



Article

Cognitive Load Increases Spoken and Gestural Hesitation Frequency

Simon Betz ^{*}, Nataliya Bryhadyr, Olcay Türk and Petra Wagner 

Department of Linguistics, Phonetics Workgroup, Bielefeld University, 33615 Bielefeld, Germany

^{*} Correspondence: simon.betz@uni-bielefeld.de

Abstract: This study investigates the interplay of spoken and gestural hesitations under varying amounts of cognitive load. We argue that not only fillers and silences, as the most common hesitations, are directly related to speech pausing behavior, but that hesitation lengthening is as well. We designed a resource-management card game as a method to elicit ecologically valid pausing behavior while being able to finely control cognitive load via card complexity. The method very successfully elicits large amounts of hesitations. Hesitation frequency increases as a function of cognitive load. This is true for both spoken and gestural hesitations. We conclude that the method presented here is a versatile tool for future research and we present foundational research on the speech-gesture link related to hesitations induced by controllable cognitive load.

Keywords: hesitation; gesture; pauses; card game; cognitive load

1. Introduction

1.1. Pauses and Hesitations

Speech does not work without pausing. Pauses occur naturally at phrase boundaries; they might be inserted purposefully into speech for a dramatic effect, or they might occur more or less involuntarily as hesitations. Pauses are furthermore physiologically necessary for speakers to catch breath, and they are vital for listeners to ease perception (Trouvain and Werner 2022). In this study we will focus on hesitation pauses. Hesitations are elements of spontaneous speech that temporally extend speech delivery. These delays have communicative functions. Speakers can buy time by hesitating in order to remedy speech plans or retrieve complicated lexical items (cf. (Eklund 2004) for a thorough overview). Listeners can use the extra time granted by hesitations as well to improve speech input processing, which may consequently yield further benefits such as an increased task performance (Betz et al. 2018, 2017). Traditionally, however, hesitation phenomena, especially fillers, were connoted negatively, and speakers uttering fillers were perceived as incompetent, cf. e.g., (Fischer et al. 2017). Only in the 1980s, the view shifted towards hesitations as a powerful communicative device (cf. Allwood et al. 1990; Chafe 1980; Clark 1996; Levelt 1989; Shriberg 1994). We use the term *hesitation* as an umbrella term to cover the following three phenomena: silences, fillers, and lengthenings. Our working definitions for the three phenomena are as follows.

Silences are any intervals without speech by the active speaker, which are perceived as hesitation. A common cue that leads to silences being perceived as hesitant is an interruption to the flow of speech, i.e. silences occurring within an utterance in syntactically or prosodically marked positions. The annotation of silences is usually carried out perceptually, by trained annotators, as to this day, no established, objective method to their detection exists. Eklund (2004) observed that duration is a poor cue to analyze silences and Campione and Véronis (2002) warn of using duration thresholds as they might skew the results, even though they are often employed in the automatic detection of silences.

Fillers are vocalizations of a central vowel and/or a nasal (e.g., *uh*, *uhm*, *mh*). In theory there are other forms of fillers, such as lexical items with low propositional content,



Citation: Betz, Simon, Nataliya Bryhadyr, Olcay Türk, and Petra Wagner. 2023. Cognitive Load Increases Spoken and Gestural Hesitation Frequency. *Languages* 8: 71. <https://doi.org/10.3390/languages8010071>

Academic Editors: Jürgen Trouvain and Bernd Möbius

Received: 19 November 2022

Revised: 13 February 2023

Accepted: 24 February 2023

Published: 2 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

central vowels critically attached to other words or other nonverbal vocalizations such as lip smacking, but these items made up less than 1% of the fillers in our data, so we omitted them for the study at hand.

Lengthenings are stretches of markedly elongated syllables that are perceived as hesitation, usually by comparison to the speech rate of segments surrounding it. Similarly to silences, lengthenings are typically annotated manually, as they cannot easily be differentiated from other types of lengthenings that occur in running speech. In terms of duration, hesitation lengthening is typically longer than the common phrase-final lengthening, but can be confused with accentuation lengthening. For human annotators, however, a differentiation between hesitation-related vs. non-hesitation related lengthening is possible with near-perfect agreement [Betz et al. \(2016\)](#), possibly due to pitch characteristics and position within the syllable. A remaining problem is that lengthening is frequently missed by annotators, which causes data sparsity issues [Betz et al. \(2017\)](#).

Silences and fillers have obvious connections to pauses. Silences have been referred to as silent pauses in many previous studies, but as recent research suggests, this term is unfortunate, as there might be nonverbal content within, such as breathing or lip smacking ([Belz and Trouvain 2019](#); [Trouvain and Werner 2022](#)). Fillers have been referred to as filled pauses as a contrast to silent pauses, but as [Belz and Trouvain \(2019\)](#) and [Trouvain and Werner \(2022\)](#) noted, pauses can be filled with more than just *uhs* and *uhms*, which renders the term unfortunate as well. We thus opted for the more neutral terms above, which, however, does not void these phenomena's inherent connection to the topic of pausing. Lengthenings are less obvious manifestations of pauses. In fact, lengthenings were often omitted from pause research such as ([Trouvain and Werner 2022](#)) altogether. This seems to have phonetic reasons. A silence, as per our working definition stated above, and as conceived by [Trouvain and Werner \(2022\)](#), needs to be audible as well as visible in the signal, thus inserting a pause into the speech signal. A filler also stands out markedly from the surrounding speech, as its vowel quality and pitch contour is usually different from other vowels in the language system or from the surrounding speech, respectively ([Belz 2021](#); [Belz and Reichel 2015](#); [Jabeen and Betz 2022](#)). Lengthening, however, cannot be isolated from the flow of speech. Rather, it is the result of modifying the durational *gestalt* of lexical items by stretching the articulatory movements in time. Lengthening is preferably realized on continuous segments such as vowels or nasals, during which an articulatory target state remains unaltered as long as the lengthening lasts ([Betz et al. 2018](#)). In this sense, lengthening can be regarded as pausing within the execution of the articulatory plan, but with ongoing phonation ([Eklund 2004](#)). Such lengthening is interpreted by listeners as hesitation, the extra time so generated is used by listeners just like other hesitations ([Betz et al. 2018](#)), and it serves pragmatic functions such as the expression of epistemic states like uncertainty ([Betz et al. 2019](#)). Lengthening shares many characteristics of pauses and fillers, and they are a very frequent and unobtrusive hesitation signal (cf., [Betz \(2020\)](#) for an overview), for which reasons they are treated as pause manifestations in this study alongside fillers and silences.

1.2. Hesitations and Gesture

Spontaneous bodily movements that accompany speech, i.e., gestures, are not randomly produced movements, and they have been shown to serve various functions aiding speech planning and communicative acts in speech (see [Bavelas et al. 1992](#); [Butterworth and Beattie 1978](#); [Goldin-Meadow 1999](#) among others). This suggests that speech and gesture production processes are tightly coupled ([De Ruiter 2000](#); [Kita and Özyürek 2003](#); [McNeill 1992](#)). Consequently, gesture should be sensitive to fluency disruptions and hesitation phenomena occurring in speech, especially since gesture is also highly adaptive by nature ([McNeill 2005](#)).

There are different types of movement that can accompany speech. [Novack and Goldin-Meadow \(2017\)](#) importantly distinguish *gestures* from *actions* based on the goal of movement. They state that gestures are produced to accomplish representation of

information and communication, whereas actions are produced to achieve communication external goals (e.g., object manipulation). For instance, in cases of hesitation, gestures may be used to help with lexical retrieval, but many gestures produced during hesitations are claimed to be those referring to the breakdown itself and not representing the content of the sought concept (Graziano and Gullberg 2018). These types of gestures are referred to as metaphoric gestures in McNeill's (1992) terminology (i.e., the hand creates a physical representation of a hesitation as a concept), and they too are markers of gestural hesitation.

In terms of hesitation in actions (i.e., non-referential gestural hesitation, or more generally, hesitation in the kinematics of movement), what can be considered as hesitation is broad: jerky movements, suspension of movement such as freezing at certain positions, sudden cancellations of movement, restarting movement (see Section 2.3.3). In this study, we are only interested in manual movement, and we consider all such movements in the hands/arms as hesitation (Moon et al. 2011), cf. Allwood et al. (2005) for others such as gaze aversion and frowns. Within our design, it is possible to have markers of referential (e.g., metaphoric) and non-referential hesitations (e.g., pauses) (Section 2.1). Therefore, we will refer to these as gestural hesitations for the sake of simplicity from hereon.

In relation to spoken hesitations in Section 1.1, one important question then arises: "How exactly are gestural hesitations manifested during spoken hesitations?". Graziano and Gullberg (2018) sought to answer such a question in their cross-linguistic study involving a retelling task in Dutch and Italian, and their findings showed that participants gestured more during fluent speech compared to when they were hesitating. Further, they produced more gesture holds (i.e., halts in production, see Section 2.3.3 for definitions of gestural constituents) during disfluent speech than fluent. There were also more holds than strokes (i.e., the core of a gesture that carries meaning). They also reported that when speech stopped, so did the gesture. Similar findings pertaining to gesture rate and synchronization were also reported in Kosmala et al. (2019).

Seyfeddinipur (2006) explored when the points of gestural hesitations occurred in German living space descriptions (they refer to these as gesture suspensions), while also looking into their temporal relation with spoken hesitations. The most common gesture constituent accompanying disfluent speech was the stroke (cf. Graziano and Gullberg 2018), and gestural hesitations tended to occur with the stroke or with the preparation which brings the hand to the onset of the stroke (Section 2.3.3). There was no difference in gesture rates and in the frequencies of gestural hesitations between fluent and disfluent speech (cf. Graziano and Gullberg 2018). Further, gestural hesitations were found to occur at the same time as spoken hesitations or slightly precede them, which provides evidence that temporal synchrony of speech and gestures can be maintained during hesitations (McNeill 1992). This is further supported by the studies that elicited disfluent speech with the help of delayed auditory feedback, which is known to make speech slow down and disfluent. Even under such conditions, prominent points within the strokes (Loehr 2004) were still synchronized with prominent points in speech (e.g., pitch peaks) (McNeill 1992; Pouw and Dixon 2019).

Overall, the number of studies on gestural hesitations and how they interact or synchronize with speech and spoken hesitations is limited. This is partly due to the fact that gestural hesitations are considered as atypical productions and tend to be excluded from studies that focus on speech-gesture interaction (e.g., Türk 2020). Moreover, despite using similar tasks, existing studies also report conflicting findings about when gestural hesitations occur and their frequency in relation to spoken hesitations. Therefore, more studies using different tasks and methods are needed to shed light on these interactions.

1.3. Speech and Pausing under Different Levels of Cognitive Load

Cognitive load, typically defined as the amount of working memory store dedicated to an ongoing task, has been investigated for its impact on speech production for several decades (Lively et al. 1993), but mostly on heavily controlled data such as collected in laboratory dual-task paradigms or similar, in which speakers had to produce clearly defined

target words or phrases. In recent years, cognitive load has often been investigated from the perspective of its automatic perception or classification (Schuller et al. 2014), and with a less strong focus of the fine phonetic details of its acoustic-phonetic and prosodic realization. Where these aspects have been examined, they have often focused on the segmental level, searching for evidence of hyperarticulation or increased overall “tension”, also affecting suprasegmental aspects such as fundamental frequency (Dahl and Stepp 2021; Lively et al. 1993; Yap et al. 2011). However, cognitive load has also been shown to influence pausing patterns such as pause frequency and type (fillers vs. silences) (Montacié and Caraty 2014; Yin and Chen 2007). In particular, Montacié and Caraty (2014) report that pause frequency is a very effective predictor for cognitive load in human speech production. In addition, perception studies have shown that listeners interpret the production of (filled) pauses in object descriptions, expecting objects that are less familiar and hence more difficult to describe (Arnold et al. 2007). The domain of lengthening has, to our knowledge, not been investigated in relation to cognitive load.

More recently, the generalizability of existing data collections on speech under cognitive load has been called into question (Vukovic et al. 2019). Their arguments align with (Wagner et al. 2015), who ask for more diversity of speaking styles as well as more ecologically valid data sets (where appropriate) in phonetic research in general.

There are few lines of study that deal with gesture’s relationship with cognitive load, though they are not directly relevant to the present study. One pertains to cognitive rhythm theory’s view on the gesture’s role as a facilitator, functioning primarily to assist speech production (Goldman-Eisler 1967; Sweller and Chandler 1991). The general claim is that speech consists of cycles of acts of planning and production containing a high number of pauses and shorter expressions and fewer hesitations with fluent speech respectively. Here, the implication would be that the planning cycles (i.e., hesitant period) bear more cognitive load. Aboudan and Beattie (1996) examined the effects of this speech–pause ratio on gesturing. They observed that if the hesitant cycles were shorter than average, there were more gestures during fluent periods, which can be interpreted as a strategy of dealing with increased cognitive load due to less time taken for planning. Furthermore, Butterworth and Beattie (1978) found that gestures (i.e., their onsets) tended to occur during silences, preceding their lexical affiliate in speech with great delay (cf. Graziano et al. 2020 for difficulties in identifying lexical affiliates). They interpreted some of these silences as caused by difficulty in retrieving a desired lexical item (increased cognitive load), and gestures were generated to assist speech, lightening the cognitive load (Morrel-Samuels and Krauss 1992 but cf. Kita 2000; Kita and Özyürek 2003). Gesture’s role in cognitive load management has also been reported in applied and developmental areas of study (Cook et al. 2008; Goldin-Meadow et al. 2001).

To our knowledge, no study has so far investigated the *multimodal* production of pauses in relation to cognitive load in detail as described in Section 1.1. Furthermore, there is hitherto little evidence from speech productions under different levels of cognitive load that have been elicited in more spontaneous, interactive settings.

1.4. Aims of This Study

In this study, we present a novel method to elicit spoken and gestural pauses in a game setting. This method is intended to create small-scale specialized data sets featuring controlled, but still ecologically valid, productions of pauses. The game setting allows for control in the sense that it limits the conversation topic to the game domain with its own specific vocabulary, and which controls the cognitive load of the speaker, which in turn influences pause occurrence and placement. In terms of ecological validity, we can record unscripted spontaneous interaction that is still confined by the game domain. In the analysis, we can then split the game into episodes of varying cognitive load to study the effects on speech production. On the topic of gesture, our study aims to contribute to the body of studies investigating the distributions of gestural hesitations and types of gesture constituents these co-occur with (gesture phases, e.g., stroke, see Section 2.3.3).

We are also interested in how gestural hesitations co-occur (i.e., synchronize) with spoken hesitations. In our game setting, a player talks about their game movements as they play the game, which is inherently different from narratives or retelling tasks (Section 1.2) since it primarily involves object manipulation while also allowing for multimodal expressions of hesitation. Consequently, we are able to test how the frequency of these two types of hesitations are affected by varying levels of cognitive load.

Our contribution in this paper is twofold: on the one hand, we provide a detailed description of our method. We expect the game created for this study to be reusable in future studies investigating cognitive load, speech and co-speech movement. On the other hand, we present the first empirical study conducted within our game setting, in which we address one general research question and two hypotheses.

Our research question asks whether we can provide a proof of concept for our game setting: We expect our setup to be cognitively demanding, which should give rise to a large amount of pausing behavior, comparable to scenarios like map tasks, e.g., [Anderson et al. \(1991\)](#) or the various methods deployed in the DUEL corpus ([Hough et al. 2016](#)). We particularly expect large amounts of hesitations, namely silences, fillers and lengthenings. We will analyze hesitation type distribution and hesitation frequencies, and compare them to results of previous studies, in order to determine whether or not our framework provides a suitable methodological framework for studying hesitations. Furthermore, we investigate the functionality of our game as a framework to analyze the spoken and gestural correlates of cognitive load in a controlled environment. Our hypotheses are:

1. Game situations with higher cognitive load will elicit more spoken *and* gestural hesitations than those with lower cognitive load.
2. Given the tight coupling of gesture with speech (cf., Section 1.2 and [Wagner et al. 2014](#)), we predict that gestural hesitations will mostly co-occur with spoken hesitations (see Section 2.3.3 for limitations on annotations). Moreover, we predict that spoken and gestural hesitations in synchrony will show form-related similarity. For instance, when speech goes into a halt, so will gesture, resulting in more gestural hesitations in the form of gestural pauses (i.e., holds, cf., Section 2.3.3).

2. Materials and Methods

For this study we decided to use a game to systematically elicit multimodal pausing behavior as a function of different levels of cognitive load. Strategic card games appear to be excellent candidates for such an endeavor, as their basic rules often can be learned within minutes, as is necessary in laboratory settings, but can still reach high levels of complexity: A good example for this is Magic: The Gathering, which has been claimed to be the most complex real-world game ([Churchill et al. 2019](#)) and takes years to be fully understood, as a result of the high number of unique cards that can be played. To manipulate the level of game complexity, we decided to recycle the basic concept of Magic and related card games. The game is played with cards that have certain costs and effects. Unlike traditional card games, in which cards tend to have only parameters, such as color, image or value, in this game, players have to pay for each card they play with an in-game resource, which in turn can be generated by some cards in the game. For cognitive load management, we restricted our game to nine different cards with partly overlapping effects. That way, we ensured a relatively quick mastery of the game for our participants. We chose a space setting for the game, which makes it clear that it is an abstract game with little to no real-world connection. We created a small backstory of the player being stranded with a spaceship with the task to restart the engines. The cards represent tools, configurations, and environmental features that can be used for this. The goal is achieved once the player has drawn the entire deck of cards. In the following sections, we provide a detailed description of the game and the experimental setup as well as the recordings, data annotations, and statistical analyses.

2.1. The Game

We created a simple resource-management card game for this study. The game is played solitaire-style and the goal is to win the game by drawing the entire deck of 21 cards, ideally in one turn. In this experiment, the player is asked to comment their own gameplay to the experimenter, who sits on the opposite side of the table as a listener. The experimenter only interacts when a violation of the game rules is apparent or when the player addresses a question to them.

2.1.1. Cards

The cards are the key elements of the game. Their layout resembles that of cards in established strategic card games or board games, such as Magic, Hearthstone, or Wingspan. For visual appeal, public domain artwork is used as the card background. The functional elements of a card are (1) a box containing the name of the card (2) symbols indicating the energy costs (resources that must be spent) to play a card and (3) a box that describes the effect of a card (what happens when this card is played). The cards come in three different types (colors) of resources: red, green, and blue. White energy symbols indicate that any resource color can be used to play the card. The effects of the cards are summarized in Table 1 and visual examples are provided in Figure 1.

Table 1. Components of the card game, with respective complexity and amount within the deck. Cards marked with * start face up outside the deck as “jokers” players can use at any time.

Cost	Effect	Complexity	Amount
one colored resource	generate three resources of one other particular color	0	6
two different colored resources	draw three cards	0	6
zero	generate one resource of any color *	1	3
one resource of any color	generate one resource of any color, then draw a card	2	4
two resources in any combination of colors	generate two resources in any combination of colors, then draw a card	4	4



Figure 1. Example cards. 1: Name; 2: Cost; 3: Effect. Costs explained: Starlight costs one green resource, Configure costs two resources of any color, Nebula costs one red and one green resource.

2.1.2. Card Complexity

For this study, we assign a complexity level to each card, which is determined by the amount of choices the card inherently offers. Cards with higher complexity are expected to create increased cognitive load, which in turn is expected to be reflected in pausing behavior. In Table 1, the different types of cards and their complexity are displayed. Cards with the lowest complexity level (0) are those which have clearly defined costs and clearly defined effects. Cards “Starlight” and “Nebula” in Figure 1 would be examples of such cards: The player needs to have the required resources available and has to execute the corresponding game moves, but has to make no choices. To the opposite extreme, a card with the highest complexity level would be “Configure” in Figure 1. For that card, as indicated by the two white energy symbols, the player has to decide which two resources they want to spend to play the card. The effect then adds another two energy units in any combination, which means that the player has a total of four choices to make while playing the card, which gives the card the complexity level 4. Correspondingly, there are cards of complexity levels 1 and 2, meaning they each imply one or two choices by the player. Descriptions of these cards can be found in Table 1 as well. During gameplay, players are confronted with more levels of complexity than the ones defined by the card being played, such as how many other cards are available to be played, which resources are available at that time, or what the stage of the game is. However, card complexity is expected to have the dominant influence on the cognitive load related to every game move, as it constitutes the central element of the game.

2.1.3. Board Layout and Game Moves

Figure 2 shows a schematic layout of the game board. Although it is essentially a card game, some assistive elements are required. In the bottom left there is the deck of cards from which players draw cards whenever a card instructs them to do so. In the bottom middle, players place their hand cards. They need to be placed on the table in order to be captured by the camera. When playing a card, players pay the costs, execute the effect and then move the card on a “played pile” located over the deck. As indicated in Figure 2, there is a supply of tokens for counting energy resources on the bottom right of the board. In our game, these are white felt balls of about 2 cm in diameter, which can be placed noiselessly. The counting works by putting tokens into colored boxes which are located on the farthest side of the game board. If a card instructs a player to *generate three red energy units*, the player would have to take three tokens from the supply and put them into the red box. If a card costs one red and one green energy unit, players would have to take one token out of the red box and one token out of the green box and return those to the supply. The middle of the game board is left empty in order to create distance between the different game areas, which forces players to move their arms and hands to execute game moves. As it is required to comment the gameplay, these movements are likely to become co-speech movements on which gestural hesitations are likely to be observable. The movement of resource tokens into the colored boxes is expected to be most relevant for the study at hand, as we expect people to hesitate verbally and gesture-wise while considering which resources to generate. For example, a player playing a card of complexity level 4 as described in Section 2.1.2 might utter a phrase like *“I am generating one greennnnn: ... and one red energy”* while moving game tokens from the supply to the target boxes. Simultaneously, players might also slow down or halt their hand movement while uttering the hesitation around the word “green”. For the experiment, the schematic layout was printed on a piece of colored cardboard that covers the entire table, in order to make sure every player was using the same setup. A snapshot of a participant playing the game can be seen in Figure 3.

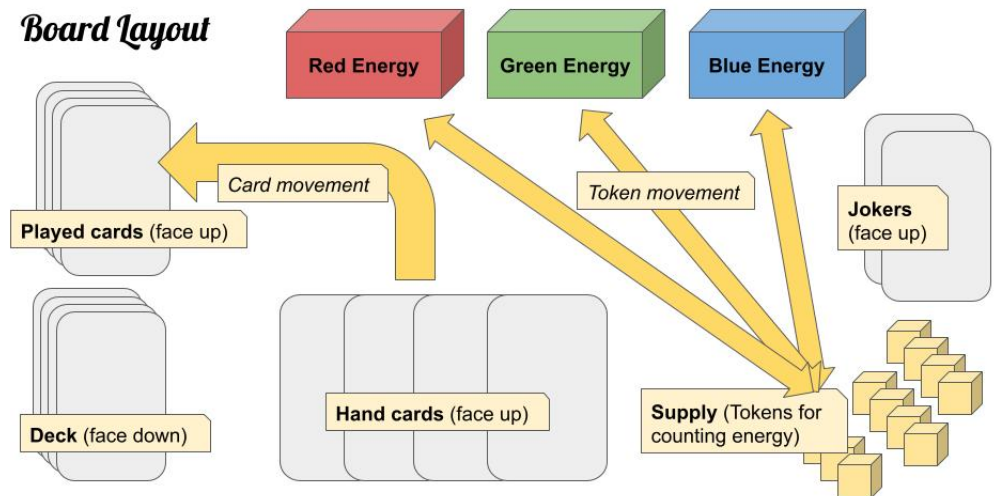


Figure 2. Board layout.



Figure 3. Game scene.

2.2. Participants

2.2.1. Recruitment and Payment

Twenty participants were recruited via social media and participated in our study. The entire process of instruction, prior testing, and recording during gameplay lasted a maximum of 30 min, usually less. Participants were paid 5 EUR each for their effort. One participant was removed from the recordings for technical reasons, yielding data of nineteen participants (seven female, twelve male; median age = 30). All participants spoke German as their mother tongue, some were bilinguals. All participants had an academic background, being students or scientists. We gathered additional data to ensure participants' qualification to partake; a detailed evaluation of this data is out of scope for this study.

2.2.2. Instructions

Before starting the game, participants received instructions on how to play the game. Instructions were all given by the experimenter, who was in all cases the first author of this

paper. Instructions were formulated freely but followed a script. Participants were first told how to use cards and resources to achieve the goal of drawing the entire deck. This was illustrated by four example cards, the same for each participant, which would then become the four opening hand cards. The four cards correspond to one item from each row in Table 1, except for a card of complexity 4. Two more cards of complexity 1 were placed face up on the side of the game board, and participants were told they are allowed to move them to their hand as jokers whenever they needed it. The remaining 17 cards were placed face down as a deck to draw from on the board. The order of the cards in the deck was the same for each participant, but not known to the participant. Cards were ordered in a way that it was possible to reach the goal of drawing the entire deck in one turn. Explaining the game took a maximum duration of 5 min.

2.2.3. Training, Learning, Fatigue

After instructions, the game and recordings started immediately. There was no training phase for the participants, as it was deemed unnecessary due to the simplicity of the game. Learning effects and accelerations (or possibly, fatigue) over the course of the game are expected and will be taken into account in the analyses (cf. Section 2.4).

2.2.4. Technical Setup

The recordings took place in a lab environment at Bielefeld University. The lab is furnished like an apartment and equipped with (deactivated) smart-home technology. Recordings took place in the apartment's living room-like setting, to ease immersion into the game. On the table was the game setup depicted in Figures 2 and 3. Opposite the participant, a shotgun microphone (Stage Line ECM-2001) was placed on the table. The experimenter sat behind the microphone, facing the participant. Behind the experimenter, a camera Sony HDR-CX 690 was mounted on a tripod, filming the entire game scene and the participant (1920 × 1080 at 25 fps) (see Figure 4). This setup enabled to not distract the participant with headsets or wires and still ensures an overall good video and audio quality.

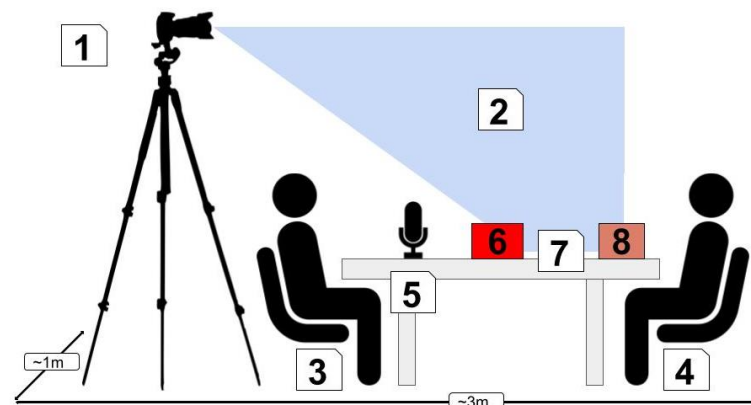


Figure 4. Filming setup. 1: Camera 2: Camera angle 3: Experimenter 4: Participant 5: Shotgun Microphone 6: Colored resource boxes 7: Game space 8: Resource token supply.

2.3. Annotation

2.3.1. Hesitation and Pause Annotation

Audio data was annotated starting with automatic speech recognition using the BAS Webservices (Kisler et al. 2017) to first create a raw transcript which was further annotated with Praat (Boersma and Weenink 2014). A trained annotator then perceptually annotated each file for hesitation and pausing phenomena on a separate annotation tier. The labeling instructions were as follows:

- **Silences (SIL):** Mark any interval as silence if it is devoid of speech and if you perceive this interval as hesitant when listening to the utterance(s) around it.

- **Fillers (FIL):** Mark any interval as filler if you perceive this hesitation to be vocalized. Fillers usually consist of a central vowel only (*uh, er*), or of a central vowel + nasal (*uhm, erm*), or in rarer cases of a nasal only (*mmm*).
- **Lengthenings (LEN):** Mark any interval as lengthening if you perceive a given segment as vocalized with markedly slower speech rate than the surrounding segments, *and* if you perceive this segment as hesitant in the context of the surrounding utterance. The interval should span the entire word that contains the lengthening, and a colon is to be used to mark the segment where the lengthening is, e.g., *then:* if the final nasal in that word is lengthened.

All annotations were ultimately based on perception for a lack of an appropriate alternative method. As we describe in Section 2.3.4, there was substantial agreement between annotators. Note that the level of analysis for lengthening in this study is the word, which is a simplification. Lengthening does manifest on segment level, but for our analysis we are merely interested in the binary information of lengthening being present or not, and not in fine phonetic detail, which makes it sufficient to classify words as being lengthened or not.

2.3.2. Complexity Annotation

The complexity annotation is based on the cards that are being played. We added a new annotation tier “game move”. A game move in our case is constituted by a card being played, which consists of the player selecting the card, moving it to the table, paying the resources required for it (which involves moving tokens), and executing the effect of the card, which means drawing new cards or managing resource tokens. After a card’s effect has fully been executed the annotator adds a boundary on the game move tier. If the player continues the game uninterrupted, this marks the beginning of the next game move. In case the player interrupts the game, e.g., to ask questions to the experimenter, the next game move begins at the end of this interlude. Each interval on the game move tier that contains a game move has a numerical value for complexity. The complexity is a direct mapping of the number of choices the card that is played within this game move offers, cf. Section 2.1.2. As speakers do not always verbalize which card they play, the video recording is used as support for annotation. In some cases it is possible that players think about multiple cards within a game move, which may become apparent from verbalizations, e.g., “I could play A or I could play B...”. In that case, the complexity values are added up, but capped at 4, our maximum complexity level for individual cards. Game moves involving multiple cards made up less than 1% of the corpus. Intervals that do not constitute game moves, such as requests, do not have complexity values and are ignored in the analysis. Furthermore, these cases are rare; less than 3% of intervals on the game move tier.

2.3.3. Gesture Annotation

A subset of the data (11 participants) was annotated for gestures using ELAN (ELAN 2019). The annotation task involved the annotation of manual gestures accompanying speech only, thus limiting our analyses to these contexts as well. Without any access to speech, three independent annotators annotated complete gesture phrases (a single complete and meaningful unit of bodily action, Kendon 2004). The annotation task was twofold: (1) segmentation of gesture phrases into their constituents, i.e., gesture phases (G-phases from here on); (2) identification of gestural hesitations if any.



Figure 5. G-phases in a single gesture phrase.

The segmentation of G-phases followed [Kita et al. \(1997\)](#), making use of changes in direction and in velocity profile. However, we use [McNeill’s \(1992\)](#) terminology to refer to these phases (see Figure 5). In a typically constructed gesture, the first phase is the *preparation* phase, in which the hand departs from a resting position (Figure 5a,b) or the end of the previous gesture to enable the execution of *the stroke*. The stroke is the core of a gesture as it is executed with maximum effort, and it carries the meaning of the gesture (if communicative) (see Figure 5c). The stroke can be preceded and followed by *holds* where the hands are frozen in location (typically a few video frames and often with slight drifts) (see Figure 5d). In the *retraction* phase, the hand returns to a resting position (Figure 5e). We added another category, *jerks*, to this classification to capture a wider range of phases that may express hesitation. These appear as flinches in movement where the hand often pulls back or to the side during the execution of one of the phases described here. Jerks can lead to the cancellation or restarting of the interrupted phase, which is interpreted as hesitation (see Section 1.2).

The final annotation step involved the annotation of gestural hesitation. The annotators qualitatively determined whether or not any of the annotated G-phases were perceived to contain one of the hesitation related phenomena we describe below:

- **Cancellation:** G-phases can be canceled halfway through their execution, which leads to the cancellation of gesture to be re-planned or restarted. Jerks always lead to cancellations, but cancellations do not necessitate a jerky gesture.
- **Pause:** Pauses in the execution of a gesture are classified as gestural hesitation. There is a natural overlap of these with hold phases which are essentially pauses in the movement of the hand. There are plenty of holds with fairly short durations occurring in many gestures (esp. after the stroke, 3–4 frames in length). The annotators were tasked to filter these out and only annotate a pause if there was a perceptually salient one that can be considered as a time-gaining strategy.

- **Slow down:** The execution of G-phases can appear to be slowed down (relative to the other G-phases in their environment), which is a strategy for planning/re-consideration.
- **Stall:** A player can perform stalling movements often through manipulation of game objects. For instance, the player can move a game card left to right on the table, a move not related to the game, while preparing for their next move.
- **Metaphors:** It is possible to have metaphoric gestures signaling hesitation/uncertainty. In our context, these are also not necessary for playing the game and are different from game-related gestures in terms of their form (McNeill 1992). For instance, they might occur as circular motions of the hand or palms-up hands shaking left to right to indicate uncertainty while saying "I don't know" as the player is thinking aloud.

The G-phase segmentation and identification of gestural hesitations were carried out for both hands on different tiers in ELAN, which resulted in approximately 45 min of gesture annotation.

2.3.4. Annotator Agreement

To validate the annotations for hesitations in speech, a second expert annotator annotated data for four speakers which were randomly selected (ca. 20% of the corpus in total). We calculated Cohen's kappa separately for lengthening on the one hand and for fillers and silences on the other hand. For lengthening, the level of analysis is the word (which can be lengthened or not), and for fillers and silences, the level of analysis is intervals between words, which may be hesitant due to silences or fillers, or non-hesitant due to a lack of those. We reach substantial agreement for both cases, ($\kappa = 0.824$, $N = 1830$) for lengthening, ($\kappa = 0.828$, $N = 406$) for silences and fillers.

To test annotator reliability for gesture annotations, three annotators blindly annotated the data of one participant (~10% of the number of recordings annotated for gesture). Fleiss' kappa was then run to determine if there was agreement between the annotators' decisions on the placement of G-phase and gestural hesitation boundaries, and the types of G-phases and gestural hesitations annotated. The timewise agreement on the boundary placement was tested through categorizing cases where the annotations overlapped less than 60% as disagreement (even if they had the same type annotation). Fleiss' kappa showed global substantial agreement for G-phase annotations ($\kappa = 0.614$, $N = 35$) and for gestural hesitation annotations ($\kappa = 0.675$, $N = 135$) (Landis and Koch 1977).

2.4. Analysis

For the analysis of spoken pauses, in Section 1.4, we posed the question of whether our setup elicits a large enough amount of hesitations. We will use descriptive statistics to report frequencies of individual hesitation types and compare the measures to previous studies.

For the analysis of gestures, in Section 1.4, we first stated that the study aims to explore the synchronization of gestural hesitations with spoken hesitations. Here, the term synchronization is used in its broader sense to refer to overlaps of these in time. Namely, if one type of annotation overlapped with another, we considered them as synchronized. We extracted these overlaps as intervals on a separate tier in ELAN, which constituted our data for our first investigation presented in Section 3.2. It must be noted that we excluded overlaps with spoken hesitations shorter than 150 ms from our data. These overlaps occurred near the edges of annotated units, and therefore, one annotation overlapped with another annotation for a very short duration, creating spurious data points for statistical analyses. Overall, this led to the omission of 388 instances (out of 1985) from the statistical analyses.

In Section 1.4, we predicted that there would be more spoken hesitations and gestural hesitations accompanying game moves with high cognitive load. We statistically tested these two hypotheses relating to speech and gesture separately using two linear mixed-effect regression models in R (Bates et al. 2015; R Core Team 2015). For each, a full model was first fit with all possible predictors of the frequency of hesitation phenomena. These included (1) card complexity (0 to 4); (2) type of spoken hesitation (e.g., silence); (3) number

of game moves in the progression of the game (e.g., the 11th game move); (5) log duration of game moves; (6) log duration of words or G-phases containing hesitation; (7) G-phase type (e.g., stroke; only used in gestural analysis). We did not include gestural hesitation type as a predictor because not every gestural production was accompanied by one, and therefore, the missing values generated by this would bias the models (since the absence of hesitation type would naturally predict frequency). Finally, the participant was also included in the full models as a random effect. In the next step, non-significant effects in these initial models (one for speech and one for gesture) were removed from the models using backward elimination with the *ranova()* function in the *lmerTest* package (Kuznetsova et al. 2017) for random effects and *step()* function in the *stats* package for fixed effects. After the elimination, the remaining significant effects were fitted into final models to be interpreted.

3. Results

3.1. Spoken Hesitations

3.1.1. Phenomena

We expected our setup to elicit large amounts of hesitations constituting pauses (cf. Section 1.4). This assumption is confirmed, silences, fillers and lengthening are abundant in our recordings. As can be seen in Table 2, these three phenomena amount to a total of 1648 in a relatively small corpus of 110 min or 8141 words, which corresponds to a hesitation rate of 14.9 per minute, or per 20.2% of words.

Table 2 summarizes the distribution of hesitation types: Silences are most frequent, followed by lengthenings and fillers. As we will discuss in Section 4, this distribution differs somewhat from the general expectation.

Table 2. Spoken hesitations constituting pauses in our data.

Hesitations	N	%	Corpus Specifics	
Silence	901	54.7	Words in corpus	8141
Lengthening	498	30.2	Corpus duration	110 min
Filler	249	15.1	Words affected by hesitation	20.2%
total	1648		Hesitations per minute	14.9

3.1.2. Effects of Cognitive Load on Spoken Pausing Behavior

In Section 1.4 we hypothesized that cognitive load, operationalized as card complexity, will lead to an increased amount of hesitation. We tested this hypothesis using mixed-effects regression modeling detailed in Section 2.4. We included all the predictors listed in Section 2.4 in the full model as shown below:

$$Frequency \sim Complexity + Move\ number + Spoken\ hesitation\ type + Hesitation\ unit\ duration + Move\ duration + (1|Participant)$$

The elimination method (cf. Section 2.4) exhibited a significant random effect for participant with the intercept between -2 and 2 across participants. An analysis of individual variation is out of scope for this paper, but we kept the random effect in the final model to account for it. The elimination process showed significant effects for all five fixed effects, so the final model was kept the same. Table 3 shows the significant effects matrix of the model.

Table 3. The matrix of significant fixed effects on spoken hesitation frequency.

	Sum Sq	Mean Sq	NumDF	DenDF	F Value	Pr(>F)
Complexity	1888.843	629.614	3.000	1627.960	111.671	0.000
Move number	539.412	539.412	1.000	1627.495	95.672	0.000
Log move duration	3264.814	3264.814	1.000	1621.792	579.058	0.000
Log hesitation duration	25.118	25.118	1.000	1632.857	4.455	0.035
Hesitation type	49.076	24.538	2.000	1627.519	4.352	0.013

We observe a clear effect of cognitive load on spoken hesitation production. The frequency of hesitations increases as a function of card complexity (cf. Figure 6a). This effect is highly significant (cf. Table 3).

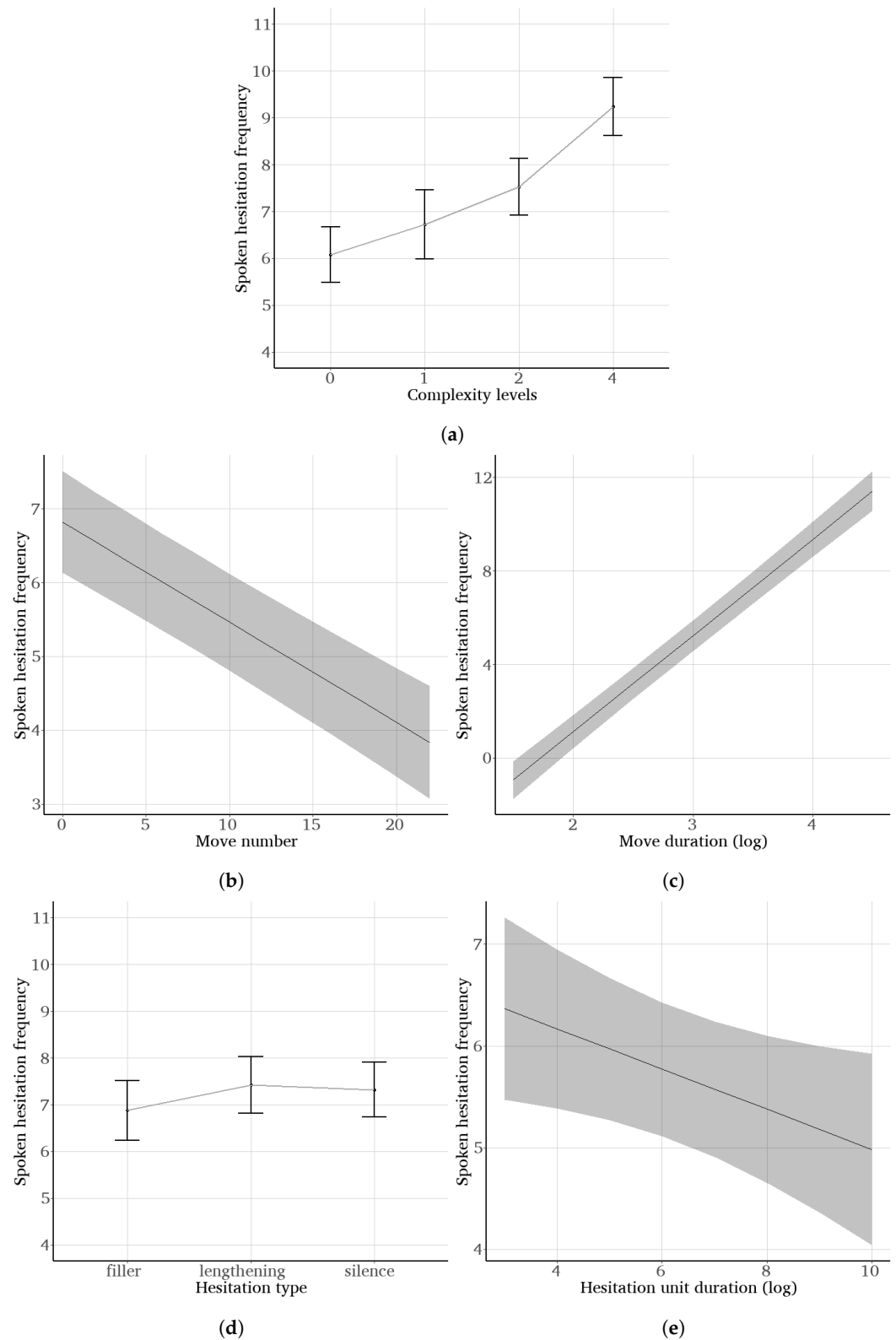


Figure 6. The estimated means of significant complexity levels (a), move number (b), log move duration (c), hesitation type (d), and log hesitation unit duration (e). Error bars and grey shaded areas indicate 95% CIs.

Move number and move duration also have significant effects on hesitation frequency: The higher the move number, indicating the progress of the game, the lower the frequency of hesitations (cf. Figure 6b) and the longer the game move lasts, the higher the frequency of hesitations therein gets (cf. Figure 6c).

Finally, hesitation type and hesitation unit duration have a significant impact on hesitation frequency: Lengthening appears to be mostly responsible for large amounts of hesitations within one game move, followed closely by silences and then by fillers (cf. Figure 6d, indicating that when hesitations occur in multiples within a game move, lengthenings are mostly involved (despite they are not the most frequent phenomenon overall). Further, there is a trade-off relationship between hesitation unit duration and hesitation frequency: the longer the individual hesitation phenomena, the fewer hesitations there are within every game move.

3.2. Gestural Hesitation

3.2.1. Phenomena

Table 4 shows the number and the type of occurrences of spoken hesitations, G-phases, and gestural hesitations in the subset of data annotated for gesture. First of all, it can be seen in Table 4a that the distribution of spoken hesitations in the subset parallels the main dataset’s. Unsurprisingly, Table 4b also shows that the most frequent G-phases annotated in the data are strokes, followed by preparations and holds (see Karpiński et al. 2009; Türk 2020 for similar distributions). Interestingly, however, jerks, movements often associated with hesitation (Section 1.2), are observed in as few as 23 instances in our data.

Table 4. The number of gesture phases, gestural hesitations, and spoken hesitations in the subset of data annotated for gesture.

a. Spoken Hesitations		
	N	%
Silence	558	56.1
Lengthening	283	28.4
Filler	154	15.5
Total	995	
b. G-Phases		
	N	%
Stroke	1258	42
Prep	733	24.5
Hold	601	20.1
Ret	369	12.3
Jerk	34	1.1
Total	2995	
c. Gestural Hesitations		
	N	%
Pause	283	32.7
Slow	187	21.6
Stall	173	20
Cancel	166	19.2
Meta	56	6.5
Total	865	

In general, ~29% of G-phases carry gestural hesitation (Table 4c). The most common type of these in our data is pauses. The remaining types have roughly equal number of observations, except for metaphoric expressions of hesitation which are the least common

type of multimodal hesitation markers. This is also predictable given that the participants are engaged in object manipulation instead of a conversation with a partner in our game setting (see Section 4).

The most common G-phase type that contains some form of gestural hesitation is the stroke (~49%, N = 424) followed by holds (~37%, N = 318). Preparations and retractions have a low number of gestural hesitations (~10%, N = 100) which are mostly slowed down phases or cancellations. This shows that gestural hesitations tend not to co-occur with peripheral gesture constituents but rather with core constituents (in terms of their linear ordering, see McNeill 1992).

Table 5. V-hesitations overlapping with G-hesitations.

	Cancel	Lengthen	Meta	Pause	Stall	n
Filler	10%	15%	11.7%	56.7%	6.7%	60
Lengthening	14.7%	31.3%	2.7%	43.3%	8%	150
Silence	16.6%	24.6%	4.7%	44.8%	9.2%	337

In our data, only ~55% of spoken hesitations (N = 547) are synchronized with gestural hesitations (and ~63% of the gestural hesitations are synchronized with spoken hesitations). Only 53% of the time the onsets of gestural hesitations precede the onsets of spoken hesitations. Table 5 shows a two-way frequency table of how these types of hesitations are synchronized. All types of spoken hesitation (i.e., spoken pauses, see Section 1.1) chiefly overlap with gestural pauses, which highlights a form-related relationship between multimodal hesitation productions (i.e., halts in production in both modalities). Similarly, although the relationship is not substantial, lengthenings also show a slightly increased number of co-occurrences with slowed down G-phases which can be considered to be similar to lengthening in terms of form (i.e., halts in articulatory movements with sustained phonation go with halts in manual movement with sustained hand shape). Overall, these findings suggest that spoken and gestural productions go hand-in-hand when signaling hesitation in a game setting. In Section 3.1.2, we already examined how various factors related to this setting, especially cognitive load, influence the frequency of spoken hesitations. The parallelism in the distributions presented so far suggests that these factors can also be used to predict the frequency of gestural hesitations (Section 1.4). Next, we present the findings of this investigation.

3.2.2. Effects of Cognitive Load on Gestural Hesitation

In Section 1.4, we predicted that factors affecting the frequency of spoken hesitations can also affect the frequency of gestural hesitations. We tested this prediction using mixed-effects regression modeling detailed in Section 2.4. For the analysis, we included all the predictors listed in Section 2.4 in the full model as shown below:

$$Frequency \sim Complexity + Move\ number + Spoken\ Hesitation\ type + Gphase\ type + Hesitation\ unit\ duration + Move\ duration + (1|Participant)$$

First, the elimination method (see Section 2.4) determined a significant participant random effect where the intercept estimates range between -2 and 2 across participants. An analysis of individual variation is out of the scope of this paper. However, the participant random effect was included in our final model to account for this effect (as in Section 3.1.2). Next, the elimination process was able to reduce the number fixed effects to four: complexity, move duration (log), G-phase type, and move number. A final model was then fitted with these fixed effects (and the participant random effect). Table 6 shows the matrix of significant effects in the model.

Table 6. The matrix of significant fixed effects on gestural hesitation frequency.

	Sum Sq	Mean Sq	NumDF	DenDF	F Value	Pr(>F)
Complexity	2569.155	856.385	3.000	1495.482	76.901	<0.001
Log move duration	6139.755	6139.755	1.000	1437.191	551.331	<0.001
G-phase type	119.927	29.982	4.000	1491.984	2.692	0.030
Move number	166.836	166.836	1.000	1498.862	14.981	<0.001

In Figure 7, the estimates of game move duration and game move number, or game progress, are plotted. The effects of these terms are similar to the ones presented in Section 3.1.2. Namely, the frequency of gestural hesitations is reduced as the game progresses further (Figure 7a), which implies a learning effect. In contrast, the frequency of gestural hesitations is increased along with the duration of the game move (Figure 7b).

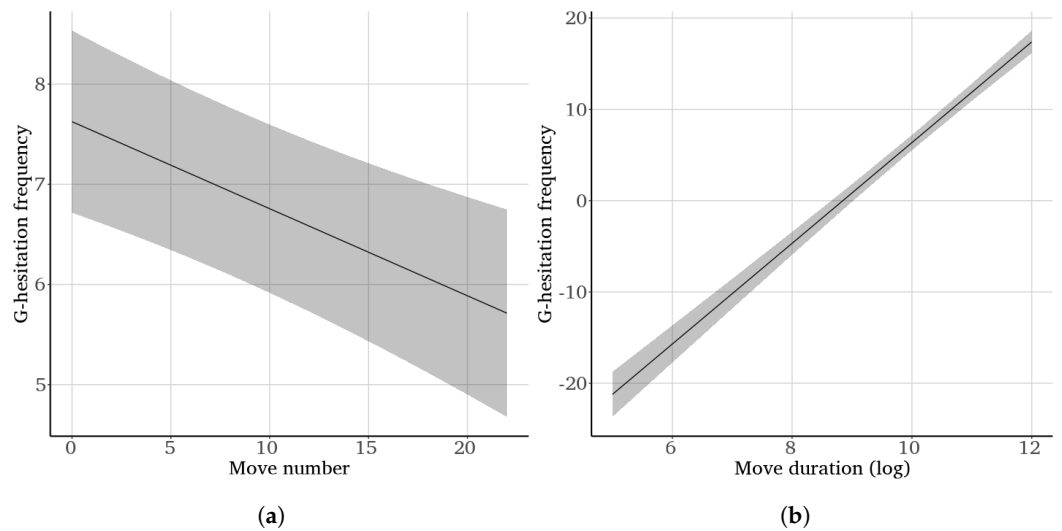


Figure 7. The estimated means of significant (b) Move duration and (a) Move number terms in the final model with their CIs at 95%.

Figure 8 shows the estimated means of complexity and G-phase type. It can be seen in Figure 8a that complexity has a similar effect on the frequency of gestural hesitation as it has on that of spoken hesitation. That is, the participants produced more gestural hesitations at higher complexity levels. Post-hoc pairwise comparisons (adjusted with Tukey’s method) showed that the estimates at complexity level 4 and level 0 were significantly different from each other ($t(1499) = -14.3, p < 0.001$) as well as from the intermediate levels 1 and 2 (levels 1 and 2 were not significantly different from each other, $t(1496) = 2.22, p > 0.117$). Overall, this is in line with our general hypothesis that high cognitive load associated with high complexity of game moves predicts the frequency of gestural hesitations as well as that of spoken hesitations.

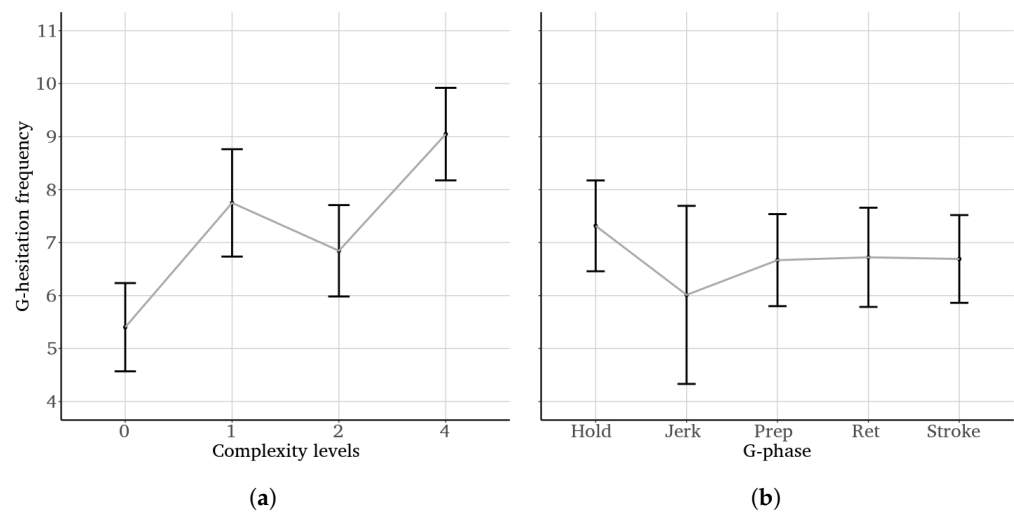


Figure 8. The estimated means of significant (a) complexity and (b) G-phase type terms in the final model with their CIs at 95%.

The estimates of the final significant effect, G-phase type, are plotted in Figure 8b. It shows that the frequency estimate of gestural hesitations was slightly higher when there were holds compared to the other types of G-phases. Note that gestural pauses were the most common type of gestural hesitations (Section 3.2.1), and they require the annotation of holds by definition (see Section 2.3.3) although there is no one-to-one correspondence (i.e., not every hold is a pause, compare Table 4b,c). Therefore, this effect might be interpreted as a by-product of these factors. However, in pairwise comparisons, the only significantly different pair is the stroke-hold contrast ($t(1494) = 2.92, p < 0.029$). Accordingly, although strokes occurred more frequently and attracted more hesitations (Section 3.2.1), the frequency of gestural hesitations was predicted to increase with holds.

From the presented distributions and models so far, it can be seen that spoken hesitation and gestural hesitation phenomena are intertwined, and their frequencies can be predicted using a similar set of variables. To add further evidence to this finding and to quantify the relationship, we performed a correlation analysis for the frequencies of these hesitations per game move. Figure 9 shows the output of this analysis. It shows that the two frequencies are moderately correlated ($r(1592) = 0.67, p < 0.001$), supporting our findings and interpretations presented in this section.

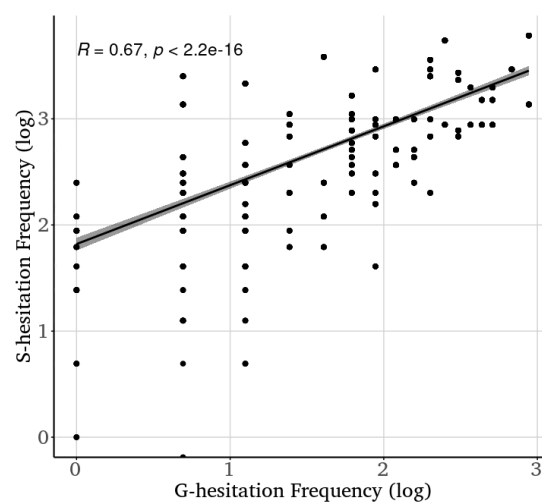


Figure 9. Pearson correlation between the frequencies of spoken hesitations (S-hesitations) and gestural hesitations (G-Hesitations).

4. Discussion

We have presented a study revolving around a specifically designed card game to elicit ecologically valid hesitations while being able to control for cognitive load. The game-based elicitation functioned very satisfactorily. The explanation of the game took less than five minutes per participant, all participants managed to play the game through in one turn, and despite the game being easy to understand, all participants thought more about the more complex game moves and revealed this in their commentaries. The mapping of complexity to hesitations was equally straightforward, which leads us to suggest to build upon this paradigm for future studies. However, there are still adjustments that we feel are recommended. It could be beneficial to balance the amount of cards of different complexities, which would simplify gathering sufficient data across complexity levels, and would potentially work for studies with fewer participants. It could further be worthwhile to adjust the overall goal of the game from “draw all cards” to “play all cards”, which would balance the amount of game moves players have to take. Another aspect to consider regards the design of the complex cards: several players used the complex cards as simple card draws, yielding hesitation-free moves like “I spend two resources, generate the same resources and draw a card”. This could be modified by forcing players to choose different resource combinations for input and output. For our purpose, the size of the deck and the resulting game duration was adequate, neither causing fatigue nor resulting in data sparsity. As the results regarding game moves in Section 3.1.2 indicate, there is a learning effect, as participants become faster over the course of the game. This suggests that our approach of having a short explanation immediately followed by the game without any further training phase was a useful approach for our target, namely to have participants being cognitively engaged, yet allow for a smooth gameplay.

With regard to spoken hesitations, we can answer our research question (Section 1.4) affirmatively, as we observed large amounts of hesitations. We observed 1648 hesitations in a 8141 word/110 minute corpus, corresponding to 20% of words being affected by hesitation, or 14.9 hesitations per minute. In relation, [Shriberg \(1994\)](#), [Eklund \(2004\)](#) and [Fox Tree \(1995\)](#), among others, stated that hesitations occur on or around 5–10% of words in everyday speech. ([Lickley 2001](#)) reports on 43% of utterances being disfluent when speakers are burdened with a task such as instructions. Compared to that, in our data, 67% of utterances contain hesitations. We have shown that our method generally works, not only that control over cognitive load can be achieved, but also that generally sufficient amounts of hesitations result from the game setting.

Another rather unexpected aspect is the distribution of hesitation types in our data. In earlier corpus studies, fillers typically outnumbered lengthenings ([Eklund 2004](#)). The rate of lengthening is also much higher than in our own previous studies that dealt explicitly with lengthening: As summarized in [Betz et al. \(2017\)](#), we found the lengthening rate to be 0.57 per minute in a corpus of free spontaneous speech (GECO) ([Schweitzer and Lewandowski 2013](#)), and 1.6 per minute in a task-oriented corpus (DUEL) ([Hough et al. 2016](#)). In this experiment, we observe 498 lengthenings in 110 min, or 4.52 per minute.

The reason for this might be the task at hand. Lengthening requires speech material to be applied to. Our setup constantly provides speakers with new material as they play a game, and they constantly have to communicate about it. In addition, speakers have to think about the details of what exactly to do within the game while they are processing the complex cards. This combination of factors seems to be fertile ground for the spawning of lengthening. By comparison, a spontaneous conversation between friends might yield more fillers than lengthening, because the content of the conversation may be freer which may cause empty thinking pauses where lengthening cannot manifest as easily and might be replaced by hesitations that can stand alone.

As lengthening research has often suffered from a data sparsity problem (cf. [Betz et al. 2017](#)), we believe it is a promising result that the method deployed here not only yields high hesitation rates overall, but also lengthening frequency five to ten times higher as compared to previous studies.

When it comes to gestural hesitations, we observed that almost two thirds of them were directly synchronized with (i.e., overlapped with) spoken hesitations. We did not analyze the cases where these hesitations did not overlap but simply occurred near each other; in our qualitative observation, there were many such cases. Overall, it can be said that in the majority of cases whenever a spoken hesitation was occurring there was a gestural hesitation occurring nearby (although not necessarily perfectly synchronized), confirming our predictions (Section 1.4).

As for the distributions related to G-phases and gestural hesitations, our findings both differ from and confirm those of earlier studies. In our data, the most common G-phase type was the stroke and most gestural hesitations tended to occur with these. This is in line with [Seyfeddinipur \(2006\)](#) except for preparations which they also reported to co-occur with gestural hesitations. There were few co-occurrences with these in the data. Unlike [Graziano and Gullberg \(2018\)](#), there weren't more holds than strokes in total, however holds were more likely to bear gestural hesitations (~50% of annotated holds) than strokes (~30% of annotated strokes). This is in line with [Graziano and Gullberg \(2018\)](#) where they claimed there were more holds during disfluent speech than fluent. Further, the most common gestural hesitations that were synchronized with all types spoken hesitations were pauses. Since spoken hesitations can be considered as pauses (Section 1.1), this suggests coordination between speech and gesture which is that when speech goes into a pause, so does gesture, as predicted in Section 1.4).

We did not observe many jerky movements in our data. This was surprising as they are associated with hesitation leading to cancellations and restarts of gesture, and because of this, they can be considered as indicators of stronger and more abrupt hesitations. There were also few examples of metaphoric gestures signaling hesitation in the data although they were reported to be common ([Graziano and Gullberg 2018](#)). Our interpretation is that these variations were due to the nature of the task. In our setting, the participants were primarily engaged in object manipulation (i.e., game) instead of a narrative (cf. [Graziano and Gullberg 2018](#); [Seyfeddinipur 2006](#)). The element of object manipulation potentially allows for fewer referential gestures (as the hands are busy), and there is no critical need to express hesitation in a separate referential gesture without a conversational partner (except for habitual uses). Similarly, the participants could plan their game moves at their own pace without having to manage a partner's expectations, and therefore, stronger cues to hesitation, i.e., jerks, were unnecessary. Overall, this implies that hesitation phenomena in gesture are sensitive to contextual constraints.

We hypothesized that increased card complexity leads to increased cognitive load, which in turn causes an increase in hesitation rates. We could confirm this hypothesis not only for spoken hesitations but also for gestural hesitations using a similar set of predictors, in line with our hypothesis in Section 1.4. This result is not unexpected, since previous studies have established a link between cognitive load and hesitations as well as a tight link between speech and gesture (Section 1.3). The novelty of our contribution is that speech and gesture perform similarly to cue hesitation uniformly - they both tend to pause (i.e., formal interaction) and at comparable rates (i.e., temporal interaction). Moreover, these phenomena can be elicited with a simple card game setup in a more ecologically valid way instead of using settings with stronger interventions (e.g., delayed auditory feedback).

The other effects that we observed include a clear learning effect, i.e., spoken hesitation frequency drops as the game progresses. This needs to be taken into account for future studies that the game in the current shape should not be simply expanded, e.g., by means of increasing the deck size, to obtain more data. The data points per time unit decrease the more acquainted participants become to the game. Furthermore, hesitation frequency increases linearly with game move duration. This seems to be a mutual relationship: a complex game situation leads to increased cognitive load, which leads to more hesitations and more time spent on that game move. However, the game was not controlled for how long it actually takes to physically perform the required moves, therefore, it is difficult to determine the source of the increase in the game duration. As a result, we decided to keep

it as a predictor in our models (although the effect may seem theoretically obvious), and the model converged without any warnings of collinearity.

There is a negative correlation between the duration of individual hesitation phenomena, dubbed “unit duration”, and hesitation frequency. This seems to be a straightforward trade-off - speakers either hesitate by producing many instances of hesitation or they produce fewer but longer instances. This aspect might deserve some attention in future studies; it is possible that inter-speaker variation is key for explaining it.

For the frequency of gestural hesitations, we observed an effect of G-phase type in which holds increased this frequency when compared to strokes. These two are the most common G-phase types observed in our data. We interpret this as a by-product of gestural pauses being the most common hesitation type, as they require hold annotations by definition (Section 2.3.3). Regardless, the significantly increased rate of occurrence of gestural hesitations with holds instead of strokes (of which there were more in the data), support our hypothesis on speech and gesture operating in tandem in expressions of hesitation.

For spoken hesitations, we found a significant effect indicating that the type of hesitation has an influence on how many hesitations there are per game move. Interestingly, this effect cannot be observed for gestures. We argue that this is evidence to our claim that lengthening is to be regarded as functionally the same as fillers and silences - as pauses, in this case attributable to cognitive load.

Overall, we have shown that it is possible to elicit large amounts of different types of pause phenomena in an ecologically valid, though still laboratory-based setting, and to simultaneously monitor the amount of cognitive load on a speaker. We believe that game settings have a large potential for research in phonetics and linguistics, because they have participants engage with the task at hand, but allow for a relatively flexible level of control (with the help of adjustable game rules).

We believe the potential of this methodological approach is strengthened by our findings on the effect of cognitive load on hesitations, which replicated previous findings based on more controlled laboratory settings. The replication of results across different methodological designs actually shows that the effect of cognitive load on speech delivery is a robust one, which may potentially be less context dependent than other, more speaking style related effects (Wagner et al. 2015). However, as our design was not strongly interactive, it may not have triggered functions of hesitations which are related to audience design, and which may have led to a different distribution of hesitations, including cross-modal aspects.

Generally, our results provide further evidence for the strong cross-modal link existing in speech production planning, which we have shown to be evident as well in the realization of hesitation related pauses. Lastly, the results corroborate our initial claim that lengthenings are indeed a special type of pause realization, in addition to silences and fillers.

Author Contributions: Conceptualization, investigation, analysis, writing—original draft, S.B.; annotation, research, writing—original draft, N.B.; investigation, analysis, writing—original draft, O.T.; supervision, writing—review and editing, P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021–438445824.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Bielefeld University (protocol code 2022-189 on 7 July 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Access to Datasets and game materials can be requested via mail to the corresponding author.

Acknowledgments: We are very thankful to our student assistants: Dominique Hofmann for helping us with gesture annotation and to Moritz Wackerbarth for helping with spoken hesitation annotation.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Aboudan, Rima, and Geoffrey Beattie. 1996. Cross-cultural similarities in gestures: The deep relationship between gestures and speech which transcends language barriers. *Semiotica* 111: 269–94. [CrossRef]
- Allwood, Jens, Elisabeth Ahlsén, Johan Lund, and Johanna Sundqvist. 2005. Multimodality in own communication management. Paper presented at the Second Nordic Conference on Multimodal Communication, Göteborg, Sweden, April 7–8, pp. 1–20.
- Allwood, Jens, Joakim Nivre, and Elisabeth Ahlsén. 1990. Speech management—on the non-written life of speech. *Nordic Journal of Linguistics* 13: 3–48. [CrossRef]
- Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, and et al. 1991. The hrc map task corpus. *Language and Speech* 34: 351–66. [CrossRef]
- Arnold, Jennifer E., Carla L. Hudson Kam, and Michael K. Tanenhaus. 2007. If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33: 914. [CrossRef] [PubMed]
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]
- Bavelas, Janet, Nicole Chovil, Douglas A. Lawrie, and Allan Wade. 1992. Interactive gestures. *Discourse Processes* 15: 469–89. [CrossRef]
- Belz, Malte. 2021. *Die Phonetik von äh und ähm: Akustische Variation von Füllpartikeln im Deutschen*. Berlin and Heidelberg: Springer.
- Belz, Malte, and Uwe D. Reichel. 2015. Pitch characteristics of filled pauses in spontaneous speech. Paper presented at the 2015: Disfluency in Spontaneous Speech, Edinburgh, Scotland, August 8–9.
- Belz, Malte, and Jürgen Trouvain. 2019. Are 'silent' pauses always silent? Paper presented at the 19th International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia, August 5–9.
- Betz, Simon. 2020. Hesitations in Spoken Dialogue Systems. Ph. D. thesis, Universität Bielefeld, Bielefeld, Germany. [CrossRef]
- Betz, Simon, Birte Carlmeyer, Petra Wagner, and Britta Wrede. 2018. Interactive hesitation synthesis: Modelling and evaluation. *Multimodal Technologies and Interaction* 2: 9. [CrossRef]
- Betz, Simon, Jana Voße, Sina Zarriß, and Petra Wagner. 2017. Increasing recall of lengthening detection via semi-automatic classification. Paper presented at the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden, August 20–24, pp. 1084–88.
- Betz, Simon, Petra Wagner, and Jana Vosse. 2016. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In *Tagungsband Der 12. Tagung Phonetik Und Phonologie Im Deutschsprachigen Raum*. Munich: Ludwig Maximilian University of Munich, pp. 19–23.
- Betz, Simon, Sina Zarriß, and Petra Wagner. 2017. Synthesized lengthening of function words—The fuzzy boundary between fluency and disfluency. In *Proceedings of the International Conference Fluency and Disfluency*. Edited by Liesbeth Degand. Stockholm: Royal Institute of Technology (KTH), pp. 15–19.
- Betz, Simon, Sina Zarriß, Éva Székely, and Petra Wagner. 2019. The green tree—Lengthening position influences uncertainty perception. Paper presented at the 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019, Graz, Austria, September 15, pp. 3990–94.
- Boersma, Paul, and David Weenink. 2014. Praat: Doing Phonetics by Computer [Computer Program]. Available online: <http://www.praat.org/> (accessed on 7 July 2022).
- Brugman, Hennie, and Albert Russel. 2004. Annotating Multimedia/ Multi-modal resources with ELAN. Paper presented at the LREC 2004, Fourth International Conference on Language Resources and Evaluation, Nijmegen, The Netherlands, May 26–28. Available online: <https://archive.mpi.nl/tla/elan> (accessed on 7 July 2022).
- Butterworth, Brian, and Geoffrey Beattie. 1978. Gesture and silence as indicators of planning in speech. In *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*. Edited by R. N. Campbell and P. Smith. New York: Plenum, pp. 347–60.
- Campione, Estelle and Jean Véronis. 2002. A large-scale multilingual study of silent pause duration. Paper presented at the Speech Prosody 2002, International Conference, Aix-en-Provence, France, April 11–13, pp. 199–202.
- Chafe, Wallace. 1980. Some reasons for hesitating. In *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. Berlin: Walter de Gruyter, pp. 169–80.
- Churchill, Alex, Stella Biderman, and Austin Herrick. 2019. Magic: The gathering is turing complete. *arXiv* arXiv:1904.09828.
- Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.
- Cook, Susan Wagner, Zachary Mitchell, and Susan Goldin-Meadow. 2008. Gesturing makes learning last. *Cognition* 106: 1047–58. [CrossRef]
- Dahl, Kimberly L., and Cara E. Stepp. 2021. Changes in relative fundamental frequency under increased cognitive load in individuals with healthy voices. *Journal of Speech, Language, and Hearing Research* 64: 1189–96. [CrossRef]
- De Ruiter, Jan Peter. 2000. The production of gesture and speech. *Language and Gesture* 2: 284–311.

- Eklund, Robert. 2004. Disfluency in Swedish Human–Human and Human–Machine Travel Booking Dialogues. Ph. D. thesis, Linköping University Electronic Press, Linköping, Sweden.
- Fischer, Kerstin, Oliver Niebuhr, Eszter Novák-Tót, and Lars C. Jensen. 2017. Strahlt die negative Reputation von Häsitationsmarkern auf ihre Sprecher aus? Paper presented at the 43rd Annual Meeting of the German Acoustical Society (DAGA), Kiel, Germany, March 6–9, pp. 1450–53.
- Fox Tree, Jean E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34: 709–38. [[CrossRef](#)]
- Goldin-Meadow, Susan. 1999. The role of gesture in communication and thinking. *Trends in Cognitive Sciences* 3: 419–29. [[CrossRef](#)] [[PubMed](#)]
- Goldin-Meadow, Susan, Howard Nusbaum, Spencer D. Kelly, and Susan Wagner. 2001. Explaining math: Gesturing lightens the load. *Psychological Science* 12: 516–22. [[CrossRef](#)] [[PubMed](#)]
- Goldman-Eisler, Frieda. 1967. Sequential temporal patterns and cognitive processes in speech. *Language and Speech* 10: 122–32. [[CrossRef](#)] [[PubMed](#)]
- Graziano, Maria, and Marianne Gullberg. 2018. When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology* 9: 879. [[CrossRef](#)]
- Graziano, Maria, Elena Nicoladis, and Paula Marentette. 2020. How referential gestures align with speech: Evidence from monolingual and bilingual speakers. *Language Learning* 70: 266–304. [[CrossRef](#)]
- Hough, Julian, Ye Tian, Laura de Ruyter, Simon Betz, David Schlangen, and Jonathan Ginzburg. 2016. DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. Paper presented at the 10th edition of the Language Resources and Evaluation Conference, Portoroz, Slovenia, May 23–28; pp. 1784–88.
- Jabeen, Farhat, and Simon Betz. 2022. Hesitations in Urdu/Hindi: Distribution and Properties of Fillers & Silences. *Interspeech* 2022: 4491–5. [[CrossRef](#)]
- Karpiński, Maciej, Ewa Jarmołowicz-Nowikow, and Zofia Malisz. 2009. Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. *Speech and Language Technology* 11: 113–22.
- Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.
- Kisler, Thomas, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–47. [[CrossRef](#)]
- Kita, Sotaro. 2000. How representational gestures help speaking. *Language and Gesture* 1: 162–185.
- Kita, Sotaro, and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language* 48: 16–32. [[CrossRef](#)]
- Kita, Sotaro, Ingeborg van Gijn, and Harry van der Hulst. 1997. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *International Gesture Workshop*. Berlin and Heidelberg: Springer, pp. 23–35.
- Kosmala, Loulou, Maria Candea, and Aliyah Morgenstern. 2019. Synchronization of (dis) fluent speech and gesture: A multimodal approach to (dis) fluency. Paper presented at the 6th Gesture and Speech in Interaction Conference, Paderborn, Germany, September 11–13.
- Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82: 1–26. [[CrossRef](#)]
- Landis, J. Richard, and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33: 159–74. [[CrossRef](#)]
- Levelt, William J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.
- Lickley, Robin J. 2001. Dialogue moves and disfluency rates. Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech, Scotland, UK, August 29–31.
- Lively, Scott E., David B. Pisoni, W. Van Summers, and Robert H. Bernacki. 1993. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America* 93: 2962–73. [[CrossRef](#)]
- Loehr, Daniel P. 2004. *Gesture and Intonation*. Ph. D. thesis, Georgetown University, Washington, DC, USA.
- McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.
- McNeill, D. 2005. *Gesture and Thought*. Chicago: University of Chicago Press.
- Montacié, Claude, and Marie-José Caraty. 2014. High-level speech event analysis for cognitive load classification. *Interspeech* 2014: 731–5. [[CrossRef](#)]
- Moon, A. Jung, Chris A. C. Parker, Elizabeth A. Croft, and H. F. Machiel Van der Loos. 2011. Did you see it hesitate?-empirically grounded design of hesitation trajectories for collaborative robots. Paper presented at the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, September 25–30; pp. 1994–99.
- Morrel-Samuels, Palmer, and Robert M. Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18: 615. [[CrossRef](#)]
- Novack, Miriam A., and Susan Goldin-Meadow. 2017. Gesture as representational action: A paper about function. *Psychonomic Bulletin & Review* 24: 652–65.
- Pouw, Wim, and James A. Dixon. 2019. Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cognitive Science* 43: e12721. [[CrossRef](#)] [[PubMed](#)]
- R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

- Schuller, Björn, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive and physical load. *Interspeech 2014*: 427–31. [\[CrossRef\]](#)
- Schweitzer, Antje and Natalie Lewandowski. 2013. Convergence of articulation rate in spontaneous speech. Paper presented at the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France, August 25–29; pp. 525–29.
- Seyfeddinipur, Mandana. 2006. Disfluency: Interrupting Speech and Gesture. Ph.D. thesis, Radboud University Nijmegen, Nijmegen, The Netherlands.
- Shriberg, Elizabeth Ellen. 1994. Preliminaries to a Theory of Speech Disfluencies. *Ph D. thesis*. University of California, San Diego, CA, USA.
- Sweller, John, and Paul Chandler. 1991. Evidence for cognitive load theory. *Cognition and Instruction* 8: 351–62. [\[CrossRef\]](#)
- Trouvain, Jürgen, and Raphael Werner. 2022. A phonetic view on annotating speech pauses and pause-internal phonetic particles. In *Transkription und Annotation Gesprochener Sprache und Multimodaler Interaktion: Konzepte, Probleme, Lösungen*. Tübingen: Narr Francke Attempto Verlag, vol. 55.
- Türk, Olcay. 2020. Gesture, Prosody and Information Structure Synchronisation in Turkish. Ph. D. thesis, Victoria University of Wellington, Wellington, New Zealand.
- Vukovic, Maria, Vidhyasaharan Sethu, Jessica Parker, Lawrence Cavedon, Margaret Lech, and John Thangarajah. 2019. Estimating cognitive load from speech gathered in a complex real-life training exercise. *International Journal of Human-Computer Studies* 124: 116–33. [\[CrossRef\]](#)
- Wagner, Petra, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57: 209–232. [\[CrossRef\]](#)
- Wagner, Petra, Jürgen Trouvain, and Frank Zimmerer. 2015. In defense of stylistic diversity in speech research. *Journal of Phonetics* 48: 1–12. [\[CrossRef\]](#)
- Yap, Tet Fei, Julien Epps, Eliathamby Ambikairajah, and Eric H. C. Choi. 2011. Formant frequencies under cognitive load: Effects and classification. *EURASIP Journal on Advances in Signal Processing* 2011. [\[CrossRef\]](#)
- Yin, Bo, and Fang Chen. 2007. Towards automatic cognitive load measurement from speech analysis. In *Human-Computer Interaction. Interaction Design and Usability*. Edited by Julie A. Jacko. Berlin and Heidelberg: Springer, pp. 1011–20.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.