal educational research that explored meanings expressed in various forms of communication and the interrelationship between them (e.g. Djonov et al., 2021; He, 2021; 2023; Jewitt, 2009; Lim, 2021; Lim et al., 2012; Ngo et al., 2021; Tseng and Djonov, 2023; Unsworth, 2008; Wu, 2022, 2024a, 2024b, 2025; Wu and Ravelli, 2021).

Intersemiosis, the process of coordinating and integrating different semiotic resources to create a meaningful text or event (Ravelli, 1995), has become a crucial concern in multimodal education studies informed by SFL since the late 1990s (e.g. He, 2023; Lemke, 1998; Lim, 2021; Matthiessen, 2009). Existing studies (e.g. Lemke, 1998; Matthiessen, 2009) already recognized that for a multimodal text to 'hang' together as one piece, there must be a certain degree of coordination among diverse semiotic resources. Matthiessen (2009) further noted that different combinations of semiotic resources might operate with different intersemiotic principles because, as Baldry and Thibault (2006: 4) cautioned, different semiotic resources adopt different organization principles to make meaning.

Early SFL-informed multimodal educational studies (e.g. Lemke, 1998) largely focused on static multimodal texts, so much attention was given to the interplay of image-text relationships that drew on spatiality (composition and layout) as the meaning-making mechanism. In contrast, temporality was somewhat backgrounded in their discussions. Recently, multimodal educational scholars informed by SFL (e.g. Bateman et al., 2017; He, 2023; Lim, 2021; Tseng and Djonov, 2023; Zhao, 2010) increasingly recognized the need to consider both time and space when analyzing intersemiosis in dynamic video data. However, a continuing challenge for multimodal educational scholars is to develop theoretical and methodological tools to account for the complexity and the emergent meaning when multiple semiotic resources interact in classroom interaction.

The co-deployment of speech and embodied actions in classroom interaction results in temporal co-emergence and creates intersemiosis. Drawing on examples from a corpus of video recordings of face-to-face classroom interaction in a tertiary setting, this paper aims to demonstrate how rhythm, as an integrative principle, coordinates the interaction of speech and embodied actions in the classroom to create a coherent and meaningful lesson. The rhythm model, which is a theoretical framework that focuses on the temporal coordination of speech and embodied actions, is employed as a way to understand how these semiotic resources interact in classroom communication. Embodied actions in this paper refer to the physical movements of the body that accompany speech, including embodied movement as a transition in space, gestures, nods and shifts in gaze and body orientation. Additionally, the paper develops transcription methods to capture and visualize the patterns of multimodal rhythmic interaction in the classroom.

A few SFL-informed multimodal educational studies have examined the intersemiotic relationship between speech and embodied actions in classroom interaction (e.g. Amundrud, 2017; Lim, 2011, 2021; Hao and Hood, 2019; Ngo et al., 2021). For instance, Lim (2021), who drew on ideas of language–image relations (Lim, 2004), proposed that speech and gesture could either formulate a co–contextualization relation whereby speech and gesture semantically converged or a re-contextualization relation whereby speech and gesture semantically diverged. Ngo et al. (2021) theorized embodied action as paralanguage that depended on language to make meaning and noted that paralanguage coordinated with the prosodic features of language. These studies contribute important insights to inquiries of the speech-action synthesis.

Given that speech and embodied action use temporality as the organization principle for making meaning (Deppermann, 2013; Deppermann et al., 2010; Kress, 2010; Lim, 2021; Mondada, 2016, 2018), this paper demonstrates how rhythm functions as an integration principle to organize the synthesis of speech and embodied action in classroom interaction at different time scales. Temporality, in this context, refers to the way in which time is used and structured in communication, particularly in the coordination of speech and embodied action. Speech and embodied actions are considered independent semiotic resources that are rhythmically coordinated in classroom interaction to make meaning together. Drawing on social semiotic studies of rhythm (Van Leeuwen, 1992, 2005; Martinec, 2000, 2018), the paper conducts nuanced multimodal rhythm analyses of video clips documenting classroom interaction. The paper finds that the rhythmic coordination between speech and embodied actions can occur across time scales and speaker turns in classroom interaction. These multimodal rhythmic patterns are semantically motivated and contextually conditioned.

This paper has both methodological and practical value. Methodologically, a rhythm model is particularly useful in investigating intersemiotic patterns across speech and embodied actions in dynamic classroom interaction at multi-scalar temporality, which is still descriptively challenging and thus largely under-explored in existing multimodality educational studies informed by SFL. Pedagogically, multimodal rhythm analysis of classroom interaction would enable an understanding of how interactions of complex semiotic resources in pedagogic practices facilitate a coherent lesson experience for the teacher and students.

2. A social semiotic account of rhythm informed by SFL

This paper largely draws on Van Leeuwen's (1992, 2005) social semiotic theorization of rhythm to explore how rhythmic patterns of speech and embodied actions interact at different time scales to fuse meaning together and create a more or less coherent lesson. The paper also draws on Martinec's (2000, 2018) ideas to examine how teachers and students jointly produce rhythms in their interactions in the classroom, which contributes to an understanding of how teachers and students coordinate with each other to play different roles in construing pedagogic experience.

Following Van Leeuwen (2005: 182), the essence of rhythm is a repeating alternation between two polar states: an up and down, a tense and lax, a loud and a soft, a night and a day, an ebb and a flow, and so on. Rhythm is not an alternation between

'steady states', but a wave-like motion (Van Leeuwen, 2005: 182). Similar to Halliday and Greaves' (2008) modelling of intonation, Van Leeuwen (1992, 2005) argues that rhythm plays a vital role in realizing information structure. That is, rhythm can realize textual meaning pertaining to the organization of discursive flow and the creation of cohesion (Halliday and Matthiessen, 2004: 30). Rhythmic patterning enables the speaker to anticipate what needs to be focused on and what carries the semantic weight, thus facilitating a successful understanding of the message (Van Leeuwen, 1992). Martinec (2000, 2002, 2018), who follows Van Leeuwen (1992, 2005), proposes a hierarchical model of rhythm whereby rhythm exists in monologue and dialogue. In addition to establishing prominence, Martinec (2002) finds that rhythmic patterns are related to different social relationships among participants and that rhythm can be jointly produced whereby rhythmic chains are extended across speaker turns.

A social semiotic account of rhythm includes rhythmic accentuation and rhythmic juncture (Van Leeuwen, 2005). Rhythmic accentuation is made more prominent and 'attention-catching'. It can be realized by diverse means, either in a single manner or a combined manner, such as increased loudness, pitch or duration, or, in the case of embodied action, some other form of increased force (Van Leeuwen, 2005: 189). The accentuation plays a key role in articulating meaning, because it foregrounds the sounds or movements that carry the key information, which helps to get the message across (Van Leeuwen, 2005: 183). Rhythmic juncture is concerned with the segmentation or boundary in the flow of time and is marked by a momentary interruption in the spacing of the accents. It can be realized diversely, such as a pause in the speech, a rallentando (slowing down) in a bodily action or some other discontinuity (Van Leeuwen, 1985). Rhythmic juncture creates a time frame for communicative acts (Van Leeuwen, 2005: 184). As the rhythmic grouping level goes up, the boundary becomes stronger.

For the purposes of this paper and the analysis below, each semiotic resource will be taken to have its own rhythm. Multimodal classroom interaction is thus a site of multiple rhythms, whereby diverse rhythms co-exist and interact. In this paper, only dynamic motions of different body parts with speech (spoken English) are analyzed. Based on repeated observations of data, it is these dynamic bodily actions that are rhythmically coordinated with the speech in classroom interaction. These include the movement of the head as nods and shifts in gaze, the movement of the hand and arm as gestures, the movement of the torso as shifts in body orientation, and the movement of the whole body as the embodied movement. By contrast, the analysis does not include the static state of the body, such as bodily posture and positioning in space.

Different semiotic resources have different ways to realize rhythmic accentuation and rhythmic juncture. In spoken English, which is foot-timed, rhythmic accentuation is realized by stressed syllables (Van Leeuwen, 2005; Halliday and Matthiessen, 2004). A stressed syllable can be made extra-prominent via a significant jump in pitch or an increase in duration or loudness. Each interval between two stressed syllables constitutes a foot, which acts as a unit of rhythm analysis in English (Halliday and Matthiessen, 2004). Feet are organized into tone groups that structure discourse into information units, with each information unit comprising the functions of (optional) Given and (obligatory) New. The tone groups play a crucial role in structuring information, with each unit carrying a specific function. A tone group is a unit of speech that has a tonic syllable carrying the main pitch movement: the main fall, rise or change of direction, and consists of stressed syllables and unstressed syllables that follow it (Halliday and Matthiessen, 2004: 89).

Fig. 1 presents a sample rhythm analysis of speech in this paper. There is one tone group (marked with '//') with one tonic syllable '***who***' (marked in italics and bold), as well as four feet (marked with '/') with one stressed syllable '*ward*' (marked in italics) and two silent beats (marked with '^'). As indicated in Fig. 1, a foot can start with a silent beat that maintains the rhythm even when an expectant beat is not articulated, just as in music. The tonic syllable '***who***' carries the major pitch movement and is made prominent and attention-catching at the level of the tone group. The stressed syllable '*ward*' is made prominent at the level of the foot. These two syllables carry the major semantic weight and communicate the key message in speech.

//^And/***who*** did the/^ erm a/*wards* one?

Fig. 1. A sample rhythm analysis of speech.

Embodied movement refers to the physical relocation of the whole body in space and is constituted by one moment of motion and two moments of stasis forming a promenade (McMurtrie, 2017). Following McMurtrie (2017), who further draws on van Leeuwen (2005), the repeating alternation of stasis and motion in a promenade creates rhythm. The second stasis is the prominent point in the promenade because the motion pauses, forms the boundaries in the promenade, and marks the point of arrival, which is analogous to New in speech. The transformation from stasis to motion at the beginning of the promenade creates the point of departure, which is analogous to Theme in speech. The steps in the promenade are also rhythmically timed (Wu, 2024a), akin to the downbeats in music or the stressed syllables in speech.

Fig. 2 illustrates a sample rhythm analysis of embodied movement. In this scenario, a promenade unfolds during a classroom interaction. This promenade, lasting 4 seconds and comprising four steps at 1-s intervals, results in four beats. The final step, marked with a bold '+', is the most significant as it represents the culmination of the entire promenade, akin to the tonic syllable in speech.

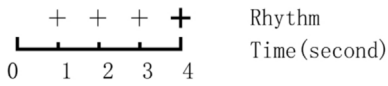


Fig. 2. A sample rhythm analysis of embodied movement.

Other embodied actions, including gestures, nods, and shifts in gaze and body orientation, are rhythmically organized with speech (Hood, 2011; Ngo et al., 2021). However, the realizations of prominence and juncture for these embodied actions have not been mapped out in existing multimodal educational studies informed by SFL. Thus, this paper annotates the occurrence of these bodily actions, aiming to understand how they are rhythmically coordinated with speech at different time scales. It is worth noting that the practice of transcribing embodied actions alongside the rhythm of speech is a common research practice in existing multimodal rhythm analysis (e.g. Hood, 2011; Norris, 2009), and this paper will follow this methodological approach.

3. Data and method

3.1. Video data of classroom interaction

The exploration of multimodal rhythmic interaction draws on examples from a corpus of video recordings of authentic classroom interaction situated in so called 'Active Learning Classrooms', as a part of a larger project that explores a multi-faceted understanding of space (i.e. the built environment) in the context of pedagogic practices in an Australian university (Wu, 2022). Video filming of classroom interaction is supported with full ethical approval (HC190413) and written consents from teachers and students of relevant classes. In order to reduce intrusion, the researcher only places one camera in the back corner of the classroom. This camera has a fish-eye lens, which enables a full capture of the whole classroom. The collection of classroom interaction in video format enables the researcher to zoom in on the video for specific interaction details. This paper conducts multimodal discourse analyses of speech, nod, gesture, shift in gaze and body orientation and embodied movement. Based on repeated observations of the data, these semiotic resources often operate together when the teacher and students interact with each other in the classroom. A multimodal rhythm analysis of speech and diverse bodily actions thus enables a nuanced investigation of dynamic multimodal gestalts.

Two teachers of film studies lessons, under the alias names of John and Emma, are selected in this paper for analysis and complement each other in terms of their pedagogic styles. While both teachers have more than 10 years of teaching experience, based on classroom observations, these two teachers manifest different pedagogic styles in their lessons. There is also a variation of student participation and teacher-student interaction in their lessons. The selection of different teachers with different pedagogic styles provides the possibility to explore how multimodal resources are taken up by different teachers and students to facilitate their pedagogic practices, which consequently results in diverse multimodal rhythmic patterns in the classroom.

Three clips of these two lessons are selected for detailed rhythm analyses. The first clip is conducted by Emma (female) and is part of an exercise – *Join the Dots*, whereby the teacher and the students discuss what they have written on the whiteboards to connect all the films they have studied so far. This clip is selected to demonstrate four types of multimodal rhythmic patterns produced in the classroom interaction. The second clip is conducted by John (male) and is part of a structure exercise – *Discussion of Structure Exercise,* whereby the teacher and the students discuss the answer for a structure exercise together. This clip is selected to demonstrate how different actions in classroom interaction can synchronize with the rhythm of speech at different timescales and how these synchronies are semantically motivated. The third clip is conducted by Emma again and is part of an exercise – *Nebraska and Indie films*, whereby the teacher and students describe the film features of *Nebraska* together and discuss the distinguishing feature of diegetic sound. This clip is selected to demonstrate how multiple actions in a pedagogic practice can be rhythmical in themselves but not integrated with the rhythm of speech in classroom interaction, conditioned by the situational context including the nature of the lesson activity and the design of the spatial environment.

3.2. Transcription methods

For the following multimodal analyses, multimodal transcripts are valuable because they guide the analyses, help develop insights, and provide verifiable evidence in developing an argument for the audience (Wu, 2022). Multimodal transcription of video data is an interpretative and representational process whereby the researcher needs to make complex decisions in terms of what is transcribed and how it is transcribed (Wu, 2022). Existing multimodal studies provide conventions for transcribing video data (e.g. Bezemer and Mavers, 2011; Cowan, 2014; Mondada, 2018). Elan is also a valuable tool for transcribing video data in different settings, but multimodal rhythm analysis is complex and time-consuming and cannot be adequately automatized. Additionally, existing conventions and software are not specifically designed for multimodal rhythm analysis, so it is quite challenging to visually demonstrate the multimodal rhythmic interactions if these conventions are used. Thus, this paper develops its transcription methods to capture the features of different semiotic resources and to visually present their interaction with the rhythm of the speech, drawing on principles of musical scores. It is worth noting that since

speech formulates the temporal references for embodied actions in the selected data in this paper, the proposed transcription methods only apply to situations in which speech is what Van Leeuwen (2005: 184) has called the 'guide rhythm', that is, speech is the primary resource and the main ongoing activity that guides the pace and structure of the interaction.

The representation of time is a crucial aspect of the multimodal transcription of video data because temporality is a fundamental organization principle for multimodal interaction (Deppermann, 2013; Kress, 2010; Mondada, 2018; Van Leeuwen, 2005; Wu, 2022). It is thus essential for the researcher to consider the temporal arrangements of different semiotic resources in the transcripts. The designed template in this paper allows readers to visually 'see' the points of alignment and the temporal unfolding of these interactions. It arranges temporality horizontally, with different semiotic resources and body parts detailed and separated on the vertical axis. Semiotic actions on the vertical line occur simultaneously, while everything on the horizontal line occurs consecutively. Multimodal transcripts are meant to be read from left to right and top to bottom.

Instead of representing time as a simple line axis, as seen in existing studies (e.g. Cowan, 2014; Deppermann, 2013; Mondada, 2018), this paper labels the time information at the bottom line at 1-s intervals, a standard unit of annotation for embodied actions in SFL-informed multimodal studies (e.g. Baldry and Thibault, 2006; Lim et al., 2012; Wu, 2024a, 2024b). This design enables a calculation of the duration for each occurrence of the semiotic resources in the interaction, which is a relevant factor in the following rhythm analysis at different time scales in Section 4.2. The concrete timing can also be verbally described in Section 4 in a reader-friendly manner, functioning as temporal deixis and helping readers quickly locate the points of multimodal interaction and match the verbal descriptions in the text with the visual transcripts.

The annotation shows several horizontal lines organized vertically. These lines are numbered consecutively from one to seven and cover the following forms of expression. The horizontal information is aligned vertically according to time. The first line represents speech rhythm and presents the prosodic features of speech. The second line represents the occurrence of nods. The third line represents the occurrence of gestures. The fourth line represents the occurrence and duration of gaze shift. The fifth line represents the occurrence and duration of the body orientation shift. The sixth line represents the movement rhythm and the duration of the movement, and the bottom line represents the time of the interaction. A new set of numbers is added in the annotation on the horizontal line when time is not continuous in the selected clips of annotations.

Unlike existing conventions that provide details of the embodied actions in the transcripts (e.g. Mondada, 2018), annotations of embodied actions in this paper are limited to their occurrence with no specification of their embodied features (e.g. gesture shape, shift target in gaze and body orientation, movement direction, etc.) for two reasons: on the one hand, multiple semiotic resources are transcribed in this paper, so a precise and comprehensive transcription of the embodied features in a limited space would amount to an unreadable multimodal transcript; on the other hand, the simplification of embodied actions in the transcripts enables a focused display of their interaction with the rhythm of speech as alignment or disalignment, which is the key objective of the analysis in this paper. These transcription designs ensure that the annotations are relevant to the analysis and maintain the readability of the transcripts.

Close-up photos of the selected video are embedded in the transcripts to recover situational context in the interaction and to provide relevant and specific features of the visual records, such as the positioning place, the interactive participants and the details of embodied actions. These photos play a crucial role in supporting and enriching the symbolic annotations in the multimodal transcripts, thereby enhancing the readers' understanding and knowledge. They also enable the readers to understand and verify the claims made by the researcher (Mondada, 2018). The exact moment in the video the photos refer to is specified through the spatial alignment of the photos with the line of speech, indicated by a symbol '#' on the lines of the speech and the photo (Mondada, 2017). Sometimes, several photos are displayed consecutively to demonstrate the trajectory of the embodied actions. Circles and arrows are also annotated in the photos to highlight the relevant details and facilitate understanding of the photos (Mondada, 2017). More precisely, a green arrow represents gaze shift, yellow a body orientation shift, and blue an embodied movement. In contrast, a red circle highlights the occurrence of hand gestures and nods, with their quantity showing their frequency. However, instead of inserting the photos between the horizontal lines (e.g. Deppermann, 2013; Mondada, 2018), this paper places them right above the speech line because speech formulates the temporal references for embodied actions in the selected data. This spatial arrangement can also clearly demonstrate how embodied actions coordinate with the rhythm of speech without any image getting in the way between the speech and embodied actions.

The identity of the participants in the interaction is specified in the multimodal transcripts. Initials are used to represent the source of the speech. More particularly, TE represents teacher Emma, TJ represents teacher John, and S1, Ss, and S2 represent the specific students involved in the interaction. These initials are highlighted in bold and placed before each utterance, a critical step that supports the reader in identifying the speaker and forming boundaries between different utterances. The participant in the embodied actions is implicitly linked back to the speaker's identifications in the speech track. The participants' initials are also marked up in the photos to help the reader identify the source of the embodied actions.

The embodied movement is annotated as a transition in space and involves the whole body's movement (McMurtrie, 2017). It should be noted that only the teachers move their whole body as a transition in space in the selected data, so there is no movement of students' whole torso. During the annotation, the promenade is considered complete if the motion stops and the body remains positioned in one space over 2 seconds. Two seconds is used as the reference point because, based on repeated observations of the data, the teacher sometimes slows down during their promenade, and one step can take about 2 seconds before the enactment of the next step. By contrast, if the teacher remains positioned over 2 seconds, there is often no further enactment of steps until the next promenade.

If there is no transition in space but movements of the torso that change the body's posture, this is identified as a shift in body orientation (Kendon, 1990). If a head movement changes the target of the gaze, it is identified as a shift in gaze (Kendon, 1967). Otherwise, it is identified as a nod. The shift is considered complete if the motion during the shift in gaze or body orientation stops. The gesture is identified as the movement of the hand and arm (Kendon, 2004). Following Kendon (2004), gestures can be segmented as emerging, shaping, and withdrawing. However, as stated earlier, the annotation of gestures in this paper is limited to their occurrence to focus on rhythmic interactions. Additionally, the gestures in the selected data progress quickly in a limited duration. So, a gesture is annotated only when the researcher recognizes its shape without making any distinction between the shape itself.

When it comes to representing visual details in a transcript, the use of initials and symbols is key. This approach allows the researcher to communicate speech, nods, gestures, and shifts in gaze and body orientation efficiently despite the space limitations.

For the rhythm of spoken English, this paper follows the transcription conventions in SFL (Halliday and Greaves, 2008) as introduced above. A double forward slash '//' represents tone group boundaries, and the tonic syllable is formatted in bold and italics. A single forward slash '/' represents a foot boundary, and the stressed syllable is formatted in italics. If the stressed syllable is made extra-salient, it is represented by an arrow '↑'. A caret symbol '^' represents a silent beat.

For the rhythm of embodied movement, a symbol '+' represents the steps that create beats in the promenade. This symbol, in bold '**+**', represents the prominent point of at the promenade level. A symbol '〰' represents the duration and boundary of the promenade.

Given the frequent occurrence of nods and gestures in a short interaction, their duration is not annotated. Instead, their frequencies are indicated by the number of initials. TN represents the teacher's nods, SN the student's nods, TG the teacher's gestures, and SG the student's gestures.

Shifts in gaze and body orientation take a relatively longer time than nods and gestures in the selected data, so their duration is represented. In terms of gaze shift, TGS stands for the teacher's gaze shift, SGS stands for the student's gaze shift, and T/SGS stands for both the teacher's and the student's gaze shift. A symbol '——' represents the duration and boundary of the gaze shift. In terms of body orientation shift, TBOS stands for the teacher's body orientation shift, SBOS stands for the student's body orientation shift, and T/SBOS stands for both the teacher's and the student's body orientation shift. A symbol '----------' represents the duration and boundary of the shift in body orientation.

In order to enhance the validity of the transcription, the approach described in this paper fostered a collaborative environment by inviting two experienced scholars in SFL and one music scholar without any linguistic background to listen, watch, and annotate the three selected video clips together with the researcher. The three scholars are not involved in the project, but their diverse expertise and perspectives enrich the research. Given the complexity of video data and classroom interaction, the selected videos are watched and annotated multiple times in a thorough process. The four scholars first listen to the audio of the data several times to annotate the prosodic features of speech. The software Praat is used in the transcription process to facilitate the auditory analysis. Then, the scholars watch the video footage to identify and annotate the features of embodied actions, starting with embodied movement, then shifts in gaze and body orientation, and finally, nods and gestures. After that, they re-watch the videos to merge the embodied actions' features with the speech's prosodic features on a single timeline. Finally, the annotations are cross-checked and revised until agreement is reached.

4. Multimodal rhythm analysis of classroom interaction

This section conducts a detailed multimodal rhythm analysis of the three selected video clips that document classroom interaction, with each video clip as the unit of analysis. The section begins by identifying four types of multimodal rhythmic patterns in classroom interaction, showcasing how rhythms coordinate across participants and semiotic resources. It then delves into the intricate exploration of how diverse embodied actions interact with the rhythm of speech at various time scales.

4.1. Four types of multimodal rhythmic patterns in classroom interaction

This subsection demonstrates four ways teachers and students produce multimodal rhythms in their classroom interaction. In this example of classroom interaction, the teacher, Emma, asks the students whether the films they have studied so far have all won an Oscar. One student (S1, female) replies that not all the films have won an Oscar but they are all at least nominated. Emma elaborates further that these films are all profit-making and critically recognized in some way. The overall interaction lasts 31 s.

The speaker, Emma, enacts embodied actions such as movement, gaze and body orientation to synchronize with her own speech. The speech's tempo rhythmically structures her embodied action. For instance, in Excerpt 1, from the 1s to the 5s, Emma articulates, 'who did the, erm, awards one? Who was doing the awards?' (line 1). At about the 1.5s, Emma enacts a shift in gaze and body orientation to synchronize with the tonic syllable '***who***' in the speech (line 1, lines 4–5, Fig. 1.1). These shifts continue to the 4s and are in sync with the verbal articulation – 'who did the, erm, awards one'. The temporal trajectories of the embodied actions largely overlap with that of the speech (line 1, lines 4–5). At about the 4s, Emma enacts another shift in gaze and body orientation to synchronize the tonic syllable '***who***' in the speech (line 1, lines 4–5, Fig. 1.2). From the 4s to the 5s, Emma speeds up in her second sentence by articulating similar words in just one second (line 1). Her second shift in gaze

and body orientation also speeds up to synchronize with her speech, resulting in similar temporal trajectories between the shift in gaze and body orientation and the speech (line 1, lines 4–5). At about the 4s, Emma enacts a promenade to move to the classroom front (line 6, Fig. 1.2 and Fig. 1.3). This promenade continues to the 6.5s, and the first two steps in the promenade are in sync with the two stressed syllables in the speech – '***who***' and '*ward*' (line 1, line 6).

#figure 1.1

#figure 1.2

#figure 1.3

1	speech	**TE**: //^And/***who*** did the/^erm a/*ward*s one? //***Who*** was/*do*ing the a/*ward*s? //^You/***guy***s were /*do*ing the a/*ward*s//***weren't*** you?
2	nod	SN
3	gesture	SG
4	gaze	TGS TGS S/TGS
5	body orientation	TBOS TBOS
6	movement	+ + + +
7	time	1S 2S 3S 4S 5S 6S 7S

Excerpt 1

Alternatively, the speaker can construct a single rhythm of speech with no other embodied actions to synchronize with it. In Excerpt 2, from the 7s to the 8s, Emma verbally articulates, 'Did any film not win an Oscar?' (line 1). Emma remains positioned in the classroom front and enacts no embodied actions to synchronize with her speech (lines 2–6, Fig. 2.1). Similarly, from the 21s to the 24s, Emma verbally articulates, 'Just nominated, but they are all at least in the awards season, weren't they?' (line 1). No embodied action is enacted to synchronize with her speech (lines 2–6). She remains positioned in the student pod centre and looks at S1 (Fig. 2.2). At this point, the speech primarily undertakes the semiotic labour to communicate the information at stake.

#figure 2.1

#figure 2.2

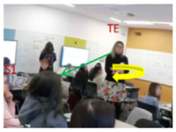
1	speech	**TE**: //*Did* any film/*not* win an/***O***scar?		1	speech	**TE**: //***Jus***t nominated. //*But* they are /***all*** at least/*in* the a/*ward*s season/↑*weren't* they?			
2	nod			2	nod				
3	gesture			3	gesture				
4	gaze			4	gaze				
5	body orientation			5	body orientation				
6	movement			6	movement				
7	time	7S	8S	7	time	21S	22S	23S	24S

Excerpt 2

The teacher and the student can also *jointly* produce rhythmic patterns in their interaction. In other words, multimodal rhythms can be produced across speaker turns. One manifestation of the joint production of rhythms in the interaction is that another speaker enacts embodied actions to synchronize with one speaker's speech. In other words, the tempo of speech by one speaker rhythmically coordinates the embodied action of another speaker. For instance, in Excerpt 3, from the 16s to the 20s, Emma enacts one promenade, largely in sync with S1's speech – 'Oh, no. Someone, er, some films just nominated' (line 1,

line 6). This promenade has six steps, and other than the first step, the remaining five steps are in sync with the five stressed syllables in S1's speech, including '*oh*', '*some*', '***just***' and '*no*' (line 1, line 6, Fig. 3.1). Additionally, from the 19.5s to the 20.5s, the teacher, Emma, enacts one shift in gaze and body orientation to synchronize with S1's speech – 'just nominated' (line 1, lines 4–5, Fig. 3.2). At the 19.5s, the shifts in gaze and body orientation are in sync with the tonic syllable '***just***' in the speech (line 1, lines 4–5). The speech of S1 rhythmically structures the embodied actions of Emma.

#figure 3.1

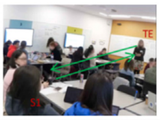


#figure 3.2

1	speech	**TE**: //*Did* they/↑*all* win/***O***scars? **S1**: //oh, no. // *Some*one/^ er /*some* films /***just*** *no*minated.
2	nod	
3	gesture	SG SG SG SG SG
4	gaze	TGS SGS SGS TGS
5	body orientation	TBOS
6	movement	+ + + + + +
7	time	14S 15S 16S 17S 18S 19S 20S

Excerpt 3

Another way to co-produce rhythms in the interaction is when both the teacher and the student enact embodied actions to synchronize with the speech at stake. For instance, in Excerpt 4, from the 5s to the 7s, Emma articulates, 'You guys were doing the awards, weren't you?' (line 1). At about 5.4s, Emma and S1 enact one shift in gaze to synchronize with the tonic syllable '*guys*' in Emma's speech (line 1, line 4, Fig. 4.1). Also, from the 9s to the 10s, during S1's speech – 'erm, yeah' (line 1), both S1 and Emma enact one shift in gaze to synchronize with the speech, resulting in similar temporal trajectories between the gaze shift and the speech (line 1, line 4, Fig. 4.2). This joint synchrony by the speaker and the audience might indicate a shared participation and engagement in the ongoing conversation.



#figure 4.1

#figure 4.2

1	speech	**TE**: //^You/***guy***s were doing the a/*wards*//***weren't*** you?	1	speech	**S1**: //^err, yeah.
2	nod	SN	2	nod	
3	gesture	SG	3	gesture	
4	gaze	S/TGS	4	gaze	S/TGS
5	body orientation		5	body orientation	TBOS
6	movement	+ +	6	movement	
7	time	5S 6S 7S	7	time	9S 10S

Excerpt 4

The multimodal synchronies in the interaction appear crucial in construing a sense of rhythmic coordination across different participants and different semiotic resources in communicative practice. This finding aligns with Norris (2009), Deppermann (2013) and Mondada (2018) who argue that rhythms can be produced both within and across turns. In the analysis of daily conversation, Norris (2009) demonstrates how the embodied hand-arm movement of another speaker can reproduce the rhythm of speech by one speaker. This paper finds further multimodal rhythmic patterns in classroom interaction; that is, another speaker can silently participate in the ongoing conversation by synchronizing their embodied actions, as a transition in space or as a shift in gaze and body orientation, with the speech rhythm of one speaker, as shown in Excerpt 3 and Excerpt 4. The multimodal analysis also demonstrates that while each semiotic action might operate with

different temporalities, they are finely coordinated and synchronize to different degrees with the rhythm of speech to constitute a higher level of pedagogic practice and formulate a coherent lesson.

4.2. Multimodal synchronies at different levels and semantic motivations

This section demonstrates how multiple actions in classroom interaction are hierarchically organized and can synchronize with the rhythm of speech at different time scales. On this basis, the paper discusses how these synchronies are semantically motivated.

In this example, the teacher, John, discusses the placement of a referencing sentence in an essay with the students collectively. John and the students (Ss, the students as a whole) affirm that this type of sentence should be placed in the paragraph. After the affirmation, John elaborates on the importance of placing some referencing sentences that refer back to the argument in the paragraph of an essay, somewhere between the introduction and the conclusion. The overall interaction lasts 26 s, and it is John who primarily speaks during the interaction. In addition to speech, John uses different body parts to communicate with the students in the interaction. These includes the movement of his hand, head, torso, and the whole body, resulting in different levels of semiotic actions. The students remain seated throughout the interaction.

4.2.1. Larger time scales, larger bodily synchronies

Following Norris (2009), each of these semiotic actions in the classroom interaction has its own rhythms and might operate with different time scales. My analysis finds that the tempo of the speech rhythmically structures the embodied actions at different time scales and that the larger the time scale is, the larger the bodily synchrony is involved.

At a small time scale and roughly at the interval of each articulation of a word (less than one second), John's communication becomes a symphony of speech and bodily actions. His gestures and nods create rhythmic beats that mirror the pace of his speech, and the number of these beats corresponds to the number of words in his speech. For instance, in Excerpt 5, from the 11s to the 13s, he verbally articulates five words – 'not necessarily every single paragraph', and enacts five nods and five gestures separately to synchronize with the speech (lines 1–3, Fig. 5.1). Similarly, from the 14s to the 18s, he verbally articulates 12 words – 'but there should be some references at some points in your essay', and enacts 12 nods and 12 gestures separately to synchronize with the speech (lines 1–3, Fig. 5.2). This synchronization between the frequency of the occurrence of nods and gestures and the number of words continues from the 11s to the 26s when John verbally elaborates that a referencing sentence should be placed in a paragraph between the introduction and the conclusion of an essay (lines 1–3).

#figure 5.1

#figure 5.2

1	speech	TJ://↑*Not*necessarily/*e*very/↑*sin*gle /*pa*ragraph. //*But*there/↑*should* be some /*re*ferences at some/*po*ints //*in* your/*e*ssay. //*How* the /*po*ints
2	nod	TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN
3	gesture	TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG
4	gaze	TGS TGS TGS
5	body orientation	TBOS TBOS TBOS
6	movement	+ + +
7	time	11S 12S 13S 14S 15S 16S 17S 18S 19S 20S

1	speech	/*you* are/↑*ma*king are re/*la*ting/*back* to that/↑*que*stion/*also*the/*task*. //*May*be halfway/*po*int be/*tween in*tro/*du*ction and conclusion/↑*some*thing //*like* that.
2	nod	TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN
3	gesture	TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG TG
4	gaze	TGS TGS TGS
5	body orientation	TBOS TBOS
6	movement	+ + + + + +
7	time	21S 22S 23S 24S 25S 26S

Excerpt 5

Additionally, his gestures and nods adjust their beats to synchronize with the tempo of his speech. From the 11s to the 13s, John articulates five words (line 1); from the 14s to the 18s, John articulates 12 words (line 1); from the 19s to the 22s, John articulates 15 words (line 1); from the 23s to the 26s, John articulates ten words (line 1). John speeds up his speech in his second and third articulation and then slows down during his final articulation. His nods and gestures also fasten during the second and third verbal articulation and then slow down during the final articulation.

It is worth noting that rapid hand gesture aligning with the tempo of the speech is a common observation in SFL-informed paralanguage studies (e.g. Hood, 2011; Ngo et al., 2021). However, my research extends these studies by providing concrete transcription methods to visualize their rhythmic alignments. Besides hand gesture, my research shows that nods can also rhythmically align with speech and is a common semiotic resource used by John in his lesson. Based on my observation in situ, John often synchronizes his hand gesture and nods with the tempo of the speech when he elaborates on a specific knowledge point.

At a larger time scale in Excerpt 6, John shifts his head and torso to synchronize with his speech in a more extensive duration. These shifts in gaze and body orientation (lines 4–5) often takes a longer time than the occurrence of gesture and nods (lines 2–3). For instance, from the 11s to the 12.5s, John enacts one shift in gaze and body orientation to synchronize with his speech – 'not necessarily every single' (line 1, lines 4–5, Fig. 6.1). These shifts last over one second and correspond to four words in the speech (line 1 and lines 4–5). Similarly, from the 17.5s to the 20s, John enacts one shift in gaze and body orientation to synchronize with his speech 'in your essay, how the points' (line 1, lines 4–5, Fig. 6.2). This synchronization between the shift in gaze and body orientation and the speech lasts over 2 seconds and corresponds to six words in the speech.

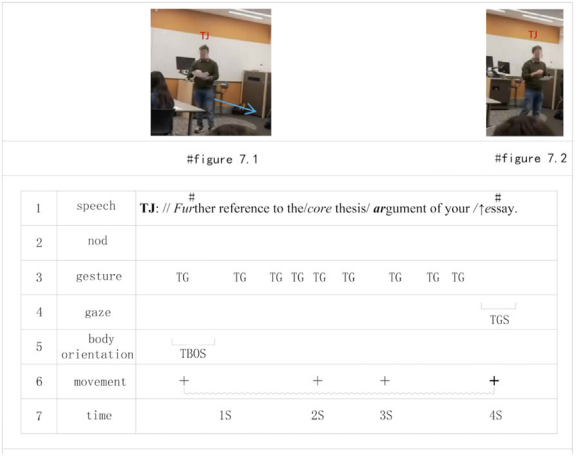
#figure 6.1

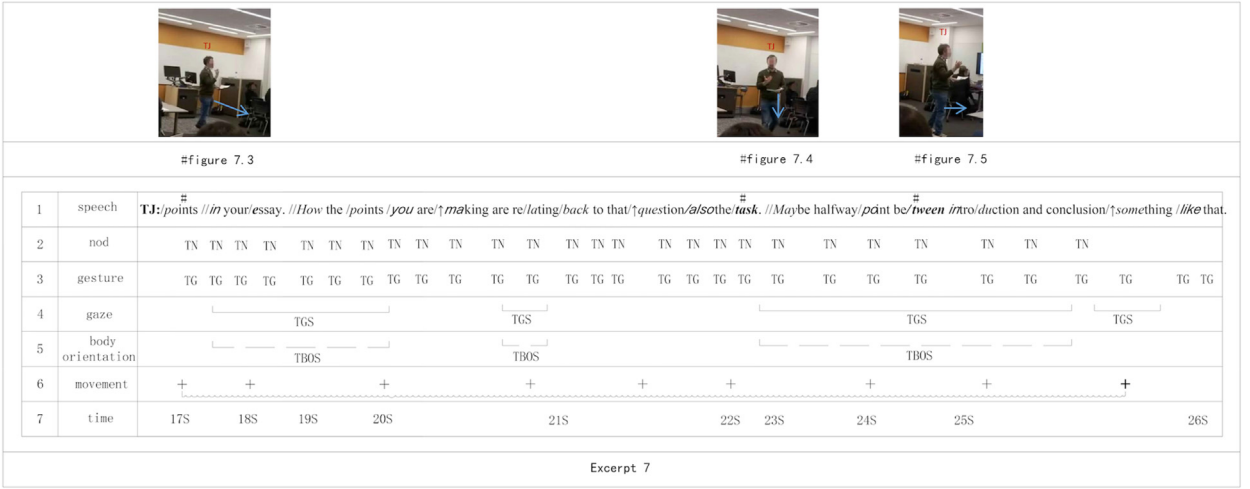
#figure 6.2

1	speech	**TJ:** //↑*Not* necessarily/*e*very/↑*sin*gle /*pa*ragraph.					1	speech	**TJ:** /*po*ints //*in* your/*e*ssay. //*How* the /*po*ints					
2	nod	TN	TN	TN	TN	TN	2	nod	TN	TN	TN	TN	TN	TN	TN
3	gesture	TG	TG	TG	TG	TG	3	gesture	TG	TG	TG	TG	TG	TG	TG
4	gaze	TGS					4	gaze	TGS						
5	body orientation	TBOS					5	body orientation	TBOS						
6	movement						6	movement	+		+				+
7	time	11S	12S			13S	7	time	17S		18S		19S		20S

Excerpt 6

At an even larger time scale in Excerpt 7, John relocates his whole body and moves himself to different places in the classroom during the interaction, and these promenades synchronize with the speech rhythmically. For instance, from the commencement of the interaction to the 4s (4-s duration), John enacts one promenade. This promenade has four steps, each enacted at 1-s intervals (line 6). During this promenade, John moves from the classroom front to the student pod centre (Fig. 7.1 and 7.2). The whole promenade is in sync with the speech – 'Further references to the core thesis argument of your essay' at the level of tone group, and the four steps in this promenade are in sync with the four stressed syllables '*fur*', '*core*', '***ar***', and '*e*' at the level of the foot (line 1, line 6). From the 17s to the 26s (about 9-s duration), John enacts another promenade (line 6). This promenade has nine steps, each not regularly timed at 1-s intervals but rhythmically aligned with the speech. These nine steps in the third promenade are in sync with the nine stressed syllables in the speech – '*po*', '***e***', '*you*', '*la*', '*ques*', '***task***', '*po*', '*du*', '*some*' (line 1, line 6). During this promenade, John moves from the classroom front to the student pod centre (Fig. 7.3 and Fig. 7.4) and then moves around the student pod centre to face the students at different pods (Fig. 7.5).





The above finding aligns with the viewpoints of Martinec (2000), Norris (2009) and Ngo et al. (2021); that is, rhythms of different semiotic actions are linked and rhythmic interaction can occur at different levels of social interaction. This paper also extends their research by specifying and demonstrating that the level of multimodal synchrony might relate to the portion of the body and the duration of time in bodily actions. In particular, the analysis demonstrates that in the selected data, at a small time scale, speech can rhythmically coordinate with small bodily actions such as nods and gestures; at a larger time scale, speech can rhythmically coordinate with larger bodily actions such as a shift in gaze and body orientation; at an even larger time scale, speech can rhythmically coordinate with whole bodily action such as movement.

4.2.2. Semantic motivations for multimodal synchronies at different levels

The above analysis demonstrates how speech in classroom interaction rhythmically coordinates embodied actions at different time scales. This section discusses that these synchronies are not randomly enacted in classroom interaction but are meaningful choices the participants make for successful communication; that is, these multimodal synchronies at different levels are semantically motivated.

In order to discuss the semantic implications of multimodal synchronies, one notion in SFL, metafunction, is used in this section. Following Martin and Rose (2007) who further draw on Halliday (1978), language simultaneously construes three types of metafunctional meaning: the ideational meaning that construes human experience in terms of entity and activity sequence; the interpersonal meaning that models interactions as an exchange between participants that assigns them different roles; and the textual meaning that organizes the text as waves of information. In addition to language, metafunction has also been used as the meaning mechanism in embodied actions (e.g. Lim, 2021; Ngo et al., 2021).

Ideationally and textually, the steps in the first and third promenades in Excerpt 7 are often in sync with the content words in the speech that construe different entities in experience, which, as noted by Ngo et al. (2021), assigns textual prominence to the ideational meaning at stake. For instance, in Excerpt 7, at the 3s and the 4s, the two steps in the first promenade are in sync with the two semiotic entities – 'argument' and 'essay' in the speech (line 1 and line 6), which highlights the information in

the speech at stake (Wu, 2024a). Similarly, at the 17s, the 18s, the 20s, the 21.5s, the 22s, the 24s, and the 25s, and the 25.6s, the steps in the last promenade are in sync with the semiotic entities – 'point', 'essay', 'you', 'point', 'question', 'task', 'point', 'introduction', and 'something' and the activity entity – 'relating' (line 1, line 6), which highlights the noticeability of the information at stake – the placement of a reference sentence should be a halfway point between the introduction and conclusion in an essay (Wu, 2024a).

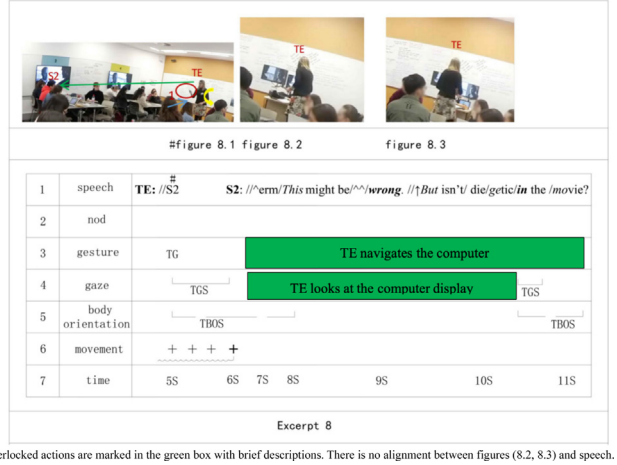
Interpersonally, the multimodal synchronies among speech, embodied movement, and shifts in gaze and orientation can reinforce an authorial presence of the teacher and enact multiple teacher roles simultaneously in classroom interaction. For instance, in Excerpt 7, from the 1s to the 4s, John's speech – 'Further reference to the core thesis argument of your essay' (line 1), realizes a verbal demand for information from the students and establishes the teacher as an authorial figure in the classroom (Martin and Rose, 2007). At the 4s, he shifts his gaze from the document to the student (line 4), which realizes a visual demand of the students (Kress and Van Leeuwen, 2006) and monitors the students' response. The synchrony between gaze and speech (line 1, line 4) in this example not only resonates with the interpersonal meaning made by speech but also reinforces the demand, resulting in a stronger authorial presence of the teacher (Wu, 2024a). During his speech (line 1), he also moves away from the classroom front (line 6, Fig. 7.1), an authoritative space that connotes teacher authority (Lim et al., 2012; Wu, 2024b) to position himself in the student pod centre (line 6, Fig. 7.2), an interactional space that encourages student participation (Lim et al., 2012; Wu, 2024b). As such, the coordination of multiple resources in the interaction in this instance enables the teacher to enact different pedagogic roles simultaneously: as a lecturer, encourager and monitor, which facilitates the performance of the complex pedagogic activities at stake.

4.3. Interlocked rhythms conditioned by the situational context

This subsection demonstrates how embodied actions can dis-align with the rhythm of speech in classroom interaction and how such dis-alignment is conditioned by the situational context including the nature of the lesson activity and the design of the spatial environment. In other words, in classroom interaction, embodied actions can be rhythmical themselves but not rhythmically integrated with speech, enacting what Lomax (1977: 27) terms interlocked rhythms whereby multimodal semiotic resources co-occur in the interaction but perform different semiotic work without interfering with one another.

In this interaction instance, the teacher Emma first asks one student the difference between diegetic sound and non-diegetic sound. Then the student (S2, male) provides the answer – diegetic sound is in the movie, whereas non-diegetic sound is outside of the soundtrack[1]. Finally, Emma affirms the answer. The interaction occurs in the middle of the lesson and lasts 15 s.

In Excerpt 8, from the 5s to the 6s, Emma is in the middle of a promenade during her speech – 'S2?', whereas the students remain seated throughout (line 1, line 6). This promenade has seven steps and the last three steps between the 5s and the 6s are not in sync with any speech (line 1, line 6). At the 5s and during the fourth step in the promenade, Emma enacts one hand gesture and one shift in gaze and body orientation (Fig. 8.1). However, between the 5s and the 6s, these shifts in gaze and body orientation continue in silence and not rhythmically in sync with the speech. During this promenade, Emma moves herself from the student pods to the lectern. She turns her back at the students and looks at the computer screen at her lectern (Fig. 8.2). From the 7s to the 11s, S2 (male) articulates, 'This might be wrong but isn't diegetic in the movie?' (line 1). Emma remains positioned at the lectern and is busy with navigating the computer display (Fig. 8.3). Her embodied actions of navigating the computer and looking at the screen are not in sync with the speech rhythm of S2.



Note: interlocked actions are marked in the green box with brief descriptions. There is no alignment between figures (8.2, 8.3) and speech.

[1] Strictly speaking, this is not quite correct. Presumably what the student means (and the teacher affirms), is that diegetic sound is part of the represented action of the movie (e.g sounds made by the actions of the characters), whereas non-diegetic sound is external or additional to it (e.g mood music).

Although from the 5s to the 11s, speech and embodied actions rhythmically dis-align with each other, these multimodal rhythmic patterns are not accidental but motivated. These rhythmic patterns are conditioned by the situational context, that is, the nature of the lesson activity and the design of the spatial environment. On the one hand, because of the spatial design of the classroom whereby the computer for display is placed at the lectern, Emma, who is positioned in the student pod, needs to relocate and position herself at the lectern to navigate the computer and select the display content – some film examples of diegetic sound and non-diegetic sound. On the other hand, the specific moment of her embodied action is conditioned by the nature of the lesson activity: it is when Emma and S2 discuss the difference between diegetic sound and non-diegetic sound that motivates Emma to navigate the screen for display. The nature of the knowledge at stake – diegetic sound is part of the action represented in the movie, whereas non-diegetic sound is external to it – needs to be demonstrated through some film examples and prompts Emma to navigate the computer.

The interlocked rhythm analysis provides complementary insights to existing multimodal rhythm studies (e.g. Martinec, 2018; Norris, 2009) that emphasize the synchrony of different resources in interaction. More particularly, Martinec (2018) pinpoints that the extension of rhythmic coordination across speaker turns is contingent upon the interrelationship between the participants, so jointly produced rhythms extend more consecutive turns in casual conversations than in political interviews because the participants are discursively framed as equal and cooperative in a casual conversation. This paper adds further insight to this point and notes that the situational context, including the nature of lesson activity and the design of the spatial environment, also affects the extension of rhythmic coordination across turns and across semiotic resources, resulting in rhythmic dis-alignment. The demonstration of interlocked rhythms adds empirical evidence to Deppermann's (2013) hypothesis that participants can perform one action in one semiotic resource, while in another semiotic resource, they are already oriented to some other business. Thus, each semiotic resource has its own distinct place in the ongoing production of interactional structure. The overall analysis in this section demonstrates how multimodal rhythmic patterns adjust as alignment or dis-alignment in the unfolding of the lesson activity, which empirically confirms Mondada's (2016) claim that multimodal gestalts are deeply embedded in the specific ecology of the activity.

5. Discussion

Investigating intersemiosis of speech and embodied action in classroom interaction has been a key concern in multimodal educational studies informed by SFL (e.g. Hood, 2011; Lim, 2021; Ngo et al., 2021). This paper demonstrates how rhythm can function as an integration principle that coordinates the interaction across semiotic resources and participants at different time scales. Multimodal analyses in this paper demonstrate that while each semiotic resource might operate with different temporalities, they are finely coordinated and synchronize to different degrees to constitute a pedagogic practice and formulate a coherent lesson. More particularly, in the selected data of this paper, the tempo of the speech in classroom interaction can rhythmically structure the embodied actions at different time scales, and the larger the time scale is, the larger the bodily synchrony involved. These multimodal synchronies at different time scales are semantically motivated. Multiple pedagogic roles can also be simultaneously enacted for the teachers to teach knowledge and manage the classroom during the intersemiotic process in the interaction. Additionally, multiple semiotic resources co-occurring in the classroom interaction can be rhythmical in themselves but not rhythmically integrated, with each semiotic resource performing its own semiotic work in the ongoing pedagogic practice. Taken together, multimodal rhythmic patterns in classroom interaction are contextually conditioned and semantically motivated.

Modelling multi-scalar temporalities through the lens of rhythm coincides with a continuing challenge for scholars in multimodal educational studies informed by SFL (e.g. Amundrud, 2017; Lim, 2021; Ngo et al., 2021) to develop theoretical and methodological tools to deal with the complexity and the emergent meaning when speech and diverse embodied actions interact in the classroom. The multimodal rhythm analyses in this paper demonstrate temporality is an inherent organization principle for speech and embodied action in interaction, which complements Lim's (2021) modeling of speech-gesture intersemiosis based on ideas of image–text relations that draws on spatiality as the meaning-making mechanism in the intersemiotic process.

This paper also demonstrates that rhythm can coordinate social interaction at multi-scalar temporalities, whereby rhythmic patterns of different time scales are necessarily interlinked. This finding aligns with Norris' (2009) rhythmic model of lower-level and higher-level actions and Martinec's (2002) hierarchical model of rhythm. However, the paper also extends their research by providing comprehensive and concrete transcription methods to capture and visualize the interactive patterns of multiple resources as multimodal alignment and dis-alignment. These methods, showcased through nuanced case analyses, also reveal how interactions at different time scales are semantically motivated. They attend to the holistic and synesthetic nature of multimodal interaction and complement existing conventions (e.g. Cowan, 2014; Mondada, 2017). They are specifically designed for multimodal rhythm analysis when embodied actions are subordinated to the time of the speech, thus allowing for coherent and systematic rhythm analysis of multimodal gestalts (Deppermann, 2013; Mondada, 2018; Wu, 2022).

Finally, this paper demonstrates that rhythm in classroom interaction is inherently dialogic and contingent in an unfolding pedagogic practice; that is, rhythmic coordination can extend across speaker turns and vary as the lesson unfolds. This point aligns with the finding of Deppermann (2013), Mondada (2016, 2018), Martinec (2018) and Norris (2009): rhythmic patterns

in the interaction are emerging and changing, depending on the situational context, the ecology of the spatial environment, and the social relationship between the participants. The paper also enriches their studies by demonstrating that multimodal rhythms in classroom interaction can be produced across speaker turns when one speaker's tempo of speech rhythmically coordinates the embodied action of another speaker.

Understanding the intersemiotic relation between speech and embodied action through the lens of rhythm has pedagogic implications. The multimodal analyses show that the lesson is a site of rhythmic assemblage, whereby multiple rhythms co-exist and are finely coordinated at different temporal scales in a motivated manner. In the pedagogic context, rhythmic alignment and dis-alignment amongst different semiotic resources serve pedagogic functions to facilitate the teaching and learning practices at stake.

6. Conclusion

While limited to a few video clips of classroom interaction, this paper provides fine-grained multimodal rhythm analyses of a wide range of semiotic resources, including speech, embodied movement, nods, gestures and shifts in gaze and body orientation, and focuses on the intersemiotic mechanism between these resources. The paper demonstrates how rhythm can contribute at various time scales to the coordination and synchronization of meaning-making dependent on the kind of rhythm-making resources employed. It also presents a new form of annotation and description of data that supports the study of rhythm. Hopefully, the theoretical and methodological techniques devised in this paper can be applied to other multimodal interactions whereby embodied actions are coordinated to the time of the speech. This would enable future researchers to zoom in at specific moments of a dynamic communicative practice to analyze and discuss the moments of intersections of multiple semiotic resources in detail.

These rhythm transcription and analysis methods can also enable future research to empirically test several interesting hypotheses in existing rhythm studies. Future research can explore whether speakers can plan and anticipate the upcoming message when speech and embodied actions are timed regularly (Van Leeuwen, 2005; Martinec, 2018). Furthermore, Cohen and Faulkner (1984) suggest that multimodal rhythmic synchronies of speech and embodied action over a short duration can help segment the knowledge contents into small chunks of the message, which might enable the teachers and students to collate messages extracted from preceding multimodal discourse and store that message in memory that must be retained for further processing at the next higher level of information collation. Future research can thus explore whether teachers and students can amplify the prominence of knowledge by adjusting multimodal rhythmic patterns. For instance, if a message at a higher level is considered more prominent and attention-catching than those at lower levels (e.g. Halliday and Greaves, 2008; Martinec, 2000; Van Leeuwen, 2005), will the teacher amplify the prominence of a certain knowledge point by putting it at a higher rhythmic level in the speech?

CRediT authorship contribution statement

Xiaoqin Wu: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

This work was supported by Humanities and Social Science Fund of Ministry of Education of China, under grant [24XJC740007] and [24YJC740021], Southwest University Educational Reform Grant, under grant [2023JY081], Fundamental Research Funds for the Central Universities, under grant [SWU2309714], and Guangdong Planning Office of Philosophy and Social Science, under grant [GD24YWY05].

Declaration of competing interest

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgement

I want to extend my sincere gratitude to Professor Louise Ravelli, Professor Theo van Leeuwen, Associate Professor Helen Caple, Dr. Janina Wildfeur and Dr. Yufei He for their insightful comments on an early draft of this manuscript. I am also grateful to Editor-in-Chief Sune vork Steffensen and the three anonymous reviewers whose constructive comments significantly improve the quality of this paper.

Data availability

The authors do not have permission to share data.

References

Amundrud, T., 2017. Analyzing Classroom Teacher-Student Consultations: A Systemic-Multimodal Perspective. Macquarie University, Sydney (PhD thesis).
Baldry, A., Thibault, P., 2006. Multimodal Transcription and Text Analysis. Equinox, London.
Bateman, J., Wildfeuer, J., Hiippala, T., 2017. Multimodality: Foundations, Research and Analysis–A Problem-Oriented Introduction. de Gruyter, Berlin.
Bezemer, J., Mavers, D., 2011. Multimodal transcription as academic practice: a social semiotic perspective. Int. J. Soc. Res. Methodol. 14 (3), 191–207.
Cowan, K., 2014. Multimodal transcription of video: examining interaction in early years classrooms. Classr. Discourse 5 (1), 6–21.
Cohen, G., Faulkner, D., 1984. Memory for text: some age differences in the nature of the information that is retained after listening to texts. In: Bouma, H., Bouwhuis, D.G. (Eds.), Attention and Performance X: Control of Language Processes. Lawrence Erlbaum, London, pp. 501–514.
Deppermann, A., 2013. Multimodal interaction from a conversation analytic perspective. J. Pragmat. 46 (1), 1–7.
Deppermann, A., Schmitt, R., Mondada, L., 2010. Agenda and emergence: contingent and planned activities in a meeting. J. Pragmat. 42 (6), 1700–1718.
Djonov, E., Tseng, C.I., Lim, F.V., 2021. Children's experiences with a transmedia narrative: insights for promoting critical multimodal literacy in the digital age. Discourse, Context & Media 43, 100493.
Fitzgerald, A., Hackling, M., Dawson, V., 2013. Through the viewfinder: reflecting on the collection and analysis of classroom video data. Int. J. Qual. Methods 12 (1), 52–64.
Flewitt, R., 2006. Using video to investigate preschool classroom interaction: education research assumptions and methodological practices. Vis. Commun. 5, 25–50.
Goodwin, C., 2000. Action and embodiment within situated human interaction. J. Pragmat. 32, 1489–1522.
Halliday, M., 1978. Language as Social Semiotic: The Social Interpretation of Language and Meaning. Arnold, London.
Halliday, M.A.K., Greaves, B., 2008. Intonation in the Grammar of English. Equinox, London.
Halliday, M.A.K., Matthiessen, C.M.I.M., 2004. An Introduction to Functional Grammar, third ed. Arnold, London.
Hao, J., Hood, S., 2019. Valuing science: the role of language and body language in a health science lecture. J. Pragmat. 139, 200–215.
Hannula, M.S., Haataja, E., Löfström, E., Moreno-Esteva, E., Salminen-Saari, J.F., Laine, A., 2022. Advancing video research methodology to capture the processes of social interaction and multimodality. ZDM–Mathematics Education 54 (2), 433–443.
He, Y., 2021. Towards a stratified meta functional model of animation. Semiotica 239, 1–35.
He, Y., 2023. The role of rhythm in science-animated videos: construing entities and bridging across different semiotic modes. Vis. Commun. https://doi.org/10.1177/14703572221112680.
Heath, C., Hindmarsh, J., Luff, P., 2010. Video in Qualitative Research: Analyzing Social Interaction in Everyday Life. Sage, London.
Hood, S., 2011. Body language in face-to-face teaching: a focus on textual and interpersonal meaning. In: Dreyfus, S., Hood, S., Stenglin, M. (Eds.), Semiotic Margins: Meaning in Multimodalities. Continuum, London, pp. 31–52.
Jacobs, J.K., Kawanaka, T., Stigler, J.W., 1999. Integrating qualitative and quantitative approaches to the analysis of video data on classroom teaching. Int. J. Educ. Res. 31 (8), 717–724.
Jewitt, C., 2009. The Routledge Handbook of Multimodal Analysis. Routledge, London and New York.
Kendon, A., 1967. Some functions of gaze-direction in social interaction. Acta Psychol. 26, 22–63.
Kendon, A., 1990. Conducting Interaction: Patterns of Behavior in Focused Encounters. Cambridge University Press, Cambridge.
Kendon, A., 2004. Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge.
Kress, G., 2010. Multimodality: A Social Semiotic Approach to Contemporary Communication. Routledge, London and New York.
Kress, G., van Leeuwen, T., 2001. Multimodal Discourse: The Modes and Media of Contemporary Communication. Arnold, London.
Kress, G., van Leeuwen, T., 2006. Reading Images: The Grammar of Visual Design, second ed. Routledge, London and New York.
Lemke, J.L., 1998. Multiplying meaning: visual and verbal semiotics in scientific text. In: Martin, J.R., Veel, R. (Eds.), Reading Science. Routledge, London, pp. 87–113.
Lim, V.F., 2004. Developing an integrative multi-semiotic model. In: O'Halloran, K. (Ed.), Multimodal Discourse Analysis: Systemic Functional Perspectives. Bloomsbury, London, pp. 220–246.
Lim, V.F., 2011. A Systemic Functional Multimodal Discourse Analysis Approach to Pedagogic Discourse. National University of Singapore, Singapore (PhD thesis).
Lim, F.V., 2021. Investigating intersemiosis: a systemic functional multimodal discourse analysis of the relationship between language and gesture in classroom discourse. Vis. Commun. 20 (1), 34–58.
Lim, V., O'Halloran, K., Podlasov, A., 2012. Spatial pedagogy: mapping meanings in the use of classroom space. Camb. J. Educ. 42 (2), 235–251.
Martin, J.R., Rose, D., 2007. Working with Discourse: Meaning beyond the Clause. Continuum, London.
Martinec, R., 2000. Rhythm in multimodal texts. Leonardo 33 (4), 289–297.
Martinec, R., 2002. Rhythmic hierarchy in monologue and dialogue. Funct. Lang. 9 (1), 39–59.
Martinec, R., 2018. Linguistic rhythm and its meaning: rhythm waves and semantic fields. Ling. Hum. Sci. 14 (1–2), 70–98.
Matthiessen, C.M.I.M., 2009. Multisemiosis and context-based register typology: registerial variation in the complementarity of semiotic systems. In: Ventola, E., Guijarro, A.J.M. (Eds.), The World Told and the World Shown: Multisemiotic Issues. Palgrave Macmillan, London, pp. 11–38.
McMurtrie, R., 2017. The Semiotics of Movement in Space. Routledge, London and New York.
Mondada, L. 2017. Multimodal transcription conventions. Last accessed at https://franz.unibas.ch/fileadmin/franz/user_upload/redaktion/Mondada_conv_multimodality.pdf.
Mondada, L., 2011. The interactional production of multiple spatialities within a participatory democracy meeting. Soc. Semiotic. 21 (2), 289–316.
Mondada, L., 2012. Talking and driving: multiactivity in the car. Semiotica 191, 223–256.
Mondada, L., 2016. Challenges of multimodality: language and the body in social interaction. J. SocioLinguistics 20 (3), 336–366.
Mondada, L., 2018. Multiple temporalities of language and body in interaction: challenges for transcribing multimodality. Res. Lang. Soc. Interact. 51 (1), 85–106.
Ngo, T., Hood, S., Martin, J.R., Painter, C., Smith, B., Zappavigna, M., 2021. Modelling Paralanguage from the Perspective of Systemic Functional Semiotics: Theory and Application. Bloomsbury, London.
Norris, S., 2009. Tempo, Auftakt, levels of actions, and practice: rhythm in ordinary interactions. J. Appl. Ling. 6 (3), 333–355.
Ravelli, L., 1995. Intersemiosis: the constraints and potential of verbal-visual interaction. In: Paper Presented to the International Systemic Functional Congress Beijing.
Tseng, C.I., Djonov, E., 2023. Children's comprehension of time in audiovisual narratives: a multimodal discourse and empirical approach. Ling. Educ. 73, 101144.
Unsworth, L. (Ed.), 2008. Multimodal Semiotics: Functional Analyses in Contexts of Education. London and New York: Continuum.
Van Leeuwen, T., 1985. Rhythmic structure of the film text. In: van Dijk, T.A. (Ed.), Discourse and Communication – New Approaches to the Analysis of Mass Media Discourse and Communication. de Gruyter, Berlin.
Van Leeuwen, T., 1992. Rhythm and social context. In: Tench, P. (Ed.), Studies in Systemic Phonology. Frances Pinter, London.
Van Leeuwen, T., 2005. Introducing Social Semiotics: An Introductory Textbook. Routledge, London.
Wu, X.Q., 2022. Space and Practice: A Multifaceted Understanding of the Designs and the Uses of *Active Learning Classrooms*. University of New South Wales PhD thesis, Sydney.
Wu, X.Q., 2024a. Embodied movement as a stratified semiotic mode: how movement, gaze and speech mean together in the classroom. Text Talk. https://doi.org/10.1515/text-2023-0164.

Wu, X.Q., 2024b. Spatial pedagogy: exploring semiotic functions of one teacher's movement in an active learning classroom. Semiotica. https://doi.org/10.1515/SEM-2024-0017.
Wu, X.Q., 2025. A Multimodal Framework of Pedagogic Practices in Space. Routledge, New York.
Wu, X.Q., Ravelli, L., 2021. The mediatory role of whiteboards in the making of multimodal texts: implications of the transduction of speech to writing for the English classroom in tertiary settings. In: Diamantopoulou, S., Orevik, S. (Eds.), Multimodality in English Language Learning. Routledge, London, pp. 161–175.
Zhao, S.M., 2010. Intersemiotic relations as logogenetic patterns: towards the restoration of the time dimension in hypertext description. In: Bednarek, M., Martin, J.R. (Eds.), New Discourse on Language: Functional Perspectives on Multimodality, Identity and Affiliation. Continuum, London, pp. 195–218.

Xiaoqin Wu is Lecturer at Southwest University, Advisory Editor of *Visual Communication*, and Principal Investigator of three research grants. Her research interests include multimodal discourse analysis informed by social semiotics and systemic-functional linguistics. Her most recent publications include a monograph with Routledge and several journal articles in *Digital Scholarship in the Humanities*, *Semiotica*, *Text & Talk*, *Journalism*, *New Media & Society*, *Journal of Pragmatics*, *International Journal of Speech, Language and the Law*, *Visual Communication, Multimodality & Society*, etc.

报告编号：202502-131

检 索 报 告

项目名称：论文被收录引用情况证明

委 托 人：西南大学外国语学院　伍小琴

日　　期：2025 年 02 月 25 日

认证单位：教育部科技查新工作站 N08

二 O 一九年制

<table>
<tr><td>检索项目名称</td><td colspan="4">委托人提交论文被 SSCI 收录引用情况</td></tr>
<tr><td rowspan="2">查新机构</td><td>名　称</td><td>教育部科技查新工作站 N08</td><td>名　称</td><td>400715</td></tr>
<tr><td>地　址</td><td>重庆市北碚区西南大学图书馆</td><td>地　址</td><td>023-68253283</td></tr>
<tr><td>联系人</td><td colspan="4">伍小琴（联系电话：15228800415）</td></tr>
<tr><td>委托文献目录</td><td colspan="4">1.Wu, XQ , Rhythm as an integration principle for modeling speech-action intersemiosis in classroom interaction: a social semiotic perspective, LANGUAGE SCIENCES,2025,108,101704 .</td></tr>
<tr><td>检索的数据库范围</td><td colspan="4">1. Social Sciences Citation Index (SSCI)
2. Journal Citation Reports
3. 中国科学院文献情报中心期刊分区表</td></tr>
<tr><td>检索要点</td><td colspan="4">论文被 SSCI 收录和影响因子及分区情况</td></tr>
<tr><td>检索结论</td><td colspan="4">经检索，委托人提交的 1 篇论文被 SSCI 收录。检索结果详细情况见附件。

检索人（签名）：魏小璐　　职称：馆员

教育部科技查新工作站 N08

2025 年 02 月 25 日</td></tr>
<tr><td>备注</td><td colspan="4"></td></tr>
</table>

附件：SSCI 收录情况

序号	排名	题名	检索号	影响因子	出版时间	语种	出版商
1	第一作者 通讯作者	Rhythm as an integration principle for modeling speech-action intersemiosis in classroom interaction: a social semiotic perspective	WOS:001402444300001	IF_{2023}=1.3	2025 年	English	国外

Social Sciences Citation Index (SSCI)：收录 1 篇。

第 1 条，共 1 条

标题: Rhythm as an integration principle for modeling speech-action intersemiosis in classroom interaction: a social semiotic perspective

作者: Wu, XQ (Wu, Xiaoqin)

来源出版物: LANGUAGE SCIENCES 卷: 108 DOI: 10.1016/j.langsci.2024.101704 出版年: MAR 2025

入藏号: WOS:001402444300001

语言:English

文献类型:Article

ISSN:0388-0001

eISSN:1873-5746

地址: [Wu, Xiaoqin] Southwest Univ, Coll Int Studies, 2 Tiansheng Rd, Chongqing 400715, Peoples R China

通讯作者地址: [Wu, Xiaoqin] (corresponding author), Southwest Univ, Coll Int Studies, 2 Tiansheng Rd, Chongqing 400715, Peoples R China

电子邮件地址: xiaoqinwu415@gmail.com

2023 年该刊在 JCR 分区：

Categories	Rank	Quartile
LINGUISTICS	97/296	Q2
LANGUAGE & LINGUISTICS	56/392	

2023 中科院分区信息(升级版)：

小类学科：LANGUAGE & LINGUISTICS 语言与语言学;分区：3 区 ；

小类学科：LINGUISTICS 语言学;分区：4 区 ；

大类学科：文学;分区：2 区 ;
