The Shrink Point:

Audiovisual Integration of Speech-Gesture Synchrony

Dissertation

zur Erlangung des akademischen Grades

Doktor der Philosophie

an der Fakultät für Linguistik und Literaturwissenschaft

der Universität Bielefeld

vorgelegt von

Carolin Kirchhof

Gutachter:

Prof. Dr. Petra Wagner

Prof. i.R. Dr. Dafydd Gibbon

Bielefeld, 2016

Tag des Kolloquiums: 21.10.2016

Gedruckt auf alterungsbeständigem Papier °° ISO 9706

Danksagung

Mein größter Dank gilt Dafydd und Petra: Ihr habt mich aus der Verzweiflung geholt und mir wieder Mut und Hoffnung gegeben, und das kurz vor knapp. Der tatsächliche Abschluss dieser Dissertation wäre aber auch ohne meine Prüfer Martina, Horst und Stefan sowie meine großartigen Schreibgenossen Oskar, Leni, Linda, Helena und Svenja nicht möglich gewesen, die sich mit mir und unseren Schreibprojekten die Tage und Nächte um die Ohren geschlagen haben. Ich bin zudem unfassbar froh über den Support von Papa Erhard und den meiner Freunde und Kollegen aus Rudel und Büro, die PhoPhos und die PsyKlis, die mich haben machen lassen, aber mir gelegentlich auch den passenden Tritt verpasst haben. Danke ich für all die Unterstützung über die letzten 12 Jahre, die ich an dieser wundervollen Fakultät verbringen durfte. Ich bin dann jetzt fertig hier.

I am deeply grateful to Sue Duncan, who sparked my interest in gestures so many years ago at Berkeley, and then gave me the opportunity to learn and study in Chicago. I am also thankful for the great conversations I had there with David McNeill, AT, and the lab gang – I learned so much! The same goes for the ISGS crew! I should also be grateful to JP de Ruiter, who gave me the opportunity to research what fascinated me in the first place. Finally, I want to express my gratitude to Adam Kendon, who has been giving me very helpful comments on my work from the beginning.

That’s your responsibility as a person, as a human being – to constantly be updating your positions on as many things as possible. And if you don’t contradict yourself on a regular basis, then you’re not thinking.

(MALCOLM GLADWELL, 2014)

Table of Contents

Index of Figures

Index of Tables

List of Abbreviations

A:	audio
AV:	audio before video
AVI:	audiovisual
CV:	consonant-vowel
c.p.:	concentration problems
EEG:	electroencephalography
ERP:	event-related potentials
FBI:	Federal Bureau of Investigation
G:	gesture
GP:	Growth Point
gphr:	gesture phrase
gphs:	gesture phase
GS:	gesture before speech
L:	listener
LTM:	long-term memory
MU:	minimal psychological unit
NP:	noun phrase
S:	speech
S':	speaker
SFM:	speech-focused movements
SG:	speech before gesture
SP:	Shrink Point
ToT:	Tip-of-the-Tongue
V:	video
VA:	video before audio
VS(O):	verb-subject(-object)
v.s.:	videre supra
WM:	working memory

Terminology

Audiovisual integration; also **AVI**: The mental process of connecting signals from different modalities as belonging to the same signal.

Cause-and-effect signals: Non-speech audiovisual signals involving sounds caused by a known source, for example knocking or clapping. This categorization explicitly excludes speech-only utterances.

Gestures: Throughout this dissertation, the word ‘gesture’ will imply idiosyncratic, spontaneous movements of the arms and hands that co-occur with spontaneous, natural speech in a conversational setting. Self-adaptors, for example scratching as well as gestures of other body parts, for example phonological gestures, are not included when not explicitly mentioned. If not otherwise stated, gesture phrases (**gphr**s) will encompass all gestural motion between two resting positions (“equilibrium position”; Butterworth & Beattie, 1978), that is, their onset, stroke, (apex,) and retraction phases. The discrimination of different gesture types used is based on the widely semantic categorizations by McNeill (e.g., 1992; 2005), but does not strictly adhere to it in every aspect. The following types of gestures are to be distinguished:

Beat gestures (beats): Gestures of the rhythmical variety without semantic content.

Deictic gestures (deictics): Pointing gestures of one or more fingers, hands, or arms as well as tracing of shapes or trajectories.

Emblematic gestures (emblems): Codified gestures such as “thumbs up” or “the middle finger” that can function in place of a spoken word or phrase; NOT cherological items of any sign language (cf. Stokoe, 1960/2000).

Iconic gestures (iconics): Gestures which show shape, size, or movement features that resemble aspects of objects or actions characteristic of objects or actions that are being referred to in the speaker’s speech. . This includes the metaphoric gestures of McNeill’s terminology (see also de Ruiter, 2000; p. 285) and pantomime, but not gestures used for turn management.

Growth Point (GP), unpacking of the: The temporal interval during which speech and gesture overlap with the most intense semantic power (see Chapter 3.2).

Information uptake: The process between perception and comprehension.

Modality: The means by which a speaker relates information, here: speech and/or gestures in particular.

Perception: The sensing of audio and/or visual signals through eyes and/or ears. This process does not involve comprehension, but alongside comprehension is part of the signal reception process.

Percept: The perceived signal resulting from integrating the perceived modalities.

1 Introduction

Consciously or not, we communicate with every means available to us (Peirce, 1894/1998; de Saussure, 1972/1983). Verbal utterances can be wisely planned or produced spontaneously and unplanned, as can, for instance, facial expressions (e.g., Ekman, 2003). The general outward appearance as well as body posture and limb movements also express a lot from about speakers or listeners. Regarding these many layers of communication, a plethora of advisory literature on body language has been published since the late 1960s (e.g., Morris, 1967; 1982; 2002; Fast, 1971; McNeill, 2015), and it has enjoyed a continuous popularity on the market ever since. With promises of enhanced power and success (e.g., Trautmann-Voigt & Voigt, 2012; Latiolais-Hargrave, 2008) or of FBI-Agent-like abilities in reading people (Navarro & Karlins, 2008; also Morris, 2002), authors capture the minds and bodies of their readership. A large area of these explorations into the universe of body language is taken up by gestures – of the hands, of the head, of the feet. But apart from the psychological insights some advertise to be gained from such gestures, these movements can also relate communicative content alongside verbal utterances and even alone.

The exploration into the intricate connection between speech and gestures was initiated and influenced by, for example, Efron (1941), Kendon (1978; 1980; 2004), Schegloff (1984), and McNeill (1985; 1992; 2005; 2012). Spontaneous speech and semiotically related gestures are produced roughly simultaneously (e.g., Kendon, 2004), and it has long been agreed upon that gestures can support or add to the content related through speech alone (e.g., Krauss, Morrel-Samuels & Colasante, 1991; Melinger & Levelt, 2004; McNeill, 2005; Holler, Shovelton & Beattie, 2009). There has been a major focus on the lexico-semiotic connection between spontaneously co-produced gestures and speech in gesture research. Due to the rather precise timing between the prosodic peak in speech with the most prominent stroke of the gesture phrase in production, Schegloff (1984) and Krauss et al. (1991; also Rauscher, Krauss & Chen, 1996), among others, coined the phenomenon of *lexical affiliation* (see also Chapters ;). There are various issues with this fixed interpretation of speech-gesture-interlocking, least of all the general lack of lexicalization of non-emblematic gestures (see, e.g., Kendon, 2004). De Ruiter and Wilkins (1998) as well as de Ruiter (2000) suggested that the semiotic connection between co-produced speech and gesture is relating to a whole utterance rather than only to the point at which speech and gesture coincide the strongest (cf. the *Growth Point* theory, e.g., McNeill, 1985). By following Krauss et al. (1991), the *Conceptual Affiliation Study*, as the first empirical study of this dissertation will investigate the nature of the semiotic relation between speech and gestures.

Not only regarding temporal factors, the focus in gesture research has long been on the *production* of speech-accompanying gestures and on how speech-gesture utterances contribute to communication. An issue that has mostly been neglected is in how far listeners even *perceive* the gesture-part of a multimodal utterance. Since there is no cause-and-effect relation between the modalities, as there is in lip-motion and airflow and speech, the synchrony of speech and gesture in production cannot be fully explained with physical articulatory mechanisms. Whether this synchrony is relevant for perception or comprehension and in what way the two modalities are linked in the production process has been under constant review (e.g., de Ruiter, 2007; Kita & Özyürek, 2003; Krauss, Chen & Gottesman, 2000), and will be further discussed within the scope of this dissertation. Additionally, how synchrony itself needs to be understood will be explored in the context of speech-gesture production, in particular regarding temporal overlaps within multimodal utterances (e.g., Allen, 1983). A unanimous understanding of this central concept is essential to a detailed analysis of multimodal signals.

For researchers in the field of speech-lip perception, perceived synchrony has long been an area of focus. It is, for instance, a common phenomenon that the dubbing of foreign-language films often does not match the lip movements of the original to the point. Depending on the language pairs and the viewer-listener's familiarity with dubbing, the resulting speech-lip asynchrony will be noticeable to differing degrees. But, depending on the language translation pairs and the money and motivation available to them, translators, voice actors, and technical staff can make dubbing just barely noticeable. When speech and lip movements diverge too far from the original production synchrony, this can be highly irritating to the viewer, even when audio and video stem from the same original recording (e.g., Vatakis, Navarra, Soto-Faraco & Spence, 2008; Feyereisen, 2007) – there is only a small temporal window of audiovisual integration (AVI) within which viewer-listeners can internally align discrepancies between lip movements and the speech supposedly produced by these. What happens when listeners realign speech-lip signals with slight asynchronies has been prominently investigated by McGurk and MacDonald (1976), among others.

Several studies in the area of psychophysics (e.g., Nishida, 2006; Fujisaki & Nishida, 2005) found that there is also a time window for the perceptual alignment of non-speech visual and auditory signals. These and further studies on the AVI of speech-lip asynchronies by Massaro, Cohen, and Smeele (1996; also Massaro & Cohen, 1993; Vatakis et al., 2008) have inspired research on the as-of-yet sparsely dealt with perception of speech-gesture utterances. A pioneer approach to whether listeners attend to speech-accompanying gestures was made by McNeill, Cassell, and McCullough (1994; Cassell, McNeill & McCullough, 1999), who discovered that listeners take up information even from artificially combined speech and gestures. This approach using semantically mismatched signals was, among others, adopted by Goldin-Meadow (e.g., Goldin-Meadow, Kim & Singer, 1999), particularly in the classroom context. More recent studies researching the AVI of speech and gestures have employed event-related potential (ERP) monitoring as a methodological means to investigate the perception of multimodal utterances, also taking into account temporal relations (e.g., Gullberg & Holmqvist, 1999; 2006; Özyürek, Willems, Kita & Hagoort, 2007; Habets, Kita, Shao, Özyürek & Hagoort, 2011).

While the aforementioned studies from the fields of psychophysics and speech-only and speech-gesture research have contributed greatly to theories of how listeners perceive multimodal signals, there has been a lack of explorations of natural data and of dyadic situations. This dissertation will investigate the perception of naturally produced speech-gesture utterances. For this purpose, a corpus of spontaneous dialogical speech and gestures was gathered to create stimuli for the different studies on speech-gesture perception conducted within this dissertation.

The synchrony between speech and gestures is prominent during speech production, and multimodal synchrony is essential for speech-lip utterance perception. Accordingly, one aspect to investigate will be the perception of audiovisual synchrony as well as of asynchronies between naturally co-produced speech and gestures. Two sets of studies will apply two different methodologies to create an encompassing picture of in how far listeners perceive different degrees of speech-gesture asynchronies:

The *Perceptual Judgment Task* will inquire on as how natural listeners perceive different degrees of audio advances and delays in speech-gesture utterances as well as in physical cause-and-effect stimuli (Chapter 7). These studies will reassess the windows of AVI previously observed in non-speech and speech-only audiovisual signals a well as those approximated by the ERP studies by Özyürek et al. (2007) and Habets et al. (2011). Using a slider interface, the *Preference Task* will have listeners re-synchronize temporally manipulated stimuli similar to those tested in the Perceptual Judgment Task (Chapter 8). The results of these studies will provide insights into whether listeners perceive asynchronies when not presented with a set of asynchronies to choose from, as well as how the timing of speech-gesture production relates to what listeners prefer for perception. Connecting the discoveries about the conceptual affiliation between speech and gestures with how listeners perceive variation in the temporal alignment of the two modalities in face-to-face conversation will shed light on the connection between production synchrony and its relevance for the listeners.

The temporal interval during which speech and gesture are the most co-expressive is known as the unpacking of the *Growth Point* (GP; e.g., McNeill, 1985; 1992). The perceived essence of speech-gesture utterances would be the counterpart of the GP in the speaker, the GP unpacked by the speaker during the multimodal utterance and then audiovisually integrated by the listener to recreate the idea the speaker wanted to relate. The result of the perceptual repacking of speech-gesture information as a conceptual phenomenon would be the *Shrink Point*[1] (SP) (see Figure 1): The speaker S' will produce a speech-gesture utter-

[1] The seeds for the "Shrink Point" hypothesis were first planted by JP de Ruiter in an unpublished research proposal draft in 2010.

ance containing the GP, which is unpacked during this utterance and then perceived and integrated by the listener into the SP.

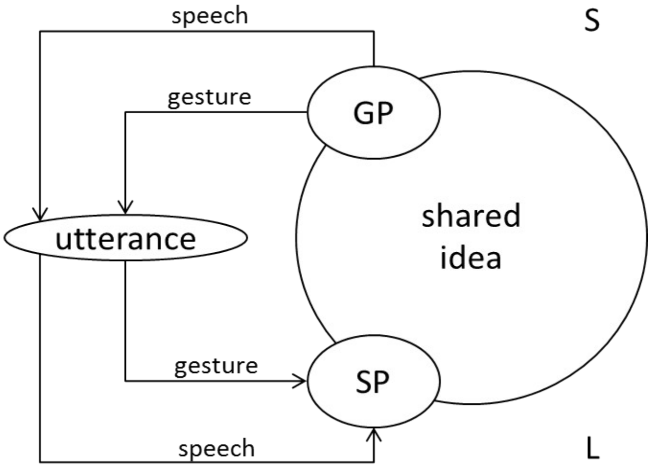


Figure 1: GP-SP transmission cycle (basic draft).

Drawing, for example, from speech-gesture production models based on Levelt's (1989) model of speech production (e.g., de Ruiter, 1998; 2007; Krauss et al., 2000; Kita & Özyürek, 2003), a model draft of a possible transmission cycle between GP and SP will be proposed. Based on the results and analyses thereof of the studies conducted for this dissertation and their analysis before the provided theoretical background, the model draft will be expanded to include the temporal and semantic alignment of speech and gestures in production and their audiovisual and conceptual integration during perception based on experimental data (Chapter 9). The successful transmission of a compressed idea unit via speech and gestures will be telling with regard to the degree of communicative efficiency of speech-gesture synchrony and its overall relevance for the perception of multimodal language signals.

1.1 Thesis Structure

After an introduction to major topics in gesture research, the recurring debate about the communicative function of speech-accompanying gestures will be addressed in Chapter 2. A portrayal of the major research foci regarding the connection between speech and gestures in language production from the speaker's perspective, that is, their temporal synchrony and lexical or conceptual affiliation will lead to a discussion of McNeill's GP theory. Several researchers have proposed production models drawing from this research. These will be discussed conclusively to connect the different temporal and semantic features of speech-gesture production and to define which mechanisms need to be present on the production and perception sides of the GP-SP transmission cycle.

Chapter 4 will explore the theories behind signal perception, that is, AVI. Following an overview of how methodological as well as theoretical foundations laid by psychophysics have been applied in the research on speech perception, several studies from the area of speech-gesture perception will be discussed. Connecting the findings form these different approaches to multimodal signal perception, the SP hypothesis will be formed and an extended model draft of the GP-SP transmission cycle will be proposed in Chapter 4.5.

A set of hypotheses regarding the GP-SP transmission cycle in general, and the semantic and temporal affiliation between co-produced speech and gestures specifically, will be put forward in Chapter 5. Methodologies will be presented for three differing approaches to these hypotheses, one contesting the methodology of Krauss et al. (1991) to research lexical speech-gesture affiliation, one investigating how listeners perceive asynchronies in speech-gesture production by means of the Perceptual Judgment Task, and one having listeners realign desynchronized speech-gesture stimuli into what they believe resembles production synchrony. The processes of data collection and coding regarding the corpus created for these studies will conclude this chapter.

The Conceptual Affiliation Study on the semiotic affiliation between speech and gestures will be presented in Chapter 6, testing the lexical versus ideational connection between speech and gestures. The conceptual affiliation between the two modalities will be explored by having participants choose those parts of utterances they believe to be semantically correlating with the concurrent gesture phrases. A semantic connection between speech and gestures beyond gesture strokes and select lexical items will be proposed.

Proceeding from semantic to temporal synchrony, Chapters 7 and 8 will explore the listeners' perception of spontaneous speech-gesture utterances in their original synchronies as well as with either modality preceding the other in temporal steps based on the research discussed in Chapter 4. Physical cause-and-effect stimuli will provide the baseline for the speech-gesture stimuli. The Perceptual Judgment Task (Chapter 7) will explore which degrees of (a)synchrony are perceived as more natural by the listeners. The Preference Task (Chapter 8) will then approach the perception of speech and co-produced gestures from the production side. By combining the results of the two tasks, statements about the preferred as well as the acceptable temporal windows of AVI for speech and gestures will be made.

The implications of the windows of AVI for speech-gesture utterance will be embedded into the context of conceptual transmission in Chapter 9. The transmission cycle from GP to SP will be modified on the basis of the results of the three sets of studies and then be expanded into a working model of this cycle. Finally, possible implications of the SP hypothesis for gesture theory and other areas of research will be discussed after readdressing the central hypotheses.

2 Theories of Speech-Gesture Production

2.1 Introduction

During the onset of speech-gesture research, Efron (1941/1972) studied the cultural foundations of facial and manual expressions in Eastern Jews and Southern Italians residing in New York City – two population groups well known for their multitude of conventionalized gestures. His observations opened up a field of language studying concerned with more than speech that soon expanded beyond emblems. Efron put an explicit focus on hand gestures, and also included head and trunk movements, but not facial expressions, posture or gaze. He introduced the categorical description of gestures, cartographing features such as motion radius, form, hand shape and position, involved body parts, tempo as well as linguistic, referential properties. Kendon (1967, p. 57), who throughout his career has continuously been concerned with Sicilian gesticulations (e.g., 1995; 2004), also attributed "a somewhat context-independent meaning (as shaking the first is a gesture of anger)" to gestures (p. 57). Shortly after, Ekman and Friesen (1969) further expanded on the categorization of gestures, particularly on speech-accompanying gesticulations without codified meaning, terming them *illustrators*, what Efron had (1941/1972) considered *physiographic* hand gestures.

Expanding on the explorations of his predecessors, and following the semiotic model proposed by de Saussure (1972/1983), McNeill (1985) ascribed a signifying function to gestures similar to that of speech (p. 352; also Schegloff, 1984). Expanding on the seminal work by Efron (1941/1972) and Kendon (1967; 1985), McNeill (1985) aimed at demonstrating the immediate interconnectedness of speech and accompanying gestures as arising from the same psychological plan and sharing computational space. The interval in production where speech and gesture temporally and semiotically overlap the most he deemed the unpacking of a Growth Point (e.g., McNeill & Duncan, 2000, discussed further in Chapter 3.2), a phenomenon widely represented in gesture research up to today. McNeill (1985) began his endeavor by expanding on the categorization of gestures proposed by his precursors. First, he singled out emblems from speech-accompanying gestures as gesticulations potentially independent from speech but depending on social constructs (p. 351; cf. Kendon, 1967). The close interconnectedness of gestures with speech is further evident in the distinction McNeill (1985) made between beats and conduit gestures, both taking over meta-narrative, or rather extra-narrative functions. While the former are used by speakers to emphasize words or features, for example in political speeches, the latter can bridge between utterances or speech units. McNeill (1985) demonstrated further parallels between the linguistic and gestural domain, namely between iconic gestures and onomatopoeia; another level of this are metaphoric gestures. These *iconix* stand in a direct propositional relation to speech, unpacking with the utterance to complete a sign. More detailed categorizations that are still commonly used, subsumed under the *Kendon Continua* of gestures, go from gesticulations to sign languages and, including deictic, beat, discourse, emblematic, iconic, metaphoric, and path gestures, were further established by McNeill in 1992 (also 2005; see Terminology).

Following decades of research in various areas of speech-gesture communication, the anthology *Language and Gesture* (McNeill, 2000) offers an encompassing snapshot of major issues that are prevalent to date. The constantly expanding field of speech-gesture research, regularly intersecting with other research areas, can be distributed roughly into the following thematic groups:

The **communicative function** of co-produced speech and gestures, which includes topics such as gestures as discourse markers, for example for grounding, alignment, floor-distribution, and perlocution, etc. (e.g., Krauss et al., 1991; de Ruiter, 2000; Alibali, Heath & Myers, 2001; Melinger & Levelt, 2004; Holler, Shovelton & Beattie, 2009). A large sub-field of this is concerned with **sign languages**, for example regarding non-lexicalized gestures in those languages (e.g., Stokoe 1960/2000; Hoiting & Slobin, 2007).

The **co-production** of speech and gestures includes the general issue of **production synchrony** as well as the functional interaction of the two modalities in the areas of, for example, speech facilitation, lexical access, and thinking-for-speaking (e.g., Krauss et al., 2000; McNeill, 1985; McNeill & Duncan, 2000; Kita &

Özyürek, 2003). Further sub-fields are concerned with language development in mono- and multilingual contexts as well as with the role of gesture production in educational settings (e.g., Sekine, Stam, Yoshioka, Tellier & Capirci, 2015; Goldin-Meadow & Alibali, 2013). That **primates** and other **non-humans** also use gestures is another recurring topic (e.g., Pika, Liebal, Call & Tomasello, 2005).

Several researchers have been engaged with **modeling** speech-gesture production, coding and implementing systems of speech-gesture interplay to understand the production process better (e.g., de Ruiter, 1998; 2000; 2007; Krauss et al., 2000; de Ruiter, Bangerter & Dings, 2012; Kita & Özyürek, 2003; Neff, Kipp, Albrecht & Seidel, 2008; Bergmann, Kahl & Kopp, 2014). Some researchers within this field are concerned with the construction of grammars of gestures (e.g., Kok, Bergmann, Cienki & Kopp, 2016; Rossini & Gibbon, 2011; Hassemer, Joue, Willmes, Mittelberg, 2011; Fricke, 2012; 2008; Gibbon, Hell, Looks & Trippel, 2003) and with facilitating the programming of speech-accompanying gestures into robots or virtual agents (e.g., Wheatland, Wang, Song, Neff, Zordan & Jörg, 2015; Srinivasan, Bethel & Murphy, 2014; Sowa, Kopp, Duncan, McNeill, & Wachsmuth (2008).

Research on the **neurological** mechanisms behind speech and gestures has often been closely intertwined with that on the production process. Key aspects include the connection of gestures with, for example, cognition, emotions, and clinical linguistics (e.g., Trofatter, Kontra, Beilock & Goldin-Meadow, 2015; de Ruiter & De Beer, 2013; Hogrefe, Ziegler, Wiesmayer, Weidinger & Goldenberg, 2013; Kipp & Martin, 2009; Ekman, 1992).

The **comprehension** of speech and gestures then is a natural counterpart to their production. Next to overlaps with the general communicative function, research foci include, for example, information-uptake from gestures, particularly in instructional situations (e.g., Goldin-Meadow et al., 1999; Gullberg & Kita, 2009; Nobe, Hayamizu, Hasegawa & Takahashi, 2000).

What nearly all of the above-mentioned research areas presuppose, especially that of comprehension, is the perception and integration of the co-produced speech and gestures. They deal with the synchrony of speech or other communicative signals and gestures in production in one way or the other, yet the perception of the multimodal utterances has only been addressed by few (e.g., Habets et al., 2011; Özyürek et al., 2007; Gullberg & Holmqvist, 1999; 2006).

This dissertation is aimed at expanding the research of speech-gesture perception. From the numerous areas of research listed above, a selection relevant to the later analysis of speech-gesture perception will be discussed in more detail in the remainder of this chapter. To what degree multimodal utterances have a communicative capacity has been under discussion over the years in the gesture research community. An overview of various standpoints on this topic will be given in Chapter 2.2, concluding that speech-gesture utterances are indeed communicative via both modalities. Chapter 2.3 will then focus on the co-production of speech and gestures from a temporal point of view. The temporal overlap between gestures and certain parts of verbal utterances has, for example, inspired research on whether gestures play a role in lexical access (Chapter 3.1). McNeill's GP theory also attempts at explaining the temporal connection between speech and gestures before and during their co-utterance, bridging the gap between the planning and execution phases of multimodal utterances; the GP theory will be discussed in detail in Chapter 3.2 as a central theoretical concept for the development of the SP hypothesis.

Others have suspected more concrete lexical affiliations between co-produced speech and gestures, that is in a many-to-many relationship between lexical items and gestures. Various viewpoints on this will be discussed in Chapter 3.3, giving reasons for experimentally exploring whether a conceptual rather than a lexical relationship between the modalities is reasonable to assume; the empirical study resolving this dichotomy will be presented in Chapter 6. Combining the differing investigative angles on speech-gesture production, several researchers have proposed models formalizing and explaining the interplay between speech and gestures in production (Chapter 3.4). Analyzing some of these models will provide insights into the speakers', and hence the listeners', language systems, allowing a glimpse of how multimodal messages will be perceived and then processed toward comprehension.

2.2 Communicative Function

While it is generally agreed that gestures are communicative (e.g., Mead, 1938), the questions of in how far and for whom is still debated. While there is plenty of research supporting gestures' benefit for the speaker and listener studying the processes underlying speech-gesture *production* or *comprehension* will naturally have a focus on either interlocutor. This dissertation is concerned with the *perception* of co-expressed speech and gestures because the author believes in an exchange of information between speakers and listeners via both modalities for reasons that will be expanded on in the remainder of this chapter.

At times, non-codified speech-accompanying gestures were regarded by some, for example by Feyereisen and Seron (1982) or Butterworth and Hadar (1989) as mere byproducts of speech production. What has been ascertained by now is that speakers cannot help but gesture when speaking, even over the phone, albeit with a lower word-gesture ratio (Bavelas, Gerwing, Sutton & Prevost, 2008; Butterworth, Hine & Brady, 1977; cf. Alibali et al., 2001). This has also been observed for monolog speech (e.g., Butterworth & Beattie, 1978; Beattie & Aboudan, 1994): Every speaker will have an addressee in mind, even if it is themselves (cf. McNeill, 2000, pp. 23f.), making every instance of speech inherently communicative, and potentially gesticulatory.

As has been outlined above, semiotic properties have been assigned to different kinds of gestures (e.g., by Argyle, 1975; Schegloff, 1984; McNeill, 1985). These properties were discovered by studying speech-gesture production and then categorized by only a small number of observers who subjectively interpreted them. Since no communicative intent was recorded, the agreement of the gesture interpretation by the observers with what the gesturing speaker intended to communicate did not factor in the determination of the gestures' meanings. And this is a general crux with spontaneously produced communication that may never be resolved conclusively: Whether speakers intentionally said X or gestured Y (cf.

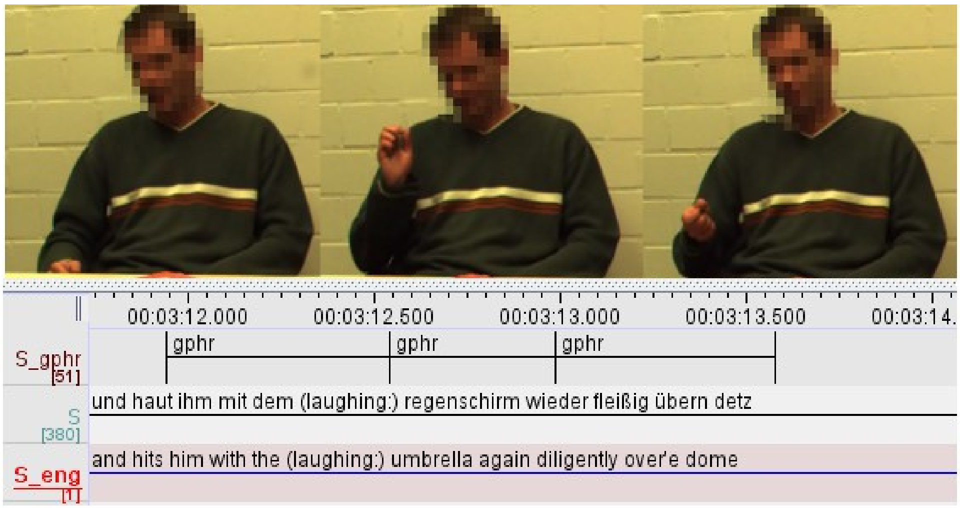
Melinger & Levelt, 2004). The only cases where one can be certain of the gestural communicative intention of the speaker are those (beat) gestures used for emphasis, that is, on a meta-narrative level, or with those deictic gestures made to indicate positions or directions not uttered in speech[2] - they are produced consciously and strategically to complement the verbal utterance. The communicative properties of iconic gestures have been extensively addressed by de Ruiter (2003), who argues that “a) gesture is a communicative device, and b) gesture and speech are mutually compensating for difficulties in the other channel” (p. 340). To determine the amount of communicative benefit from any speech-accompanying gesture, the information uptake in both speaker and listener would have to be, and has been measured by, for example, Gullberg and Kita (2009; see Chapter 4.2 on comprehension). An account of the communicative potential of co-expressed speech and gestures will be given in the following as it is prerequisite for researching the communicative gains from gestures.

2.3 Production Synchrony

The phenomenon of temporal overlaps between speech and gestures, of partial synchrony, has given rise to many hypotheses on the semantic synchrony of speech and gestures (e.g., McNeill; Kendon; v.s.). The interplay between temporal and semantic synchrony has lead to the assumption that listeners require production synchrony between speech and gestures to achieve the largest possible information uptake. Not only the multimodal production synchrony, but also the general information gain from speech-gesture utterances is a crucial issue in the analysis of whether and how listeners perceive and integrate (a)synchronies between co-produced speech and gestures. Chapter 4.2 will address this as well as other factors influencing the comprehension of speech and gestures, providing the psycholinguistic foundations for the studies presented in Chapters 6, 7 and 8. Beforehand, the nature of temporal synchrony in production between speech and gestures will be discussed in the following.

[2] In this context, emblems are a special case. They are codified, lexicalized gestures and will be used deliberately by a speaker to communicate, regardless of presence of speech.

Kendon (2004) and others (e.g., McNeill, 2005; Gebre, Wittenburg & Lenkiewicz, 2012) have divided gesture phrases into several parts to facilitate a more precise analysis of their timing and meaning. Conventionally, gestures will start and end at a resting position or transition point that frames a gesture phrase (gphr). The motion or set of motions in between will consist of phases, much like the syllable structure of onset, nucleus and coda (see also McNeill, 2005, pp. 30ff.): The onset, or *preparation phase*, will bring the hand(s), and possibly the arms, to the position where the core gesture is to be executed, for example by raising a hand in order to "slam" it down to support a point made in speech. The stroke then, in this case the slamming motion, is deemed the meaningful part of the gesture. Depending on the researcher and interpretation, each *stroke* will receive singular attention, others take repetitive strokes to be part of one stroke phase. In the corpus created as part of this dissertation, repetitive strokes were treated as singular gphr because the aim was to detect prominent gestures in general. An instance of this is the triple hitting motion accompanying S''s speech in (gphr 801-803), the first gphr of which can be seen in Figure 2:



und haut ihm mit dem (laughing:) regenschirm wieder fleißig übern detz

and hits him with the (laughing:) umbrella again diligently over'e dome

Figure 2: The displayed stroke is repeated three times within each gphr..

Concluding the gesture or series of gestures will be the *retraction* phase, in which the participating body parts will either go back to their resting position or transfer into another preparation phase; it is possible that the retraction phase is skipped in cases of immediate stroke-preparation transitions. After having identified the different gesture phases, Kendon (2004) reports the following on how to determine how speech and gestures synchronize:

> 3 The principal feature in this organization that we noted is how what is distinguished as the *stroke* of the gesture phrase is performed in close temporal proximity to that part of the associated tone unit that expresses something that can be regarded as semantically coherent with it. The *nucleus* of the gesture phrase, that is, the stroke and any hold that may follow it, tends to be performed in such a way that it is done at the same time, or nearly at the same time as the pronunciation of the word or word cluster that constitutes the nucleus, in a semantic sense, of the spoken phrase. This means that, by coordinating temporally the nucleus of the gesture phrase (i.e., the stroke and any post-stroke hold) with the semantic nucleus of the spoken expression, the speaker achieves a *conjunction* of two different modes of expression. . . . The precise way in which a coincidence is achieved between a gesture phrase and that part of the tone unit to which it is related semantically appears to be variable. (pp. 124f.)

Applying the phase-structure during the analysis of speech-gesture utterances revealed that gesture strokes usually preceded or ended at the prosodic peak of an utterance, at the sentence stress (Kendon, 1972; 1980). In the corpus used in this dissertation (see Chapter 5.3), these findings were confirmed for a correlation between stroke onset and speech intensity (green), but not for pitch accent (F0 contour, blue), as can be seen, for instance, in the PRAAT visualization shown in Figure 3 (see also Figure 5). The subjective perception of the prosodic peak aligns with the assumption of the temporal co-occurrence, though:

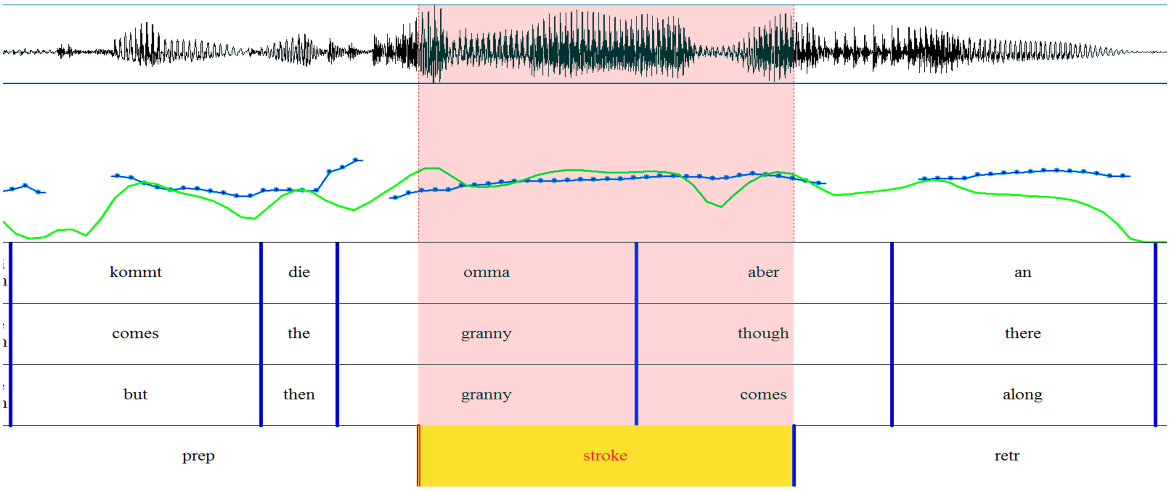


Figure 3: PRAAT analysis of stroke-pitch accent correlation (gphr 129).

Building on the temporal synchrony of prosodic peaks and gesture strokes, the interval during which gestures support speech in meaning was expanded to a time span “synchronized with linguistic units” (McNeill, 1985, p. 351). Regardless, research has often focused solely on the rather restrictive interval during utterances where peak and stroke coincide to look for a semantic connection, particularly in the context of perceptual analyses (e.g., Habets et al., 2011; McClave, 1994; Morrel-Samuels & Krauss, 1992). Kendon (2004), for instance, put an emphasis on semantic coherence, noting that temporal coincidence between the two modalities “appears to be variable” (p. 126). Further research showed that a gesture stroke usually does not follow the stressed syllable in speech (McNeill, 1985); Nobe (1996) added that already the gesture onset can precede the sentence stress, which again supported the “phonological synchrony rule”, as it has been called by McNeill (1992, referred to in de Ruiter, 1998, p. 29), giving more weight to the semantic substance at the point of peak-stroke synchrony. While the focus in the analysis of speech-gesture utterances has been broadened from punctual synchrony to a wider temporal span, the rather restrictive idea of 'lexical affiliation' (de Ruiter, 2000) still prevails: Chapter 3.3 will discuss lexical as well as other views on the the semantic and temporal affiliation of speech and gestures in more detail.

From the collection of opinions and findings on the temporal synchrony between speech and gestures summarized above it becomes apparent that there is no unified understanding of which parts of the co-produced speech and gesture are to be synchronous in production, that is, prosodic peak and gesture stroke onset, the whole stroke phase and semantically affiliated speech, or whole speech and gesture phrases. One reason for this might be that there is no consensus among the gesture community on what synchrony is exactly, that is, whether and which verbal and manual parts of an utterance have to synchronize from start to end, or whether a verbal utterance is rather a temporal container into which the gestural phrase is embedded. De Ruiter (1998) states on adding a temporal factor to his Sketch Model (see Chapter 3.4) of speech-gesture production that

> [f]irst of all, synchronization should be defined in such a way that it is possible to locate the affiliate of any iconic gesture unambiguously. Second, synchronization should be defined carefully. (p. 19)

Particularly the definition of synchrony is highly relevant for a model that includes utterance production as well as perception such as the GP-SP transmission cycle (see Figure 1) to be developed within this dissertation. Not only is the timing of the modalities relevant for comprehension, but divergences from certain degrees of asynchrony can potentially result in a breakdown of AVI (Massaro et al., 1996) and cause failures in communication.

How different events can be temporally related has been explored, among others, by Golani (1976) in the context of animals' limb coordination, and most prominently by Allen (1983), who chose a more encompassing approach (see also Gibbon, 2009). Both Golani and Allen proposed a collection of interval-based temporal relationships between two events, noticing that intuitively, succeeding events often do not do so with exact start-end fixation points but rather overlap to a certain degree. Golani (1976) put forward a set of 13 possible temporal relations between two limb movements (p. 87). In an unrelated "attempt to characterize the inferences about time that appear to be made. . . during a dialogue" (p. 834), Allen (1983) formulated an algebraic calculus based on temporal relations. His model of an interval-based temporal logic that should be expressive as well as computationally effective also contains 13 theoretically possible temporal relations between two intervals and is applicable to a wide range of scenarios, reaching from language production over economic processes up to historical scales.

According to Butterworth and Hadar (1989), who at this point refer to Golani (1976), "[o]f these 13 relations, 9 would satisfy McNeill's (1985) rather minimal condition of temporal overlap. . .", regardless of the onsets and offsets of gphr (p. 170). Readdressing the issue, Hadar and Butterworth (1997; see also de Ruiter, 1998) suggest that those of the relations that involve absolute synchrony of the onsets of speech and gestures are highly improbable and thus can also be neglected. In the case of speech-gesture utterances, the temporal overlap can be regarded on the phase level, but also in more detail, for example the temporal relations of the stroke phase of the gesture and certain lexical items, or of gesture apex, that is, the climax of the stroke, and prosodic peak. Taking these restrictions into account, Thies (2003) lists the following six possible temporal relations between speech (S) and gesture (G) intervals, that is, verbal utterances and gesture phrases (gphr), which are easily transferable onto any annotation system using tiers (Figure 4; the numbers reference the enumeration by Allen (1983):

> 6: G contains S, hence also anticipates S;
> 7: S contains G, hence precedes G;
> 8: S overlaps G;
> 9: G overlaps S;
> 12: S occurs before G, that is, S and G are temporally disjunct;
> 13: G occurs before S, that is, the G is finished before S starts. (pp. 53f.)

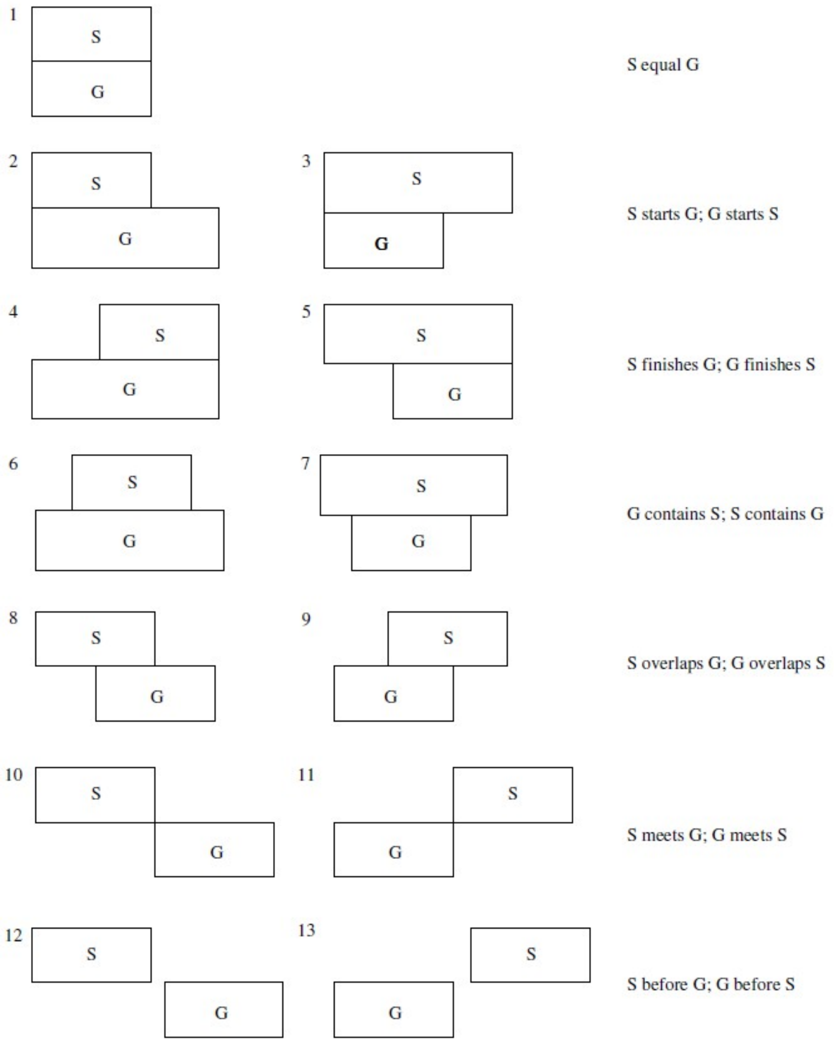


Figure 4: Temporal relations of speech and gestures based on Allen (1983; Thies, 2003).

It is important to note here that, counter to common observations of production, speech can also precede a gesture (12). Naturally, there will be speech before, during and after speech-accompanying gestures, but this fact is often neglected because focus is put on the analysis of synchronously produced signals; this will be discussed in more detail in the context of studies on the perception of speech-gesture synchrony in Chapter 4.4. In Figure 3, for instance, when considering the stroke phase as an instance of a gesture interval, and the interval of highest intensity in speech, the temporal relation of S and G would, at first glance, fall under category (2) S starts G. Since “locating the beginning and end of gestures (even if restricted to the stroke) is often problematic” (de Ruiter, 1998, p. 19), however, and the interval can be broken down into more detailed levels such as syllables and gesture phases, the peak in the intensity of the speech is rather cradled by the gphr (6).

As has been mentioned above, the subjective understanding of speech-gesture synchrony in the literature is manifold, and it has to be specified to be used as a factor in a model of speech-gesture production (and reception). Such a model should be defined widely enough to explain any occurrence of speech-gesture co-production as well as the different assumptions of affiliation between the modalities. This is only feasible when including semantic as well as temporal factors. Intervals of overlap will, on a higher level, be treated as co-*produced*, the full multimodal utterance as co-*expressive*. Within the scope of this dissertation, 'speech' and 'gesture' as used from hereon will include the following:

Speech:

- sentential units governed by a theme-rheme structure (see Chapter 6.5);
- within these units: intervals terminated voluntarily, for example through repairs/rephrasing or self-interruptions, or involuntarily, for example tip-of-the-tongue (ToT) states, interruptions by the listener, or outside events.

Gesture:

- gphrs without instant repetitions, not taking into consideration superimposed beats;
- within these phrases: intervals terminated voluntarily, for example through repairs/rephrasing or self-interruption, or involuntarily, for example self--adaptors, interruptions by the listener, or outside events.

Following these definitions, the experiments on the perception of speech-gesture asynchronies will not use stimuli desynchronized from either prosodic peak or gesture stroke as anchor points but shift the modalities in relation to the whole utterance. This way, asynchronies will be comparable across stimuli and naturally occurring temporal overlaps will be reconstructable.

The hypothesis that a semantic connection between gestures and speech would already exist pre-utterance and the observation that parts of the gesture will precede certain parts of speech in production has also inspired some to suspect a speech-facilitating function of gestures. An overview of research on this topic will be given in the following Chapter 3.1, while Chapter 3.2 will expand on the GP theory, which encompasses aspects regarding temporal as well as semiotic synchrony.

3.1 **Lexical Access**

Iconic gestures in particular overlap with the speech they are co-produced with semantically as well as temporally (v.s.). The fact that the onset of the gesture stroke often precedes that of the most strongly semantically relatable parts of speech has been taken by some to indicate a facilitatory function of the gesture toward the speech (e.g., Butterworth & Beattie, 1978; Morrel-Samuels & Krauss, 1992). At times, speakers gesture *instead* of speaking, for example they use emblems like “thumbs-up” or they “gesture” toward somebody to speed up their argument or walk. These kinds of gestures are produced deliberately to communicate something to the addressee. As emblems, they are culturally specific, non-verbal signals that are comprehensible without disambiguating speech. Another type of gesture that occurs without speech can be observed when more or less fluent speech is crucially disrupted from the speaker's side, for example through ToT states (see, e.g., Beattie & Coughlan, 1999), and speakers signal to their interlocutor that they are searching for a word or at least wanting to hold the floor. In this case, the gesture would be discourse-regulating to a certain degree, but it might at the same time be narrative, that is, when it semantically coincides with the word the speaker is looking for. In this case, the function of the gesture would be 'layered', which “means that single gestures convey content on the discourse and narrative levels simultaneously” (McNeill, 2005, p. 172). According to Cassell and McNeill (1991) and McNeill (1992), layering in (a series of) single complex gestures has three sub-categories, that is, a *paranarrative*, a *metanarrative*, and a *narrative* one (cf. McNeill, 2005, pp. 172f.). When the speaker is spiraling their flat hand like winding yarn up a spool with their extended fingers, this gesture is considered by some to have a meta-narrative function (e.g., Chen, 2002; Beattie & Coughlan, 1999; McNeill, 2005); pointing at the listener, for instance, would include a paranarrative function of the gesture. The sub-category of narrative gestures encompasses mostly iconic gestures that might also be co-produced with the speech they relate to but can also be produced instead of speech, that is, gestural counterparts to the idea a speaker wants to express. These kinds of gestures are believed by some to facilitate lexical access (e.g., Butterworth & Hadar, 1989; Hadar & Butterworth, 1997; Morrel-Samuels & Krauss, 1992). In the following, an overview of experimental approaches toward testing this hypothesis will be given, bridging the gap from temporal production synchrony towards the theory of lexical affiliation between speech and gestures (Chapter 3.3).

Butterworth and Beattie (1978) pioneered with a collection of studies on the possible speech-facilitating functions of gestures. Their methodology and results have been replicated and probed by themselves and others many times, most prominently by Beattie and Coughlan (1999), the latter with a focus on iconic gestures. Butterworth and Beattie (1978) observed that delays in speech production indicate planning processes such as lexical selection (e.g., Goldman-Eisler, 1958; cf. Gahl, Garnsey, Fisher & Matzen, 2006), and that speech-focused movements of the hands and arms are rhythmically, and often also semantically, timed with speech. They hypothesized that if gestures were involved in speech planning processes, they should be affected by speech delays. Referring to Henderson, Goldman-Eisler, and Sharbek (1966), who noticed rhythmical differences between the planning and execution phases in speech, Butterworth and Beattie (1978) conducted two experiments analyzing dialogical and monolog speech with regard to these differing phases. In a third experiment, they connected the results of the first two experiments with the originally co-produced gestures recorded during speech elicitation. The methodology of the speech-only experiments will be summarized below to form a basis for a broader discussion of the third experiment and of relation between gestures and lexical planning according to Butterworth and Beattie (1978).

The authors recorded dyadic conversational arguments of strong, speaker-picked propositions for experiment 1. Analyzing the recordings, they hypothesized that temporal cycles of hesitant and fluent speech should coincide with initiations of “well-understood linguistic unit[s]” like sentences or clauses (Butterworth & Beattie, 1978, p. 349). This was be confirmed for 32 out of 42 'cycle transitions' by means of a pen-oscillograph analysis. To avoid preexisting constraints such as clauses in the further analysis, eight judges then divided the speech transcripts into ideas, or ideational units, with a 50% agreement quota (p. 350). These ideas coincided with clause boundaries most of the time. Butterworth and Beattie interpreted this to be indicative of a cognitive rhythm of idea planning and execution “which will be realized linguistically as several (surface) clauses” (p. 350). For experiment 2, the participants were instructed to give monological descriptions of loosely connected things, for example of five rooms (low cohesion condition) as well as to describe relations between parts of objects or event sequences, for example consecutive actions of a single male at the discotheque (high cohesion condition). In both conditions, the participants were instructed not to mention the direct connections between these things. The pen-oscillograph output showed fewer pauses at the start and end of idea units and most pauses between idea units, among other things. Butterworth and Beattie (1978) tentatively hypothesized these findings to be indicative of lexical or idea planning, or even of higher idea planning.

Using a monologue from experiment 2 and three additional recordings from dyadic academic conversations, Butterworth and Beattie (1978) expanded their analysis to include various types of hand and arm gestures in experiment 3. The authors identified the following types of gestures (p. 352):

(1) SFMs – "speech-focussed movements": hands and arms, including beats, gestures, non-gestures, etc., except self-adaptors;

(2) gestures[3] (sic!): more complex movements with semantic relation to speech components;

(3) changes in the resting position ("equilibrium position").

Inter-coder agreement was achieved by jointly rechecking the data. Following the hypothesis of idea expression, "the exact time between the initiation of the gesture and the first phone of the word with which it was associated was noted" (Butterworth & Beattie, 1978, p. 352) by the coders. This assumption about a 1-1 word-gesture relationship has been criticized by McNeill (1989), among others, not only for the variation in duration and overlap between the modalities outside word boundaries. An alternative approach to the semantic relation between speech and gestures will be discussed in Chapters 3.3 and 6.

Butterworth and Beattie (1978) found that SFMs (1) were about three times more frequent in pauses during execution phases, that is, descriptions, than during planning phases, that is, Introductions, across speakers; gestures (2) were about five times more frequent in pauses during execution phases than during planning phases. Additionally, gestures were about three times more frequent in pauses during execution phases than in phonation periods of the execution phases across speakers. For the residual class of "SFM-gestures"[4], no difference in frequency was found between the pauses in the execution or planning phases; they were most common during the phonation periods of the planning phases. The distribution of SFMs, which may or may not include type (2) gestures, over the initial or concluding phases of ideas was not consistent across the six fluent participants (p. 354), giving the other observations a tentative flavor. The analysis of variance

[3] During the course of their paper, Butterworth and Beattie (1978) switch between treating type (2) as standalone and as a sub-type of (1) (cf. p. 354f.).

[4] This category is not further explained by the authors, but it can be assumed that it includes self-adaptors and other hand and arm movements not fitting in (1), (2) or (3).

(ANOVA) with phase as factor reported SFMs to be “more frequent per [idea] unit time” during the execution phases than during the planning phases of the recorded utterances. It further revealed a significant main effect of speech fluency on SFMs: SFMs were more frequent during phonation in the planning phases, while they were more frequent during hesitations within execution phases. Additionally, significantly fewer SFMs in speaker-change pauses were observed than in planning or execution phase pauses (p. 355).

From the distribution of non-beat gestures across utterance phases, Butterworth and Beattie (1978) assumed a functional relation of these SFMs to lexical planning but not to idea planning (p. 355). This assumption was corroborated by the authors through the lexicosyntactic classes associated with these gestures (more on the subject in Chapter 3.3), that is, nouns (41.3%), verbs (23.8%), and adjectives (15.9%). The authors related this distribution to the number of unpredictable lexical items in these word classes. That any kind of semiotic co-signal will most likely be related to content words rather than to function words was not discussed (cf. Kirchhof, 2010, Chapter 2.2.4; also Lutzeier, 2006, p. 80, on lexicosemantic fields). Another argument for the relation of gestures to utterance planning put forward by Butterworth and Beattie (1978) was the timing of gesture onsets in that “the initiation of gestures usually precedes, and never follows, the words they are associated with. The mean delay being around .80 secs., with a range of .10 secs. to 3.5 secs. . . .” (p. 355; cf. Chapter 2.3). Butterworth and Beattie (1978) hypothesized that the temporal delay in production between speech and gestures might be explained by the differing sizes of their respective lexicons – lexical items existing in a far larger number than hand or arm configurations. In relation to this, they referred to McNeill (1975), who described gestures as a “semiotic extension” of what Butterworth and Beattie (1978) termed ideational units (p. 359), which goes in line with the conceptual substance of content words and the gestures coinciding with these. The authors concluded that lexical selection might not be part of ideational planning. They related lexical selection to gestures while recognizing this to not be a “sufficient condition for the occurrence of [g]estures” (p. 358). While the results discussed and conclusions made by Butterworth and Beattie (1978) are mostly tentative, their experimental methodology showed unique insights into speech-accompanying gestures and their relation to lexical access. Butterworth and Hadar (1989), for instance, used some of the results to develop a computational model of speech production (cf. Chapter 3.4).

Beattie and Coughlan (1999) partially replicated Butterworth and Beattie's (1978) experiments. Drawing on the findings by Goldman-Eisler (1968) on the temporal continuum of lexical access from spontaneous to well-rehearsed speech, they further analyzed how gestures might be connected with lexical retrieval. Beattie and Coughlan (1999) suspected gestures to be involved in lexical retrieval due to their temporal occurrence alongside speech in relation to word familiarity. They referred to Butterworth and Beattie (1978) and Butterworth and Hadar (1989; cf. Chapter 3.4), who also found that gesture onset precedes the onset of the semantically related speech segments in production. This temporal relation as well as, for example, observations of “a compensatory increase in the frequency of gestures per word in aphasic patients” (Feyereisen, 1983; in Beattie & Coughlan, 1999, p. 37) lead Beattie and Coughlan (1999) to test the influence of gestures on resolving induced ToT states. Their aim was to “test experimentally the Butterworth & Hadar theory that iconic gestures have a functional role in word retrieval”, using a more informed and focused approach than the pioneer experiment by Butterworth and Beattie (1978). They conducted a study in two parts by investigating (1) iconic gestures in connection with single word retrieval of unpredictable lexical items, and (2) the relation of gesturing in general toward resolving ToT states.

(1) was tested by reading out definitions of high imageability target words to participants, that is, of words that are likely to evoke a rather extensive image, to elicit target words after inducing ToT states. After a certain period of time without resolution, participants were presented with a cue. While participants free to gesture resolved 66.8% of ToT states, participants bound from gesturing by folding their arms was 72.4%, a difference failing to reach significance (Butterworth & Beattie, 1978, p. 46). The total number of resolved ToT states was higher with the gesturing group, though (p. 46). These results did not give significant support for gestures' facilitatory functions during lexical access or for the resolution of ToT

states, also because the study participants who did gesture also encountered more ToT states.

Part (2) of the study by Beattie and Coughlan (1999) intended to analyze the connection between iconic gestures and resolving ToT states. To determine the degree of lexical relation between the target words and their co-produced or preceding gestures, the authors showed recordings of these speech-gesture utterances to 18 judges, with the speech muted. The judges then had to select the words the participants were searching for from a list in which the original target word was included. They had an inter-rater agreement of 87.8%, and a “Chi-square analysis on the correct and incorrect scores revealed that the judges performed significantly better than chance (χ^2 (1) = 80.49, $p < .005$)” (p. 43). A major fallacy of this assessment of lexical affiliation between speech and gestures is that iconic gestures need speech to disambiguate their meaning. Hadar and Butterworth (1997) comment on the sentential ambiguity of iconic gestures issue as follows:

> The meaning of an iconic gesture is typically vague in itself. Whilst iconic gestures often have recognizable physical features. . ., their meaning can seldom be derived from their form with any degree of certainty. . . .(p. 148)

Without context, the identification of the actually co-produced utterance is next to impossible. Spivey and Tanenhaus (1998), for instance, who examined the effects of referential context on ambiguity resolution, found that information provided by the previous discourse were used to resolve temporal ambiguities and to reduce processing difficulties. What then remains from the methodology of Beattie and Coughlan (1999) to have judges decide on the gestures' meanings is that speech and iconic gestures are semantically connected when analyzed by an observer. This makes a direct relation between imagetic gestures and word retrieval less probable. The specific problematics of this methodology will be further discussed in relation to the possible lexical affiliation of speech and gestures in Chapter 3.3 (cf. e.g., Morrel-Samuels & Krauss, 1992). As has been touched upon before, the lexicosemantic properties of non-iconic gestures are debatable. Another difficulty with the methodology of Beattie and Coughlan (1999) might be that occasionally “a combination of gestures occurred in . . . ToT state[s], that is, iconic gestures, beats and self-adaptors" (p. 45). Some take this to mean that gestures help with lexical access in non-fluent situations (e.g., Butterworth et al., 1977; Beattie & Coughlan, 1999), while others assume a broader context of bidirectional compensation between speech and gestures (e.g., de Ruiter et al., 2012).

What cannot be detected by the naive observer, who sees semantic, even lexical connections between certain gestures and speech segments, is which internal processes leading to their co-utterance. One hypothesis is that their lexicons are intertwined, that there is a lexical affiliation that leads to temporal alignment in production (e.g., Krauss et al., 1991). Others have proposed a broader affiliation of the modalities on an ideational, or conceptual level (e.g., Vygotsky, 1987; Kirchhof, 2011). The most prominent theory of how speech and gestures interact before and during utterance planning is the GP theory by McNeill (1985) on which the SP hypothesis will be largely based (Chapter 4.5). The following Chapter 3.2 will provide a detailed discussion of the GP theory and how it incorporates ideational units.

3.2 The Growth Point

McNeill (1985) proposes that a gesture as a "global-synthetic image can itself be regarded as the verbal plan at an early stage of development" (p. 367). The statement that "there is no system break between thinking and speaking" (p. 370) suggests a linear production process for speech, and that at some point there is a junction to gesture production. McNeill (1985) holds the proposition of a shared computational stage by reporting on the collective rise and fall of speech and gestures in the counter-directional processes of language acquisition and aphasia (pp. 362ff.). This linkage in regression, albeit in connection with idiopathic Parkinson's disease, is further investigated by, for example, Duncan, Galati, Goodrich, Ramig, and Brandabur (2004) and Duncan (2008; 2009).

McNeill (1985) draws from Vygotsky (1987) in that he presupposes a "minimal psychological unit" (MU) containing a perfect match of imagery and linguistic means in the speaker's mind that they want to express. Depending on the physiological and lexical constraints the speaker is under, including mechanisms of thinking-for-speaking, they will attempt to express the most explicit version of this MU. The ideational unit that contains this maximal content and how it can be expressed is termed “Growth Point” (GP) by McNeill (e.g., 1985) – from the point onwards and during the interval when speech and gesture interact the most, like a flower, the intended utterance will grow to full bloom. In the following, the construct of the GP will be determined in more detail. Chapter 5.2 then will formally connect the GP with the SP as its perceptual counterpart, developing a methodology of testing the connection between the two ideational units.

Historically, a variation of terminologies has been used to refer to the division of sentences into one more and one less informative part, often in the context of different theoretical frameworks. Two of the most prominent terminologies are those of *“psychological subject – psychological predicate* (von der Gabelentz 1869, Paul 1880) [and] *theme – rheme* (Ammann 1928: Thema-Rhema, Mathesius 1929, Prague School (Dane, Firbas), Halliday 1967b)” (von Heusinger, 1999, pp. 101f., emphases in the original).

> [V]on der Gabelentz (1869) . . . compared the sequence of thoughts or psychological concepts with the sequence of linguistic expressions in a sentence. He then distinguished two levels: the grammatical level and the psychological level of composition. Von der Gabelentz defines the psychological subject as “that about which the hearer should think”, and the psychological predicate as “that what he should think about”.
> (von Heusinger, 1999, p. 110)

Coming from these psychological contrasts, Paul (1880), and later Ammann (1928) transferred the psychological dichotomy to communication, re-terming it into theme and rheme. This distinction, then, is made with respect to topical aspects, that is “informational units are described as the part the sentence is about and the part what is said about it” (von Heusinger, 1999, p. 102), with a focus on grammatical structures. Categorizing parts of an utterance either as psychological subject and predicate or theme and rheme is not mutually exclusive, but rather varies in focus. Approaching language from a psychological viewpoint, Vygotsky (1978) applied the informational sentence dichotomy of psychological subject and psychological predicate to his concept of minimal psychological units. He gives the following example:

> [In the] sentence “The clock fell.” . . . [,] “the clock” is the subject, and “fell” is the predicate. . . . [T]his sentence uttered twice in a row . . . [,] the clock was already in my consciousness, the clock is the psychological subject, which the speech is about. The notion that the clock fell emerges second. In this case “fell” is the psychological predicate, that which is said about the subject. In this case the grammatical and psychological segmentation of the sentence coincide. . .. (p. 272)

Had the clock not been the topic of previous conversation, and the noise of the clock falling would be perceived, “The clock fell.” would still be uttered in relation to said event. However, “fell” would be the psychological subject, the essence of the utterance, “the clock” only taking secondary meaning.

In summary, the psychological predicate as proposed by Vygotsky (1978) is the newsworthy content of an utterance (cf. e.g., McNeill, 2015; 2005; Kirchhof, 2011), consisting of any number of lexical items. Factors differentiating the psychological predicate from its context are form and timing. Whether the utterance is speech only or composed of speech and gesture adds another level of timing, that is, when gestures will contribute their expressive features to the utterance. An example given by McNeill (2005) was taken from narrations of participants having watched Canary Row (Freleng, 1950), a series of cartoons starring Sylvester the cat and Tweetybird: In one scene, Sylvester tries to reach Tweety by sneaking up through a drain pipe attached to a multistory building. In a later scene, he chooses to climb up outside the windpipe (for more details on the context of elicitation see Chapter 5.3.1). Following Vygotsky (1978), the psychological predicates are the “drain pipe” in the first attempt, and “inside” in the second one, interiority versus exteriority representing the essential information distinguishing the latter from the former. Participants described in McNeill (2005) made this distinction not only in their verbal narrations, but also in their gestural expressions (pp. 109ff.). By means of gestures, the participants distinguished the newsworthy information from the context, emphasizing the change in the expressed psychological unit. One participant failed to make the distinction between “inside” and “outside” in their speech, and the gesture also failed to express this distinction. McNeill (2005) interpreted this to support the strong connection between MU, speech, and gesture because the psychological predicate of “interiority” was not present in the MU, so it was not expressed in either modality (cf. Chapter 3.4) and a GP containing the location of Sylvester on the drain pipe had not been not formed.

To test the GP hypothesis, Duncan, Parrill, and Loehr (2005) incepted specific GPs in the cartoon narrations (v.s.) by changing the order of Sylvester's drain pipe attempts at catching Tweety. This way, the newsworthiness within the narrations would change, and hence the psychological predicate. The authors discovered that when the “inside” clip was shown 15 clips before the “outside” clip, no gestures expressing this interiority were produced, but the participants still used expressions such as “inside” or “through”. When the “outside” clip was presented 15 clips before the “inside” clip, the participants did differentiate the location of Sylvester verbally as well as manually (cf. McNeill, 2005, p. 111). When no attempt involving the drain pipe was shown to the participants before the “inside” attempt, the psychological predicate would be the most distinguishing feature contrasting the current attempt from the previous attempt, for example the drain pipe itself. McNeill (2005) interpreted these findings to indicate that the psychological predicate of “interiority”, that is, the newsworthiness of Sylvester's methodology, was co-expressed by the participants through speech and gestures only when the pipe had already been a newsworthy item before, “making room” for a new one.

McNeill (2005) proposed that gesture and speech “choose” psychological predicates, adapting to processes and changes in discourse. The two modalities are timed and formed in such a way to best enable the differentiation of the predicate from the context. The GP then is an ideational unit containing imagery as well as linguistic encoding. It comes into existence through constant adaptation to discourse and context. McNeill (2005) describes this ideational unit on the grounds of Vygotsky's MU:

> By a unit we mean a product of analysis which, in distinction from elements, possesses all the basic properties of a whole. Further, these properties

> must be a living portion of the unified whole which cannot be broken down further. (McNeill, 2005, p. 9)

What is crucial to note here is that neither MUs nor GPs are a sum of their imagistic, lexical, and other parts, but rather their product. One might say that the GP is a specific variation, or sub-unit, of the MU, as it pertains to speech and not to writing – which would contain other possible means of physical expression. It is is the mental representation of imagery fused with linguistic competence (*langue*; de Saussure, 1972/1983). This mixture of modalities contains syntactic and categorical constraints onto which imagery has to be mapped. As one of the expressive means externalizing the GP, gesture embodies the imagistic part of the ideational unit. Gestures have, as McNeill (2005) proposes, global as well as synthetic properties which they bring into the GP: *global* in that they are holistic, similar to the rhetorical figure pars pro toto, expressing various features of an ideational unit at once; *synthetic* because they can express meaning that is otherwise spread across an utterance due to the iterative, syntactic structure of speech. And yet, the two semiotic modalities of speech and gesture embody the same idea within a GP. The imagistic/global (gesture) and syntactic/linear (speech) channels form a co-expressive dialectic. As McNeill (2005) writes, the GP is a somewhat unstable mixture of "inherently oppositional" semiotics and modalities (p. 18), changing their configuration depending on the immediate context. This leads to constant instability, to a dynamic, that ever adapts to context, intention, and other factors.

For now, the GP is an ideational unit, waiting so to speak for its expression at the right moment during an utterance. McNeill (2005) metaphorically deems the GP to be a package, containing imagistic, linguistic, and other parts of a potential utterance. When speech and gesture are co-produced, the GP is "unpacked" and the ideational unit is exposed. During the interval that speech and gesture are co-produced, they express the maximum of the ideational unit – leaving out one modality would express less of its contents or the utterance would take much longer to give the same information (cf. de Ruiter et al., 2012). Kendon (1988) comments on this that "we can have the impression of completeness of information without the gesture, even though the gesture does add to the total meaning of the utterance” (p. 135; see also Bavelas et al., 2008). Regulated through the intentions and intuitions of the speaker, “[a] surface linguistic form emerges that cradles the GP in stable and compatible form” (McNeill, Quaeghebeur & Duncan, 2008, p. 14). Syntactic constraints determine where the unpacking of the GP can initiate.

Thinking-for-speaking, or rather thinking-while-speaking correlates with the GP here because language competence cooperates with cognitive imagery in order to be communicative (McNeill & Duncan, 2000). As can be seen, for example, in a comparison of English and Chinese, different languages show different speech-gesture synchronies. From this it can be deduced that the GP formation process differs and, thus, “thinking”. It has to be noted, however, that languages allow for more than one way of unpacking the GP, depending, for example, on grammatical focus. Chinese, for instance, can have subject focus, and English can have topic focus, but this is usually resolved with syntax (McNeill & Duncan, 2000). How speech and gesture are timed, that is, where in the utterance the GP is unpacked, depends (1) on the psychological predicate of the utterance, and (2) on syntactical constraints. Figure 5 shows a transcription of the example already mentioned in Figure 3:

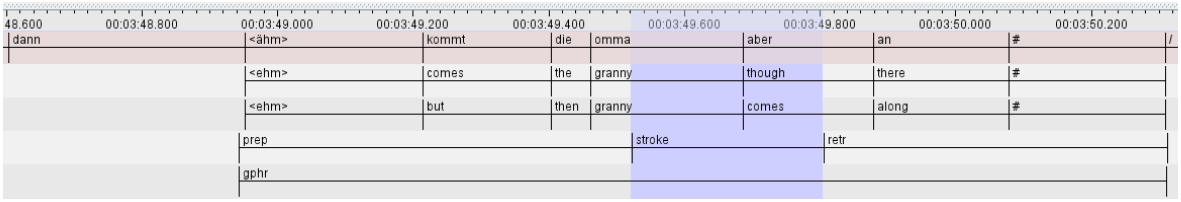


Figure 5: Example of a GP unpacking (gphr 129).

While we cannot be fully certain about what the current MU related by S' is, we know about the context of the utterance: S' describes the recurring theme of the granny hitting and chasing Sylvester with an umbrella, after various of Sylvester's attempts at catching Tweety have already been told, which is a context L and S' shared during the recording. Within the more or less immediate context of the utterance, Sylvester's attempts are already known and the actions of the granny are newsworthy, that is, “die omma” is the psychological predicate (1). The utterance is further governed by rules of German syntax (2) because “dann” needs to be followed by VS(O), and “aber” can only occur before or after the NP “die omma”. In fact, the gesture stroke, S' pretending to hit someone or something once with grabbed object held in the right hand, temporally overlaps with the verbal expression of the psychological predicate of the utterance. With or without the gesture, the utterance would express the newsworthy information of the granny appearing, but only through the gesture does L know that the granny holds the aforementioned umbrella in her hand and attacks Sylvester with it. As we are aware of just about what image S' had in mind by knowing the cartoon stimulus, we can draw conclusions from this image and the utterance produced as to what MU S' wanted to relate: a fused idea of various scenes from Canary Row. The GP then would be an information package containing the maximal content of the MU expressible through a combination of speech and gesture. Using gesture to express part of the action-containing information in combination with speech appears to be more efficient than either describing this in speech or gesture only.

As has been discussed with regard to this example, the onset of the gesture stroke phase does not coincide with the speech pitch accent but with a peak in speech intensity (Figure 3). Following the definition of synchrony agreed upon in Chapter 2.3, the gphr cradles the speech (temporal relation 6) on the utterance level while on the phase-level the stroke apex is framed by the word “omma” (temporal relation 7). It is important to note, however, that the complete multimodal utterance is required to fully express what S' intended to relate to L. While the overlap between speech and gesture during the stroke phase of the gesture can be regarded as an interval of co-production, the scope of the gesture is the whole verbal utterance, making it co-expressive throughout. For the hitting example (Figure 3), the *maximal* co-expressivity initiates with the beginning of gesture stroke, that is, when meaning of the utterance experiences a point of growth – the moment the GP “pops open”, and then fades out toward the end of the utterance; the general co-expressivity of speech and gestures lasts throughout the utterance of the psychological predicate.

In summary, McNeill's GP is an ideational unit involved in utterance planning, externalized through a semantically and temporally coordinated co-utterance of language and gesture from a certain point in time onwards: At the onset of the second modality, be it speech or gesture. The unpacking of the GP occurs, due to the nature of its expressive means, during an interval of multimodal overlap - the terminology of Growth *Point* is somewhat misleading here, since a point is usually tiny, and quite possibly of only brief temporal duration. As has been touched upon in Chapter 2.3 in the context of different ways of speech-gesture synchrony in production, “[t]here is no reason to assume. . . that these 'endpoints' are truly zero-width points rather than intervals small enough so that they appear to be instantaneous” (Allen, 1983, p. 841). McNeill (2011) comments on this that normally, “one GP-cycle [lasts] about 1~2 secs. Then the content changes and. . . information can be lost.” (Ch. 4.4.1). It has been established that the gestural part of the utterance usually precedes its co-expressive speech, but not always, depending on the perspective. Regardless of exact temporal coordination, “the time limit on growth point asynchrony is probably around 1~2 secs., this being the range of immediate attentional focus” (McNeill, 2012, p. 32). This estimation coincides with the average duration of multimodal utterances containing gphr in the corpus (Chapter 5.3.5), namely 0.9917 s (SD = .574) – the mere presence of a gesture in an utterance, regardless of its onset, might allow for an unpacking of a GP. Another temporal constraint for the unpacking of the GP would be that, according to Levelt, Schriefers, Vorberg, Meyer, Pechmann, and Havinga (1991), “if the gesture is physically delayed later than 300 ms. before the apex would normally have occurred, speech cannot adapt anymore. . . [because p]honological encoding has an estimated duration of around 300 ms” (cited in de Ruiter, 1998, p. 18). How asynchronies in production might influence the listener's comprehension would assume that production synchrony is actually noticed by them. Whether listeners perceive this synchrony as well as divergences from it will hence be experimentally investigated in Chapters 7 and 8.

The period of the unpacking of the GP, or the time during which speech and gesture overlap, is a phenomenon well noted by numerous researchers (e.g., McNeill, 1985; 1992; 2005; Kendon, 1979; 1980; 2004; Krauss et al., 2000). It is also one of the reasons lexical affiliation between the two modalities has been suspected by some. The direct lexical connection between one or more lexical items to a co-expressed gesture has had persistent usage in the research on gesture production and comprehension. It has not only been the basis for production models

(Chapter 3.4), but also for research undertaken in connection with gesture perception (Chapter 4.4). In the following, the roots and growth of this often presupposed lexical connection between speech and gestures will be discussed, laying the theoretical grounds for expanding on their semantic connection in the studies presented in Chapter 6.

3.3 **Lexical vs. Conceptual Affiliation**

Schegloff (1984) proposed that “various aspects of the talk appear to be 'sources' for gestures affiliated with them” (p. 273; v.s.), an idea relating to lexical access as well as to the co-expressivity of speech and gestures. McNeill (1992, pp. 37f.) gave a comprehensive summary of Schegloff's elaborations, describing a lexical affiliate as “the word or words deemed to correspond most closely to a gesture in meaning”; that is, that gesture and speech do not relate synonymous meaning but that they are co-expressive (see Chapter 3.2). There is a general lack of more concrete specifications regarding the nature of lexical affiliation, for example whether a lexical affiliate must only be phonological in form and synchronous with the gesture (cf. Chapter 2.3 on temporal relations), whether and to what degree it includes pragmatic information, and whether it is governed by any grammatical constraints. It seems helpful at this point to look at the details of the affiliation proposed by Schegloff (1984) and others, that is, on which level the lexical connection with the gestures is assumed. Intuitively, “affiliation” is understood as an economic relationship, for instance between a company and a CEO, or as a societal association. These kinds of affiliations are marked by the often strategically chosen profit that at least the associate gains from the affiliation, and at best the institution as well. Transferring this interpretation of affiliation onto the relation between co-produced speech and gestures, the lexical items belonging to the lexical affiliate would then profit from the gesture – similar to what is described within the GP theory. Hence, all lexical items included in the newsworthy part of the verbal utterance would be lexically affiliated with the gesture.

As has been discussed above (Chapter 3.2), the unpacking of the GP reaches its semiotic peak during the time interval in which speech and gesture stroke overlap, while semantic co-expressivity lasts throughout the utterance. The question now is whether lexical affiliation is indeed meant solely as the temporal overlap of the gesture stroke with certain lexical items of the co-produced speech, a standpoint taken, for instance, by Schegloff (1984) or Harrison (2013). Due to the temporal factor, the lexical affiliate would be in phonological form only, regardless of syntactical boundaries or pragmatic strategies. If this affiliation can transcend co-production and reach co-expressivity on a holistic, conceptual level (e.g., de Ruiter, 2000; McNeill, 2005; Kirchhof, 2011), in that the affiliate(s) of the gesture *need* not be explicitly present in the multimodal utterance at hand, will be explored in this chapter. “[W]hether a gesture completely encompasses its verbal affiliate, or whether speech and gesture overlap only partially” (Thies, 2003, p. 53) will be discussed before the speculation that temporal coordination is a factor, not a condition, for co-expressivity. With the help of semiotic correspondence, the rheme, or psychological predicate, of an utterance will be identified even outside syntactical or utterance borders. In Chapter 5.2.1, a methodology for the transcendence from the traditional, rather fixed definition of lexical affiliation toward conceptual affiliation will be proposed and tested in Chapter 6. For this purpose, previous research into the temporal, lexical, and semiotic relations between speech and gestures will be analyzed in the following.

According to Schegloff (1984), taking gestures as indicative of new content in speech is plausible because the gestural counterpart “– both its onset and its acme or thrust – precedes the lexical component it depicts” (p. 276). While Schegloff's formulation is unspecific regarding the “lexical component”, that is phonemes, words, phrases or whole utterances, it allows for a general embedding of a gphr in a verbal utterance or for an overlap of a gphr with co-produced speech. By narrowing down these temporal possibilities, the idea of a direct semantic affiliation between a gesture stroke and the lexical item it precedes in onset or stroke developed:

> There is general agreement that gestures anticipate speech: Gesture and speech are coordinated temporally such that gesture initiation typically precedes speech onset of the *lexical affiliate*, the word or phrase that accom-

> panies the gesture and seems related to its meaning.
> (Morrel-Samuels & Krauss, 1992, p. 615; after Schegloff, 1984)

While the term “lexical item” is often implicitly treated as single or succeeding words, several researchers have conducted research on how to find *the* lexical item affiliated with gestures within instances of utterances produced (e.g., Krauss et al., 1991; Goldin-Meadow et al., 1999). Morrel-Samuels and Krauss (1992), for instance,had participants select lexical affiliates for a number of gestures from provided speech transcripts (p. 618). The authors had decided beforehand against an extended co-expressivity of speech and gestures by restricting lexical affiliation to single words or compounds (cf. Krauss et al., 1991), not allowing for an extended interpretation of lexical affiliates. Adopting this definition, the corresponding lexical item is easy to find when analyzing deictic gestures, for example in someone saying “Look at *that*” and simultaneously pointing at the referenced object. With increasing imagistic complexity of the information conveyed through a multimodal utterance, finding a direct word-gesture relation will become increasingly difficult, and often impossible. An example of the complexity of detecting a lexical connection between speech and gesture was given by Kirchhof (2011): Speaker A says, “The yard looked so beautiful,” while making a motion like flicking water downward with her right hand. Intuitively, one might interpret the gesture to semiotically relate to the “yard”, placing it in conversational space.[5] Directly asking A what she intended to express with the gesture, however, revealed the context of drizzling rain while the sun was shining – background information that had not yet been introduced into the conversation. This context placed the semantic relation between the speech and gesture on rain drops on grass stalks and plants in the yard. Assuming a direct lexical affiliation would have been too narrow to fully interpret the message, and yet, experimental designs exploring the semantic connection between speech and gestures often still leave out context in their analysis (e.g., Bergmann et al., 2011; Krauss et al., 2000).

It seems reasonable to suppose that the unrefuted co-expressivity of the two modalities, as it is present in the GP theory, is more fundamental than how the different utterance parts are connected in any n-to-m relationship (e.g., McNeill,

[5] I am grateful to Dafydd Gibbon for pointing this out.

1992; 2005; McNeill & Duncan, 2000; de Ruiter, 2000). One could say that a gesture and its lexical affiliate stand in a 1-n relation: A gesture may correspond semantically to one or more lexical items inside an utterance. When the kinship in meaning is obvious, the context of the utterance would indeed not influence this relationship. The lexical affiliate could even trigger the gestural counterpart because of the idea they share (see Chapter 3.1), a process recurring whenever a gesture matches a lexical equivalent. One shortcoming of this interpretation is its onesidedness, that is, that the interpreting side stops matching the production side when looking for the closest match for a gesture in the speech it synchronizes with. Looking for synonymy in words, within sentence boundaries, will not produce the full picture, which is why the two concepts of speech-gesture “semiosis”, lexical affiliation and co-expressiveness, have to be set apart clearly. McNeill (1992) wrote on this matter that

> [a] lexical affiliate does not automatically correspond to the co-expressive speech segment. A gesture, including the stroke, may anticipate its lexical affiliate but, at the same time, be synchronized with its co-expressive speech segment. (p. 37)

This follows the temporal definition of lexical affiliation as also put forward by, for example, Krauss et al. (1991) and Goldin-Meadow et al. (1999). Following McNeill (1992), lexical affiliates can be regarded as a subset of co-expressive speech, a definition that would also encompass the “yard” example given above: The complete multimodal utterance is co-expressive, gesture stroke and possibly its onset precede or overlap with one or more stressed verbal items (Chapter 2.3). For speech and gesture to be co-expressive, a combination of speech signals can share meaning with a gesture, and they need not be uttered consecutively without other lexical items between them. Rather, they might be distributed across an utterance, or beyond utterance borders, and still stand in an n-1 relationship with the gesture.

The characteristic that gesture-speech co-expression sets the rheme apart from the context is another important distinction from lexical affiliation, which has been discussed in Chapter 3.2 with regard to the GP. Finally, the stroke-peak observed in *production* synchrony is not as relevant for co-*perception* (e.g., Efron, 1941/1972; Cassell et al., 1999). Gesture and speech can still share meaning when they are not produced in full synchrony. From the viewpoint of perception this further supports co-expressivity above direct lexical affiliation. A wider temporal scope for analyzing bi-modal expressions would also be helpful in finding shared meanings of gesture and speech. This is the case with the utterance in Example 1[6] (Kirchhof, 2011, p. 3), which was produced describing another one of Sylvester's attempts at catching Tweety: S' describes the scene in which Sylvester is dressed up as a bell hop to get into the hotel room Tweety is currently in.

so n[e **ro**te mit goldenen **knöp**fen]

such a red one with golden buttons

Example 1: Co-expression vs. lexical affiliate.

The “speaker [S] traces the position of the buttons on a double button row in a zigzag motion. The palms of his clawed hands face the chest” (Kirchhof, 2011, p. 3). Within a narrow definition of lexical affiliation, “knöpfen” would be directly connected to the gesture, as S' traces the button positions. In Example 1, the gesture indeed begins before and ends with this lexical affiliate, the second prosodic stress put on on “knöp”. The indexical “so ne” is the trigger of the rheme, so to speak, announcing a more detailed description of the uniform; the stroke phase of the gesture begins with “ne”, and everything from “rote” to “knöpfen” is the rheme. The gesture that overlaps in time with the speech phrase is fully co-expressive to the image conveyed: The bellhop uniform. Disregarding the context that Sylvester dressed up in a uniform, the co-expressivity hypothesis would not work, while that of lexical affiliation would. Since both S' and L will naturally have this context, this is not a problem. By sharing one communicative space, both can grasp the full image. For other instances, such as the “yard” example discussed above, a naive observer would not be able to come to conclusive results, or, for that matter, an interlocutor that had not inquired about the gesture.

[6] Bold print indicates prosodic stress, square brackets the gesture stroke phase.

Krauss et al. (1991) hypothesized, as has been touched upon above, that the semantic affiliation between speech and gestures is “a post-hoc construction deriving primarily from the listener-viewer's comprehension of the speech and bears no systematic relation to the movements observed” (p. 744). The authors conducted five experiments to examine “the information that conversational hand gestures convey to naive observers” (Krauss et al., 1991, p. 744). The three of those related to the ad hoc interpretation of gestures in communication will be discussed in the following; focus will be put on the methodology used by and the contribution of Krauss et al. (1991) to the issue of lexical versus conceptual affiliation to lay the grounds for the methodology proposed in Chapter 5.2.1. They narrowed down the temporal and semantic scope of the gesture from linguistic units to adjacent words or compounds before conducting their examinations. Through agreement by 10 judges, lexical affiliates between speech and gestures in videotaped photo descriptions were defined post hoc, which restricted the choice of affiliates for the participants to a controlled minimum (cf. Morrel-Samuels & Krauss, 1992; Beattie and Coughlan, 1999; Chapter 3.1). The subjectively rated affiliate pairs were mixed with random speech-gesture pairings and presented to naive participants. In the first two perception tests, participants in groups of four cooperatively chose the lexical item(s) they felt closest to the potential meaning of the accompanying gesture in muted videos. Krauss et al. (1991) reported

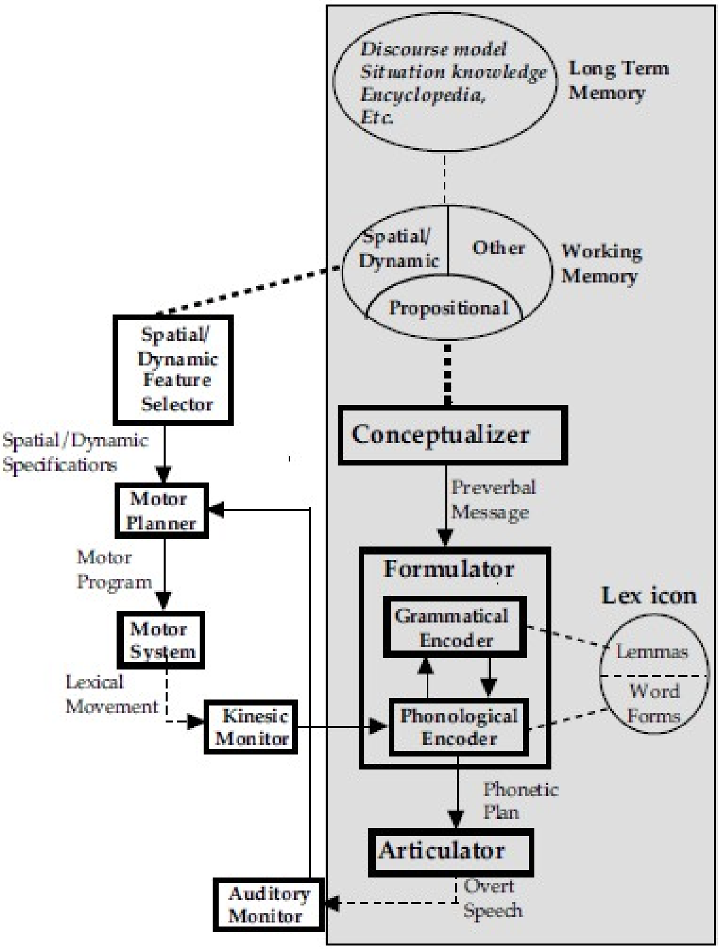


Figure 6: Speech-gesture production model as proposed by Krauss et al. (2000).

that “[f]or 93% of the gestures, a majority of participants selected the *correct* lexical affiliate; on nearly half of them, at least 90% of the judges made the correct choice” (p. 745; emphasis added); the authors admitted that measuring the contribution of gestures to the meaning of an utterance in percentages was not methodologically sound. After the two tests on subjective perception, the researchers grouped the gestures and their selected verbal affiliates from the photo descriptions into the semantic categories of 'description', 'object', 'action', 'location', regardless of their pairings (p. 746; cf. Kirchhof, 2010, on the restrictions of semantic categories); re-analyzing the results by these categories demonstrated a 73% accuracy for actions (p. 747). This lead Krauss et al. (1991) to conclude that gestures were indeed co-expressive and not fully tantamount to or redundant with speech and they refuted their former assumption of unilateral communicativeness toward the listener. In a third experiment, the authors tested whether “perceived gestural meanings derive mainly from the meaning of their [preselected] lexical affiliates” (p. 749), presupposing these to be the major source of gestures' semantic content. Participants were instructed to identify the select semantic categories in speech-only and gesture-only stimuli. In one condition, the judgments were solely based on speech transcripts, while in the other conditions participants were presented with either speech-gesture, speech-only or, gesture-only stimuli.

Krauss et al. (1991) interpreted the results to suggest that “the association between the semantic category assigned to the gesture and the semantic category of the lexical affiliate is greater when the coder can hear the sound” (p. 750). Recognizing that the “four unordered categories” (p. 750) were not suitable for this task, the authors took their findings to imply that speech will give a gesture another interpretation than a gesture alone would trigger. Kendon (1972), among others, has long called this phenomenon emblematicity (see Terminology and Chapter 2.1). Regarding the disambiguating function of speech toward gestures, the question arises whether the two modalities can actually differ in their semantic categories. Expanding the semiotic focus of a gesture to more than a lexical affiliate would allow for this (see Chapter 6). Eventually, Krauss et al. (1991) concluded that gestures helped with resolving ambiguity in speech when no cross-utterance context was given (p. 751). That the semantic content of both modalities was recognized by the authors, however, is based on rather fuzzy results, not only due to the preselection of “correct” lexical affiliates.

As has been commented on above, the methodology of Krauss et al. (1991) was problematic in parts. First, presenting participants with pre-defined lexical affiliates does not contribute to general assumptions on gesture perception or comprehension, but only allows for conclusions about specific communicative settings. The restriction to democratically selected affiliates excluded the possibility that further co-expressive speech might add to the content of the stimuli. Second, the distribution of lexical items into semantic categories, even when restricted to the narrow context of photo descriptions, is a tedious task that encompasses many inter-rater differences (see Kirchhof, 2010). Categorizing *gestures* into semantic categories ensues additional issues, least of all their ambiguity without co-uttered speech – the categories 'description', 'object', 'action', 'location' would, for instance, all accommodate an upward motion of the right hand. Third, the visibility of lips and facial expressions in the video stimuli could have influenced the judges as well as the participants in making their decisions. These shortcomings will be considered in the methodology to investigate the semantic affiliation between speech and gestures proposed in Chapter 5.2.1.

De Ruiter (1998) approached lexical speech-gesture affiliation from a different angle. Instead of focusing on temporal synchrony, he concentrated on the semantic relationship between co-produced speech and gestures. As de Ruiter (personal communication) stated, “[i]f you use the temporal definition, the gesture stroke can only definitionsweise[7] be synchronous with the affiliate”, which is a valid statement. Instead, de Ruiter and Wilkins (1998) suggested that the speech-affiliate of a gesture would be “the word or phrase with which the gesture is semantically and pragmatically linked” (p. 605), that is, be co-expressive with. As has been mentioned above, deictic gestures have rather explicit affiliates in speech, and are temporally constrained. De Ruiter (1998) tested this in a pointing experiment with “the first hypothesis […] that the lexically stressed syllable will provide the synchronization point of the gesture in one-word utterances […, that is, the] primary stressed sylla-

[7] by definition

ble within the word" (p. 30). Participants were to name pictures of objects and their definite determiners while pointing at them when an LED light lit up next to an object; including the determiner provided syllable space for the utterance pitch accent. The pointing gestures synchronized with the onset of the nouns (p. 36), from which the author concluded that the speech adapted to the gesture because the latter necessitates greater (physical) preparation. As the stroke was the meaningful part of the gesture in this specific context, the phonological synchrony rule was reaffirmed.

In a second experiment, de Ruiter (1998) expanded the methodology to include contrastive stress (p. 37). Participants were triggered to produce utterances such as "The **gree**n car, not the blue one". In the speech-gesture utterances, while pointing at the referents, the gesture onset adapted to that of the stressed word – "if the contrastive stress was on the adjective, pointing was initiated 23 ms earlier than when it was on the noun" (p. 44). When the stressed syllable came later in the emphasized word, the stroke hold was slightly longer. While these findings supported the phonological synchrony rule, due to the rather short duration of pointing gestures in general, the gphr did not necessarily synchronize with the full lexical item it related to. De Ruiter (1998) commented on this that pointing gestures are usually of short duration, making "the phonological synchrony rule. . . more a kind of *constraint* than a synchronization principle" (p. 36). Longer utterances might have differed, but this could not be tested with the methodology used in the experiment. While de Ruiter (1998) more or less confirmed the role peak-stroke synchrony plays for lexical affiliation, he only did so for pointing gestures, which are, next to emblems, semantically closest to lexical items (McNeill, 2005, p. 7). Similar information on, for example, iconic gestures is still lacking.

What is crucial to de Ruiter's (1998) experiments is that they analyzed naturally produced language, which is often not the case in gesture research, especially in that on lexical affiliation. This aspect should definitely be kept in mind for further research on this topic. Using the results from the pointing experiments, de Ruiter (1998) proposed the "Sketch Model" for speech-gesture production, or rather for information processing. Several researchers have designed such models in order to understand the co-production of speech and gestures more deeply by formalizing the production processes into testable constructs. In the following chapter, the most influential ones of these models will be discussed, laying further methodological grounds for the analysis of speech-gesture affiliation and perception. Chapter 5 will then introduce methodologies aimed at circumventing, among other issues, the shortcomings of Krauss et al. (1991) and propose more appropriate approaches toward finding perceptual counterparts of co-produced speech and gestures. These methodologies will be applied in a study on the lexical versus conceptual affiliation between speech and gestures in Chapter 6.

3.4 Production Models

Developing models, that is, modularizations sketching processes or chains of processes, is an established way of exploring language processing. Abstracting the complexities involved in language production will provide insights into the bigger picture, as information processing models are “essential theoretical tools for exploring the processing involved in gesture and speech” (de Ruiter, 2000, p. 285); note that the models discussed in the following do not contain explicitly formulated, programmable processing modules that can be tested in a computer application by entering speech-gesture data. Rather, they should be regarded as precursory stages to such computational models, and their comparison as a pre-selection process. Discussing the speech-gesture processing models based on Levelt's (1989) model for speech production by de Ruiter (1998), Krauss et al. (2000), and Kita and Özyürek (2003) will identify the crucial factors for designing a speech-gesture processing model for the listener (Chapter 4.5). De Ruiter (2000) discussed the compatibility of such models with, for instance, McNeill's GP theory and lexical access (cf. Krauss et al., 2000), aspects of which are highly relevant to modeling the GP-SP transmission cycle. On the theoretical foundations of lexical affiliation as well as Levelt's (1989) speech processor, Krauss et al. (2000) proposed a production model for lexical gestures accompanying speech. Another highly discussed model for speech and gesture production, the Interface Hypothesis, was developed by Kita and Özyürek (2003). Due to its strong relations with thinking-for-speaking and with how langue will influence utterance production, this model will be discussed subsequently. Contrasting Kita and Özyürek's perspective of gestures as a window to the mind (cf. also Goldin-Meadow, 2003), de Ruiter (2007) proposed that gestures are rather postcards from the mind, drawing from conceptual transmission (Chapter 3.3) versus communicative intention. However, whether gestures are communicatively intended or are “only” of communicative content that either supports speech or is redundant with it is not a factor for the modeling at hand: All gestures produced will have to be explained by such a model, because the listener will be able to perceive them (cf. Chapter 4.2), and all gestures will have originated in an MU that provided the source for both speech and gesture.

De Ruiter (2007) groups speech-gesture processing models into three architectures, that is, the *Window Architecture*, the *Language Architecture*, and the *Postcard Architecture*. The grouping of these models into architectures is helpful for tracing the imagistic persistence from the MU to the multimodal externalization in the different models. Essentially, a speech-gesture production model – or architecture – will have to fulfill the following requirements to be mirrored and modified into a model of speech-gesture perception: It should (1) recognize that speech and gestures *originate in the same mental image* or idea unit or MU, (2) implement *feedback loops* between the production process and the context and between the motor and linguistic formulating modules, that is, incorporate “coordinative structures” (McNeill, n.d., p. 55), and (3) be able to explain a *temporal coordination* of the two modalities to allow for the GP to unpack as it has been observed in spontaneous utterances. This chapter will give an overview of the architectures as proposed by de Ruiter (2007), and exemplary models thereof, and analyze their respective properties fit for transference into a model of a GP-SP transmission cycle regarding requirements (1) through (3).

Previous research suggested that gestures provide a window into the mind (e.g., Beattie, 2003; Goldin-Meadow, Alibali & Church, 1993; McNeill, 1992; McNeill & Duncan, 2000). McNeill (1992), for instance, proposed that speech and gestures are separated in computation but fuse again in production when unpacking the GP. This would make the GP a package delivered more or less directly from the mind, making the GP theory (Chapter 3.2) belong to the Window Architecture (Figure 7). As de Ruiter (2007) remarks, “[t]he whole point of the Window Architecture is that linguistic processing is bypassed, which is why it provides us with a window into the mind” (p. 35).

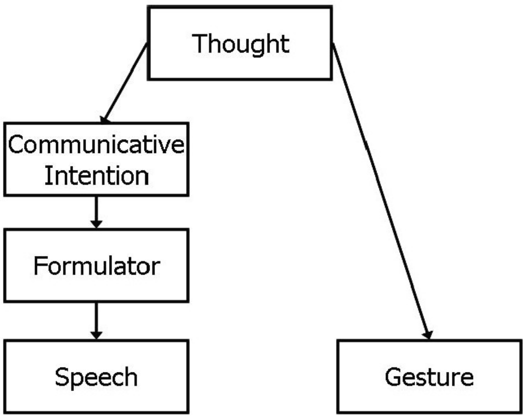


Figure 7: Window Architecture (de Ruiter, 2007).

According to McNeill, “[g]estures exhibit images that cannot always be expressed in speech, as well as images the speaker thinks are concealed” (McNeill, 1992, p. 11). De Ruiter (2007) expressed doubts on this by stating that “most of the communicative signals that we produce in interaction are not consciously planned, and this holds for speech as well as for gesture” (p. 32). Here, one has to differentiate between communicative intent and the content of an idea chosen to be conveyed. While not all gestures might support conveying what the speaker wants to relate, they do always express content from the idea, or GP, the speaker is expressing (p. 32). One theoretical construct supporting this differentiation is the occurrence of speech-gesture mismatches (see, e.g., Goldin-Meadow et al., 1999), which will be discussed in more detail in Chapter 4.2.

De Ruiter (1998) presupposed that gestures are semiotically linked to speech in production, which is supported by their shared semantics (see, e.g., Chapter 3.2). He expanded Levelt's (1989) model for speech production into the Sketch Model (Figure 8), which includes all types of gestures, except for beats, in the utterance planning stage. “[I]n this model, iconic and metaphoric gestures as defined by McNeill (1992) are indistinguishable. Both types of gestures are generated from spatio-temporal representations in working memory” (p. 22). The Sketch Model assumes that (1) gesture and speech have a communicative function, that (2) both modalities originate from the same communicative intention (cf. GP theory), that (3) the conceptualizer distributes the communicative load over the speech and gesture channels, and that (4) speech and gesture will compensate for shortcomings in the other channel while (5) both utterance planning units operate independently from each other except for occasional mutual checking. “This is the so-called *Mutually Adaptive Modalities* assumption (de Ruiter, 2006), later also called the *Trade-off Hypothesis*” (de Ruiter & de Beer, 2013, p. 8, emphasis in original; cf. de Ruiter et al., 2012). The way the Sketch Model works is that the communicative intention of the speaker will be split by the conceptualizer into packages ('units') processable by the gesture planner and the formulator. A combination of speech-gesture units is then externalized in a linear fashion. Within the Sketch Model, the conceptualizer, containing the GP (de Ruiter & de Beer, 2013, p. 8), will initiate separate planning processes for speech and gesture, drawing from information from long term memory (LTM) and working memory (WM). While the gestural part of the utterance is sketched, like an image, the language part goes through grammatical and phonological encoding to form a pre-verbal message. Both parts of the

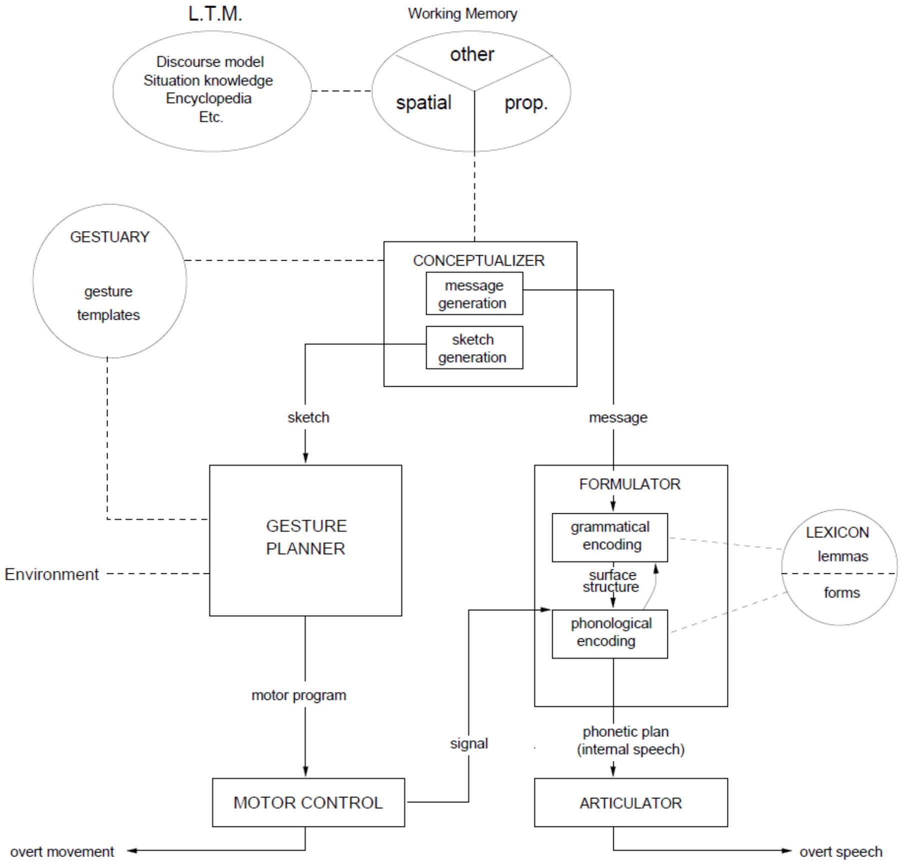


Figure 8: Sketch Model (de Ruiter, 1998).

concept are thus trimmed by the physical restrictions of the motor control and articulator and prepared for externalization (cf. Levelt, 1989). This process will result in a timely coordinated co-utterance of overt movement and speech; feedback circles between the modules constantly adapt to the communicative situation (de Ruiter & de Beer, 2013, p. 8; see also Indefrey, 2011, pp. 10f.).

De Ruiter and de Beer (2013) tested the Sketch Model in the context of non-fluent aphasia, exploring its adaptability to communicative changes. Analyzing spontaneously co-produced speech and gestures, they found a lower rate of gestures per time unit in aphasic speakers than in non-impaired speakers, but a higher rate of gestures per number of words. Non-fluent aphasia will affect lexical planning and the conceptualizer will hence deliver smaller packages to the speech formulator to make utterance production more processable (p. 10). The motor control of the speaker would not be affected by the non-fluent aphasia directly, but the same concept would still have to be externalized by both modalities. Through bilateral checking between the speech formulator and the motor control module, GP unpacking would function just as it would with fluent speech. Due to the holistic nature of gestures, the ultimate utterance-gesture ratio in non-fluent speech would be similar to that in more fluent speech, but fewer words would be produced (de Ruiter & de Beer, 2013, p. 10). Since both Broca's and Wernicke's area are involved in the production as well as in the perception and comprehension of language and gesture, the Sketch Model should be adaptable to information processing by the listener.

De Ruiter (2007) suggested that commonalities between speech and gesture are still present at the conceptualizing stage (cf. Levelt, 1989), but that the production stages are separate. He thus proposed the Postcard Architecture (Figure 9), implying that gestures are rather postcards from than windows to the mind. “The Postcard Architecture implies that information to be communicated is dispatched into gesture and speech channels by a central process.” (de Ruiter, 2007, p. 25), which allows for gestures to express content not contained in speech. It also permits cross-channel compensation and trade-off, in contrast to, for example, the Window Architecture (Figure 7). According to de Ruiter (2007), “[a]n utterance is a carefully crafted postcard from the mind, providing the interlocutor with both text (speech) and the accompanying visual illustration (gesture) in the same multimodal message" (pp. 25f.). Following the Postcard Architecture, gestures cannot provide a full or direct representation of the mental image.

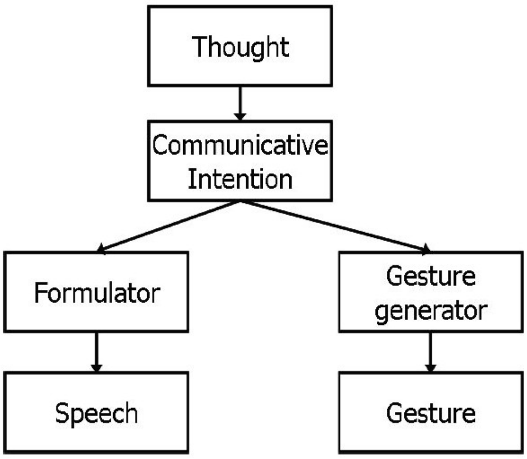


Figure 9: Postcard Architecture (de Ruiter, 2007).

> The . . . statement that 'speech is a window into the mind', would be either trivial, in the sense that we obviously gain information about the speaker's mind from their speech, or very wrong, in the light of the complex processing necessary to transform a communicative intention into articulatory behavior. The transformation of a thought into an overt gesture is different from, but not necessarily less complex than, the processes that transform communicative intentions into speech, and that these transformations prevent gesture from being a window into the mind. The fact that listeners can interpret gestures with relative ease (if they have access to the speech as well) is precisely why they cannot be windows into the mind.
>
> (de Ruiter, 2007, p. 35)

In other words, saying that gestures (or speech) were windows into the mind would assume gestures to be clearer, more direct representations of highly complex thought processes, which counteracts the less complex and efficient expression of utterances. The postcard metaphor reduces this assumption to incomplete overlaps between mental representations and explicated expressions, but also allows for intramodal redundancy and compensation.

The Postcard Architecture further assumes all information expressed in gesture and speech to be communicative in the sense that it is produced as part of the speaker's communicative intent (p. 26). Krauss et al. (2000) are skeptical with regards to the communicative intent of gestures, stating that "if gestures originate in the speech processor, gestural information would consist exclusively of information that was part of the communicative intention" (p. 272). As has been discussed earlier, communicative value differs from communicative intent, and intent cannot be indirectly tested for and should hence not be assumed. The MU will contain more information than is intended to be communicated, and gestures might also express parts of the MU that do not further this intent, in the form of postcards, so to say. The Sketch Model (de Ruiter, 1998) allows for this differentiation, which makes the Postcard Architecture an abstraction of this model.

Another model of speech-gesture production, also based on Levelt's (1989) speech processing model is the Interface Hypothesis by Kita and Özyürek (2003). In this model, communicative intent is fully shared by both modalities through an internal coordination of production (Figure 11). Similar to other hypotheses on lexical affiliation (Chapter 3.3), the Interface Hypothesis is sorted into the models of Language Architecture by de Ruiter (2007). By definition, gestures in the Language Architecture are not generated directly from the mental concept but are rather engaged in supporting the verbal message (Figure 10). Models within this architecture are solely language-driven.

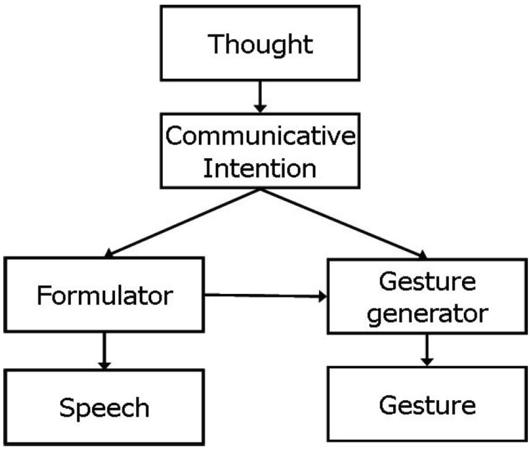


Figure 10: Language Architecture (de Ruiter, 2007).

Kita and Özyürek (2003) aimed at specifying how the content of representational gestures is determined by studying speakers of different languages. The authors found that the Sketch Model (de Ruiter, 1998) as well as the speech-gesture production model by Krauss et al. (2000) assumed that gestures were generated before linguistic planning and that this predicted "that the information encoded in a gesture is not influenced by how the information could be verbally expressed" (Kita & Özyürek, 2003, p. 17). Others, for instance Butterworth and Hadar (1989) or Schegloff (1984), assumed lexical affiliates to be the source of iconic gestures, the problematics of which have been discussed in Chapter 3.3. Kita and Özyürek (2003) proposed the Interface Hypothesis as an alternative to what they call the "Free Imagery Hypothesis" (de Ruiter, 1998; Krauss et al., 2000) and the "Lexical Semantics Hypothesis" (e.g., Butterworth & Hadar, 1989; p. 17). The Interface Hypothesis assumes that gestures originate from the interface between spatial thinking and speech, referring to what Slobin (1987) termed "thinking-for-speaking".

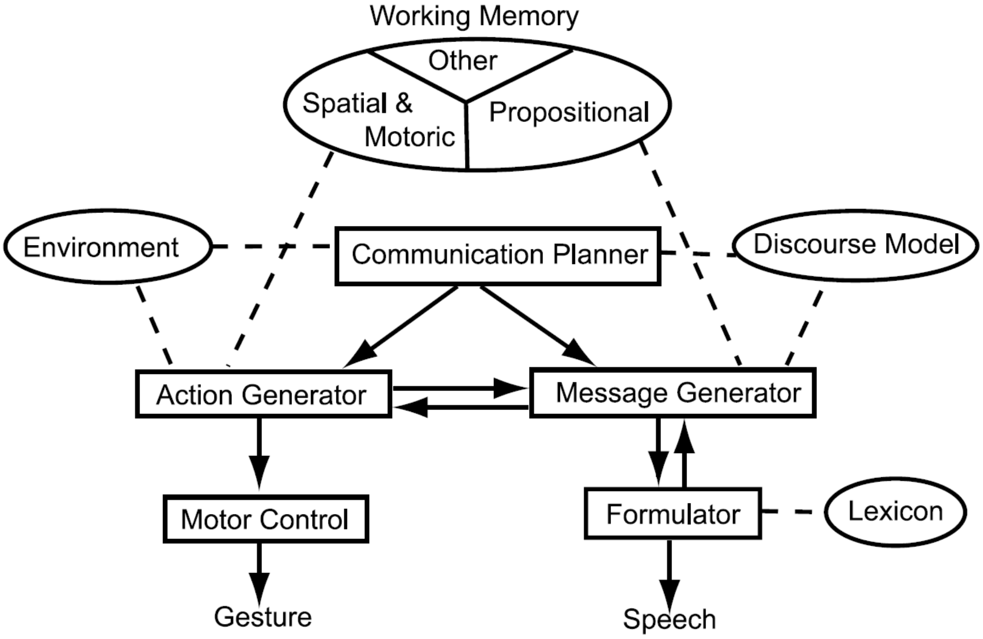


Figure 11: Model of the Interface Hypothesis proposed by Kita and Özyürek, 2003.

Within the Interface Hypothesis, the imagistic properties of the GP are simultaneously processed from WM through (1) the message generator, producing the most efficient expression in speech, and through (2) the action generator, which handles "the spatio-motoric properties of the referent (which may or may not be verbally expressed)" (Kita & Özyürek, 2003, p. 18). In contrast to the models previously discussed, here is a direct coordination between the gesture and speech production modules instead of occasional feedback processes; temporal synchronization is not included (p. 27). As with the Sketch Model (de Ruiter, 1998), speech-gesture coordination takes place internally and does not rely on external sensual feedback (cf. Krauss et al., 2000). Kita and Özyürek (2003) split Levelt's conceptualizer into a communication planner, which forms the communicative intent, and a message generator, which makes the Interface Hypothesis more detailed at this stage, at least graphically. The “gestural content is not fully specified in mechanisms dedicated to communication, such as Levelt's Conceptualizer, but rather in a more general mechanism that generates actions (Action Generator)” (p. 28).

Having tested the GP theory in a comparison of English, Spanish and Chinese, McNeill and Duncan (2000) found that different languages showed different interval positions of speech-gesture synchronies while the GP was efficiently unpacked in all. They argued that processes of thinking-for-speaking correlated with the GP because language competence cooperates with cognitive imagery in order to be communicative. Kita and Özyürek (2003) attempted at implementing these findings in their Interface Hypothesis (Figure 11) and tested it against former models like the Sketch Model (de Ruiter, 1998) in various language contexts. Following up on McNeill and Duncan (2000; see also Duncan, 2001/2006), Kita and Özyürek (2003) focused on attempts at Tweety by Sylvester within the Canary Row series (Freleng, 1950) that contained path and manner. For example, the scene in which Sylvester is kept from reaching Tweety because the bird throws a bowling ball into the pipe he is climbing up was chosen – Sylvester swallows the ball and rolls downhill into a bowling alley (see Appendix 11.1 for further details). Speakers of Turkish and Japanese were expected to differ from speakers of English due to the way manner and trajectory are usually expressed in these languages (p. 23). As predicted by Kita and Özyürek (2003), the two groups of non-English speaking participants often expressed either manner or trajectory in their gestures (pp. 24f.) and rarely merged both features. The authors argued that this was due to grammatical structure as well as vocabulary, which fits with the findings by McNeill and Duncan (2000) and Duncan (2001/2006) for Chinese and Spanish versus English. Kita and Özyürek (2003) concluded that “the data . . . support[ed] the Interface Hypothesis, but they [were] not compatible with the Free Imagery Hypothesis and the

Lexical Semantic Hypothesis" (p. 27). In fact, the data elicited from the participants in Kita and Özyürek (2003) could be explained by all three hypotheses, since they were expressed widely enough to capture a large variety of co-produced speech and gestures (cf. de Ruiter, 2007, p. 34).

The Interface Hypothesis was formulated somewhat more explicitly with regard to the interchange between the motor and speech planning modules, making it more applicable to the specific variations found by the authors and, for example, Duncan (2001/2006). Kita and Özyürek (2003) stated that their model, in which gestures are generated in the action generator, would contrast the position that gestures are solely produced due to communicative intent (cf. Chapter 2.2). In conclusion, the Interface Hypothesis does not differ significantly from the previously discussed models except for that it incorporates greater roles for pre-utterance exchange between the formulators for speech and gestures and communicative intent.

Like the Interface Hypothesis, the model proposed by Krauss et al. (2000) would also be subsumed under the category of Language Architecture (Figure 10) because speech and gesture are generated in separate processes while the speech controls the externalization of the gestures. Similar de Ruiter (1998), Krauss et al. (2000) approached speech-gesture production on the basis of Levelt's speech processor (1989). Krauss et al. (2000) aimed at researching the origins and functions of gestures by introducing them into Levelt's model. As Krauss and Morrel-Samuels (1991) did, Krauss et al. (2000) started from the viewpoint that the communicative value of gestures is optional (p. 262) and approached speech-gesture modeling from the angle of lexical access and retrieval. While many researchers have studied whether gestures might facilitate lexical access (Chapter 3.1), Krauss et al. (2000) rightly bemoaned that "none of the writers who have suggested this possibility has described the mechanism by which gestures affect lexical access" (p. 265). The authors assumed that a speaker had a "source concept" in mind that would form in working memory, depending on the communicative situation or intent, similar to Vygotsky's (1984) MU, but they did not refer to this or other previous research regarding this matter.

Krauss et al. (2000) considered symbolic gestures (e.g., emblems), deictic gestures, motor gestures (e.g., beats), and lexical gestures (i.e. mostly iconic gestures with differing degrees of representation, p. 276) in their processing model. While the authors had gestures run through utterance planning in a side process, they reconnected this side process with the lexical planning stages at various intersections, describing cooperating systems of production for the two modalities (p. 265). In the model proposed by Krauss et al. (2000), speech processing, as in Levelt (1989), consists of three major modules (Figure 6): After the speaker has formed a source concept in their working memory, (1) the conceptualizer will concretize the planned utterance with regard to context and what the speaker wants to relate. As a preverbal message, the information will be processed by (2) the formulator with regard to grammatical and phonological restraints, back-channeling with the lexicon. From the preverbal message, a phonetic plan is formed, which is then executed by (3) the articulators. During this verbalization process, the spatial-dynamic features from the “source concept” that were not chosen to be verbalized are processed by the motor planner, the gestural equivalent of the speech formulator. The motor planner has two functions, namely (1) to start a lexical gesture[8], or (2) to end a lexical gesture (p. 268). Which of the two possibilities applies will be regulated through cross-modal kinetic and auditory monitoring. In other words, the motor planner will keep gestural information on hold until the phonological encoder requests its actions, for example in lexical access or manual pointing (p. 269, “cross-modal priming”).

Krauss et al. (2000) assumed that gestures will be initiated directly from WM, and any information they might convey would not be part of speaker's intention and, hence, not communicative. As has been touch upon above, communicative intent is a delicate issue, which makes it hard to incorporate in abstract models. As can be seen in Figure 6, the authors divided WM into three sections, namely 'spatial/dynamic', 'propositional', and 'other'. In their proposed model, speech and gestures separate at the stage of WM. In contrast to the Sketch Model (de Ruiter, 1998), the conceptualizer is solely relevant to speech processing here. Krauss et

[8] Figure 6 uses “lexical movement” instead of “lexical gesture” because it has been used “[i]n previous publications (Chawla & Krauss, 1994; Krauss, 1995; Krauss, Chen & Chawla, 1996)” by the authors (Krauss et al., 2000a, p. 3).

al. (2000) argued that in order to prepare a pre-verbal message, any information forwarded from WM to the conceptualizer needed to be in propositional form. This propositional form would include communicative intent as well as a pre-selection of what the speaker would *verbally* want to convey from their mental image in their utterance, much like jotted down notes. The authors explained the splitting off of the spatial-dynamic information at this point as follows: “[I]f gestures originate in the speech processor, gestural information would consist *exclusively* of information that was part of the communicative intent” (p. 272). By employing the auditory-kinetic feedback loop, Krauss et al. (2000) allowed for gestures to be part of communicative intent, but only for specific situations. Unintentional gestures were not explained with this model proposal, but Krauss et al. (2000) tried to “consider some of the implications of assuming that such gestures are communicatively intended” (p. 266). To clarify their choice of splitting the modalities before the conceptualizing stage, the authors refer to Kendon's (1980) “cake”-example:

> Recall Kendon's previously described example of the speaker saying “… with a big cake on it…” accompanied by a circular motion of the forearm. Although it may well have been the case that the particular cake the speaker was talking about was round, ROUND is not a semantic feature of the word *cake* (cakes come in a variety of shapes), and for that reason ROUND was not be [sic] part of the speaker's communicative intention *as it was reflected in the spoken message*. (Krauss et al., 2000, p. 266; emphases in original)

While Krauss et al. (2000) admit that the different representational formats from WM are occasionally translatable into each others' forms, the authors strongly support a preference to verbalize communicative intent rather than express it multimodally. This might be a valid point, but proposing that “round” is not a semantic feature of the word “cake” is anti-semiotic. Considering that the authors saw the origins of an utterance in a source concept from working memory, one cannot help but draw parallels to semiotics. Depending on which semiotic model one prefers, the discussed “cake” will either be the signified or concept (de Saussure, 1972/1983) or the object that should be described (Peirce, 1894/1998). A speaker will have a subjective memory or idea of a cake in mind – either a portmanteau of previously seen or otherwise perceived cakes, or a concrete image of the cake that is currently part of the conversation. Depending on the imaginative cake, “round” might indeed be a feature of said cake. And, in other models, such as the Sketch Model, the roundness of the cake might, intentionally or unintentionally, be expressed by the speaker, because of the cooperative and compensatory relationship between the verbal and motor articulators. However, the model by Krauss et al. (2000) has been strongly based on lexical affiliation (see Chapter 3.3), and it grants gestures an assisting function at best. Regarding the “round”-gesture discussed above, the authors comment on its expression that

> [b]ecause gestures reflect representations in memory, it would not be surprising if some of the time the most accessible features of those representations (i.e., the ones that are manifested in gestures) were products of the speaker's unique experience and not part of the lexical entry's semantic.
>
> (p. 273)

Krauss et al. (2000) propose that only gestures directly related to parts of the verbal utterance should be produced through the feedback loop between the phonological encoder and the motor control – a round gesture would not help retrieve the lexical item “cake”. The authors did ascribe intentional expression to emblems and certain iconic or pantomimic gestures, at least (p. 274). For other gestures, they entertained the possibility of rare cases where there might be communicative intent (p. 273), but lack of evidence kept them suspicious (N.B.: there is a general lack of evidence for either position on this issue). Krauss et al. (2000) supported the argumentation for this selection of gesture types by pointing out that producing gestures without communicative intent would go against Clark's (1996) concept of collaborative language use or joint accomplishments respectively (p. 274; cf. Grice, 1975). Also regarding the communicative intent of gestures, Krauss et al. (2000) first differentiated between mental concept (working memory) and intention, which is a valid measure (Chapter 3.2). They then excluded the communicative potential of gestures from the conceptualization stage, but without sound reasoning for or against their decision and failing to explain unintentional gesturing. The crux here is that the authors made, or rather did not make, a differentiation between the communicative intent behind and the communicative value of gestures. Krauss et al. (2000) recognized and remarked that their proposal for a model of speech-gesture production was flawed throughout their paper, saying that is was "tentative and highly speculative" (p. 277). Krauss et al. (2000) also mixed up properties of semiotics and semantics, suggesting that intentional gestures would have other origins than non-intentional ones (p. 274).

Regarding the listener side of the speech processing models based on Levelt (1989) discussed above, not much was explicated by the authors of any of the models discussed above. Since all speakers are also listeners, however, some assumptions can be made regarding which aspects of the models might be transferred from production to perception. The feedback loop between the kinetic and auditory monitor in the model by Krauss et al. (2000) is of interest here, for example. If these monitors were at work internally for the speaker during utterance production, they would also be present in the listener during perception. The authors connected the feedback process to the tight temporal coordination between speech and gestures: In case the formulator "allowed" for a gesture to be produced, the gesture would be initiated simultaneously with the lexical affiliate. Then, the auditory monitor could signal to motor control when a lexical affiliate had terminated, and the gesture would retract. Such a process would give speech the control over the duration of gestures (cf. Krauss & Morrel-Samuels, 1992). If this were true, *and* every gesture produced had communicative intent, listeners would (a) be able to identify the lexical affiliate for each and every gesture, and (b) include every gesture in their comprehension process. Both of these speculations are invalid, as has been and will further be discussed later in this dissertation (e.g., Chapters 3.3 & 4.2). For now, suffice to say that a model of speech-gesture processing should be constructed widely enough to include all natural occurrences of speech-gesture production. If it is not able to do this, it should offer arguments based on sound data for why it would or could not, and not based on "tentative and highly speculative" intuitions. Either way, the model proposed by Krauss et al. (2000) might not lend itself well as a basis for a counterpart in the listener because it is, in parts, too exclusive or not conclusive at all. This, as well as will be analyzed in more detail below before the requirements defined at the beginning of this chapter.

Looking back at the beginning of this Chapter, the following requirements for a speech-gesture production model – or architecture – will have to fulfill the following requirements have to be met by a language processing model to be mirrored and modified into a model of speech-gesture perception:

> It should (1) recognize that speech and gestures originate in the same mental image or idea unit or MU, (2) implement feedback loops between the production process and the context and between the motor and linguistic formulating modules, . . . and (3) be able to explain a temporal coordination of the two modalities to allow for the GP to unpack as it has been observed in spontaneous utterances. (p. 45)

De Ruiter (2007) has already discussed that (1) is fulfilled by models belonging to the Postcard Architecture, because the architecture "implies that information to be communicated is dispatched into gesture and speech channels by a central process." (p. 25). This central process is situated in the conceptualizer in those models that have modified Levelt's (1989) speech processor by adding the gestural component, that is, in the Sketch Model (de Ruiter, 1998) and the Interface Hypothesis (Kita & Özyürek, 2003), albeit to varying degrees. The model proposed by Krauss et al. (2000) separates speech and gesture production between the MU and the conceptualizer, treating gesture as an occasional additive rather than a constant addition to speech. Accordingly, this model violates requirement (1) and cannot be considered for the model of the GP-SP transmission cycle.

Both the Sketch Model and the Interface Hypothesis integrate feedback processes between the gesture and speech generation processes. While in de Ruiter (1998), the motor control module signals to the phonological encoder that, for example, the gesture is ready to be performed, Kita and Özyürek (2003) split the conceptualizer into three modules: The communication planner, which is in constant exchange with both an action generator and a message generator, which are also back-channeling to each other. While the Sketch Model facilitates post-formulating feedback between motor control and the message formulator, the Interface Hypothesis allows for no further exchange after the "conceptualizer" has initiated the separate execution of speech and gesture. As has been described in de Ruiter and de Beer (2013), among others, there needs to be an exchange at this later stage of utterance production due to any problems the formulator might encounter, be it non-fluent aphasia, slips of the tongue, or ToT states, to name just a few. The Interface Hypothesis would possibly re-initiate the whole utterance generation process at this point, while the Sketch Model would go into bug-fixing mode. Through this, the Sketch Model would be more time-efficient while also being able to explain more phenomena holistically, but both models could cope with this. We will sideline the Interface Hypothesis for a moment for requirement (2), and proceed to checking both models for the temporal coordination requirement (3).

Temporal coordination between speech and gesture externalization is necessary in any model designed to explain (a) the temporal synchrony of gestures with phases of the speech they are co-produced with, as it has been observed for spontaneous utterances, and (b) their rhythmical interplay (e.g., Gibbon, 2009; Loehr, 2004). While de Ruiter (1998) had originally intended the Sketch Model to explain semantic synchrony, which it does, with it he also also reaffirmed the phonological synchrony rule, at least for deictic gestures. The constant feedback processes between the motor control and the phonological encoding unit ensure a temporally arranged execution of speech and gesture. In this, the "intended" unpacking of the GP is also facilitated. In addition, the fact that the Sketch Model is also applicable to speech produced by individuals with language impairments provides various points of intersection with how listeners deal with temporal asynchronies between speech and gestures. In the Interface Hypothesis, temporal coordination might be initiated in the communication planner, but Kita and Özyürek (2003) were not explicit on this. In case of challenges with the full utterance execution, such as spatial restrictions through a passing person or the miscalculation of the distance to a piece of furniture, would quite possibly disrupt an utterance. This will occasionally happen, but usually a speaker will continue speaking, with or without a pause. Due to the feedback loops at a lower level, the Sketch Model is able to also explain such happenstances while the Interface Hypothesis is not. For this reason, the Sketch Model fulfills requirement (3) to full satisfaction, as might any model belonging to the Postcard Architecture, while the Interface Hypothesis only does so inadequately.

In conclusion, the Sketch Model is well suited for being mirrored as well as transformed into a model for the perception and processing of co-expressed speech and gestures, such as the GP-SP transmission cycle, as far as semantic and temporal synchronization is concerned. This provides the first third of what such a model should encompass: the cognitive ability to take up the information provided by the speaker. The second third then will be based on the physical ability of the listener to actually perceive and integrate the signals from speech and gestures, that is AVI. Whether listeners will have this ability will be discussed in Chapter 4.4 and then experimentally tested in Chapters 6, 7, and 8. The last third regards comprehension – whether *it* listeners will perceive multimodal information *and* have the cognitive structures to process them they will also uptake information from both speech and gestures. The "speech comprehension system" module in the production model by Levelt (1989) provides a rough sketch of the perceptional procedure involved. Since the general communicativeness of gestures has already been agreed upon (Chapter 2.2), comprehension will only be discussed briefly in Chapter 4.2. By distinguishing comprehension and perception more clearly, that is, considering perception to be a gateway between production and comprehension, the theoretical foundations of how to investigate the perception of speech and gestures will be further specified.

4 Theories of Multimodal Signal Perception

4.1 Introduction

Mead (1938) described perception as an "active search for stimuli related to the impulse" (pp. x-xi), and Watling (1950) put forth the causal theory of perception. In analogy to the observation that one cannot not communicate (e.g., Watzlawick, Helmick Beavin & Jackson, 1967), we cannot not perceive our surroundings, except for when using certain meditative strategies or due to clinical conditions. In contrast to Mead's approach to perception, the merely physical process is rather passive, but (re)cognition then leads to more conscious processes like integration, information uptake, and comprehension. For instance, we will *perceive* parts of conversations or other occurrences around us, but in direct interaction we will *process* them further to converse. Within the scope of this dissertation, the focus lies on the perception and integration of multimodal utterances, not on their comprehension.

As has been discussed above, timing is a crucial factor in how we produce and integrate utterances. Several studies in the area of psychophysics (e.g., Nishida, 2006; Fujisaki & Nishida, 2005) have found a temporal window for the perceptual alignment of visual and auditory signals, the so-called window of AVI. Drawing from non-speech scenarios with or without cause-and-effect relations between audio and video channels, studies by McGurk and MacDonald (1976) and Massaro et al. (1996; also Winter & Müller, 2010; Vatakis et al., 2008; Massaro & Cohen, 1993) have been concerned with the perception of speech-lip asynchronies. This is a phenomenon occurring, for example, with dubbed movies or video streaming. Chapter 4.3 will provide insights into how listeners deal with speech-lip asynchronies, providing further intuitions on how listeners might integrate asynchronies in speech-gesture utterances. The findings on speech-lip asynchronies have already inspired some research in the gesture field, among others by McNeill et al. (1994), Cassell et al. (1999), and Habets et al. (2011). Seyfeddinipur & Kita (2003), for instance, discovered that while strong asynchrony between speech and gestures during speaking prompts the speaker to repair their utterance through self-monitoring, the listener is expected to disregard or internally align the smaller asynchronies to ensure proper comprehension. How this realignment is achieved will be discussed, along with other factors regarding speech-gesture integration, in Chapter 4.4. Ultimately, the theoretical grounds will be determined for investigating the AVI of desynchronized speech and gestures in the studies presented in Chapters 7, and 8. Before a well-founded discussion can take place on multimodal integration, however, there is a crucial distinction to be made between perception and comprehension. This will be discussed further in the following Chapter 4.2.

4.2 Comprehension vs. Perception

Multimodal signal reception, and language reception in particular, is a complex system involving various processes. While some, for example Lewandowski (1985), see language reception as a holistic system of constant interaction between language recognition, perception, and comprehension, others, for example Kitsch (1998), suspect these processes to be iterative and successive. As has been put forth by Rickheit, Sichelschmidt, and Strohner (2002), there are two major approaches to reception models, that is, bottom-up and top-down approaches. Kitsch (1998) is one example of the bottom-up approach, describing perception as analytical and iterative. Not only for reading reception, he understood comprehension as a step-wise extraction of meaning from an utterance (Rickheit et al., 2002, p. 399; cf. Hielscher-Fastabend, 1996, pp. 82ff.). This data-driven view of language reception appears to be rather time-consuming and complex, particularly before the background of online processing. The top-down approach to language reception as proposed by, for example, Laird (1989), understands reception as a synthetic process driven by the aim to construct an encompassing representation of the facts and circumstances related by the speaker (Rickheit et al., 2002, p. 399). The listener wants to gain a maximal representation of the speaker's MU, so to say, matching mental schemata to what is perceived. The two approaches are by no means exclusive, but rather coexist, even interact, depending on the situation, medium, and quite possibly on the complexity of the message to be related.

Within the complexity of reception, there are separate but interacting processes of perception, or recognition, integration, and comprehension, much like there is a difference in production between utterance and proposition. In order to build a bridge from the MU in the speaker to the MU in the listener, one has to trace the path from production (explication of GP) via perception and information uptake or AVI toward comprehension (formation of SP). Figure 12 shows how, in abstraction, individuals might perceive and integrate speech-gesture utterances produced by a speaker, with whom they share the capacity of language production (cf. e.g., Massaro, 1989; Massaro et al., 1996) and the desire to communicatively align (e.g., Wachsmuth, de Ruiter, Jaecks & Kopp, 2013; Pickering & Garrod, 2004).

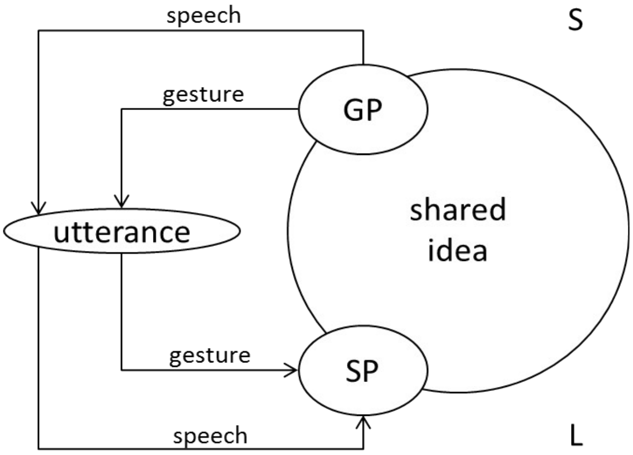


Figure 12: GP-SP transmission cycle (basic draft; same as Figure 1).

Several studies have looked at the comprehension of speech and gestures (e.g., Holler, Shovelton & Beattie, 2009; Gullberg & Kita, 2009; Gullberg & Holmqvist, 2006; Alibali et al., 2001). The crucial point is that before comprehension takes place, the listener will have to first perceive the utterance, then – consciously or unconsciously – select and combine. that is, align stimuli from the utterance to process and make sense of the utterance. In a best-case scenario, the GP unpacked by the speaker will be maximally present in the listener as the SP and take part in the formation of the listener's next MU (cf. Rickheit et al., 2002, p. 399). The model draft of the GP-SP transmission cycle (Figure 12) will later be expanded to include additional factors like context and semantic as well as temporal factors (Chapter 4.5).

While it is not the goal of this dissertation to explain in detail how gestures contribute to utterance comprehension, it has been proven that gestures are communicative (e.g., Melinger & Levelt, 2004) and that listeners are able to seize meaning from them. Yet, while “some gestures are meant to be seen” (Alibali et al., 2001) by the listeners, and hence processed, other gestures are perceived only peripherally. This is a crucial distinction to make, because if one wants to investigate when and how listeners will uptake information from gestures, it has to be ascertained a priori that they are able to perceive them. This is one of the major aims of this dissertation. Under which circumstances listeners will gain information from the speakers' gestures has been investigated within different methodological frameworks. Gullberg and Kita (2009) as well as Gullberg and Holmqvist (1999; 2006), for instance, investigated the possible correlation between gaze and information uptake in natural conversation inductively by using eye tracking. Alibali and Goldin-Meadow (1993), McNeill et al. (1994), and Cassell et al. (1999) approached gestural information uptake from a different angle – they looked at how naturally co-occurring speech and gestures or artificially created speech-gesture mismatches were integrated by listeners. Temporal asynchronies between the modalities and the effect of relative timing on comprehension have also recently been addressed explicitly (Habets et al., 2011; Özyürek et al., 2007). In the following, various approaches to the processing of speech-gesture utterances will be summarized to supply the reader with enough information on what processes occur after the perception of multimodal signals to distinguish more clearly between speech-gesture perception and comprehension. The remainder of Chapter 4 will further explore the mechanisms occurring between production and perception.

We know from eye-tracking studies on text comprehension that readers will not fixate each and every word or letter (e.g., Mayberry, Crocker & Knoeferle, 2009), but will still perceive and integrate all items – the dominance of semantics is so strong that it might occasionally be difficult to ignore parts of the content (Rickheit et al., 2002, p. 397). This reception phenomenon might as well be generalized for all multimodal perception, including that of speech-gesture utterances. Gullberg and Holmqvist (1999), i.a., hypothesized that listeners will perceive a gesture when (a) the articulation of the gesture is peripheral or (b) the speaker indicates in any way that the gesture is to be noticed (p. 35). They tested these hypotheses by having speakers (4 Swedish, 4 French) narrate a cartoon story with the goal to make the listener understand the story and punchline (p. 6); the listener's goal was to understand the story. The listeners were wearing eye-tracking equipment that recorded their visual field and fixation points. The results showed that listeners only suspended their fixation of the speakers' face in 1.6 % of the narration durations. Within these other foci, 44% of the time listeners would look at the speakers' gestures (p. 12); n.b., these results concern only active *foveal* fixations, not *peripheral* perception (p. 2). Hypotheses (a) and (b) were confirmed by the results in that listeners mostly fixated gestures outside the center, particularly in the left and right periphery (see Figure 18 for the division of gesture space according to McNeill, 1992), and even more so when the speakers themselves fixated their gesture (see also Gullberg & Kita, 2009). This observation aligns with the “active search for stimuli related to the impulse” suggested by Mead (1938, v.s.). When interpreting the results found by Gullberg and Holmqvist (1999), one has to keep in mind that “[w]hile fixations are overt physiological events, attention is a cognitive phenomenon […, an] act of directing your focus of consciousness towards the gesture (in the sense of Chafe 1994)”, which will then lead to information uptake (p. 23) and, ultimately, comprehension.

The eye-tracking methodology used by Gullberg and Holmqvist (1999), among others, is an approach to inductively measure information uptake in the listener by correlating gesture fixation with information processing. As has been mentioned above, however, whether and which information a listener gains from a perceived message varies depending on the context, their communicative goal, on their linguistic capacity (thinking-for-speaking; e.g., McNeill & Duncan, 2000; Duncan, 2001/2006), etc. Gullberg and Holmqvist (1999) instructed listeners to *understand* the story related to them, which might be considered a low-threshold goal, especially for a rather simplistic story line as in the printed cartoons used for elicitation, and their results might have varied with other instructions. In the classic McNeill lab procedure (McNeill & Levy, 1982) listeners are instructed to listen carefully to the Canary Row narration because they will have to retell it afterwards. This might lead to a higher attentiveness in general and, quite possibly, to more attention to detail as well as to gestures. Indirectly, this procedure might also be more telling with regard to the general information uptake in the listener, with or without gesture visibility (cf. Gullberg & Kita, 2009). For this reason, this elicitation procedure has been applied to create the corpus used for all studies in this dissertation, the details of which will be presented in Chapter 5.3.

As has been discussed above, there are bottom-up, that is, data-driven, as well as top-down, that is, schema-driven aspects of language reception. Regardless of the weighting of these aspects within the actual reception process, timing will be a factor:

> Die Sprachrezeption ist ein außerordentlich komplexer Vorgang des Zusammenwirkens zeitlich paralleler oder in zeitlicher Überlappung auftretender Teilleistungen, wobei diese Teilleistungen einen sehr unterschiedlichen temporalen Erstreckungsgrad besitzen.
>
> (Herrmann, 1985, p. 63)[9]

Going back to the model draft of the GP-SP transmission cycle (Figure 12), the influence of timing has not yet been considered on the perception side. Production timing, on the other hand, is regulated by feedback mechanisms between the articulators' formulators (de Ruiter, 2007). In order to later integrate timing within the reception module, the influence of multimodal (a)synchrony on perception will have to be ascertained first. This will ensure the applicability of the model for asynchronous speech-gesture signals. As with the feedback loops in the Leveltian production models, similar processes can be expected to occur during reception in the listener to accommodate changes in the utterance production by the speaker.

Multimodal signal asynchronies are well-known, naturally occurring phenomena. We perceive a causal connection between the events and sounds of clapping, of thunder and lightning, of ringing a bell, or of speech signals and lip movements. This connection we make comes from experience, from knowledge, from our LTM. Certain asynchronies between the visual and the audio component of these multimodal signals are natural and common, because light travels faster than sound

[9] Speech perception is an extraordinarily complex process of the interaction between temporally parallel or overlapping sub-efforts, with these sub-efforts being of different temporal range.

(e.g., Einstein, 1905/2005), and listeners might not even notice them. But there is variation in the perception of simultaneously produced audio and visual signals (e.g., Fujisaki & Nishida, 2005; Nishida, 2006) – at large enough asynchronies, audio and video signals will not be integrated by the listener as components of one and the same percept, especially when the listener expects a causal relation between the two. There is only a small temporal window of AVI within which listener-viewers will integrate two signals from two different modalities that are not knowingly causally related or do not seem as naturally co-occurring as they usually do. And even for causally related signals, depending on the perceived naturalness (cf. Chapter 7) of the multimodal asynchrony, AVI will vary.

Petrini, Holt, and Pollick (2010), for instance, presented expert drummers as well as musical novices with point-light displays that drummed a certain rhythm on a computer screen, and audio drums were played through headphones. The drumming lights were shown either horizontally or rotated by 90, 180, or 270 degrees, and the audio was played at steps of 66.67 ms, 133.33 ms, 200 ms, and 266.67 ms preceding or lagging the visual stimuli (p. 3). The participants were asked to judge either the audiovisual simultaneity or the temporal order of the light movements. The authors found that expert drummers performed far better than the novices at recognizing asynchronies between the lights and sounds, especially when the light bar was rotated. Petrini et al. (2010) interpreted this to imply that "the enhanced drummers' sensitivity to asynchrony probably depends on their ability to judge the temporal relationship between the auditory and visual signals, while for novices sensitivity may be based on feature matching processes (Fujisaki & Nishida, 2007, 2008)" (p. 11). The authors suggested that the professional drummers, possibly through their muscular memory, had a more deeply situated connection between drum sounds and visual signals in general – that they related the drum sounds to personal experiences of drumming and hence judged the asynchronies more accurately. The novices, on the other hand, had to rely mostly on their WM, which soon got overloaded.

In the study by Petrini et al. (2010), the participants were assisted in recognizing (a)synchronies between light and sound signals because both modalities shared a rhythm and a certain signal frequency, that is, three lights and drums in a row. In other experiments, such as by Fujisaki and Nishida (2005; also Nishida, 2006), participants were presented with singular audiovisual stimuli, for example with one beep and one blink. Again, the instructions were to determine whether the two channels were in synchrony. The participants were reportedly driven by the audio channel at temporal frequencies slightly higher than 4 Hz. According to Fujisaki and Nishida (2005), "this is the condition where visual and auditory changes are perceptible, but detection of audio-visual asynchrony is hard. Auditory driving demonstrates how the brain binds the audio-visual signals under this paradoxical situation" (p. 463).

The findings from psychophysics by Petrini et al. (2010) or Fujisaki and Nishida (2005) on the detection of audiovisual asynchronies suggest that while viewer-listeners are able to discriminate between synchronous and asynchronous stimuli in general, the circumstances under which they perceive these stimuli, for example in sets or individually, as well as task familiarity, such as by the drummers (Petrini et al., 2010), will influence their capacity of AVI. The same is to be expected for multimodal utterances in that listeners who are more firm in the language a movie has been dubbed into, will be more susceptible to desynchronized dubbing. Making use of the principles from psychophysics, the cognitive psychologist McGurk and his colleagues (e.g., McGurk & McDonald, 1976) conducted seminal research on speech-lip asynchronies, discovering what was later termed the "McGurk Effect".

4.3 Speech Perception

McGurk and MacDonald (1976) described how participants who were simultaneously presented with two different CV-syllables, for example /ga/ and /ba/, via the audio and video channels perceived the syllables as fused percepts (e.g., /da/): "98% of adult subjects gave fused responses to the ba-voice/ga-lips presentation and 59% gave combination responses to its complement" (p. 747). These findings, among others, demonstrate that the sounds of speech are not the only factors for the listener in communication, and that audiovisual synchrony plays a major role. This has also been established by Fujisaki and Nishida (2005), among others, for cause-and-effect stimuli of physical events such as light and beep signals (v.s.). Based on the McGurk effect, Massaro and Cohen (1993) tested the perception of CV-clusters and vowels at various asynchronies of up to 200 ms of the audio before the video and vice versa. The temporal range in which the bimodal stimuli were fused by the participants can be considered the window of AVI. In the following, studies by Massaro et al. (1996), van Wassenhove, Grant and Poeppel (2007), and Winter and Müller (2010) will be explored with regard to which temporal restrictions apply for the AVI of speech-lip signals. These will provide intuitions on the general integration abilities of listeners that will be included in hypotheses on integration of speech-gesture asynchronies later on.

In order to further specify this window, Massaro et al. (1996) conducted experiments with varying and slightly larger asynchronies. Next to two identification tasks, in which participants had to tell whether stimuli were in synchrony, they also used a fuzzy-logical model of perception (FLMP) that assumed the video and audio to be synchronous. The model "predict[ed] integration across different asynchronies as long as the two modalities [were] perceived as belonging to the same perceptual event", that is, to one stimulus (p. 1778, see also Fujisaki and Nishida, 2005; van Wassenhove et al., 2007). Massaro et al. (1996) used synthetic speech stimuli in addition to those from natural language. A polygon facial model displayed the randomized stimuli of modified McGurk pairs to the participants. The tested asynchronies were varied in seven steps within ± 267 ms and additionally at ±533 ms. The authors concluded from the participants' synchrony ratings that an AVI breakdown would occur at asynchronies of about ± 500 ms while integration would be optimal within a window of ± 200 ms.

Van Wassenhove et al. (2007) created syllable stimuli with fourteen steps of 33 ms in which the audio onset was put before or after the onset of the video up to discrepancies of ± 467 ms. These stimuli were used in an *identification task* as well as in a *simultaneity judgment task*, both of which were completed in succession by each participant. As in the original experiment by McGurk and MacDonald (1976), only natural speech and video recordings were used. In the identification task, the participants in Van Wassenhove et al. (2007) chose between three possible percepts in a multiple-choice fashion, that is, the actual audio signal (/ga/ above), the actual video signal (/ba/ above) or the 'fused McGurk percept' (/da/ above). In the simultaneity judgment task, participants were then asked to determine whether audio and video were in synchrony. They had to choose between "simultaneous" and "successive", regardless of order. The window of AVI as judged from the responses with the fused percept, that is, what the listener perceives from the audiovisual stimulus, reached from asynchronies of the audio 67 ms before the visual to 267 ms of the audio after the visual (range 334 ms). The participants accepted a smaller AVI window (-73 ms VA to +131 ms AV) for the stimuli in which the audio and video contained identical syllables. Van Wassenhove et al. (2007) deduced a "maximal true bimodal fusions cluster within ~200 ms" (p. 604) from this. The identification task gave results well below the estimated breakdown of AVI at asynchronies of more than 500 ms (cf. Massaro et al., 1996). Van Wassenhove et al. (2007) conclusively accepted a window of about 200 ms[10] for general alignment but assumed that "to allow the extraction of modality-specific information", tighter synchrony was necessary (p. 605).

Winter and Müller (2010) approached the AVI of speech-lip signals from a neuroscientific angle using a passive audiovisual shift detection methodology. They analyzed ERP signals for 228 audiovisual stimuli in various degrees of bidirectional asynchrony as perceived by listeners. In two experiments, 25 participants were presented with 147 syllables, words, pseudo-words, and sentences spoken by a middle-aged man, with lengths of up to 4000 ms in the following conditions ("METHOD"):

- 20 words with V only (E1);
- 20 words A only (E1);
- 20 words in AV condition, not manipulated (E1 & 2);
- 20 words in AV condition, but with female voice (mismatch, E1 & 2)

[10] No exact temporal window was given by authors, but it can be assumed to range from -73 ms (VA) to +131 ms (AV).

- desychronized by shifts of 40, 80, 120, 280 ms A before V and shifts of 40, 80, 200, 360 ms V before A (18 or 19 items per shift; E2).

In the two monitored time windows N1 (50-150 ms) and P2 (150-200 ms), amplitude and latency were analyzed separately. A repeated measures ANOVA across electrodes showed, among other results, that while no differences in amplitude were found for the A-only versus the AV-synchrony condition, speeded N1 latency was found for the original-AV condition ("RESULTS"). This supports the general assumption that bi- or multimodal communication is more effective than unimodal utterances. No differences in neural response were found among the different A before V asynchrony shifts for the original male voice, but reduced amplitude for the contrasted female voice condition was observed ("RESULTS"); the results of the V before A asynchronies are not discussed. Winter and Müller (2010) concede that "the study could [only] just partly prove a specific neural mechanism for asynchronous audiovisual speech processing" ("CONCLUSION"). However, their methodology is a step into the right direction, especially with the broad width of stimuli reaching from syllable to sentence length and including semantic as well as temporal mismatches. A deeper analysis of the gathered data might reveal interactions between these conditions. A major gain of the experiments presented in Winter and Müller (2010) for the methodology applied in this dissertation, however, is that listeners will integrate voice mismatches unproblematically. This is relevant in so far as two of the studies on speech-gesture perception will make use of stimuli in which the faces of the speakers are blurred our or blocked (see Chapter 5 on methodology) – this visual manipulation should not have too large an effect on the listeners' AVI.

The question arises whether the findings on the temporal limits of the AVI of lip-speech signals described in Massaro and Cohen (1993), van Wassenhove et al. (2007), Massaro et al. (1996), and Müller and Winter (2010), among others, also apply to the AVI of gesture and speech. It has been established that the subjectively perceived audiovisual simultaneity varies across levels of asynchrony and that the circumstances under which one is confronted with stimuli, for example in an experimental setting or in real life, are relevant to integration. Delays as well as advances of the audio or video channels can be integrated by the listener. The visual and auditory modalities of produced syllables are integrated into a fused percept between an audio advance of 30 ms and an audio delay of 170 ms (van Wassenhove et al., 2007). A general AVI of bimodal syllables is possible at asynchronies of ±150 to ±250 ms, while a significant breakdown in the perceptual alignment might be expected between ±250 ms and ±500 ms (Massaro et al., 1996). While gesture and speech have not been proven to be causally related, they are strongly connected temporally as well as semantically in production (see Chapter 2). The windows of AVI found for speech-lip asynchronies might be indicative, even if only tentatively, of which asynchronies between co-produced speech and gestures viewer-listeners will be able to integrate.

4.4 Speech-Gesture Perception

As already noted, while most research in the field of gesture focuses on production, investigating the perceived synchrony of speech and gesture has only recently garnered more attention; within about the last fifteen years there has been an increase of studies on the perception of co-speech gestures, for example by Gullberg and Holmqvist (1999; 2006) and Alibali et al. (2001). The focus in these studies has mainly been on proving that listeners are capable of information uptake from gestures, for instance by showing pictures, cartoons, or even gesture clips before or with speech stimuli to listeners and then questioning them about these. Neuroscientific methods to look into AVI as they have been applied in the context of audiovisual speech perception (e.g., Winter & Müller, 2010; Callan, Jones, Munhall, Kroos, Callan & Vatikiotis-Bateson, 2004) have also been a recent development in gesture studies (e.g., Habets et al., 2011; Özyürek et al., 2007; see Marstaller & Burianová, 2014, for an encompassing overview of the literature). In the following, examples of both groups of methodologies exploring speech-gesture perception, that is, questioning or electroencephalographic (EEG) monitoring, will be discussed with a focus on aspects applicable to investigating conceptual speech-gesture affiliation and the influence temporal asynchronies between the modalities might have on AVI.

4.4.1 Early observational studies

Eye-tracking research, for example by Gullberg and Holmqvist (1999; Chapter 4.2), has shown that listeners will perceive and even fixate speech-accompanying gestures produced by the speaker. An approach toward investigating the informational gain from speech-accompanying gestures is to analyze what happens when speech and gesture contain contradictory information, that is, when they semantically "mismatch". In the common case where both modalities communicate congruent or complimentary information, it is hard to tell whether the listener used both or just one channel to gather their desired information – the verbal channel will possibly be the dominant source in most cases (e.g., Gullberg & Holmqvist, 1999; cf. Winter & Müller, 2010). However, when speech and gesture express differing information, for example with regard to position, shape, or direction, a successive retelling by the listener can give an impression of which information they integrated more deeply. Mismatching as a research methodology is quite straightforward, and it will be discussed in more detail regarding the experiments by McNeill et al. (1994) and Cassell et al. (1999) in the following[11]. It is useful to note here that semantic speech-gesture mismatches caused by temporal shifts is by some considered a separate category of mismatches – since gestures are taken to be utterance-encompassing, holistic providers of information within the context of this dissertation, this distinction will not be made here. Rather, the impact of temporal asynchronies on the general acceptability of the multimodal utterances will be investigated on the level of audiovisual perception.

Speech-gesture mismatch experiments are aimed at showing that gestural information is not only perceived by listeners, but also that information is taken from them. In the studies by Cassell et al. (1999), participants watched video-recordings of one of the male authors retelling narrations from Canary Row elicited according to McNeill and Levy (1993). The re-retellings had been recorded twice so "that 14 target phrases accompanied by gestures were produced once with a normal gesture and once with a gesture mismatched to the content of accompanying speech" (Cassell et al., 1999, p. 8). Semantically matching gestures agreed with the co-

[11] Both publications discuss the same set of experiments and data.

produced speech, while mismatching gestures expressed contradictory information regarding the dimensions of space ('anaphor'), for example pointing in the wrong direction, perspective ('origo'), for example agent versus patient (Example 2), or manner, for example 'beckoning' versus 'grabbing' (pp. 9ff.).

Granny sees him and says “oh what a nice little monkey”. And then she

[offers him a penny].

(a) normal: left hand proffers penny in the direction of listener.
(b) mismatched: left hand offers penny to self.

Example 2: Example of origo mismatch (Cassell et al., 1999, p. 10).

Two groups of listener-viewers were asked to retell 21 utterances in three sets from the videos that included all types of matches and mismatches. In the retellings, which were elicited directly after each set of stimuli, participants conveyed information contained in the narrator's gestures that werenot mentioned in speech as well as vice versa. For mismatches, participants also tried to accommodate the semantic conflict in either or both modalities, for example by mentioning both manners, even if they were contradictory (see Example 3). Regardless of congruent or incongruent information (Cassell et al., 1999, p. 20), a high percentage of gestures was integrated into the listeners' retellings (54 % manner, 50% origo, 32% object).

narrator speech:	“and Granny whacked him one”
narrator gesture:	punching gesture
listener retelling:	“And Granny like punches him or something and you know he whacks him”

Example 3: Example of mismatch accommodation (Cassell et al., 1999, p. 20).

The stimuli used by Cassell et al. (1999) were partly produced spontaneously and partly acted out, but no manipulation of the narration videos took place. While the mismatched speech-gesture stimuli were not naturally co-produced but performed, they were also perceived and integrated multimodally by the listeners and could be used to interpret the naturally co-produced speech-gesture utterances. Gullberg and Holmqvist (1999) commented on this that the experiments by Cassell et al. (1999) provided only “a partial measure of the number of gestures that are. . . integrated into the cognitive representation” by the listener (p. 25) due to the nature of the retellings. Indeed, multiple choice or polarized questions might have been additionally informative regarding the gestural information that might have been integrated by the listeners from the stimuli. However, this would not have been informative regarding natural communicative situations, since conversations usually make do without too detailed questioning for feedback.

The fact that listeners integrated information shown in a video is also useful for further research into speech-gesture perception, because video editing allows for more finely grained stimulus manipulation than instructing actors. Holler, Shovelton, and Beattie (2009) relate to the findings of Gullberg and Holmqvist (2006) for the general visual perception of screen stimuli on varying screen sizes compared with real-life interaction. Their analysis of gaze behavior shows no greater semantics-related differences in the three conditions. Instead, Gullberg and Holmqvist do find that “[f]ewer gestures are fixated on video than live, but [that] the transition mainly affects gestures that draw fixations for social reasons” (2006, p. 76). The indirect methodology used by Cassell et al. (1999) provided first intuitions on what listeners integrate in a speech-gesture conversational setting, albeit by using quite possibly unnatural timed and displayed utterances.

Monitoring listeners with an EEG while presenting them with stimuli is another way to inductively gather information on whether and how they perceive matching or mismatching information from speech-gesture utterance. This methodology has already been briefly discussed above in relation to speech-only stimuli (e.g., Winter & Müller, 2010) and will be expanded upon with regard to speech-gesture utterances in the following Chapter 4.4.2.

4.4.2 ERP studies on gesture cognition

Özyürek et al. (2007) monitored participants for ERP while showing them videos of spoken sentences and accompanying gestures. Their methodology followed McNeill et al. (1994; Cassell et al., 1999; also Holler et al., 2009) in that the stimuli showed an actor performing previously observed iconic gestures. Özyürek et al.

(2007) created semantic mismatches of three different kinds: The verb changed but the original gesture remained, the gesture changed but the original verb remained, or both gesture and verb were changed, but were semantically congruent among themselves (p. 608). The separately recorded gestures were manually synchronized at the stroke with complementing or conflicting verbs within selected sentences “because in 90% of natural speech-gesture pairs the stroke coincide[s] with the relevant speech segment” (Özyürek et al., 2007, p. 610; after McNeill, 1992); some issues with this presumption have been discussed in Chapter 3.3 regarding lexical affiliation. In the stimuli used by Özyürek et al. (2007), the initial part of the sentence served as the prime and the paired prosodic peak (e.g., pitch accent) and gesture stroke as the target for the ERP. At the point of simultaneous exposure, the listeners showed about the same ERP-response to all target stimuli: “In all conditions, the N400 component reached its peak around 480 msec” (p. 612), with or without semantic congruency. The researchers interpreted these homogeneous results to indicate a non-sequential AVI of speech and gesture, that is, that the integration of both modalities might happen in parallel, as it has been found in speech-lip research (p. 613).

The findings by Özyürek et al. (2007) are highly relevant for researching the AVI of speech and gesture in that they further supported that listeners will perceive and process co-speech gestures using a methodology much different from by Cassell et al. (1999; McNeill et al., 1994). As with various other studies, the stimuli, which were recorded using actors, were of non-natural and deliberate speech and gestures, but even artificially incongruent speech and gestures were integrated as if they were congruent. This agrees, for instance, with the findings by Cassell et al. (1999), who deduced that gestures are not only registered by the listener but that even ‘mismatched’ information is taken from them.

Habets et al. (2011) followed up on the experimental setup and findings by Özyürek et al. (2007). They added audio offsets, that is, temporal asynchronies, to the stimuli and expanded on the matter of semantic congruency. Their stimuli representing concrete events, for example connecting, were created by combining video clips with separately recorded lone-standing verbs that had been deemed congruent or incongruent with the gestures by the authors (see Example 4). The channels were either synchronized at the prosodic peak and gesture stroke or the audio was delayed after the video (G before S).

Target Gesture	Target Words	
	Match	*Mismatch*
(1) The two fists are placed on top of each other, as if to hold a club, and they move away from the body twice.	Battering	Hurdling

Example 4: Example of stimulus construct used by Habets et al. (2011, p. 1849).

Across brain regions, the stimuli produced similar results in the participants for the synchronized condition as for when the audio was delayed by 160 ms (GS). The authors concluded from the lack of an N400 effect at an audio delay of 360 ms (GS) that “gesture interpretation might not be influenced by the information carried by speech” (p. 1852). They also claimed that “speech and iconic gestures are most effectively integrated when they are fairly precisely coordinated in time” (p. 1853). The semantic mismatches triggered significantly higher activity, quite possibly due to more complex AVI processes (p. 1851). For combinations of single words and gestures that did not naturally co-occur, the study by Habets et al. (2011) supported the findings by Özyürek et al. (2007) on incongruent speech-gesture signals. Still, the ERP results did not testify to what happens in complete, naturally co-occurring speech-gesture utterances, and the AVI window for single words with gestures might extend to somewhere between an auditory delay of 160 ms and 360 ms (GS). It is also not quite clear from Habets et al. (2011) what happens to AVI when the speech precedes the gesture. The authors deduce that “the interpretation of the gesture was fixed before the speech onset” in their study (p. 1852), which would be difficult if the channels were shifted to S before G.

Özyürek et al. (2007) showed semantic congruency was not a factor when the modalities were synchronized at prosodic peak and gesture stroke onset, even when a contextual sentence preceded the critical stimulus. This is compatible with van Wassenhove et al. (2007), who found only a minimal difference of about 30 ms between congruent and incongruent signals at which an audio advance was integrated. Habets et al. (2011) also investigated “the aspect of semantic integration of gesture and speech” (p. 1846). Since they used artificial speech-gesture pairs (p. 1848), their results can only hint at the integration of naturally co-produced utterances. Also, as in Özyürek et al. (2007), the forced synchrony of the modalities was helpful for an ERP analysis but could have made the stimuli seem even more unnatural. The cutting off of the preparation phase of the gestures could also have influenced their results.

In order to further transfer these findings onto real-life communicative situations, whether with the modalities in their original production synchrony or not, needs to investigate complete, naturally co-produced utterances. The research conducted by Cassell et al. (1999; 1994) and Holler et al. (2009), for instance, has proven the direct as well as the indirect communicative influence of speech-gesture utterances. Özyürek et al. (2007) and Habets et al. (2011) have supported that listeners perceive speech-accompanying gestures using a different methodological approach. The studies conceptualized and conducted within the scope of this dissertation will focus on investigating the relevance of timing in speech-gesture production for the perception of such utterances under the premise that the semantic cooperation of speech and gestures is independent of timing or semantic congruency. Based on the GP theory and how it fits into the cycle of speech-gesture production and reception, the SP hypothesis will be further specified before the background of the previous findings on the perception of speech-gesture utterances in the following Chapter 4.5. The details of the methodologies with which the hypothesis will be tested will be introduced in Chapter 5.

4.5 The Shrink Point

As has been discussed above, multimodal production synchrony is highly relevant for the AVI of speech-lip signals – divergences from the original temporal synchrony will lead either to different percepts, at least on the syllable level (McGurk effect: A 30 ms before V up to V 170 ms before A; van Wassenhove et al., 2007), or to a breakdown of AVI (e.g., stream lagging; A before V or V before A within

±250 ms and ±500 ms; Massaro et al., 1996). The focus in speech-gesture research on production timing and its relevance for communicative efficiency has inspired investigating the relevance of this timing for the listener's perception, that is, how comprehension is influenced by semantic or temporal divergences from the (presumed) original synchrony of speech and gestures. A central phenomenon regarding semantically and temporally coordinated speech-gesture utterances is the GP hypothesis (Chapter 3.2): During the interval of co-production between speech and gesture, speakers most efficiently communicate those parts of a current MU that they want to relate to the listener within a certain context and with a certain communicative intent. For communication to be most efficient, interlocutors need to form a common ground on the basis of shared ideas and communicative goals. For speech-gesture utterances this means that L will have to integrate what is contained in the GP package and then process it into an idea unit maximally resembling the original GP: The Shrink Point (SP). This transmission cycle is repeated at every instance of a rheme, or newsworthy information.

The GP theory has been concerned with the psychological predicate of an utterance, which is not necessarily restricted to phonologically emphasized words but rather encompasses all newsworthy information related by the speaker. Not only due to its literal meaning, the GP has often been taken as residing at the exact coincidence of the prosodic peak of the verbal utterance and the apex of the gesture stroke. This has lead to a lot of research on the so-called lexical affiliation between gesture strokes and the lexical items, that is, words or phrases, they temporally synchronize with partially or fully in production (Chapter 3.3). However, the point, or rather interval, at which apex and prosodic peak coincide is only the gateway to multimodal co-expressivity: The GP unpacks from this point onwards until the gesture retracts. Further, as has been stated by McNeill (2012), GP-unpacking might last up to 2 s, and due to their semantically and semiotically holistic nature, gestures are co-expressive with the full utterance of the current psychological predicate, or even across utterance borders (Chapter 3.3). The timing of speech and gestures in production is additionally influenced by syntactic constraints, motor planning, and other factors (Chapter 2.3). This timing is a crucial trademark of the co-production of speech and gestures but its relevance for AVI or comprehension is currently still unclear. The findings by Habets et al. (2011) that listeners in an experimental setting cannot differentiate between delays of 160ms (G before S) and manually synchronized peak-stroke stimuli, as does data from audiovisual lip-speech perception, suggest a mechanism in the listener that re-aligns and integrates audiovisual signals into their own version of the speaker's utterance. Due to, for instance, the Gricean maxims (Grice, 1975) and Clark's (1975) general idea of a common communicative goal, the listener will strive to reconstruct the MU intended to be related by the speaker in facsimile.

Since the information package of the GP is unpacked and communicated via the speech-gesture utterance, it is likely that certain mechanisms within the listener will re-assemble what they perceived into a perceptual counterpart of the GP. The interval of maximal speech-gesture co-expressivity during the unpacking of the GP is what one might call a blown-up version of the semiotic essence of the GP, which in turn is a sub-concept of an MU. This blown-up multimodal message is then perceived, canvassed, and reduced back to its essence under the influence of L's communicative goal, WM, LTM, etc. – the unpacked GP is shrunk back to its best possible mirror image in the SP. This shrinking is possible (a) through the shared ability of speaker and listener to produce speech and gestures, to communicate in general, and to form an MU, and hence a GP, (b) because of a shared communicative goal, and (c) through the capability of the listener to AVI multimodal signals that are in their original or other production synchrony up to a certain temporal window (see Chapter 4).

S' and L have the common goal of wanting to share the same idea. This might be a detailed description of a travel itinerary, of a painting, or of an experience, or it might be a location that can be pointed at (see also Bühler, 1990). In Figure 5, S' described the granny's arrival using the words “dann <ähm> kommt die omma aber an /”[12] while making a hitting motion with their right hand, which grabs an invisible, stick-like object. Through the gesture, the additional information that the granny is either hitting Sylvester with the umbrella, or pretending to do so, is added to the verbal utterance. By meeting (a), (b), and (c), L is able to integrate

[12] “<ehm> but then granny comes along /”

speech and gesture as belonging to the same utterance and then to initiate a deciphering process as to what information the multimodal utterance might hold. Since spontaneous, idiosyncratic gestures are not contained in a lexicon that is shared by S' and L like speech, their deciphering will probably take more effort than emblems or deictic gestures. Drawing from information previously related by S, and possibly from their own knowledge of the Canary Row series, after perceiving the utterance from Figure 5, L will combine all available information into a mental image of the granny arriving at the scene and swinging her umbrella in a certain manner. Much like the famous de Saussurean dog-example, S' and L will not have the exact same mental image in mind, even if they had seen the exact same stimulus (cf. Harland, 1987/2007, pp. 11ff.; also Gibbon, 2009). Factors such as those investigated by Cassell et al. (1999), for example viewer's perspective, dimensional relations, or direction, are perceived and stored differently, for instance, and S' and L will also differ in background knowledge, expectations, and so on. What will most definitely happen, however, is that L will have a concept in mind that is desired to resemble a maximal version of what S' wanted to relate. It will contain what L perceived and integrated from the multimodal utterances, which includes the GP package, as well as contextual information shared by both S' and L as well as other features from WM and LTM. While S''s communicative intention might be part of L's MU, L's own goals within a communicative setting will also influence which information they focus on during listening. The unpacking of the GP then functions as a kind of perceptual attractor that leads the listener to integrate more than just the verbal message (see also Gullberg & Holmqvist, 1999; 2006). Within L's MU, a perceptual counterpart to the GP explicated by S' is formed that is partially congruent with the GP, but that also contains information from L's experience and communicative goal: **The verbally and gesturally packaged MU of the speaker is, after the unpacking of the GP, re-packaged by the listener into a closely related, modified MU, the SP.** In case of successful transmission, L hopefully gets the point S' wanted to make.

To further expand the model draft of the GP-SP transmission cycle (Figure 12), the temporal relations between speech and gestures have to be integrated, which necessitates the implementation of a mechanisms coordinating this multimodal alignment. An expansion of Levelt's (1989) model of speech production that belong to the group of Postcard Architectures (de Ruiter, 2007) would be highly suitable to to incorporate speech-gesture perception. As has been discussed in more detail in Chapter 3.4 in the context of modeling speech-gesture production, this is, among other reasons, based on the assumption that both S' and L will have the same language processing architecture available to them. While it is not explicitly labeled in de Ruiter's (1999) Sketch Model, temporal coordination between the manual and verbal formulators is achieved by feedback loops between the formulator units. Utterances stopped verbally, for instances, are also interrupted gesture-wise (de Ruiter & de Beer, 2013). This temporal coordination function is highly relevant to the AVI of speech gesture utterances because the content-bearing part of the gphr is in most cases temporally contained in the verbal utterance (see Chapter 2.3). Yet, the gphr is co-expressive throughout the full multimodal utterance, and L's processing mechanism will have to “keep that in mind”.

Another issue regarding AVI is the automatic re-alignment of small temporal discrepancies of speech-lip signals or the mental connection made between a ringing sound and that someone must have pushed the bell button. A model of speech-gesture perception needs to be able to either sort out the modalities as belonging to the same information package or to notice that something went wrong with the production of the perceived signal. Such control mechanisms should be placed between the perception and AVI of the multimodal utterance and the SP. When the temporal re-alignment has taken place, the SP can be formed with the help of the conceptualizer. Integrating these temporal planning and processing mechanisms in the transmission cycle from GP to SP would result in an expanded model draft as shown in Figure 13.This current version of the GP-SP transmission cycle is also able to explain how participants in Cassell et al. (1999) integrated semantically incongruent information into memories of what they had seen and heard in the stimuli. The conceptualizer will attempt to “make sense” of what L perceived from both modalities using immediate and broader contextual information. Through the ongoing communication and the brevity of the existence of MUs, each GP formation and SP integration will be influenced by the situational context and recent events, among other factors. Additionally, parts of LTM will also influence the general

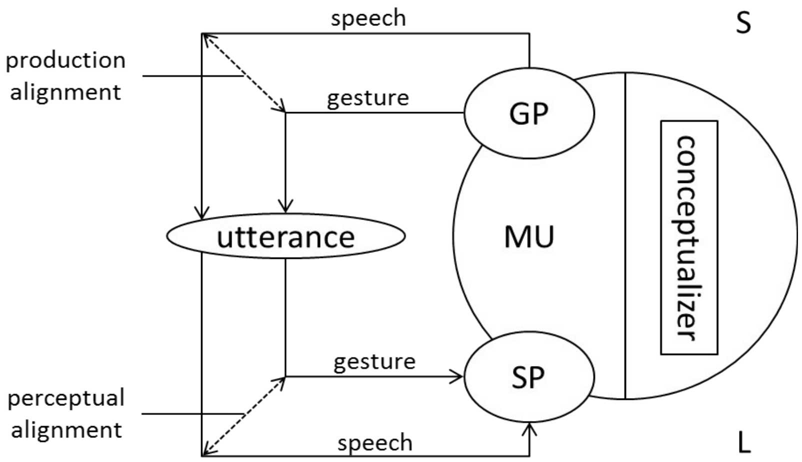


Figure 13: Model draft of GP-SP transmission cycle (including alignment processes).

course of the communicative situation, for example previous knowledge of the Canary Row series in case of the present corpus (Chapter 5.3) or other past experiences shared by the interlocutors. Finally, further uncontrollable factors in any dyadic rapport will provide further narrow and wide context of the current communication, be it linguistic, cultural, situational, etc. A working model of the GP-SP transmission cycle will have to be able to consider such external factors, at least to

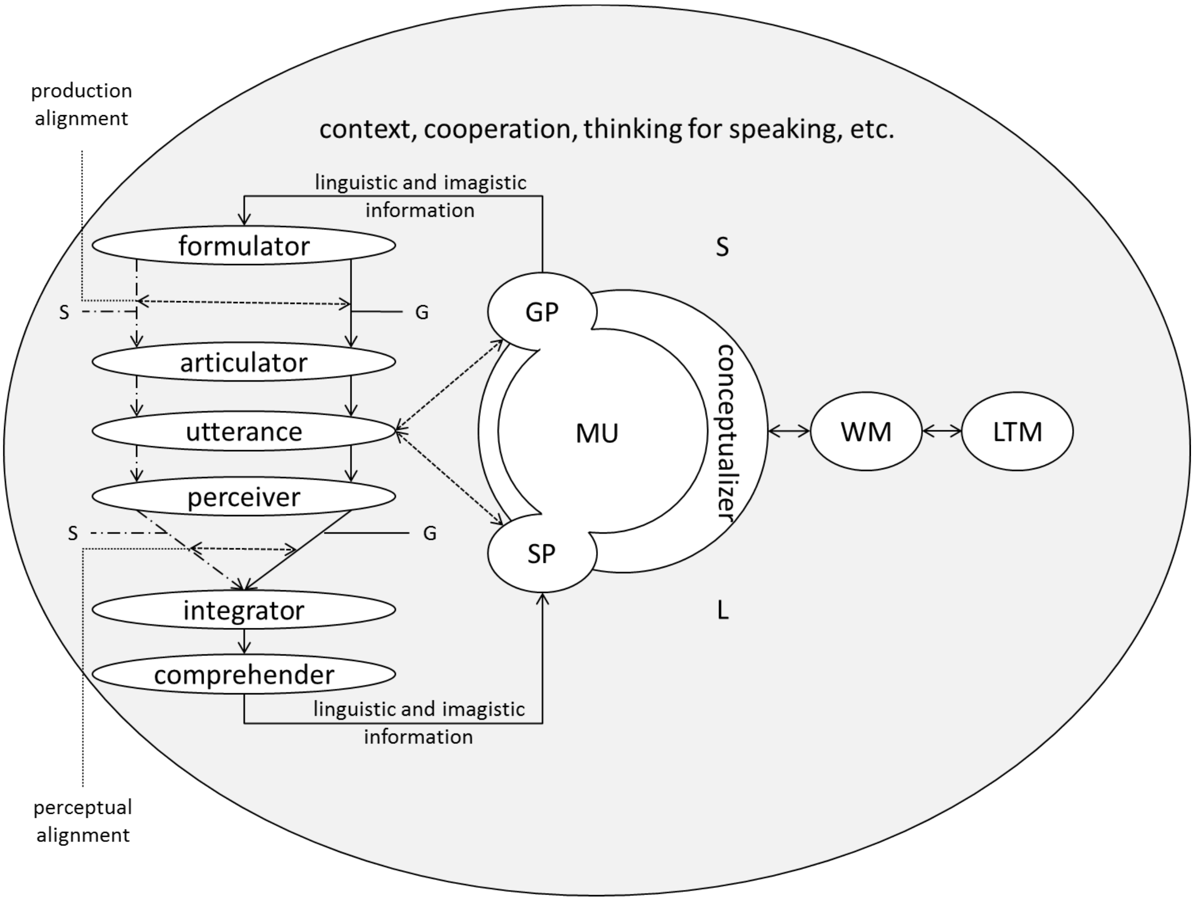


Figure 14: GP-SP transmission cycle (working model).

a certain degree in order to explain any successful or unsuccessful conceptual transfer via speech-gesture utterances. Memory, and WM in particular, is another crucial factor for the conceptual alignment of a shared idea between S' and L via the means of GP and SP. Figure 14 includes these additional factors.

This working model is to be tested for its capacity to explain the transmission of conceptual speech-gesture affiliates in naturally co-produced utterances as well as desynchronized versions of these utterances. Following the experimental examination, more concrete temporal factors regarding the optimal as well as acceptable windows of AVI for the successful integration of speech-gesture utterances can hopefully be added to the model. To what degree the language processor in the listener will be able to AVI temporal asynchronies will be the research objective of the Perceptual Judgment Task (Chapter 7) and the Preference Task (Chapter 8). The methodologies to test the extended model draft of the GP-SP transmission cycle including the factors of timing and context as well as finer parts of production models belonging to the Postcard Architecture (Figure 9) will be discussed in the following Chapter 5.

5 Methodology

Within the scope of this dissertation, the relation of speech-gesture production synchrony to the *perceived* and *preferred* synchrony between the two modalities in the listener will be examined. Regarding this, several aspects of speech-gesture production and multimodal signal perception have been discussed above. Before the theoretical background formed in the previous chapters, several hypotheses were deduced regarding the relation between the naturally co-produced and the perceived temporal and semantic synchrony of speech and gestures.

5.1 Central Hypotheses

Up to now, speech-gesture utterances have mostly been analyzed and interpreted with a focus on the point or time span during which the gesture stroke coincides with the verbal part of the utterance, and especially its prosodic peak. Taking into account the AVI capability of listeners as well as the communicative holistic nature of gestures, the following hypothesis (1) can be formed:

(1) The semiotic-semantic relation between spontaneously co-produced speech and gestures is not restricted to the lexical item(s) of the speech the gesture stroke synchronizes with but encompasses all newsworthy information given in speech.

Assuming that hypothesis (1) could be verified, the central assumption that gestures will precede its co-expressive speech most of the time, and that this is relevant to perception, should be contested. Taking into account the different kinds of multimodal synchrony, I hence propose that

(2) Listeners are able to discriminate variation in the synchrony of spontaneously co-produced speech and gestures and they will prefer a window of AVI encompassing both gestural advance and delay.

Building up on this as well as on the research discussed above, which deduced preferred windows of AVI from listener ratings on a given limited set of asynchronies, hypotheses (3) and (4) are formulated:

(3) Listeners are able to reproduce the synchronization they prefer between speech and co-produced gestures.

(4) The preferred synchrony of speech and gesture in perception will vary from that produced during spontaneous utterances.

Not only does the research reviewed above suggest that the acceptability and preference for temporal relations between speech and gestures might vary from production synchrony, but it also indicates that AVI will be different for the various gesture types, for example due to the disambiguating role of speech accompanying the gestures. It can thus be hypothesized that

(5) The preferred synchrony of speech and gestures will vary for different gesture types as well as for non-speech signals.

Finally, given the intricate connection between speech and gesture in production and their communicative power during their temporal overlap, and keeping the Leveltian speech-comprehension-system within his processing model in mind, I propose that

(6) There is a perceptual equivalent to the Growth Point (GP), that is, the Shrink Point (SP).

With the formation of a model of a GP-SP transmission cycle as a final goal, these central hypotheses will be addressed through the following methodological procedures based on theoretical foundations from the fields of of psychophysics, speech-only, and speech-gesture research.

5.2 From Growth Point to Shrink Point

There are several methodological shortcomings in previous research on the perception of speech-gesture utterances that should be avoided when studying natural communication. For one, lexical affiliation between speech and co-occurring gestures is still assumed by many, and the context in which these affiliates are identified is mostly ignored when analyzing their semantic connection. Language fragments such as syllable-only stimuli, or manually synchronized speech-gesture stimuli with our without matching meaning or speakers were analyzed. For methodological reasons, these stimuli types were fitting for their respective contexts, for example to research fused percepts or ERP responses. In order to investigate how listeners in natural conversational settings perceive audiovisual speech-gesture (a)synchrony, different stimuli are required, that is, naturally occurring and semantically complete utterances. Also, the direction and range of asynchronies between speech and gestures has been rather limited. While gestures have a tendency to begin slightly before their co-expressive speech (e.g., Morrel-Samuels & Krauss, 1992; cf. de Ruiter, 2003), a gesture beginning after speech is also possible – not only with deictic gestures, and especially when videos are played or streamed. A selection of stimuli with multimodal advances and delays will be essential to studying the perception of desynchronized speech and co-produced gestures as well as large enough asynchronies to identify an eventual breakdown of AVI. While Massaro et al. (1996), for instance, already used rather finely grained steps of asynchrony, their testing did not go beyond ± 267 ms for speech-lip stimuli. It is mere speculation whether a breakdown of AVI actually occurs in listeners. Özyürek et al. (2007) as well as Habets et al. (2011) researched the AVI of speech-gesture stimuli for gestural advance only, in few steps of asynchrony and within a small range. No information has yet been provided on audio advance before the gestures. Additionally, research has mostly been restricted to having participants “select” previously lexical affiliates or preselected asynchronies, limiting any findings to a subset of predefined meanings or temporal intervals. More specific constraints on the possible semantic affiliation between speech and gestures and on the windows of AVI for speech-gesture utterances can only be elicited by letting participants define their own preferences (Kirchhof & de Ruiter, 2012). The studies presented in this dissertation will take into account the aforementioned methodological shortcomings by using only naturally co-occurring speech-gesture utterances with sufficient context, by probing previous studies on speech-gesture affiliation, by testing an extended variation of speech-gesture asynchrony that includes audio delay and advance, and by eliciting the temporal preferences of the participants through a rating task as well as through an active resynchronization task.

The necessity of temporal speech-gesture synchrony for comprehension has often been argued for due to its omnipresence in production, particularly because of the assumed 1-n or even 1-1 affiliation of gestures and lexical items. Krauss et al. (1991) and Morrel-Samuels and Krauss (1992) are often cited on this matter. The methodology used by by both groups of authors leading to this assumption of necessity is flawed in several ways (see Chapter 3.3). Foremost, the voting of judges for correct lexical affiliates is against the aim of researching natural or subjective reception. While lexical affiliation might be identifiable without doubt for certain multimodal utterances, the majority of speech and accompanying gestures will be affiliated on a higher level, namely on the conceptual one. Assuming that both S' and L have the same language processing mechanisms, and regardless of whether and how often the formulators or articulators of the two modalities communicate between the planning and production stages, L will, when not distracted, perceive and process the utterance iteratively as well as holistically. Presumably, they will not take the gesture as accompanying only certain lexical items , but audiovisually integrate it with the the meaning of the spoken utterance as a whole. Successively, all information should be sorted out in the conceptualizer, L taking the message from the GP as an SP, integrating it in a new MU (Figure 14). This will be tested experimentally in Chapter 6 to assure that participants will be able to notice GPs in general, and, more specifically, on screen (cf. Chapter 4.4.1). Only then the possibility of a GP-SP transmission cycle can be assumed.

The Conceptual Affiliation Study, as well as the other studies presented in this dissertation, will be conducted with audiovisual stimuli created from spontaneously produced speech-gesture utterances from dyadic conversations. How the corpus used was created will be further explained in Chapter 5.3. The stimuli were later desynchronized temporally, but not mixed, matched, or mismatched semantically. To find the optimal and tolerable AVI windows of speech and gesture, more steps of asynchrony are required. The asynchronies should include delays and advances of speech in relation to gphr in order to explore more possibilities. It is paramount to determine whether listeners are at all sensitive to timing when they perceive speech-gesture utterances because otherwise experiments on the perceived and accepted synchronies by listeners might not produce reliable results.

More information is also needed on the listeners' sensitivity when it comes to the synchrony of speech and gesture in natural communication. This necessitates a methodology using natural, spontaneous language and combining identification or judgment tasks with the participants' ability to reproduce their individual preferences of simultaneity.

In two consecutive sets of studies, naturally co-produced speech and gesture fragments will be examined. The first set of studies is an online Perceptual Judgment Task (Chapter 7) in which speech-gesture stimuli as well as physical event stimuli in seven steps of asynchrony, including audio advances and delays up to ± 600 ms, were rated for their acceptability with varying degrees of head obscurity. The Perceptual Judgment Task was intended to probe the windows of AVI found in previous research and to inform us about the range of asynchronies for the stimuli in the second set of studies. In this set, the Preference Task (Chapter 8), participants had to actively re-synchronize the audio and video channels of selected speech-gesture stimuli as well as of physical event stimuli. Subjective preferences of AVI were elicited using a slider to adjust the synchrony to what they felt was correct. This combination of methodologies should be telling with regard to *acceptable* as well as *preferred* windows for speech-gesture AVI on a continuous scale instead of just ratings of preselected possibilities.

5.2.1 Conceptual Affiliation Study

In order to support the idea of lexical affiliates, that is, the semantic connection between temporally co-produced speech and gestures, Krauss et al. (1991) showed muted speech-gesture clips to participants and asked them for the gestures' meanings (cf. Chapter 3.3). One of the problems with this study was that no context was presented and, hence, the mostly unsuccessful results cannot be considered meaningful. Based on de Ruiter (2000), I propose that there is, in fact, no such thing as a lexical affiliate for every gesture. Rather, gestures should be interpreted by their conceptual affiliates, that is, the MU, GP, or a general mental concept they are semantically, but not necessarily temporally connected with. Hypothesis (1), i.e. that "[t]he semiotic-semantic relation between spontaneously co-produced speech and gestures is not restricted to the lexical item(s) of the speech the gesture stroke synchronizes with but encompasses all newsworthy information given in speech” was formed accordingly. For the purpose of testing this hypothesis, a perception study was conducted with German native speakers. They were asked to link gestures in video clips to their co-occurring speech, which was played in a separate audio clip. In contrast to, for example, Krauss et al. (1991), the participants had the liberty to pick any and all lexical items from the verbal utterance that they felt related to the video including the full gphr.

Analysis

In order to test whether participants agreed upon *the* lexical affiliate or rather on shared concepts, the results were grouped first by literal overlap between the participants' selections, then by semantic units, and finally by psychological predicates on the basis of a scheme-rheme pattern. With the help of color-coded tables, the difference between the grouping approaches were clearly visualized and preferences were calculated.

5.2.2 **Perceptual Judgment Task**

In how far listeners can discriminate synchrony variation in spontaneously co-produced speech and gestures, including gestural advance and delay (hypotheses 2 & 5), an online interface was created in which participants rated the naturalness of originally synchronous as well as of desynchronized speech-gesture utterances on a 5-point Likert scale. As has been done in previous studies (e.g., Gawne, Kelly & Unger, 2009; Oertel, 2010), an online survey was produced using SoSci Survey (Leiner, 2014) that presented the participants with re-playable Flash videos and an adjoining rating scale. Stimuli with different types of gestures were created from Canary Row narrations, which were elicited based on the procedure by McNeill and Levy (1993) and then cut, re-synchronized, and optimized for web play in Adobe Premiere Pro CS5 (Adobe Systems Inc., 2010; Chapter 5.3); non-speech stimuli containing cause-and-effect signals were tested separately using the same online interface (hypothesis 4). A trial run was slotted in before all experiments in order to train the participants and thus make the ratings more reliable.

Analysis

The Perceptual Judgment Task was conducted with three different head-visibility conditions. This way, the relevance of speech-lip synchrony or other facial cues for synchrony and speech-gesture synchrony for the perceived naturalness of the utterances was investigated. Accordingly, one of the independent variables for the statistic analysis was 'visibility', as were the different 'degrees of asynchrony' (Chapter 5.3.6). Repeated measures ANOVAs were run to analyze the effect of visibility, degrees of asynchrony, and gesture type on the perceived naturalness of the stimuli. In a follow-up study, physical cause-and-effect stimuli were rated by participants using the same methodology to provide a base line for their ability to perceive asynchronies. The results were again analyzed for the influence of degree of asynchrony using an ANOVA. Following the completion of the Preference Task, in which the same base stimuli were tested, post hoc analyses were conducted using the Statistical Package for the Social Sciences (SPSS).

5.2.3 **Preference Task**

While the Perceptual Judgment Task investigated the preferences of speech-gesture synchrony in the participants deductively, the Preference Task measured their preferences inductively, also testing hypothesis (3) regarding the listeners' ability to reproduce the preferred asynchronies. Using a synchronization interface, participants manually “reset” the speech-gesture synchrony of stimuli that had been desynchronized by more than 500 ms (hypothesized AVI breakdown, Massaro et al., 1996). Based on the results from the Perceptual Judgment Task, facial movements had been blocked in the stimuli to put a focus on speech and gestures only. The resulting subjective alignments were entered into a calculation program containing the originally entered asynchronies. By subtracting the asynchronies chosen by the participants, offsets were calculated that represented the (a)synchronies preferred by the participants in relation to the original speech-gesture production synchronies. As with the Perceptual Judgment Task, the experiment was repeated for physical-event stimuli to provide a base line. For trial purposes, the initial run of the Preference Task already included a small set of such physical stimuli in order to ensure the participants' ability to complete the task satisfactorily.

Analysis

The statistic analysis was conducted in SPSS. As for the Perceptual Judgment Task, the different 'degrees of asynchrony' (Chapter 5.3.6) were analyzed using repeated measures ANOVAs, but this time as the dependent variable. The effects of gesture type as well as the differences between gestures and physical-event stimuli were analyzed as well. Finally, the results from all sub-studies of the Perceptual Judgment Task and the Preference Task were analyzed post hoc using ANOVAs and MANOVAs with various independent variables including, for example gesture type, acceptable versus preferred audiovisual synchronies, and speech versus non-speech stimuli.

5.3 Corpus

To research how naturally co-occurring speech and gestures in their original as well as in manipulated temporal asynchrony are perceived by listeners, a corpus of natural dialog data was assembled. A large number of, preferably expansive, gestures should heighten the chances of the gestures to be perceived by the participants in the planned experiments. The procedure introduced in McNeill and Levy (1982) has proven to elicit numerous gestures and has been used in a multitude of studies (e.g., McNeill, 1985; 1992; de Ruiter, 1998; Holler et al., 2009). Also for reasons of comparability, a corpus of “spontaneous speech about a predetermined subject” (Gibbon, Winski & Moore, 1997, p. 103), that is, a series of cartoons, was collected accordingly. The technical requirements for the corpus were specific in that high quality stimuli needed to be created from the recordings to be viewed on computers and online. Another one was that the audio channels for S' and L needed to be separate or at least easily separable for later usage desynchronization.

5.3.1 Elicitation methodology

McNeill and Levy (1982) chose the cartoon Canary Row (Freleng, 1950), which is a 7-minute-long series of clips showing the cat Sylvester attempting to catch the bird Tweety using eight different strategies; both characters have certain humanoid traits. As McNeill (2005) says, this cartoon is well-suited for spontaneous speech elicitation because of the following factors: “[L]ittle speech, linear and repetitive plot line, yet varying on the surface from episode to episode, a high concentration of motion events, and brevity” (p. 261; cf. *Elicitation Protocol*). Due to the repetitive baseline of the cartoon, that is, Sylvester failing to catch Tweety eight times, as well as the fast pace with which new attempts and methods for catching the bird succeed each other, speech-accompanying gestural activity was expected to be high. A broad variety of gestures (iconic, deictic, etc.) produced by the speakers was also predicted, which would also make the data usable for a later comparison of GPs and SPs across gesture types (Chapter 8.2).

One has to keep in mind that cartoon logic applies, which is sometimes referred to by the narrators. Also, Tweety occasionally comments on Sylvester throughout the clip series; a frequent catchphrase is “I think I saw a pussycat”. This phrase as well as all other, yet minimal, speech parts in Canary Row are in English. Prior questioning of the recorded participants ensured that this would not be an issue for their comprehension or retelling of the video. The following list roughly sums up what happens in Canary Row in order of appearance; the scene titles are those used in the corpus (time is noted in the format mm:ss):

1. *intro* (00:00-00:35): The Warner Bros. intro with melody; Tweety sings song about himself and a pussycat and swings in a cage; the credits appear.
2. *bird_watchers_society* (00:35-01:23): The scenery blends over to a window with the sign “BIRD WATCHERS' SOCIETY”; Sylvester appears in the window below the sign, takes out binoculars, and zeroes in on an apartment building; in a window on an upper floor is Tweety in his cage, looking back through tiny binoculars; Sylvester tries to get into the building, but in front is a sign saying “NO DOGS OR CATS ALLOWED”; he gets kicked out.
3. *outside-pipe* (01:23-01:58): Tweety swings in his cage while singing “When Irish Eyes Are Smiling”; Sylvester climbs up a rain pipe outside the apartment building and stands next to the cage, pretending to conduct the music; Tweety cries for help and flies out of the cage into the apartment; Tweety's owner, the “granny”, kicks Sylvester out the window.
4. *bowling_ball* (inside pipe; 01:58-02:37): Sylvester scurries toward Tweety's window through the inside of the same pipe he climbed up before; Tweety dumps a bowling ball into the pipe; ball and cat collide somewhere in the middle; Sylvester comes out the pipe with the ball in his belly, rolls down the hill with his head continuously straight up, and lands inside a bowling alley; one can hear the sound of him hitting the pins, but only see the outside of the building.

5. *monkey* (02:37-03:54): Pondering about his next attempt, Sylvester notices a roller organ player with a little monkey wearing a jacket and hat; Sylvester lures the monkey around a corner using a banana as bait, beats him up, puts on his jacket and hat, and pretends to be a monkey; Sylvester climbs up outside the rain pipe again; Tweety flees inside the apartment, Sylvester follows him; when encountering the granny, Sylvester does the “monkey shtick” while searching for Tweety; the granny puts a penny in Sylvester's collecting jar and then hits him on the head with an umbrella; Sylvester tumbles out of the picture with a cony bump raising his hat.
6. *hotel* (03:54-05:04): The check-in desk of the apartment building Tweety resides in is shown; the phone rings, the clerk picks up, and we hear the granny sending for a boy to pick up her bags and bird; Sylvester listens in on the conversation from the pigeon holes behind the desk; the next scene shows Sylvester dressed up like a bellhop in front of the apartment door, knocking; the grandmother peeps out a window above the door, telling him the baggage is right behind the door; Sylvester gets inside and carries out a suitcase and the cage covered with a cloth; he throws away the suitcase and carries the cage to a back-alley; when he removes the cloth, the grandmother is inside the cage, comes out, and hits Sylvester with her umbrella, then chases him down a street continuing to do so.
7. *weight* (catapult) (05:04-05:34): In front of the rain pipe, Sylvester builds a simple seesaw out of a box and lath; he produces a large weight labeled “500 lbs”, which he then uses to catapult himself upwards; he manages to grab Tweety from inside the cage, lands back on the ground and runs away with the bird; soon, the weight hits Sylvester on the head, flattening it out.
8. *rope_swing* (05:34-05:58): Sylvester sketches excessively on a drawing board, checking Tweety's window across the street through a telescope to then readjust his measurements; he then stands on the window latch holding a rope presumably attached somewhere in the middle between the buildings; he swings across and hits the wall right next to Tweety's window, falling to the ground.
9. *streetcar* (05:58-06:52): After pondering again, Sylvester climbs up a power pole opposite Tweety's apartment building; balancing on a web of streetcar wires, he tries to get to Tweety; a streetcar driven by a male driver appears and Sylvester tries to flee from it, occasionally getting an electric shock when he connects with the tap of the streetcar; Tweety and the granny are shown as driving the streetcar, continuing to chase Sylvester along the wire.
10. *credits* (06:52-07:00): The screen zeroes into black, blending over to the classic “That's all Folks” with the Warner Bros. outro music.

5.3.2 Recordings and experimental set-up

The present corpus was recorded in the Natural Communication HD (Nat.CoMM/HD) Lab at Bielefeld University (Figure 15) in October 2010 and encompasses about 133 minutes of audio and video material. 24 pairs of S' and L were “recruited” at the university by approaching random people, posting flyers (see Appendix 11.4), and via buzz marketing.

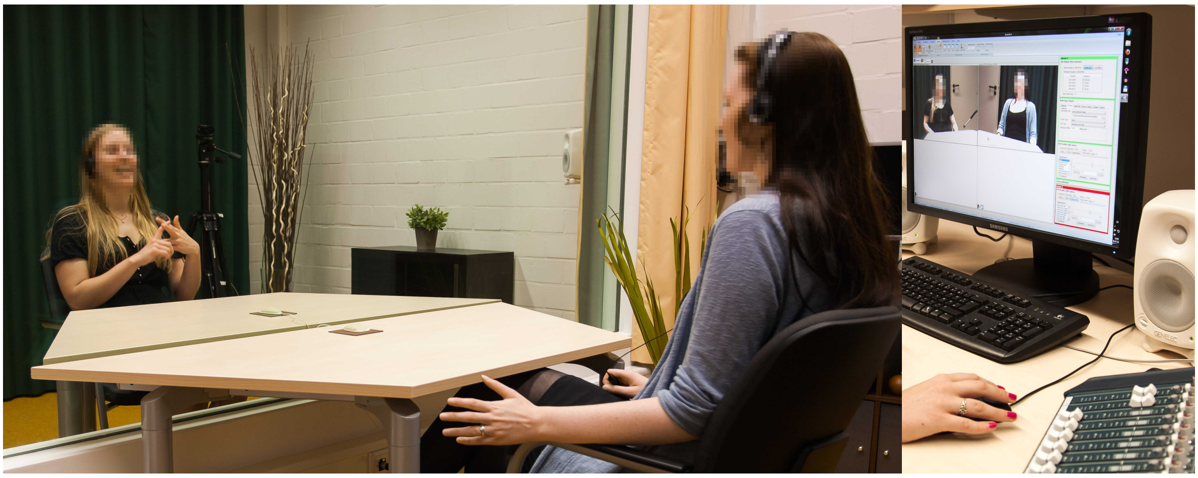


Figure 15: Trial recording at Nat.CoMM/HD lab (de Ruiter, 2012).

S' watched the series of cartoons from Canary Row twice with the instructions to later retell it to L in as much detail and as vividly as possible; S' and L had explicitly been selected as familiar and at ease with each other so that the narration would be as natural and relaxed as possible instead of monotonously monological. S' then retold the story line of Sylvester and Tweety to L, who was instructed to listen carefully, even ask questions, in order to be able to retell the story line afterwards to a third party (which never happened). Both S' and L were told that communicative efficiency would be studied and that videotaping as well as audio recording was necessary for later, more efficient analysis. S' and L were videotaped frontally so that the torso and upper limbs were visible at all times. They sat in two cabins separated by an anti-reflective, sound proof glass pane and connected via a sound system that allowed for separate channel audio recording without cross-talk. A sketch of the recording set-up can be seen in Figure 16, with speaker (S) and listener (L) separated by a plexiglass screen.

Two Prosilica GE4000 cameras (C1 & C2) with 11 megapx resolution, OnSemi KAI-11002 sensor and gigabit Ethernet port were connected to the recording system in a third room and taped both S' and L from the front (205 fps; 5395 kBit/s). The audio (1411 kBit/s; stereo) was recorded using two Shure MX393/O Microflex omnidirectional Boundary microphones (m1 & m2) directed toward each participant. Both channels were recorded using the multi-camera recording software StreamPix 4 (NorPix Inc., 2008). To be on the safe side with regard to audio recording, all cameras were equipped with additional microphones. At the beginning of each recording, S' knocked on the table in case the channels had to be synchronized later on.

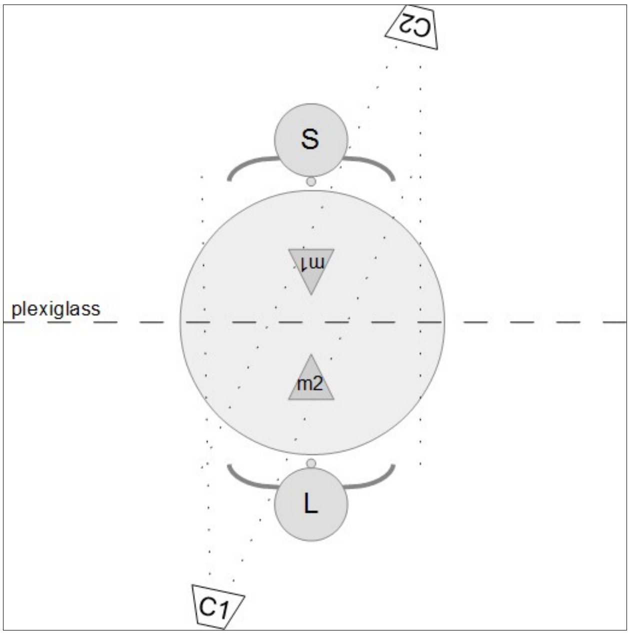


Figure 16: Recording setup.

The participants were not required to wear headphones or microphones during the elicitation, so they were not restricted by any cables or gear and could, theoretically, gesture freely. During the recordings, the participants were observed via a one-way mirror and a PC. This way, unwanted behavior as well as technical issues were detected and corrected early on. Pens and other objects were banned in the recording rooms for the same reasons after one speaker had flicked a pen continuously (recording 14.15.34.268; recordings were saved by the compiling software using the time of compiling as file names, which were kept). To get used to the unusual communicative setting, the participants would converse freely for a couple of minutes before S' was to watch the cartoon as well as afterwards. While the setting was designed to be as comfortable as possible, any effects the knowledge of being recorded might have cannot be excluded. The participants were further informed to appear in subtle, dark color clothing and not to wear scarves or long necklaces (to avoid fumbling). Both S' and L filled in a form (Appendix 11.1) in which they permitted or forbid the usage of their recordings for research and publication after the recordings were completed.

5.3.3 **Participants**

From the 24 S-L pairs, two recordings could not be processed further because of technical issues with the conversion process (10.06.57.995 & 14.01.46.033), and recording 14.15.34.268 was omitted because of the pen problem. In the remaining 21 recordings, all speakers S' (13 women, 8 men, Mage = 25.0 years, age range: 18-32 years) had German as their native language, 6 were left-handed, and 13 had a background in linguistics or the humanities. Among the potential re-tellers L were 14 women and 7 men (Mage = 24.0 years, age range: 19-29 years). All of the participants were healthy individuals without speech or aural impediments.

5.3.4 **Coding**

After converting the recordings using StreamPix 4 into file sizes fit for processing succeeding each elicitation, the video (.avi) and audio (.wav) tracks were trimmed in VirtualDub (Lee, 2010) to contain only the cartoon narrations, keeping the original audiovisual synchrony. The data were then annotated using ELAN (EUDICO Linguistic Annotator; Crasborn & Sloetjes, 2008). As with the elicitation procedure, the coding practice was adopted from McNeill (2005, pp. 262ff.; see also Beattie & Coughlan, 1999) with some modifications because it is (a) widely used across the research community and (b) has been designed for “finding” the unpacking of the GP, which is a prerequisite for investigating the SP. Speech and gestures were annotated on multiple tiers following a pre-defined annotation scheme with linguistic types and lexicons in ELAN (see Appendix 11.1.5). All annotation was done by the author of this dissertation after having been trained at and having practiced coding data stemming from Canary Row narrations in English by healthy and impaired individuals[13]. All recordings were processed through the following passes (cf. McNeill, 2005, pp. 262ff.):

(1) Watch complete narration;

(2) make orthographic, verbatim transcription[14] of speech for S' and L in small letters only, divided into short utterances, that is, units such as sentences, clauses, or intonation units; background noises were also noted (BG); non-speech sounds and meta-notations were marked as follows (cf. McNeill, 2005, pp. 273ff.):

‘/’	unfilled pause
‘<…>’	filled pause
‘#’	breath
‘%’	non-speech sound
‘{…}’	uncertain transcription
‘{… / …}’	alternative uncertain transcriptions
‘*’	speaker self-interruption
‘-’	involuntary break-off
'%ff.'	exhale
pa < a>rk	extended phonation
(…:)	certain manner of speaking, for example (creaky:) or (laughing:)

(3) label story parts in separate tier (scenes);

(4) annotate speaker gphr in additional tier (“S_gphr”)[15];

(5) review recording and coding and check for consistency and errors.

Any comments were noted in a separate tier (notes). In further annotation cycles, potential stimuli for the experiments were selected and respectively marked:

[13] During a 3-week class by S. Duncan during the LSA summer session at Berkeley in 2009, several visits to the McNeill lab in Chicago in February 2010, and during a 3-month research stay at the McNeill lab in 2011.

[14] A phonetic transcription was of no concern for later analysis within this dissertation.

[15] Earlier parts of the corpus show remnants of former annotation schemes, as, for example, described in McNeill (2005, pp. 273ff.).

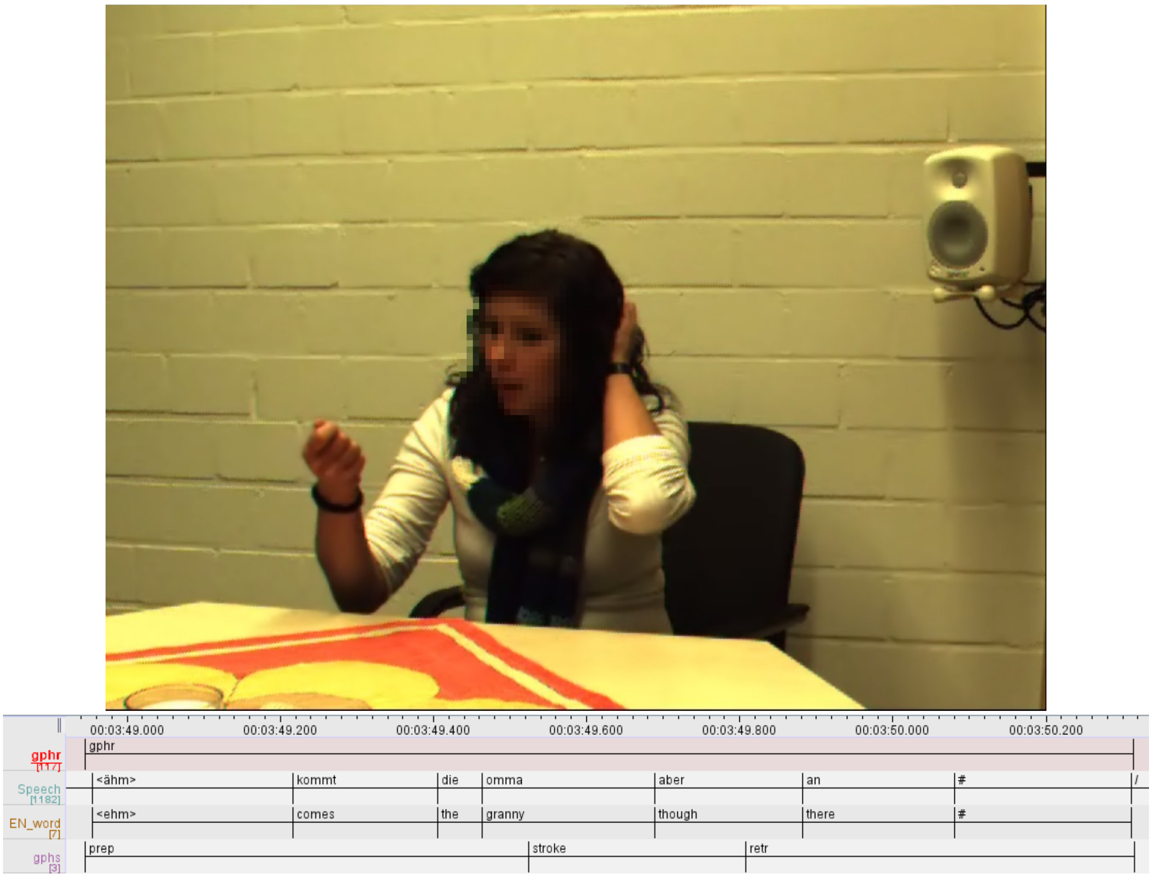
(6) label gphr suitable for experiments in additional tier (tiers: “for desync”; “for study”);

For stimuli used in presentations, publications, or exemplifications, additional annotations were made:

(7) annotate gesture phases (prep, stroke, (hold,) retraction) for selected gphr in additional tier;

(8) add English verbatim translations (S_EN_word) in additional tier (optional).

An example annotation of an extract taken from a narration of Sylvester's bell-hop attempt can be seen in Figure 17:



	00:03:49.000	00:03:49.200		00:03:49.400	00:03:49.600	00:03:49.800	00:03:50.000	00:03:50.200
gphr [117]	gphr							
Speech [1182]	<ähm>	kommt	die	omma	aber	an	#	/
EN_word [7]	<ehm>	comes	the	granny	though	there	#	
gphs [3]	prep			stroke	retr			

Figure 17: Screenshot of Canary Row narration 10.17.48.959 (gphr 132; beating with an umbrella).

5.3.5 Data description

21 of the 22 narrations had an average duration of 05:43 minutes (SD = 01:10; range = 03:43-08:16 min); the 22nd recording stopped at about 01:26 min, but the recorded material was processed regardless. While 3 S's narrated the cartoon with mostly folded hands and minimal gestural activity, the remaining S's gestured as

had been expected based on previous research. For all 22 S, 1329 gphrs were identified, regardless of type (for a list see Appendix 11.1.6). S's produced 63.24 gphrs on average (SD = 61.20), the gphr having an average duration of 0.9917 s (SD = .574) from the gesture onset to the conclusion of the retraction phase (see, e.g., Figure 17).

Since the corpus had been collected specifically for the purpose of creating stimuli for the perception experiments presented in this study, it was only annotated and analyzed in more detail selectively. Word-by-word transcription as well as gesture type identification was mostly only done for those utterances selected as stimuli for the experiments, so no gesture-word-ratio can be determined for the full corpus at the moment. A more thorough codification of the data in the future is highly desirable, as is a more detailed analysis of the data, particularly with regard to speech-gesture production synchrony in the context of rhythm in general and rhythmic syllable-stroke correlation in specific (see Chapter 7), for instance.

5.3.6 Speech-gesture stimuli

For the experiments conducted within this dissertation, parts of the narrations were extracted from the recordings using the video editing software VirtualDub, and later Adobe Premiere Pro CS5 and compressed and reformatted for the (online) experiment interfaces. As has been discussed above, “[a] large number of, preferably expansive, gestures should be elicited to heighten the chances of the gestures being perceived by the participants in the planned experiments” (p. 92). A number of stimuli was also to be used to test conceptual affiliation, so the speech and gestures had to be high in imageability. The selection of extracts to be later transformed into stimuli was chiefly made according to the following criteria:

- *position*: Gestures executed within the center-center, center, and periphery of the gesture space (Figure 18).
- *size*: Gestures executed with a certain degree of velocity that involved more motion than, for example, finger lifting, that is, a change of position of either limb of at least about 5 cm.

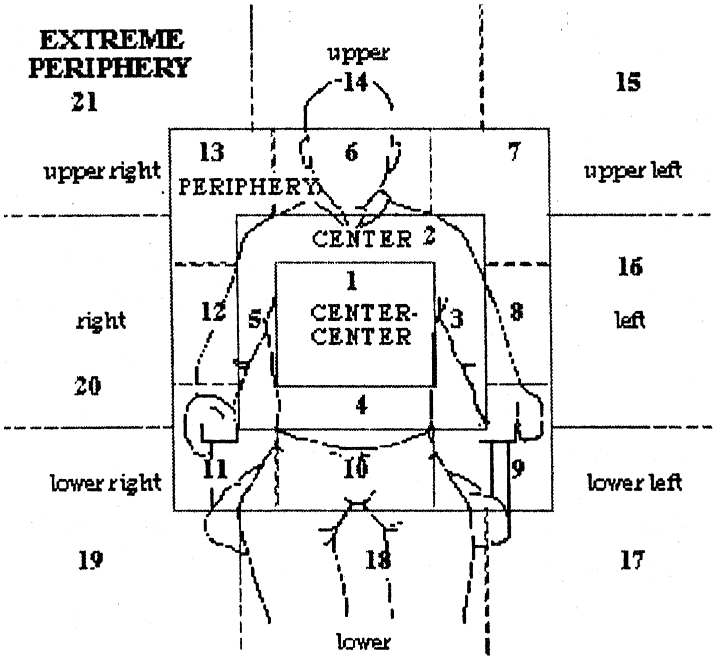


Figure 18: Division of gesture space according to McNeill (2005).

- *gesture type*: While the corpus had not been annotated with gesture types at this stage, the selection of stimulus material regarded only gestures that were potentially iconic, deictic, or emblematic. While some of the selected gestures might show superimposed beats, no pure beats were selected due to their (possiblly) deliberate nature, which contradicts the aim to research spontaneous gestures.
- *imageability*: Gestures that are high in imageability (cf. Beattie and Coughlan, 1999, on words with this trait) are highly iconic with regard to the concept they express and mostly complementing rather than redundant to the speech they are co-expressed with.
- *variety*: Next to the first four criteria, variety in form and content was important in the selection so that stimuli could be created that would be distinguishable by the participants and that also reflected the broad spectrum of gestures produced in the corpus.

Applying the criteria of position, size, gesture type, imageability, and variety resulted in the following selection (Table 1):

Table 1: List of utterances for stimulus creation.

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
1	banana_1_0	11.00.31.621	1482	230	dann lockt er den affen mit 'ner (breathy:) banane	iconic
2	bino_1_0	16.11.09.878	1635	805	halt mit'm fernglas durche gegend kuckt	iconic
3	binoculars	10.17.48.959	1870	9	als erstes nimmt sylvester einen feldstecher und kuckt #	iconic/ pantomime
4	bird_1_0	16.36.00.692	4383	835	und in dem film geht es da<a>rum dass sylvester scharf auf <äh> den vogel is	deictic/ trace
5	button_rows	14.27.42.306	1990	1282	sylvester öffnet die tür # in seiner pagenuniform vo* so ne rote mit goldenen knöpfen	deictic/ trace
6	cage_1_0	10.17.48.959	915	31	also er macht den käfig auf	iconic
7	cage_2_0	11.00.31.621	1284	272	un' dann haste den käfig da steh'n	trace
8	can_1_0	15.04.57.785	2502	681	der rennt mit de<e>r geldspendebüchse rum /	iconic
9	cat_1_0	16.36.00.692	2230	834	und sylvester is die große schwarze katze	iconic
10	catapult_1_0	14.27.42.306	3001	1304	daraufhin / wird er in die luft katapultiert	iconic
11	climb_1_0	11.00.31.621	1400	200	un klettert da ers' rau<u>f / #	iconic
12	cover_1_0	11.00.31.621	1271	290	# will dann die decke runtermachen	iconic
13	directing	14.27.42.306	1543	1244	son kleines imitiertes dirigieren /	iconic/ pantomime
14	discover_1_0	12.05.31.682	1893	421	und wie er dann<nn> tweety in seinem käfig entdeckt	deictic
15	elevator_1_0	14.27.42.306	1322	1285	dort geht es # links zum elevator /	deictic
16	everywhere_1_0	14.27.42.306	2510	1254	#%/ er fängt dann an überall nach tweety zu suchen #	iconic/ deictic
17	hat_1_0	15.04.57.785	2656	685	ja und dann zieht er den hut so höch und dann erkenntse dass (laughing:) s ne katze is	iconic
18	hit_1_0	15.04.57.785	2428	707-709	und haut ihm mit dem (laughing:) regenschirm wieder fleißig übern detz %	iconic
19	in_pipe	14.27.42.306	1180	1248	# <ähm>/ so ne rostige regenrinne	iconic/ deictic
20	in_pipe_1_0	10.17.48.959	938	76	ja ja er is in dem regenrohr	iconic
21	kicked_out_1_0	14.27.42.306	2510	1247	%<ähm>/ sylvester fliecht sofort wieder raus	iconic/ deictic

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
22	knock_1_0	11.00.31.621	2884	267	wo er dann als roomboy verkleidet is un' anklopft #	iconic
23	lift_hat	14.27.42.306	1750	1263	/%'<ähm>/%/ er lüftet dankend / den hut	iconic
24	opposite_1_0	11.17.45.463	2423	381	also das war'n so zwei hochhäuser an so / auf so ner stra<a>ße	deictic
25	organ_tot	10.17.48.959	2400	35	dann beispielsweise hört er eine <hmm> mundorgel spielen	iconic
26	penny_1_0	10.17.48.959	1551	51	dann ja hier's is 'n penny oder	iconic
27	penny_can	14.27.42.306	868	1252	der eine dose hält für die spenden	iconic/ pantomime
28	pipe_1_0	10.17.48.959	2530	19	er <n> klettert das abwasserrohr hoch	iconic
29	ring_1_0	11.00.31.621	1079	248	<mm> also das telefon klingelt	deictic
30	ring_2_0	11.00.31.621	760	370	un' dann klingelt's	iconic
31	rub_1_0	10.54.29.104	1452	143	(whispered:) da (creaky:) drüben (creaky:) isser (creaky:) endlich der leckere vogel #	iconic
32	shelf_1_0	10.17.48.959	1114	90	er sitzt im regal #	deictic/ iconic
33	sign_1_0	13.09.12.480	2559	593	draußen aufm schild steht <äh> hunde und katzen verboten #	emblem
34	sign_2_0	16.11.09.878	1416	812	steht auf so'm schild neben der tür	deictic/ iconic
35	sill_1_0	16.11.09.878	1839	815	steht halt direkt immer am fenstersims #	deictic/ iconic
36	street_1_0	15.04.57.785	2604	677	und rollt auf der bowlingkugel ne abschüssige straße runter	deictic
37	swallow_1_0	10.17.48.959	1546	62	/ er schluckt die kugel # %	iconic
38	swing_rope	14.27.42.306	1522	1315	sich rüber zu schwingen / auf <äh> tarzanmanier #	deictic
39	thumbs_up_1_0	Becker (2012)	243	n/a	klasse	emblematic
40	trace_1_0	10.54.29.104	1603	151	geht dann in das and're<e> haus rein	deictic/ trace
41	umbrella	10.17.48.959	1371	32	dann <ähm> kommt die omma aber an /	iconic
42	weight	14.27.42.306	2201	1301	und schmeißt eins von diesen / trapezförmigen <ähm>/ gewichten / auf die andere seite #/<ähm>/	iconic
43	whyever_1	16.36.00.692	923	906	warum auch (laughing:) immer %laugh	emblematic

A total of 43 multimodal utterances were chosen, including the classic “thumbs up” emblem taken from an online video. The selected utterances have an average duration of 1825.19 ms (SD = 763.39 ms), containing gphr with an average duration of 1816.02 ms (SD = 770.53 ms). They were trimmed to include enough speech for the utterance to make sense as well as only one fully executed gphr as well as enough audio and video buffer before and after the selection. Since the procedure was done in Adobe Premiere Pro CS5, the selection was not final and was expanded within the process of stimulus creation. How the extracts were manipulated to create stimuli for the different experiments will be explained in more detail in the respective materials sections of the the Conceptual Affiliation Study (Chapter 6), the Perceptual Judgment Task (Chapter 7), and the Preference Task (Chapter 8).

5.3.7 **Physical stimuli**

To create stimuli of physical cause-and-effect events to be used in the Perceptual Judgment Task (7), and the Preference Task (8), 10 videos with an average duration of 2251.67ms (SD = 954.60) were recorded (see Table 2): Snapping a book shut, a clap of the hands, clinking a class with a fork, a tap on a keyboard, knocking on a table, the plop while opening a bottle of champagne, a hammer hitting a nail, fingers snapping, hitting a bass drum, and popping a balloon with a needle. Each stimulus created from these recordings contains exactly one event with only one sound and one cause.

Table 2: Cause-and-effect events for the creation of the physical stimuli.

id	stimulus	dur (ms)	id	stimulus	dur (ms)
0	book closed	1170	7	fingers snap	5070
1	hands clap	3060	a	drum stroke	940
2	glass clinked	1080	b	balloon popped with needle	5240
3	key pressed on keyboard	2070			
4	knock on table	3110			
5	sekt pop	3020			
6	hammer hits nail	1410			

As with the speech-gesture corpus, the physical cause-and-effect event recordings had separate audio (.mp3; stereo; 216 kBit/s; 48 kHz) and video (.mp4; 25 fps; 1449 kBit/s) tracks to allow for later desynchronization in Adobe Premiere Pro CS5. Again, how they were manipulated will be discussed in the materials sections of the chapters pertaining to the experiments.

6 From Lexical to Conceptual Affiliation

6.1 Introduction

As has been discussed in Chapter 3.3, a lot of researchers have set out to find 'the' lexical item affiliated with a co-occurring gesture (e.g., Krauss et al., 1991; de Ruiter & Wilkins, 1998). While this may be easy for utterances such as “Look over there,” and simultaneous pointing at something, it gets complicated when trying to figure out what A meant by “flicking water” down with her right hand while saying, “The yard looked so beautiful” (Kirchhof, 2011). The seminal research by McNeill (1985; 1992; 2005) and Kendon (1972; 1980; 2004), among others, shows that gesture and speech are intrinsically connected. The two modalities are co-expressive not only in that they share meaning, as, for example, when we say “up” and point a finger upwards. Both speech and gesture bring individual information together at the moment of expression, at the “unpacking” of the GP (McNeill, 1992; see Chapter 3.2). However, pinpointing the meaning of a gesture at the moment of its articulation is still under debate. With deictic utterances, it might seem obvious to determine “the word or phrase with which the gesture is semantically and pragmatically linked” (de Ruiter & Wilkins, 1998, p. 605). Yet iconic gestures, that is, those re-enacting something or “painting a picture”, are often impossible to connect with just a single word or phrase. Hypothesis (1) stated that “[t]he semiotic-semantic relation between spontaneously co-produced speech and gestures is not restricted to the lexical item(s) of the speech the gesture stroke synchronizes with but encompasses all newsworthy information given in speech” (p. 85). In the following, various angles on the semiotic-semantic relation between spontaneously co-produced speech and gestures will be discussed and then tested experimentally.

In order to support the idea of lexical affiliation, Krauss, et al. (1991) showed soundless gesture clips to participants and asked them for the gestures' meanings. One of the problems with their study was that no context was presented and, hence, the mostly unsuccessful results cannot be considered reliable (cf. Chapter

3.3). Expanding on de Ruiter (2000), I propose that there is, in fact, not one single *lexical* affiliate for every idiosyncratic gesture. Rather, I suggest interpreting gestures by *conceptual* affiliates. The finding of these will take into account (1) the discourse type, (2) the discourse topic, (3) the relationship of the interlocutors, (4) the background of the interlocutors (SES), (5) the immediate context of the gesture, and (6) the type of gesture (deictic vs. iconic, for now). 10 participants decided on the lexical affiliates in Krauss et al. (1991), which were then the foundation for a set of perception experiments. The participants had transcripts at hand and the option of discussing the possible affiliations among themselves. Also, the decision process for *the* lexical affiliate for each gesture was driven by the shared goal to fully agree. This might have lead to a somewhat standardized subjective perception in the group (within-group variation was not commented on by the authors). In order to test a hypothesis of lexical affiliation, no agreement should be forced or intended from the beginning. Another aspect to be kept in mind is that temporal synchrony and lexical affiliation often go hand in hand in the interpretation of speech-gesture semiosis. Seeing lip movements and facial expressions in any stimulus, even if sound and video are separated, will lead people to look for a connection; Krauss et al. (1991) did not avoid this issue either. The present study had participants observe speech and gestures without obvious synchrony and without restrictions on the speech counterparts to be chosen for the gestures, which was intended to ensure full variety or unity in the perception of lexical affiliates. In a previous trial study, participants should identify lexical affiliates from three possible co-occurring sentences after they had watched gesture clips without sound. This turned out to be unsuccessful because (a) the situation felt too unnatural to the participants and (b) the distractor sentences were perceived as suitable as the original ones for the gesture. These results lead us to further extend the context in the stimuli further. In the study, speech and co-occurring gesture were played in direct succession. Participants noted down the lexical items that in their opinion were most connected to the meaning of the gesture.

6.2 **Participants**

18 native speakers of German (14 women, 4 men, M_{age} = 26.22 years, age range: 23-35 years), either studying or working at Bielefeld University, voluntarily took part our the study. Neither of them had c.p. and some of them had corrected eye vision. The participants were promised neither credit points nor financial reward and they could always ask for specifications; no pressure, such as time or performance requirements, was put on them. A researcher supervised the participants at home or at university throughout the study.

6.3 **Materials**

A selection of 10 fairly large iconic (imagistic[16]) gestures from the Canary Row Corpus (5.3) was made with a focus on the size and vividness of the gestures (Table 3). The stroke phases (with some ms around it for smoothness) were transformed into silent standalone DivX video clips (MPEG-4) in the dimension 640x480 px with VirtualDub. The video clips have an average duration of 1830 ms, one verb in about 8.45 words, and one gphr with an average duration of 1669.5 ms (SD = 469.38). The whole sentences or clauses that were the original co-occurring speech, including sufficient contextual information, were transformed into uncompressed wave files (16-bit PCM, 44100 Hz) with the public license software Audacity (The Audacity Team, 2011).

Table 3: Utterances used as stimuli for the Conceptual Affiliation Study.

clip	**stimulus**	**recording**	**dur (ms)**	**gphr**	**speech**	**gesture type**
1	binoculars	10.17.48.959	1870	9	als erstes nimmt sylvester einen feldstecher und kuckt #	iconic/ pantomime
2	umbrella	10.17.48.959	1371	32	dann <ähm> kommt die omma aber an /	iconic
3	organ_tot	10.17.48.959	2400	35	dann beispielsweise hört er eine <hmm> mundorgel spielen	iconic
4	directing	14.27.42.306	1543	1244	son kleines imitiertes dirigieren /	iconic/ pantomime
5	in_pipe	14.27.42.306	1180	1248	# <ähm>/ so ne rostige regenrinne	iconic/deictic
6	penny_can	14.27.42.306	868	1252	der eine dose hält für die spenden	iconic/ pantomime
7	lift_hat	14.27.42.306	1750	1263	/%'<ähm>/%/ er lüftet dankend / den hut	iconic

[16] cf. Kendon (2004, p. 99).

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
8	button_rows	14.27.42.306	1990	1282	sylvester öffnet die tür # in seiner pagenuniform vo* so ne rote mit goldenen knöpfen	deictic/trace
9	weight	14.27.42.306	2201	1301	und schmeißt eins von diesen / trapezförmigen <ähm>/ gewichten / auf die andere seite #/<ähm>/	iconic
10	swing_rope	14.27.42.306	1522	1315	sich rüber zu schwingen / auf <äh> tarzanmanier #	deictic

The verbal part of of clip pair 8 is shown in Example 5. It is the same gphr as in Example 1 (p. 39), but this time with sufficient contextual speech. The original gphr includes the preparation and retraction of the stroke (dur = 2688 ms); the cut video clip contains the stroke phase of the gesture and tolerance measures (dur = 1990 ms). The speaker traces the position of imaginary buttons on a double button row in a zig-zag motion on his chest with claw hands while his palms face the chest (Figure 19). Figure 19 shows gphr 1282 naturally co-produced with the speech from Example 5 from onset to retraction. No other gesturing happened during this utterance. The speaker had his hands folded on the table before and after.

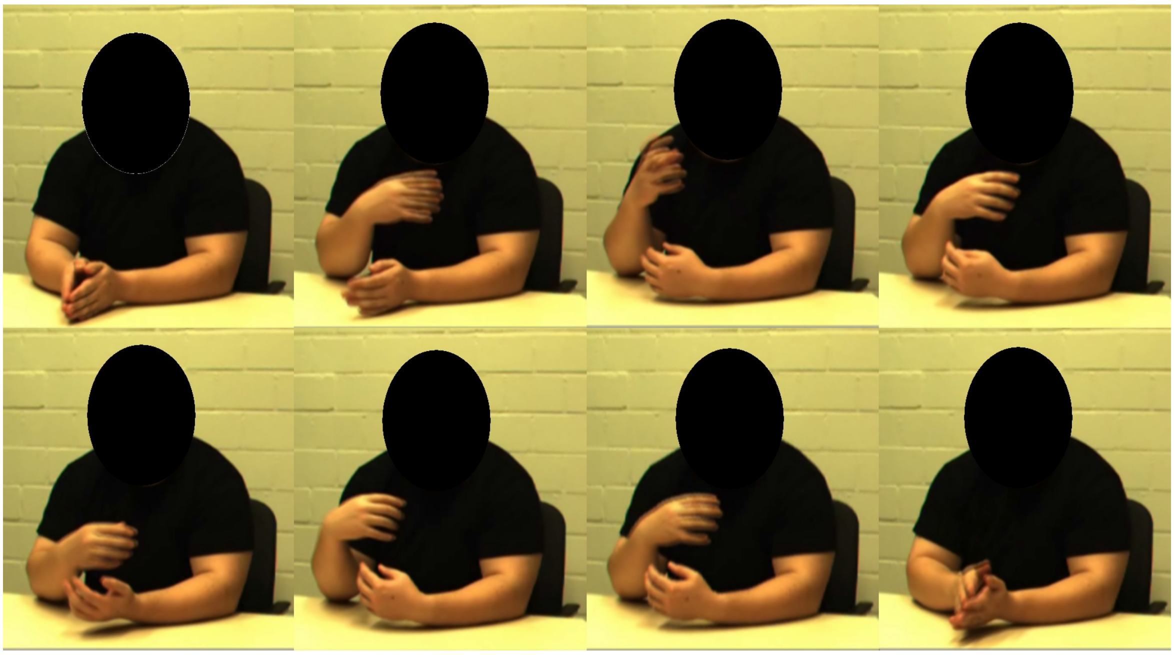


Figure 19: Gphr 1282 from Example 5.

The accompanying speech in the audio clip (dur = 664 ms) includes a breath pause “#” and the unfilled pauses “/” . The bold font in the transcript indicates sentence stress, the square brackets the position of the stroke phase. Because of the different lengths of the audio and video clip, participants could not perceive this arrangement directly.

sylvester **öf**fnet die tür # in seiner **pa**genuniform /

sylvester opens the door # in his bellhop uniform /

so n**[**e **ro**te mit goldenen **knöp**fen**]** und so /

such[a red one with golden buttons]and stuff /

Example 5: Simplified transcript of stimulus “button_rows” (clip 8).

6.4 Procedure

The participants sat in front of a notebook (1280x800 px resolution) and wore closed headphones (Sennheiser HD 201). They had mouse control over a folder containing the 12 audio-video file pairs. The participants had two contrary clip orders among them so any influence of sequentiality was balanced out in the study. They could regulate the volume but screen contrast and brightness was constant. This way, sufficient detail and visibility of the gestures were ensured. The faces in the stimuli were covered so that the participants were not able to read lips or gaze; anonymity of the recorded was given as a reason for this. The participants were informed about the source of the clips, namely Canary Row retellings, and also, if necessary, were explained the general course of events in these cartoons. They were asked to watch and listen to the clips with corresponding file names (e.g., “01.avi” and “01.wav”) as often as they liked, and they controlled the frequency themselves with the PC mouse. After having watched a stimulus pair, the participants were asked note down the word, words, phrases, or parts of words in either position of the utterance which they thought was or were connected with the gesture in meaning in a pre-numbered form (Appendix 11.2.2) that also included these instructions. In a second run, they were told to underline those parts of speech they had selected to verify their perceptions. The average session lasted about 15 minutes.

6.5 Results

Up to 14 dissimilar affiliate tokens were chosen by the participants for any clip-pair. In this context, a token is a word combination or word that does not occur in exactly the same form with another participant. Example 5, for instance, reached this maximum number. On average, 7.75 different affiliate tokens (median = 9) were noted down by the 18 participants for any stimulus pair. This variation does include minimal differences such as “trapezförmig” (trapezoid) versus “trapez” (trapeze) or “schwingen” (swing) versus “rüberschwingen” (swing across), which were counted as four separate tokens initially (cf. Table 5).

For further analysis, the tokens were grouped into affiliate types, taking into consideration word stems, optional pronouns, etc. The two example pairs from the previous paragraph would now be in two affiliate types. On occasion, one token had to be sorted into two types because they included two differing features, such as action and shape. The sorting resulted in a reduction of differing “lexical affiliates” to 4.7 affiliate types per stimulus (median = 4). For example, for the stimulus “lift_hat” (clip 7), the affiliates were reduced from 10 tokens to four types via this process. The core lexemes of these types were “dankend”, “Hut”, “lüftet”, and “lüftet den Hut”. For this stimulus, the tokens were grouped into these four types because the emphasis in the affiliates were either on thankfulness, the object of action, the action, or on all at once. This variation is, however, still far from a unison decision on lexical affiliates as presented by Krauss et al. (1991). This might be due to the participants making their associations independently in the study.

From the viewpoint of co-expressivity and the McNeillian imagery-language dialectic (McNeill, 1985; 1992), a more homogeneous grouping of the participants' speech-gesture affiliates is still possible when considering *conceptual* overlaps instead of lexical or grammatical commonalities. For instance, the lifting of the hat in clip 7 was lexically connected to either the hat or the lifting by several participants. The idea that unites them all is the action of lifting the hat – the concept that is both expressed in the speech and in the imitative gesture. Sorting all types and tokens of the respective stimuli by concept resulted in an average of 2.75 *conceptu-*

al affiliates (median = 2). Table 5 shows a distinct reduction from lexical to conceptual affiliate using the stimulus “weight” (Example 6).

/ und schmeißt **ein**s von diesen / trap**ez**förmigen ähm ge**wich**ten
/and throws one of these / trapezoid uhm weights

on[to the other side] #/
a[uf die andere **se**ite] #/

Example 6: Simplified transcript of stimulus “weight” (clip 9; gphr 1301).

In Example 6, the gesture stroke synchronizes with “auf die andere seite”. Both hands in chest height, the palms facing each other chest-wide apart throughout, the fingers fanned a bit – the hands tilt forward and freeze half way to the table. The configuration stays the same through the unfilled pauses and then the hands are folded to rest. Table 4 shows the participants' individual perceptions of the relation of the gesture to the utterance. The parts they felt were most related to the meaning of the gesture are listed in alphabetical order in the first column “What people assigned”. Instances of “/” in the table mean that the participant gave no answer. The second column gives a rough English translation of the first column. In the “different affiliate tokens”, the participants' selected affiliates are represented in a clearer form in that each minimally differing selection categorized in alphabetical order.

Table 4: From lexical to conceptual affiliate in Example 6.

What people assigned	EN equivalent	different affiliate tokens	affiliate types	conceptual overlaps
/	/	a	a	a
auf die andere Seite	onto the other side	b	b	b
auf die andere Seite	onto the other side	b	b	b
diesen	those	c	c	c
Gewicht	weight	d	d	g
Gewicht(n)	weight(s)	d	d	g
Gewichten	weights	d	d	g
Gewichten	weights	d	d	g
Gewichten	weights	d	d	g
schmeißt, Gewichte	throws, weights	e	d,j	g
Trapez	trapeze	f	g	g
trapezförmig	trapezoid	g	g	g
trapezförmig	trapezoid	g	g	g
trapezförmig	trapezoid	g	g	g
trapezförmigen	trapezoid	g	g	g
trapezförmigen	trapezoid	g	g	g
trapezförmigen Gewichten	trapezoid weights	h	d,g	g
und schmeißt	and throws	j	j	g

One participant, for example, chose “Gewicht” as related semantically to the gesture in the clip they saw. As the fourth new lexical item, it was categorized as (d) in column three. The same label was assigned to all inflections of “Gewicht”, such as its plural “Gewichten”. Any varying lexeme or combination of words was labeled differently: “schmeißt, Gewichte” (e). Here, the participant quite possibly linked object and action to the gesture differently. Answers (e) and (h) demonstrate this perceptual difference perfectly as the participants found both features noteworthy. This also results in (e) constituting a combination (d, j) in the 'affiliate type' column. Other participants chose one side of things only. The variable (b) groups those affiliates relating to position, (d) to the weight, (g) to the shape of the weight, and (j) to the action of throwing; (e) is a combination of (d) and (j) already. Looking at the groupings of the selected tokens, lexical agreement alone cannot explain a common comprehension of the gesture.

There is, however, a large conceptual overlap within all stimuli of this study (rightmost column). While participants favored one lexical affiliate over another, the image they perceived and then tried to connect to the utterance was the same: A trapezoid weight, the rheme of the utterance, the newsworthy content (cf. e.g., McNeill, 2005). When taking the missing answer (a) out of the calculation, the rest adds up to a conceptual agreement of 82.4%. This is far more than either affiliate token or type could supply. A different grouping of the original (b) with (e) and (j) would still result in a vast majority for the weight. On the other hand, when taking the influence of immediate and wider context (cf. McNeill, 1985; 1992) into account, the newsworthy information regarding this episode would be as follows: Sylvester is attempting to get to Tweety with the method “catapult” - the fact that the cat is hunting the bird had been established in the instructions.

The context given to the participants in this study was merely that of the general Canary Row scenario. This episode was either first or last in the collection of 12 stimuli, and could hence contrast with the standard cartoon plot, which would make “auf die andere Seite” just as newsworthy as the catapult. The immediate background of the stimulus sequence would be Tweety's owner beating Sylvester up with her umbrella. Accordingly, one could argue for either conceptual affiliate (c) or (g) on the basis of co-expression, newsworthiness, and the restrictiveness of lexical affiliates.

In total, the 12 stimuli had a conceptual affiliate accuracy of 80.3%. Among the twelve, there is a conceptual agreement rate of 95.88% on average (excluding non-answers). The transcripts of the deviating two samples are shown in Examples 5 and 7.

```
so ne **ro**stige **re**gen[rinne die war neben] dem **fen**ster

such a rusty rain [spout that was next to] the window
```

Example 7: Falsification 1 (clip 5).

As discussed above, a conceptual affiliate goes hand in hand with the rheme of an utterance, or its newsworthy part. Example 7 is faulty in two ways: It is lacking a verb in its theme, or main sentence, and it has no obvious rheme (“regenrinne” as an object and/or the position of the rain spout). The speaker's gesture is a slightly concave wiggling right hand that moves from central position towards the head (see Figure 18 for a map of gesture space). When knowing the plot of Sylvester's attempt described here, one can recognize the “rising hollowness” (cf. McNeill, 1985), but for the participants the presented context was insufficient. The design and position of the gesture are not interpretable without the information that Sylvester is crawling through the pipe: eight out of 18 participants could not connect the gesture to the utterance at all, 3 chose the position of the pipe and 4 the factual pipe. Also, two participants connected the gphr to “so ne” ('such a'), interpreting it as interactional rather than co-expressive. The 30% (position) to 40% (object) agreement of conceptual affiliates is distinct in contrast to the average 95.88% conceptual agreement. The fact that the utterance is not a complete sentence and has two clauses (rhemes) explains the difference in concepts participants connected with the gesture. This makes a point for the co-expressiveness of gesture in the context of themes and rhemes.

Example 5 demonstrates a further falsification of the conceptual affiliation of speech and gesture. In contrast to Example 6, the speech does not have a potential lack of themes/rhemes. Instead, there is one too many, namely (1) “Er öffnet die Tür in seiner Pagenuniform" (opening door in uniform) and (2) "so ne rote mit goldenen Knöpfen und so". The two clauses are not only separated by an unfilled pause, they also complement each other. The rheme of (1) is the opening of the door (in uniform) and (2) further specifies (1) with a description of the uniform. The gesture zig-zagging across the chest could have triggered two or even three conceptual affiliates, that is, the button design (38.8%) or the uniform in general (33.3%); four participants also included "öffnen" (opening) in some combination or other (22.2%). In contrast to these two cases, all stimuli with only one rheme were fit for a unification of the affiliates that were picked by the participants between 82.4-100% (median = 100).

6.6 Discussion

The results of the Conceptual Affiliation Study show that an iconic gesture corresponds to a rheme, and one rheme only at a time. For Examples 5 and 7 it was demonstrated that an uneven numerical distribution of gestures an rhemes causes listeners to perceive the same utterance differently. This phenomenon has been explained through *conceptual affiliation*. A speaker has an MU in mind that they want to express, and speech and gesture co-expressed in order to convey this MU through the unpacking of a GP. While speech is bound by syntax and the lexicon, hands and arms may move rather freely. Regarding this difference in flexibility, the crux of the modalities' co-expressivity was long thought to be their temporal synchrony, in particular that of stroke and peak. This factor was excluded in the experimental setup as far as possible in that the audio clips in the experiment had more than one prosodic peak in general and only one gesture phase, that is, the stroke – participants were simply not able to connect the two in unison. Additionally, the participants could decide freely which speech-gesture affiliates to pick without regulations for length, position or number. While this did not enforce the idea of the lexical affiliate, it did not exclude it either. The exact choice of words one participant made was often also picked by another, but this did not happen more than twice for either one stimulus.

The scope of lexical affiliation was widened in the analysis to include inflections and minor additions such as determiners or pronouns, which facilitated the forming of larger groups of affiliates. Still, the same stimulus seemed to trigger differing associations in the participants of the Conceptual Affiliation Study. Taking a closer look at the data gave rise to the suspicion that the instructions to pick any “word, words, phrases, or parts of words” (v.s.) might have caused participants to decide on different features of the stimuli. The action of throwing or the object thrown in Example 6, for instance, could be affiliated with various parts of the verbal utterance (see Table 2). What most answers for the stimuli had in common were the connection of the gestures with parts of the rhemes of the utterances. For Example 6, 14 out of 17 participants noted down that the weight – its existence, its shape, or its being thrown – was the part of the utterance that was related in meaning to the gesture. All participants selected features of the same *concept* for the iconic gesture that corresponded to the rheme of the utterance (a weight being thrown). This conceptual affiliate goes beyond the phonological form of “weight” and encompasses an imagined scenario, part of an MU, quite possibly the SP. Hypothesis (1), as formulated at the beginning of the methodology chapter of this dissertation, that is, that “[t]he semiotic-semantic relation between spontaneously co-produced speech and gestures. . . encompasses all newsworthy information given in speech” (p. 85), could only be falsified when a speech-gesture utterances includes two rhemes. Since the GP-SP transmission cycle operates within theme-rheme patterns, however, hypothesis (1) can be considered as working within this conceptual transmission cycle.

Grouping the lexical tokens chosen by the participants by concepts was possible for all stimuli with exactly one rheme and one gesture. Also, the finding suggests that in a model of co-occurring speech and gesture, the temporal tolerance of 1~2 s as suggested by McNeill (2012, p. 32) would not disturb comprehension (see Chapters 7 and 8). While de Ruiter's (2000) conclusion “that gestures do not have lexical affiliates but rather 'conceptual affiliates'” (p. 291) was rather exclusive, it already pointed in the right direction. Conceptual affiliation occurs across the borders of adjacent lexemes (see Example 5, “rote mit goldenen”), so “the notion of a conceptual affiliate can also explain the occurrence of the occasional gesture that seems to be related to a single word” (de Ruiter, 2000, p. 291). Finally, conceptual affiliation of speech and gesture can also explain why A combined a gesture like spraying water with “The yard looked so beautiful” – she meant the sun sparkling on drops of water distributed across the plants and grass in the yard (pp. 35; 106).

The temporal definition of lexical affiliation still matches a lot of expressions, particularly deictic ones, and could not be fully refuted with the results of the Conceptual Affiliation Study. A useful categorization of lexical affiliation would be to subsume it under conceptual affiliation, since, mostly in utterances involving iconic gestures, both kinds of affiliation can be validated. In the example of Figure 3, for instance, “omma” (granny) temporally overlaps with the gesture stroke of S' imitating the granny hitting Sylvester with an umbrella, making “omma” the lexical affiliate of the gesture in that it expresses granny's actions. The lexical items affiliated with the gesture are, however, not present in the utterance itself when taking the lexical affiliation literally, but was verbalized in a previous utterance (“und dann kommt die oma mit'm **regenschirm** wieder und **haut** ihn eins drüber #”; gphr 112). Again, choosing a holistic, conceptual affiliation of speech and gesture allows for cross-utterance semantic relations, while lexical affiliation based on temporal coinciding allows only for atomic, restricted interpretations of meaning.

All gestures discussed in this chapter can be positioned on the mandatory-speech pole of Kendon's continuum (McNeill, 2005, p. 7). Investigating how participants deal with other types of gestures would be a way to further test the idea of conceptual affiliation. Indeed, studies have been carried out investigating the recognition factor of emblematic gestures. In Gunter and Bach (2004), for instance, participants determined the meaning of emblematic and random hand postures, but without contextual speech and only from pictures. A study investigating such codified gestures in the context of conceptual affiliation should result in a high percentage of overlaps between lexical and conceptual affiliates within one cultural community. This would be in parallel to the gesture continuum because emblems can often be regarded as word-like. Additional studies are also necessary to investigate multimodal conceptual affiliation in a natural communication setting, that is, when co-produced speech and gesture are perceived simultaneously. A methodology enabling such investigation would be the kind of online survey used in the Perceptual Judgment Task, the methodology of which will be discussed in the following Chapter 7, albeit in the context of the perception of temporal speech-gesture synchrony.

7 Perceptual Judgment Task

The two sets of studies presented in this chapter will hopefully set a starting point for further investigations into the perception of asynchronies of speech and gesture. As there is a semantic bond between the two modalities, a certain window of audiovisual integration (AVI) should be expected. The Perceptual Judgment Task will test hypotheses (2), (4), and (5) of this dissertation (p. 85):

> (2) Listeners are able to discriminate variation in the synchrony of spontaneously co-produced speech and gestures and they will prefer a window of AVI encompassing both gestural advance and delay.
>
> (4) The preferred synchrony of speech and gesture in perception will vary from that produced during spontaneous utterances.
>
> (5) The preferred synchrony of speech and gestures will vary for different gesture types as well as for non-speech signals.

Based on the results of the Perceptual Judgment Task, the studies of the Preference Task (Chapter 8) will investigate the listener's ability to reproduce what they assume happens in production: While the Perceptual Judgment Task only inquired on pre-selected asynchronies, the Preference Task will allow for a fuller picture on what listener-viewers accept and prefer.

7.1 Study 1

7.1.1 Participants

141 German native speakers with mixed backgrounds completed Study 1 (100 women, 41 men, M_{age} = 24.32 years, age range: 16-67 years). They rated the perceived naturalness of 2523 stimuli created from the selection of utterances presented in Chapter 5.3.6).

7.1.2 Materials

From the 43 speech-gesture utterances in the corpus deemed large and imagistic enough to be used for perception studies, a selection had to be made to a) narrow down the number of stimuli to be rated by participants of an online experiment, to b) have stimuli comparable to those analyzed in the literature on speech-gesture production, and to c) have utterances long enough to allow for desynchronization of up 600 ms. The narrowing down resulted in the selection of 29 speech-gesture utterances listed in Table 5, with an average duration of 2460.80 ms (SD = 832.94 ms). From these, 29 clips containing full utterances with one fairly prominent, naturally co-occurring gesture each (9 deictic, 19 iconic, 1 emblematic) were created in Adobe Premiere CS5.

Table 5: Clips selected for the creation of stimuli for the Perceptual Judgment Task.

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
1	pipe_1_0	10.17.48.959	2530	19	er <n> klettert das abwasserrohr hoch	iconic
2	cage_1_0	10.17.48.959	915	31	also er macht den käfig auf	iconic
3	penny_1_0	10.17.48.959	1551	51	dann ja hier's is 'n penny oder	iconic
4	swallow_1_0	10.17.48.959	1546	62	/ er schluckt die kugel # %	iconic
5	in_pipe_1_0	10.17.48.959	938	76	ja ja er is in dem regenrohr	iconic
6	shelf_1_0	10.17.48.959	1114	90	er sitzt im regal #	deictic
7	rub_1_0	10.54.29.104	1452	143	(whispered:) da (creaky:) drüben (creaky:) isser (creaky:) endlich der leckere vogel #	iconic
8	trace_1_0	10.54.29.104	1603	151	geht dann in das and're<e> haus rein	deictic/ trace
9	climb_1_0	11.00.31.621	1400	200	un klettert da ers' rau<u>f / #	iconic
10	banana_1_0	11.00.31.621	1482	230	dann lockt er den affen mit 'ner (breathy:) banane	iconic
11	ring_1_0	11.00.31.621	1079	248	<mm> also das telefon klingelt	deictic
12	knock_1_0	11.00.31.621	2884	267	wo er dann als roomboy verkleidet is un' anklopft #	iconic
13	cage_2_0	11.00.31.621	1284	272	un' dann haste den käfig da steh'n	trace
14	cover_1_0	11.00.31.621	1271	290	# will dann die decke runtermachen	iconic
15	ring_2_0	11.00.31.621	760	370	un' dann klingelt's	iconic
16	discover_1_0	12.05.31.682	1893	421	und wie er dann<nn> tweety in seinem käfig entdeckt	deictic
17	sign_1_0	13.09.12.480	2559	593	draußen aufm schild steht <äh> hunde und katzen verboten #	emblem
18	street_1_0	15.04.57.785	2604	677	und rollt auf der bowlingkugel ne abschüssige straße runter	deictic
19	can_1_0	15.04.57.785	2502	681	der rennt mit de<e>r geldspendebüchse rum /	iconic

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
20	hat_1_0	15.04.57.785	2656	685	ja und dann zieht er den hut so höch und dann erkenntse dass (laughing:) s ne katze is	iconic
21	bino_1_0	16.11.09.878	1635	805	halt mit'm fernglas durche gegend kuckt	iconic
22	sign_2_0	16.11.09.878	1416	812	steht auf so'm schild neben der tür	deictic/ iconic
23	sill_1_0	16.11.09.878	1839	815	steht halt direkt immer am fenstersims #	deictic/ iconic
24	cat_1_0	16.36.00.692	2230	834	und sylvester is die große schwarze katze	iconic
25	bird_1_0	16.36.00.692	4383	835	und in dem film geht es da<a>rum dass sylvester scharf auf <äh> den vogel is	deictic/ trace
26	kicked_out_1_0	14.27.42.306	2510	1247	%<ähm>/ sylvester fliecht sofort wieder raus	iconic/ deictic
27	everywhere_1_0	14.27.42.306	2510	1254	#%/ er fängt dann an überall nach tweety zu suchen #	iconic/ deictic
28	catapult_1_0	14.27.42.306	3001	1304	daraufhin / wird er in die luft katapultiert	iconic
29	hit_1_0	15.04.57.785	2428	707-709	und haut ihm mit dem (laughing:) regenschirm wieder fleißig übern detz %	iconic

The aim of the Perceptual Judgment Task was to study how listeners perceive utterances in their original production synchronies as well as in different degrees of asynchrony, including speech before gesture and vice versa (see Chapter 2.3 for more details on the temporal relations between intervals of S and G). The levels of audiovisual asynchrony in previous studies were restricted to small ranges and steps with a focus on video advances: Massaro et al. (1996) included audio advances and delays of 0 ms, 67 ms, 167 ms, 267 ms, and 500 ms in their speech-lip stimuli, using small steps and a range of 1 s; Campbell and Dodd (1980) tested a large range of asynchronies, that is, 3.2 s, in large steps of 400 ms, 800 ms, and 1600 ms. For speech-gesture stimuli, Habets et al. (2011) restricted their ERP testing to gestural advances of 160 ms and 360 ms, while Özyürek et al., 2007 used only the manually synchronized combination of lexical target and gesture stroke. In order to further approximate the optimal as well as the acceptable range for speech-gesture AVI, for the Perceptual Judgment Task, a) the channels were desynchronized in both directions and b) offsets in steps of 200 ms up to ± 600 ms were selected to include and go beyond the previously tested time frames (Figure 20). Whenever offsets are mentioned with regard to the studies conducted, negative offsets will indicate the speech is in delay, after the gesture (e.g., -400 ms = GS by 400 ms), while positive offsets will have the speech in advance, before the gesture (e.g., +400 ms = SG by 400 ms).

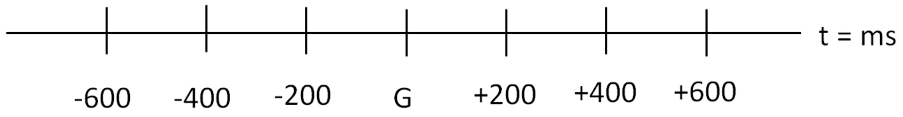


Figure 20: Scale of speech-gesture offsets.

Each of the 29 original clips was trimmed in Adobe Premiere Pro CS5 to start and end with the full gphr and to include the full utterance. First, the full audio track was shifted in steps of 200 ms from the original synchrony of S and G into SG or GS relation (see Figures 21, 22). Following this procedure, the resulting gaps of overlap between the tracks were filled with silences from the same recordings and fitting still frames of the same video (Figure 23); both channels were contained in one and the same file in video formats designed for different web browsers (.ogg, .mp4, .avi).

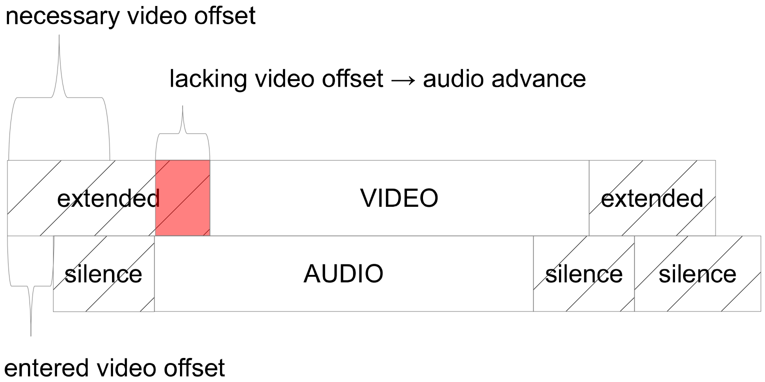


Figure 21: Shift mechanism for SG stimuli.

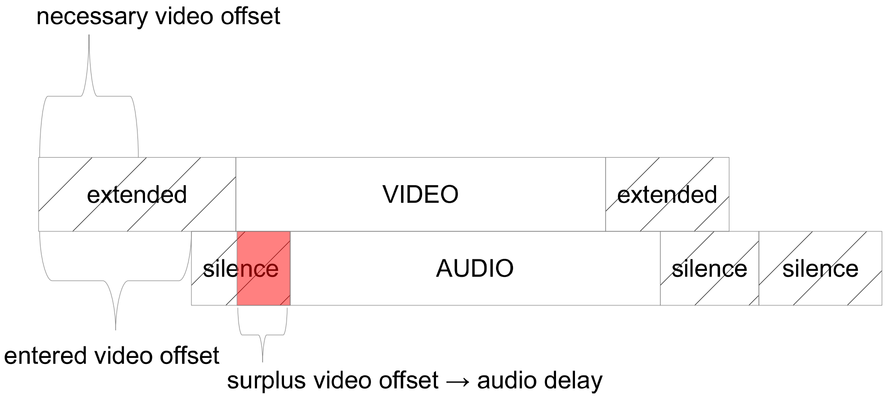


Figure 22: Shift mechanism for GS stimuli.

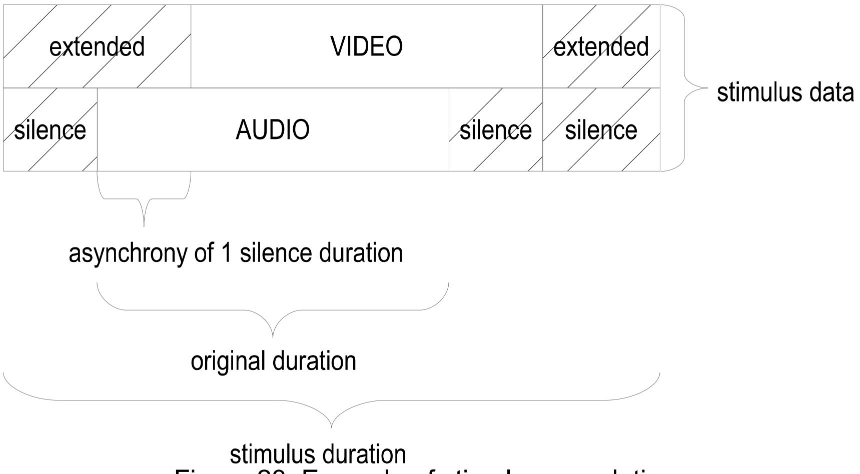
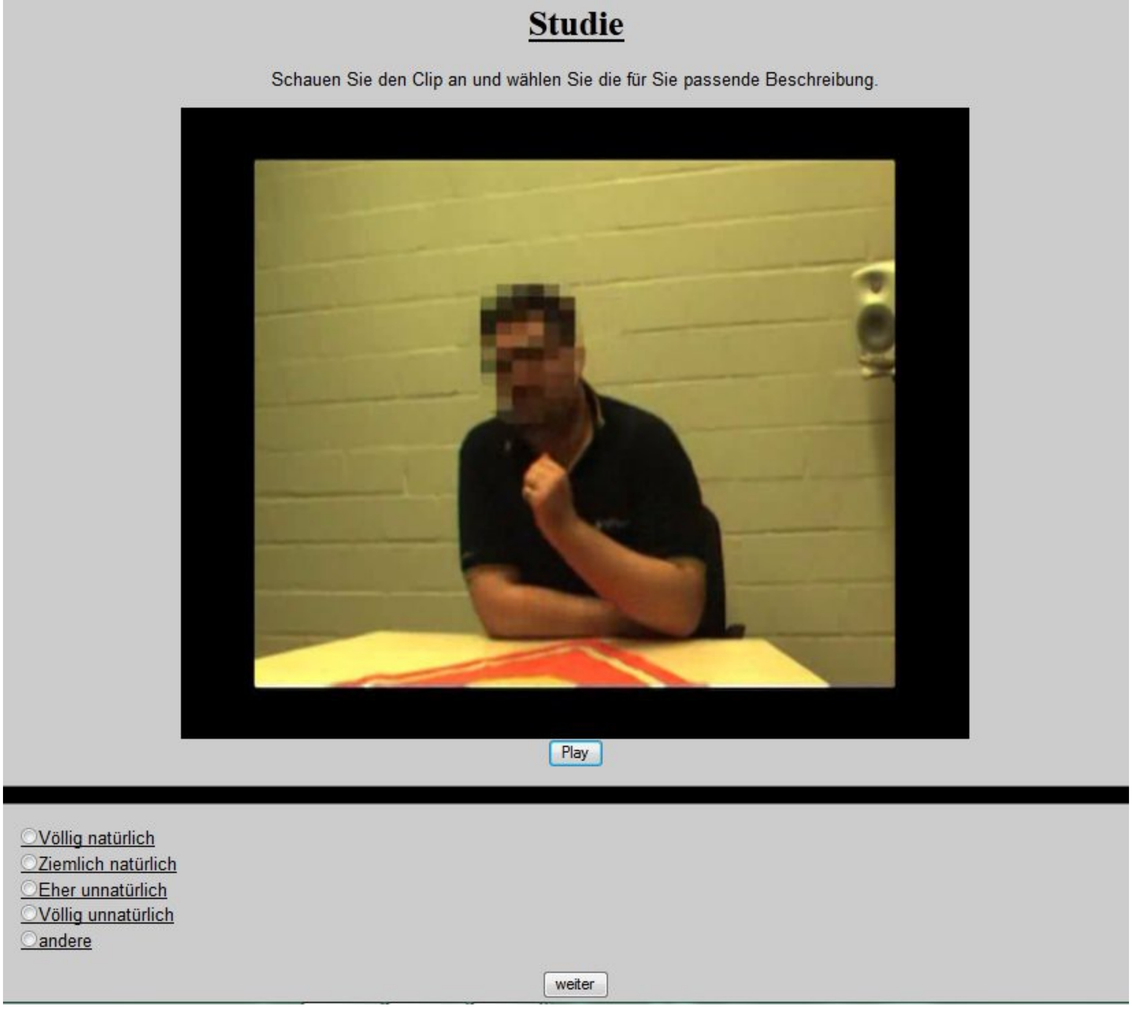


Figure 23: Example of stimulus completion.

Following the Nyquist sampling theorem, to avoid aliasing, that is, artifacts in the signal during playback, “[t]he sampling frequency should be at least twice the highest frequency contained in the signal.. . . Or in mathematical terms: fs ≥ 2fc” (Olshausen, 2000, p. 1). The video formats used for our research have a frame rate of 25 fps (25 Hz), that is, one frame every 40 ms. It can be assumed that any lagging of the video will not noticeably differ from a frame interval of 40 ms. The steps of asynchrony between the stimuli used in the Perceptual Judgment Task are of 200 ms, which makes the intervals to be 2*5 Hz and hence well within the restrictions of the sampling theorem. The audio track has a sampling rate of 44.1 kHz; with 20 kHz as the maximal audible frequency for people with 100% hearing capability, the Nyquist sampling theorem applies with 44.1 kHz>2*20 kHz for the audio track of the stimuli used. The 192 stimulus clips were then put on a local server to be accessed from the experiment interface.

7.1.3 **Procedure**

A web link to Study 1 was spread via mailing lists and social media platforms (university students, Facebook, etc.). After a biographical questionnaire, participants were informed the study would take about 15 minutes and were also strongly advised to use headphones. They were told to rate the naturalness of 28 excerpts of retellings from the Canary Row cartoon in which the video or audio had sometimes been manipulated. The participants were instructed to watch the clips as often as they liked before rating them as ‘fully natural’ (‘völlig natürlich’), ‘somewhat natural’ (‘irgendwie natürlich’), ‘somewhat unnatural’ (‘irgendwie unnatürlich’), ‘fully unnatural’ (‘völlig unnatürlich’), or ‘other’ (‘sonstiges’), the latter with an option to elaborate (see Figure 24). In a trial run with three stimuli, the participants were presented with three versions of the same original clip (gphr 151), that is, one in



Studie

Schauen Sie den Clip an und wählen Sie die für Sie passende Beschreibung.

Figure 24: Online interface for the Perceptual Judgment Task.

which the audio is 1 s before the video, one with the channels in their originally recorded synchrony, and a third in which the video is 1 s before the audio. For each participant, 28 clips were individually selected and randomized by a script in such a way that every original stimulus would occur only once and no level of asynchrony be presented twice in a row. As in the trial run, the participants could only continue to the next clip when they had selected a rating on the scale for the respective clip after having watched it as often as they felt necessary. The judgments were recorded in an Structured Query Language (SQL) database, including detailed coding of the clip variants as well as participant IDs with profiles and dropout logs. Throughout the study, a progress bar with the remaining percentage of the study was displayed.

7.1.4 Results

The gathered data were transferred from the SQL database into SPSS. Through case selection, “other”-ratings were excluded from the statistical analysis. The categorical rating variable was coded as ordinal from 0 (‘fully unnatural’) to 3 (‘fully natural’) and entered as dependent variable into a one-way univariate ANOVA with the degrees of asynchrony as independent variable. This analysis revealed a significant main effect of the degree of asynchrony on the degree of perceived naturalness (*F*(6, 2516) = 33.47; *p* < .01).

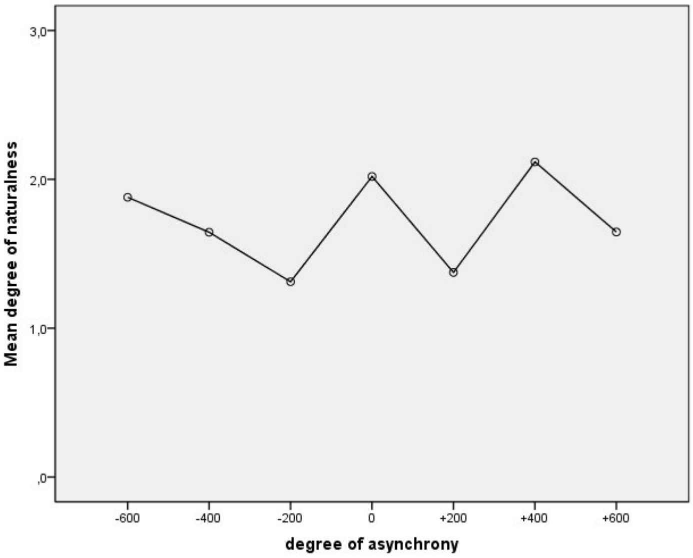


Figure 25: Mean degree of naturalness for the degrees of asynchrony in Study 1.

The mean degree of naturalness in Study 1 was M_{nat} = 2.287 (*N* = 2517, SD = 1.049). As can be seen in Figure 25, there are peaks in the perceived naturalness at 0 ms, -600 ms (GS) and +400 ms (SG) while the stimuli desynchronized by ±200 ms are least preferred by the participants. A gradual growth in acceptance occurs between -200 ms and -600 ms GS. The contrasts of the different levels of asynchrony with the original synchrony (0 ms) in the K Matrix show the correlations between degree of asynchrony and the perceived degree of naturalness to be significant at *p* < .01 for all stimuli except for those desynchronized by -600 ms (GS) and +400 ms (SG).

7.1.5 **Discussion**

The participants perceived the original condition without synchrony manipulation, an audio delay of -600 ms (GS), and an audio advance of +400 ms (SG) as most natural. The preference for the original condition, in stark contrast to the ±200 ms asynchronies, fits previous research on the McGurk effect as only asynchronies within a range of ±200 ms allow for a fused percept. The overall results of Study 1 also agree with the expected window of optimal AVI as well as with the expected breakdown of AVI, for speech-lip stimuli, beyond an asymmetric range of 500 ms (cf. Massaro et al., 1996; van Wassenhove et al., 2007). The preferred window of AVI for speech-gesture stimuli with an audio delay between -160 ms (GS) and -360 ms (GS) found by Habets et al. (2011) is not fully confirmed by these results, but the gradual growth in acceptability between the audio at -200 ms (GS) and -600 ms (GS) before the video suggests a similar tendency. The results confirm our methodology to be appropriate for researching audiovisual asynchronies by means of this instance of the Perceptual Judgment Task. The fitting of our findings with the McGurk effect further suggests that participants mostly focused on speech-lip synchrony and speech-gesture synchrony was not a major factor in Study 1.

7.2 Study 2

That the stimuli with an audio advance of +400 ms (SG) or with an audio delay of -600 ms (GS) are ranked so highly in Study 1 might be due to cues from the lip movements in the videos. Study 2 replicates Study 1 in its methodology, but the heads in the stimuli are blurred to cancel out lip visibility.

7.2.1 Participants

126 German native speakers (84 women, 42 men, M_{age} = 28.28 years, age range: 15-67 years, 42 males) participated in Study 2. They rated a total of 1812 clips for how natural they perceived those.

7.2.2 Materials

The heads of the speakers were covered in the 192 stimuli from Study 1 in Adobe Premiere Pro CS5 with a blurred layer which moved as the head moved (see Figure 24). The graphical manipulation was justified during the instruction by referring to anonymity requirements.

7.2.3 Procedure

Again, a link to the online interface was spread via social networks and mailing lists. As in Study 1, participants rated 28 stimuli following a trial run. Again, the stimuli were randomized in such a way that no utterance occurred twice across the experiment for any participant and that the same degree of asynchrony did not occur twice in a row.

7.2.4 Results

The data gathered in the SQL database was again exported to SPSS and “other”-ratings were removed after they were checked and documented. The univariate ANOVA shows a significant contrast between the visibility conditions of Studies 1 and 2 (*F*(2, 9497) = 29.982; *p* < .000). The ratings (3 = ‘fully natural’; 2 = ‘somewhat natural’; 1 = ‘somewhat unnatural’; 0 ‘fully unnatural’) were entered into a univariate ANOVA with the degrees of asynchrony as independent variable. This analysis revealed no significant main effect of the degree of asynchrony on the degree of perceived naturalness (*F*(6, 1805) = 1.46; *p* = .190).

The mean degree of naturalness in Study 2 was M_{nat} = 1.937 (*N* = 1812, SD = . 9443). The participants rated the stimuli with asynchronies of ±200 ms as most natural (see Figure 26) and rated the original condition stimuli, which were not desynchronized, nearly as low as those with an audio advance of +400 ms (SG). A stark contrast exists between the originally synchronous stimuli and those with an audio delay of -200 ms (GS) (*p* < .05), while the other degrees of asynchrony share a high similarity in their ratings for naturalness. Those stimuli with an audio advance of +400 ms (SG) are rated the most similar to the original condition (*p* = . 964).

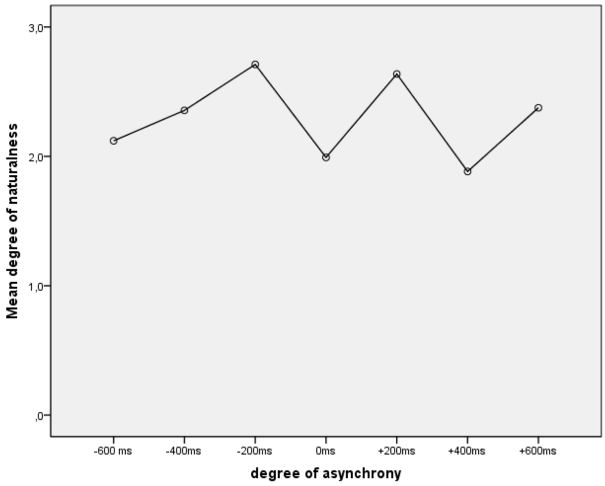


Figure 26: Mean degree of naturalness for the degrees of asynchrony in Study 2.

7.2.5 Discussion

There was a significant impact of the independent variable 'visibility' on the degree of naturalness between Studies 1 and 2 (v.s.). This confirms that lip visibility was influential in Study 1, which might have lead to the similarity of the preferred windows of AVI found in previous research and to the overall higher ratings in Study 2. The finding that in Study 2 there is no significant variation between the different degrees of asynchrony except for the participants' preference of stimuli with an audio delay of -200 ms (GS) fits well with the overall tendency of previous research that audio delay is generally preferred by listeners to audio advance (cf. e.g., van Wassenhove et al. 2007; Massaro et al., 1996). An overall precedence of gesture over speech has been equally observed in production (e.g., Thies, 2003; Morrel-Samuels & Krauss, 1992; Schegloff, 1984), at least for speech-gesture pairings with strong semantic boundaries. Despite the lack of significant contrast regarding the independent variable ‘visibility’ as well as the independent variable ‘degree of asynchrony’, the results of Study 2 will be regarded as indicative of a tendency of the participants to prefer gestures to occur in advance of speech.

7.3 Study 3

The head motion still noticeable in the stimuli of Study 2 might have influenced the participants in rating the different asynchronies. The blurry coverage of the speakers’ heads might have caused them to rate most stimuli as ‘somewhat natural’ because of the rather frequent usage of this type of anonymization in TV shows and newspapers. To avoid any and all visual prosodic indicators, no head movements at all are visible to the participants in Study 3, with gesture and speech remaining as the only independent variables.

7.3.1 Participants

325 native German speakers (240 women, 85 men, M_{age} = 24.31, age range: 17-67 years) rated the naturalness of 5165 stimuli in Study 3.

7.3.2 Materials

In Adobe Premiere Pro CS5, the heads from the speakers in the 192 stimuli from Study 1 were covered with a black rectangle which moved as the head moved; motions of neither the lips nor the head are detectable by the participants. The shoulders were left uncovered to not obscure parts of the arms gesturing. The graphical manipulation was again justified by referring to anonymity requirements during the instruction.

7.3.3 Procedure

As for Studies 1 and 2, a link to the online interface was spread via social networks and mailing lists. Participants were asked to rate 28 stimuli following a trial run. To avoid any sequential effects, the stimuli were again randomized in such a way that the same degree of asynchrony did not occur twice in a row and that no utterance occurred twice across the experiment for any participant.

7.3.4 Results

After importing the gathered data into SPSS and fitting it for analysis, univariate ANOVAs were conducted. The test of between-subjects effects regarding the visibility condition reveals a significant difference between Studies 1 and 3 (*F*(1, 7674) = 38.39; *p*.001) and 2 and 3 (*F*(1, 6963) = 8.89; *p* < .005). A univariate ANOVA shows a significant main effect of the degree of asynchrony on the degree of perceived naturalness (*F*(6, 5158) = 6.28; *p* < .01).

The mean degree of naturalness in Study 3 was M_{nat} = 1.860 (*N* = 5165, SD = . 9620), the distribution of the levels of asynchrony rated as most natural fairly flat (Kurtosis = -.847, SEM = .068). The stimuli with original synchrony (0 ms) are clearly preferred to those with audio advances of +200 ms (SG) (*p* < .001) or delays of -200 ms (GS) (*p* < .001), as can also be seen in Figure 27. No significant contrasts were found for other levels of asynchrony with the original clips.

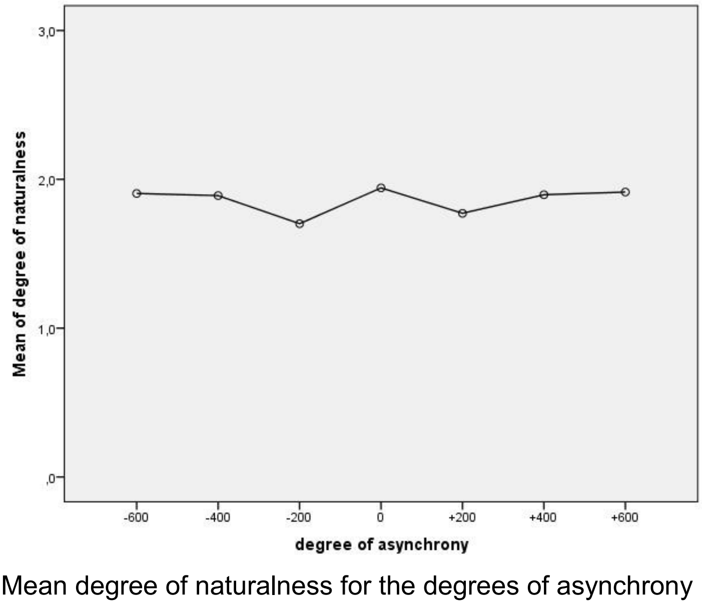


Figure 27: Mean degree of naturalness for the degrees of asynchrony in Study 3.

7.3.5 Discussion

Massaro et al. (1996) set the preferred window of AVI for syllables within maximal ranges of 150 to 250 ms of asynchrony and expected a significant breakdown in the perceptual alignment at discrepancies between 250 ms and 500 ms of asymmetric asynchrony. The participants in Study 3 clearly preferred the original synchrony (0 ms) of the stimuli to the ±200 ms asynchronies, even though about two thirds of all stimuli were rated as somewhat or fully natural. Whether this above-chance rating speaks against a breakdown of AVI or for it is debatable. This agrees with the findings by Habets et al. (2011), who hypothesize the window of AVI for single words with gestures to extend to somewhere between auditory delays of -160 ms and -360 ms (GS). Our findings expand this possible window of AVI to audio advances of up to +200 ms (SG). That all visual prosodic influence was canceled out by the head blockage is an argument supporting a wider window of AVI for speech with gestures than for speech alone.

7.4 Lab Replication

For a partial replication of Studies 1-3, a reduced number of participants rated a select group of stimuli on a desktop computer to efficiently test the methodological reliability of Studies 1 through 3 (cf. pp. 86ff.).

7.4.1 Participants

17 participants (11 women, 6 men, M_{age} = 25.0 years, age range: 22-42 years) rated a total of 255 stimuli, 85 in each visibility condition, with regard to how natural they perceived them.

7.4.2 Materials and procedure

After completing the same trial as in Studies 1-3, the participants were presented with three versions of one video clip in the lips-visible condition at -600 ms (GS), 0 ms original synchrony, and +200 ms (SG). Apart from making the lab replication more time-efficient, these degrees of asynchronies were selected because they include the suspected window of optimal AVI as well as an asynchrony at which AVI should definitely have broken down. Having watched the three stimuli several times, the participants were to choose the one most natural to them or to indicate that they were unable to decide. This procedure was then repeated for another four sets of three stimuli in the lips-visible condition (Study 1), and for another five sets of three stimuli each in the face-blurred (Study 2) and face-blocked condition (Study 3).

7.4.3 **Results**

After the data was cleaned from “undecided” ratings, a univariate ANOVA resulted in a significant main effect of visibility on the preferred degree of asynchrony (*F*(2, 208) = 2.926; *p* = .056). A post hoc Tukey test showed that the difference between the lips-visible condition and the face-blurred condition was not significant (*p* = .992). The lips-visible condition differs near-significantly from the face-blocked condition(*p* = .074) , which, in turn, does not differ significantly (*p* = .118) from the face-blurred condition.

Figure 28 displays the asynchronies as preferred by the participants. Since they had to pick one of the three options respectively, no subjective ratings of naturalness as in Studies 1 through 3 can be accounted for. The participants preferred the 0 ms stimuli and the +200 ms (SG) stimuli in the lips-visible condition while the audio delay of -600 ms (GS) was not selected at all (cf.). As can be seen from the cross-tabulation (χ²(6, 255) = 28.91, *p* = .000) in Table 6, the -600 ms (GS) asynchronies were clearly dispreferred across conditions (8.5%), while the original synchrony (46.9%) was nearly as preferred as the audio advance of +200 ms (SG) (44.5%).

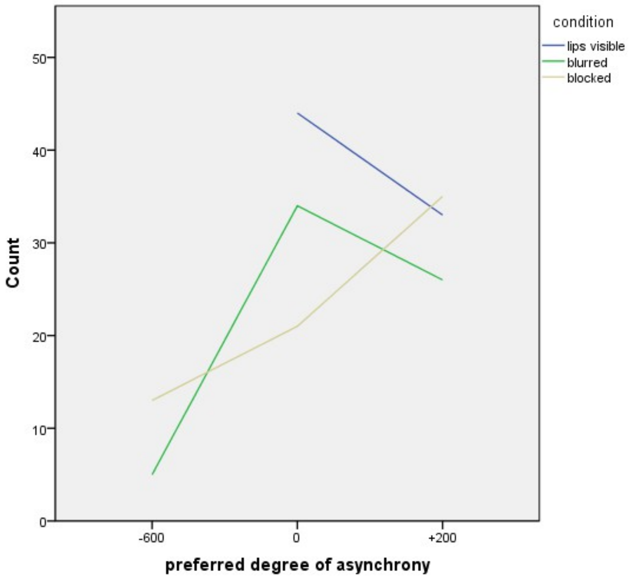


Figure 28: Preferred degrees of asynchrony in Lab Replication.

Table 6: Cross-tabulation of preferred degrees of asynchrony by visibility condition in Lab Replication.

		condition			Total
		lips visible	blurred	blocked	
preferred degree	-600	0	5	13	18
of asynchrony	0	44	34	21	99
	+200	33	26	35	94
Total		77	65	69	211

7.4.4 Discussion

The Lab Replication supports the findings from Studies 1-3. As in Study 1, the original 0 ms synchrony is slightly preferred to the audio advance of +200 ms (SG), while no participant rated an absence of audio delay as natural . As the head visibility decreases, the preference distribution among the degrees of asynchrony increases. The perceived naturalness of the original synchrony is not as accepted in the face-blocked condition (see Study 3), the audio delay of -600 ms (GS) has at least minimal acceptability. Finally, as can be clearly seen in Figure 28, the medial naturalness ratings nearly even out between the original synchrony (0 ms) and the audio advance of +200 ms (SG), which is comparable to the rather flat distribution among the naturalness ratings in Study 3 (cf. Figure 27). The general lack of variation between the face blocked and face blurred conditions also reflect the findings from Studies 1-3.

7.5 Study 4

The Lab Replication has verified the reliability of the Perceptual Judgment Task. The influence of lip synchrony and prosodic head movements has been eliminated during the course from Study 1 to Study 3 and wider windows of AVI for speech-gesture stimuli turn have become more likely than previously assumed. Study 4 will provide acceptable audiovisual asynchronies from causally and temporally fixed stimuli as a baseline to be compared to the speech-gesture ratings.

7.5.1 Participants

142 participants (95 women, 40 men, 2 other, 5 not applicable, M_{age} = 27.86 years, age range: 18-62 years) rated 2249 physical cause-and-effect stimuli in Study 4 of the Perceptual Judgment Task.

7.5.2 Materials

In Adobe Premiere Pro CS5, 10 short videos of physical cause-and-effect stimuli were desynchronized into the previously used seven degrees of asynchrony. The original clips contained exactly one instance of each of the following: A hammer hitting a nail, snapping a book shut, a clap of the hands, clinking a class with a fork, a tap on a keyboard, knocking on a table, the plop while opening a bottle of champagne, fingers snapping, hitting a bass drum, and popping a balloon with a needle. While the hammer hitting the nail functioned as a trial stimulus, the other nine sources of noise were used in the actual study.

7.5.3 Procedure

As for Studies 1-3, all newly created stimuli were uploaded onto a local server and a link to the interface was spread via university mailing lists and social networks. After completing the trial rating, the participants rated the (a)synchrony of the physical stimuli for naturalness.

7.5.4 Results

After importing the gathered data into SPSS and fitting it for analysis, a univariate ANOVA revealed a significant main effect of the degree of asynchrony on the degree of perceived naturalness (F(6, 2248) = 71.97; $p < .01$). The mean degree of naturalness in Study 4 was M_{nat} = 1.724 (N = 2249, SD = 1.1155), and already from the means graph (Figure 29) one can see a clear difference between the perceived naturalness of the speech-gesture stimuli and the physical event stimuli.

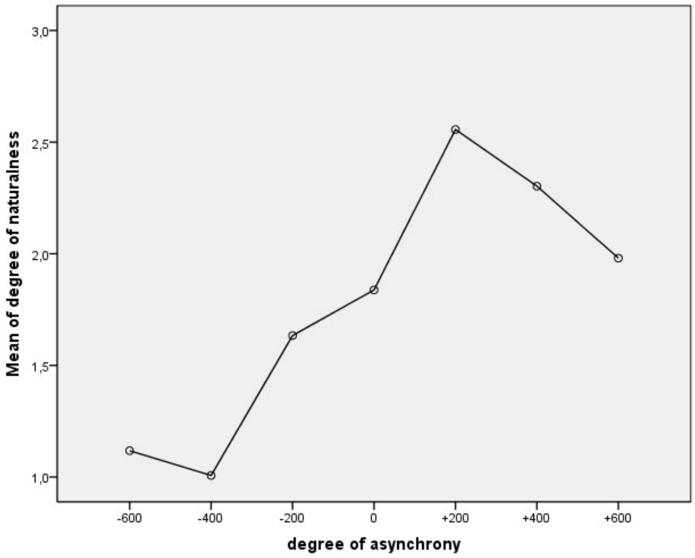


Figure 29: Mean degree of naturalness for the degrees of asynchrony in Study 4.

While in all visibility conditions, the graph peaked at different levels of asynchrony between speech and gestures, in Study 4 the participants distinctly preferred the physical event stimuli in which the audio precedes the video by +200 ms (AV). This is confirmed by a post hoc Tukey test, which showed that the audio advance of +200 ms (AV) were significantly different at *p* < .001 for all levels of asynchrony but for +400 ms (AV) (*p* = .141), and the original synchrony (0 ms), which is still significantly different at *p* < .05. The distribution of the preferred levels of asynchrony in Study 4 is barely skewed and rather platykurtic (skewness = -.271, kurtosis = -1.299, SEM = .103), with M_{nat} = 1.72 on a scale of 0 (fully unnatural) to 3 (fully natural).

7.5.5 Discussion

The window of optimal AVI for physical bimodal stimuli as determined by van Wassenhove et al. (2007) ranges from -200 ms (VA) to +200 ms (AV) around the original audiovisual synchrony, which is a slightly smaller range than the 533 ms Massaro et al. (1996) suggested for the possible window of AVI for speech-lip signals before integration breakdown. The participants in Study 4 displayed a clear preference for stimuli in which the audio precedes the video by +200 ms, which goes in line with previous research on the AVI of bimodal media in psychophysics (also see Chapter 4.1). As with the speech-gesture stimuli in Studies 1-3, the audiovisual synchrony of cause-and-effect stimuli has a significant effect on the naturalness as perceived by the participants. The well-formed distribution of the ratings provides additional support of our methodology. The results of Study 4 hence provide an excellent baseline to which the preferred asynchronies of the speech-gesture stimuli can be compared.

7.6 Discussion Perceptual Judgment Task

Next to testing hypotheses (2), (4), and (5), the aim of Studies 1 through 4 of the Perceptual Judgment Task was to find an optimal and tolerable window of AVI for audiovisual signals in general and for speech-lip and speech-gesture stimuli specifically. The findings by Massaro et al. (1996) and van Wassenhove et al. (2007), among others, suggested that the participants in the Perceptual Judgment Task would prefer audiovisual asynchronies between ±200 ms (SG or GS), while Habets et al. (2011) and Özyürek et al. (2007) found preferred windows of AVI for speech-gesture combinations between -160 ms and -360 ms (GS). To narrow down the optimal and tolerable window of AVI for co-occurring speech and gesture utterances, more extensive degrees of asynchrony as well as delay and advance of both audio and video signals in equal shares were tested. Another novelty in the methodology of the perceptual Judgment Task was the usage of naturally co-occurring speech-gesture utterances (Studies 1-3) and physical cause-and effect stimuli (Study 4) rather than artificially combined audiovisual signals in order to make more reliable predictions on natural perception.

The results of Study 4 for the physical cause-and-effect stimuli can be seen as confirmation of the judgment ability of the participants in Studies 1 through 3. Participants were able to discern between asynchronies differing by steps of 200 ms and to rate these asynchronies with regard to how natural they perceived them. Due to the full lip visibility in Study 1, this data could be compared with findings from previous research on the AVI of speech-lip stimuli (Chapter 7.1.5). Study 2 presented the participants with reduced versions of the stimuli from Study 1 – while the lips were hidden by blurring out speakers' faces, head motion was still visible. The results of Study 2 reflected those of Study 1 in terms of the ratings of naturalness for the different degrees of asynchrony; the effect of the degree of asynchrony on the degree of perceived naturalness was not significant in either Study 1 or 2 (Chapter 7.2.4. In Study 3, the stimuli with original synchrony were clearly preferred to those with audio advances of 200 ms (SG; *p* < .001) or delays of 200 ms (GS; *p* < .001), whereas no significant contrasts were found for the other levels of asynchrony. These results support hypothesis (2) of this dissertation, namely that "[l]isteners are able to discriminate variation in the synchrony of spontaneously co-produced speech and gestures and they will prefer a window of AVI encompassing both gestural advance and delay" (p. 85). This range of accepted or preferred windows of temporal asynchrony between the modalities supports hypothesis (4) as well, that is that "[t]he preferred synchrony of speech and gesture in perception will vary from that produced during spontaneous utterances" (p. 85). That participants were not able to conclusively rate the perceived naturalness of speech-gesture asynchronies beyond ±200 ms supports the findings from previous research on the optimal window of AVI for such stimuli. A breakdown in the perceptual alignment at discrepancies between ±250 ms and ±500 ms in either direction, as suggested by Massaro et al. (1996) for multimodal syllable stimuli, seems likely for speech-gesture stimuli as well. As Study 3 of the Perceptual Judgment Task demonstrated, however, "(5) [t]he preferred synchrony of speech and gestures will vary for different gesture types as well as for non-speech signals" (p. 85), and so might the acceptable window of AVI.

The results form the Perceptual Judgment Task were inconclusive regarding the occurrence of an AVI breakdown. In order to concretize the temporal alignment mechanism of the GP-SP transmission cycle, however, the temporal borders of such a breakdown need to be specified. To zero in on this as well as on the preferred window of AVI for speech-gesture utterances, an additional methodological approach is needed. The Preference Task will elicit participants' preferences for speech-gesture asynchronies by having them resynchronize former desynchronized speech-gesture utterances.

8 Preference Task

Since all stimuli with obscured heads received naturalness-ratings of more than 60% in Studies 2 and 3 of the Perceptual Judgment Task, no specific temporal window of AVI for speech and gesture can be estimated on the basis of these results. The window might go beyond the presented levels of asynchrony or it might lie somewhere in between. The Preference Task has been designed to further approximate the possible range of AVI for speech-gesture utterances. It is aimed at testing whether “(3) [l]isteners are able to reproduce the synchronization they prefer between speech and co-produced gestures”(p. 85) and whether they can specify what timing of speech and gesture they prefer without any options to choose from. Study 5 examines the stimuli from the Perceptual Judgment Task to investigate possible differences in the AVI of speech and gestures in general. Study 6 then focuses on the variation of AVI between various gesture types.

8.1 Study 5

8.1.1 Participants

20 native speakers of German took part in Study 5 (14 women, 6 men, M_{age} = 25.80 years, age range: 21-40 years). The participants were all university students and the two best performers, that is, those who got closest to the original synchronizations, were promised a 25€ voucher for a popular online retailer. This incentive was intended to make the participants more motivated in the tedious task of re-synchronizing the stimuli.

8.1.2 Materials

The test of between-subjects effects for head visibility in the Perceptual Judgment Task revealed a significant difference between Studies 1 and 3 ($F(1, 7674) = 38.39$; $p < .001$) and Studies 2 and 3 ($F(1, 6963) = 8.89$; $p < .005$). To ensure no prosodic visual cues distract participants from the gestural stimulus, the variants with the blocked-out heads from Study 3 were used for the Preference Task. 15

clips with prominent iconic gestures were selected (Table 7) and manipulated with five different initial asynchronies, dependent on the frames in Adobe Premiere Pro CS5 (277 ms, 451 ms, 607 ms, 754 ms, 1034 ms). Silences stemming from the original recordings were put in front of the audio to create fragments of equal length as the video tracks. The resulting clips were expanded before and after the fragments with silences and still frames to create space for “sliding” the channels back and forth during the resynchronization. The interface for re-aligning the audio (.mp3) with the video (.mov) was the annotation program ELAN 3.9.1 (Crasborn & Sloetjes, 2008) in its media-synchronization-mode.

In order to verify the methodology of the Preference Task and at the same time test the participants’ AVI-abilities, two physical events from Study 4 were desynchronized in the same manner as the gesture clips (clips a & b). The stimuli of a hammer hitting a nail and of fingers snapping were each manipulated to have the video precede the audio by 902 ms (VA). This strong asynchrony was selected to avoid participants accepting the desynchronized video as the original. The physical event stimuli were used as a trial and also functioned as baseline for the speech-gesture stimuli.

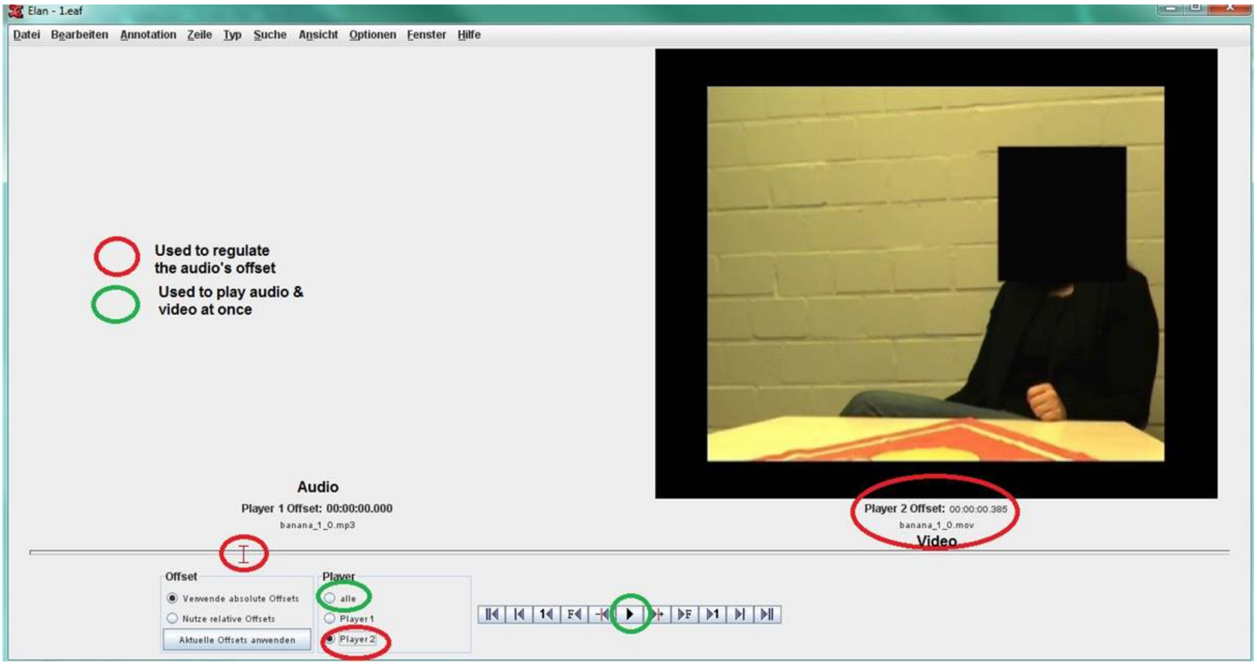
Table 7: List of stimuli used in Study 5.

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
a	hammer	How To Hammer A Nail	1410	6	hammer hits nail	physical
b	snap	How to snap your fingers Tutorial	5070	7	fingers snap	physical
1	banana_1_0	11.00.31.621	1482	230	dann lockt er den affen mit 'ner (breathy:) banane	iconic
2	cage_1_0	10.17.48.959	915	31	also er macht den käfig auf	iconic
3	can_1_0	15.04.57.785	2502	681	der rennt mit de<e>r geldspendebüchse rum /	iconic
4	catapult_1_0	14.27.42.306	3001	1304	daraufhin / wird er in die luft katapultiert	iconic
5	climb_1_0	11.00.31.621	1400	200	un klettert da ers' rau<u>f / #	iconic
6	cover_1_0	11.00.31.621	1271	290	# will dann die decke runtermachen	iconic
7	hat_1_0	15.04.57.785	2656	685	ja und dann zieht er den hut so höch und dann erkenntse dass (laughing:)	iconic

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
					s ne katze is	
8	hit_1_0	15.04.57.785	2428	707-709	und haut ihm mit dem (laughing:) regenschirm wieder fleißig übern detz %	iconic
9	in_pipe_1_0	10.17.48.959	938	76	ja ja er is in dem regenrohr	iconic
10	kicked_out_1_0	14.27.42.306	2510	1247	%<ähm>/ sylvester fliecht sofort wieder raus	iconic/ deictic
11	knock_1_0	11.00.31.621	2884	267	wo er dann als roomboy verkleidet is un' anklopft #	iconic
12	penny_1_0	10.17.48.959	1551	51	dann ja hier's is 'n penny oder	iconic
13	pipe_1_0	10.17.48.959	2530	19	er <n> klettert das abwasserrohr hoch	iconic
14	swallow_1_0	10.17.48.959	1546	62	/ er schluckt die kugel # %	iconic
15	bino_1_0	16.11.09.878	1635	805	halt mit'm fernglas durche gegend kuckt	iconic

8.1.3 Procedure

Study 5 was conducted in a quiet room on a notebook computer (1366x768 px; 15.6") with the sound coming from closed headphones (Sennheiser HD 201). The 15 stimuli were given in reversed order to half of the participants to control for sequential effects and the video size, screen contrast, and brightness were kept constant. The instructor explained the ELAN synchronization interface to the partici-


Elan - 1.eaf
Datei Bearbeiten Annotation Zeile Typ Suche Ansicht Optionen Fenster Hilfe
Used to regulate the audio's offset
Used to play audio & video at once
Audio
Video
Offset
Player
Aktuelle Offsets anwenden

Player 1 Offset: 00:00:00.000
banana_1_0.mp3

Player 2 Offset: 00:00:00.385
banana_1_0.mov

Offset: ● Verwende absolute Offsets ○ Nutze relative Offsets

Player: ○ alle ○ Player 1 ● Player 2

Figure 30: ELAN in synchronization mode as used in the Perceptual Judgment Task.

pants and showed them with the help of an example stimulus how resynchronize the channels by sliding the audio offset.

In the interface (Figure 30, cf. Figure 17), the audio and video channels are accessed through two media players. With the extended video track in fixation, the participants “slide” the audio file into place onto the video track. The participants’ task was to resynchronize the clips until they felt they were synchronized correctly.

8.1.4 Results

The temporal positions for both channels as set by the participants were entered into a spreadsheet program and the divergences from the original clip synchrony were calculated. This way, the actual preferred offsets from the original synchrony were determined. After transferring the calculated data into SPSS, descriptive statistics were elicited. The asynchronies set for the physical cause-and-effect stimuli had a range of 1420 from -978 ms (GS) of audio delay to +442 ms (SG) audio advance (excluding outliers 1 through 4: 440 ms; -154 ms GS to +286 ms SG) with M = 13.18 ms (*N* = 40, *SD* = 245.495). The asynchronies set by the participants for the speech-gesture stimuli had a range of 2662 ms from -1908 ms audio delay to +754 ms audio advance (2014 ms; -1361 ms GS to +653 ms SG excluding outliers 1-4) with M = -72.59 ms (GS) (*N* = 300, *SD* = 421.327). This difference in range and variation is clearly displayed in Figure 31 and Figure 32. The data were en-

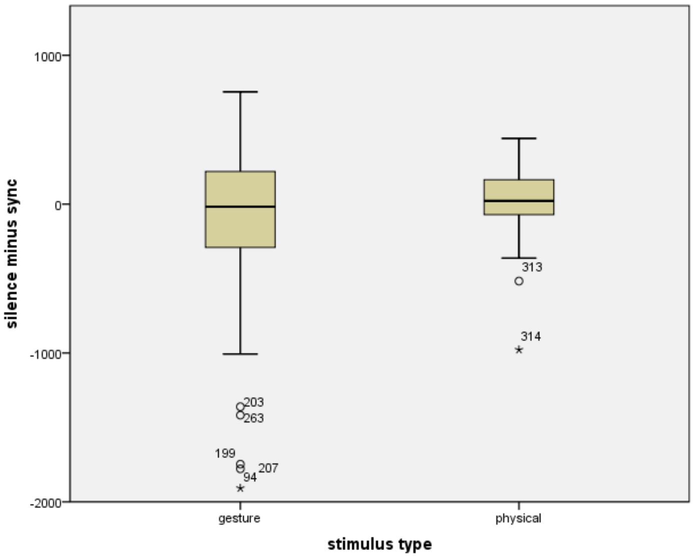


Figure 31: Range of asynchronies set for different stimulus types in Study 5.

tered into a univariate ANOVA, which revealed no significant main effect of the stimulus type (gesture vs. physical) on the preferred asynchrony of the stimuli (*F*(1, 338) = 1.58; *p* = .209).

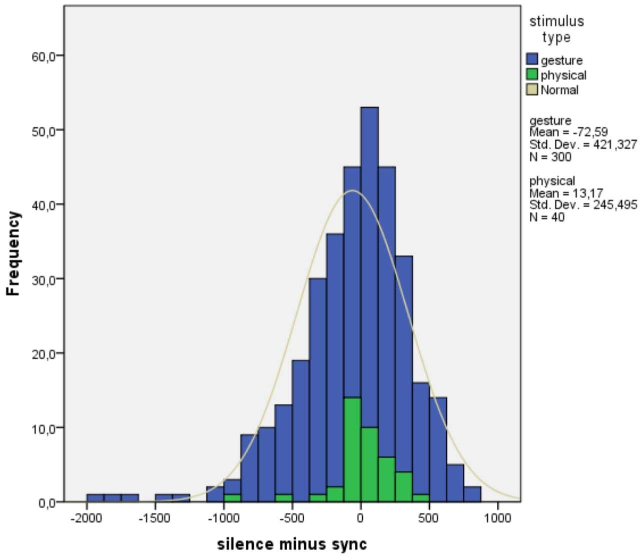


Figure 32: Histogram of range of asynchronies set for different stimulus types in Study 5.

8.1.5 Discussion

The ranges of asynchronies set by the participants in Study 5 were very different for the physical cause-and-effect stimuli (440 ms) from the speech-gesture stimuli (2014 ms). Not only is the range for the speech-gesture stimuli wider, but also is the variation within this range. However, further physical stimuli need to be tested to relativize the difference in numbers compared to the SG-stimuli (Chapter 8.2). The participants were able to approximate the preferred asynchronies for physical cause-and-effect stimuli from the Perceptual Judgment Task (> -200 ms VA and < +400 ms AV). The optimal window for AVI, that is the temporal range set by the participants, for the physical stimuli, excluding outliers, lies between -154 ms of audio delay (VA) and declines over +286 ms of audio advance (AV) towards a cut-off at +442 ms (AV).

As in the Preference Task, the participants in Study 5 displayed a wider range of acceptance for the speech-gesture than for the physical stimuli. The window of AVI set by the participants is distributed in a near-normal curve around the +200 ms (SG) mark and spreads out rather evenly between an audio delay of -1361 ms (GS) and an audio advance +653 ms (SG), excluding outliers (see Figure 32).

While Habets et al. (2011) and Özyürek et al. (2007) found preferred windows of AVI for speech-gesture combinations between -160 ms and -360 ms of speech delay (GS), the results of Study 5 clearly broaden these preferred windows. And, even though the data does not support the breakdown of AVI for discrepancies between ±250 ms and ±500 ms as suspected by Massaro et al. (1996) for speech-lip syllable stimuli, their hypothesis of a window for optimal AVI between ±250 ms still holds.

8.2 **Study 6**

Study 5 showed that the slider methodology is appropriate for eliciting the preferred audiovisual synchronization of our participants for speech gesture stimuli as well as for physical cause-and-effect stimuli. While it made use of mostly iconic and iconic-metaphoric gestures, the speech-gesture continuum McNeill (2005, p. 7) described based on Kendon (1988) suggests that a variation in temporal synchrony preference might apply for varying gesture types, namely for emblems versus other “gesticulations”. Study 6 examines this possible variation using the methodology of Study 5 for selected new stimuli.

8.2.1 **Participants**

23 German native speakers (13 women, 10 men, M_{age} = 27.91 years, age range: 20-45 years) completed the Preference Task in Study 6. They were again gathered from the university student population and an incentive was provided to enhance their motivation for precision.

8.2.2 **Materials**

6 physical cause-and-effect events not previously used as well as 13 novel speech-gesture stimuli, that is, 4 deictic, 3 emblematic, and 6 iconic gestures (see McNeill, 2005, pp. 38f.) were created from the natural data corpus using the same methodology as in Study 5. A list of the stimuli can be found in Table 8.

Table 8: List of stimuli used in Study 6.

clip	stimulus	recording	dur (ms)	gphr	speech	gesture type
1	book_with_silence	book	1170	1	book closed	physical
2	clap_with_silence	clap	3060	2	hands clap	physical
3	glass_with_silence	glass	1080	3	glass clinked	physical
4	keyboard_with_silence	keyboard	2070	4	key pressed on keyboard	physical
5	knock_with_silence	knock	3110	5	knock on table	physical
6	sekt_with_silence	sekt	3020	6	sekt pop	physical
7	climb_1_0	11.00.31.621	1400	200	un klettert da ers' rau<u>f / #	iconic
8	cover_1_0	11.00.31.621	1271	290	# will dann die decke runtermachen	iconic
9	knock_1_0	11.00.31.621	2884	267	wo er dann als roomboy verkleidet is un' anklopft #	iconic
10	bino_1_0	16.11.09.878	1635	805	halt mit'm fernglas durche gegend kuckt	iconic
11	bird_1_0	16.36.00.692	4383	835	und in dem film geht es da<a>rum dass sylvester scharf auf <äh> den vogel is	deictic/ trace
12	elevator_1_0	14.27.42.306	1322	1285	dort geht es # links zum elevator /	deictic
13	opposite_1_0	11.17.45.463	2423	381	also das war'n so zwei hochhäuser an so / auf so ner stra<a>ße %gasp	deictic
14	ring_1_0	11.00.31.621	1079	248	<mm> also das telefon klingelt	deictic
15	ring_2_0	11.00.31.621	760	370	un' dann klingelt's	iconic
16	rub_1_0	10.54.29.104	1452	143	(whispered:) da (creaky:) drüben (creaky:) isser (creaky:) endlich der leckere vogel #	iconic
17	sign_1_0	13.09.12.480	2559	593	draußen aufm schild steht <äh> hunde und katzen verboten #	emblematic
18	thumbs_up_1_0	Testbericht SCORPION EXO-500	243	n/a	klasse	emblematic
19	whyever_1	16.36.00.692	923	906	warum auch (laughing:) immer %laugh	emblematic

8.2.3 **Procedure**

The participants were presented with the same experimental setup as in Study 5, that is one ELAN interface in media synchronization mode for each of the 19 stimuli. They were instructed to resynchronize the 6 physical-event clips for means of a trial, and the 13 speech-gesture clips as the actual study.

8.2.4 Results

The resynchronization of the physical cause-and-effect stimuli resulted in a range of 1542 ms between an audio delay of -966 ms (VA) and an audio advance of +576 ms (AV) (excluding outliers 1 through 4: 881 ms; -597 ms VA to +284 ms AV) with M = -121.61 ms (*N* = 144, SD = 228.504). The range of preferred synchronization varies for speech-gesture stimuli (see also Figure 33 and Figure 34). The overall range in which the participants resynchronized the speech-gesture stimuli

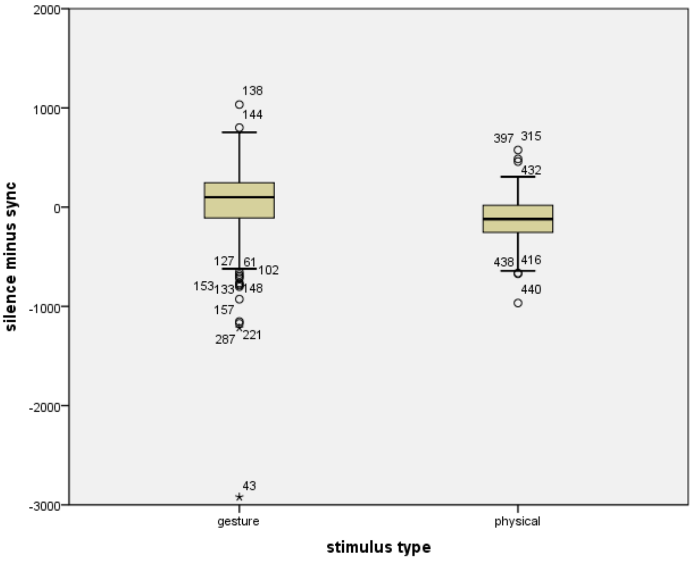


Figure 33: Range of asynchronies for different stimulus types in Study 6.

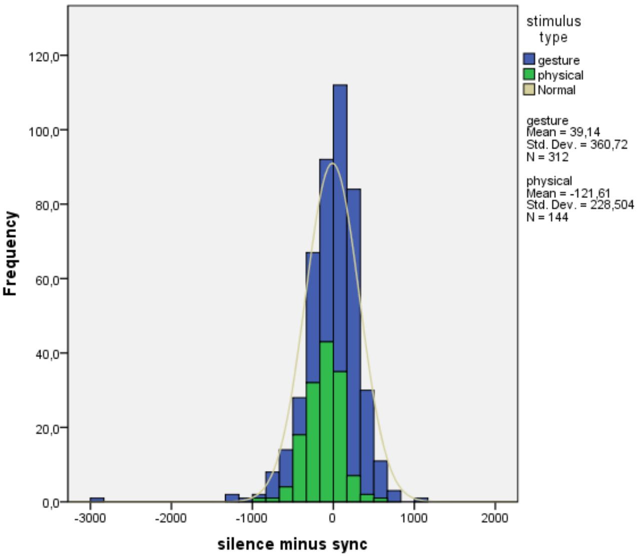


Figure 34: Histogram of range of asynchronies set for different stimulus types in Study 6.

is along 3955 ms, between an audio delay of -2921 ms (GS) and an audio advance of +1034 ms (SG) (excluding outliers 1 through 4: 1534 ms; -927 ms GS to +607 ms SG) with M = 39.14 ms (*N* = 312, SD = 360.720).

The three gesture types tested vary from this overall range as follows (see also Figure 35 and Figure 36): The iconic gestures were aligned with their co-produced speech by the participants along a range of 3955 ms between -2921 ms (GS) of audio delay and an audio advance of +1034 ms (SG) (excluding outliers 1 through 4: 1249 ms; -655 ms GS to +594 ms SG) with M = 30.17 ms (*N* = 144, *SD* = 395.233). The deictic speech-gesture stimuli were resynchronized along a range of 1622 ms between an audio delay of -1171 ms (GS) and an audio advance of

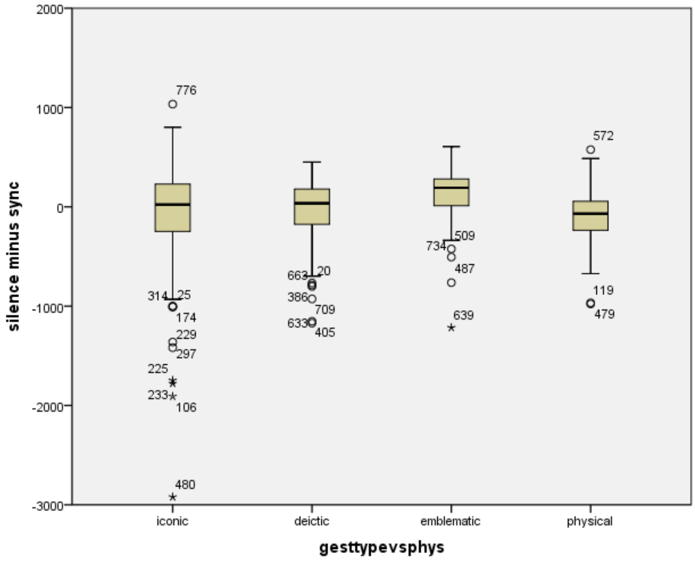


Figure 35: Range of asynchronies for gesture types and physical events in Study 6.

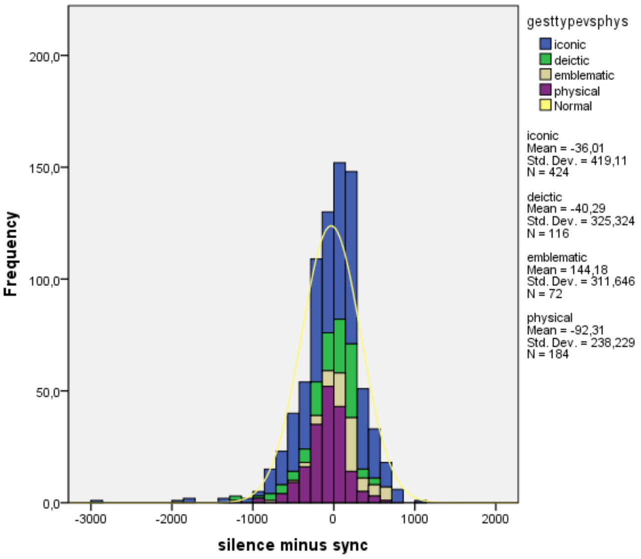


Figure 36: Histogram of range of asynchronies for gesture types and physical events in Study 6.

+451 ms (SG) (excluding outliers 1 through 4: 1141 ms; -787 ms GS to +354 ms SG) with M = -26.17 ms (*N* = 96, *SD* = 324.732). The emblematic gestures were realigned with their co-produced speech by the participants along a range of 1823 ms between an audio delay of -1216 ms (GS) and an audio advance of +607 ms (SG) (excluding outliers 1 through 4: 942 ms; -337 ms GS to +605 ms SG) with a M = 144.18 ms (*N* = 72, *SD* = 311.646).

All data from Study 6 was recoded to run an ANOVA with the gesture types versus the physical stimuli as independent variable. This analysis revealed a significant main effect of this variable on the synchrony set by the participants (*F*(3, 455) = 12.13; *p* < .01). Contrasting the gesture types with the physical events, the iconic and emblematic gestures were highly different (*p* < .01). The deictic gestures elicited slightly less yet significant temporal differences in the participant synchrony preferences (*p* < .05).

8.2.5 **Discussion**

The preferred overall speech-gesture synchrony in Study 6, that is the temporal window the participants set for the SG-stimuli, had a range of 1534 ms (-927 ms GS to +607 ms SG) while the preferred physical cause-and-effect synchrony ranged over 881 ms (-597 ms VA to +284 ms AV). Apart from the general difference in preferred synchrony, the tendency of the participants to select an audio advance in the stimuli is prominent. This fits with what Winter and Müller (2010) stated, that is, that "[i]n a natural dialog situation speech is perceived as integrated [sic] audiovisual event I which the auditory speech signal lags the visual signal approx. 3 ms per meter between talker and perceiver, though it is not temporally aligned" ("INTRODUCTION"; see also, e.g., Einstein, 1905/2005). That asynchronous production quite possibly facilitates a synchronous perception can be supported also for speech and gestures through the findings of the Perceptual Judgment Task.

The striking finding of Study 6 is the variation of preferred audiovisual synchrony for the different types of gestures with their co-produced speech. Expanding on McNeill (2005, p. 7), speech and gesture should be more closely semantically linked for iconics than for deictics, which is also reflected in the temporal synchrony in our data (iconic gestures: 1249 ms; -655 ms GS to +594 ms SG vs. deictic gestures: 1141 ms; -787 ms GS to +354 ms SG). In the same continuum of semantic synchrony, emblems are described as least semantically linked to speech since they are comprehensible without speech. In Study 6, emblematic gestures with naturally co-occurring redundant speech were examined, which resulted in the closest preferred temporal synchronies of all gestures (emblematic gestures: 942 ms; -337 ms GS to +605 ms SG). This in turn might be due to the redundant semantic relation between the modalities, which is more complementary in deictic speech-gesture combinations and mostly associative in iconic speech-accompanying gestures. The smaller window of AVI for emblematic and deictic gestures with co-produced speech is closer to the preferred window of AVI for physical cause-and-effect stimuli (881 ms; -597 ms VA to +284 ms AV). While speech is not caused by gestures, it is caused by air flow through the speech apparatus. There are certain multimodal proximity pairs expected by the listener to occur together, such as a deictic verbal expression like “over there” with a gestural one such as pointing over there alongside it. An even stronger expectation of semantic alignment, with our without temporal synchronization, might happen with gestural emblems – if they are accompanied by any speech at all, it should reinforce the gesture and hence be semantically redundant, such as a thumbs-up with a simultaneous “Well done!”.

8.3 Discussion Preference Task

8.3.1 Results of Studies 5 and 6

The range of the preferred gesture-type independent speech-gesture synchrony slightly increases when the data from Studies 5 and 6 are combined to 2172 ms, excluding outliers, (1418 ms GS to 754 ms SG). The range of preferred physical cause-and-effect synchrony also slightly increases to 995 ms (643 ms VA to 352 ms AV). The crucial difference of more than 1 s in the preferred AV synchrony by the participants remains just as clear (Figure 37). An overall main effect of stimulus type on the degree of synchrony entered by the participants was discovered (*F*(3, 792) = 7.42E8; *p* < .01).

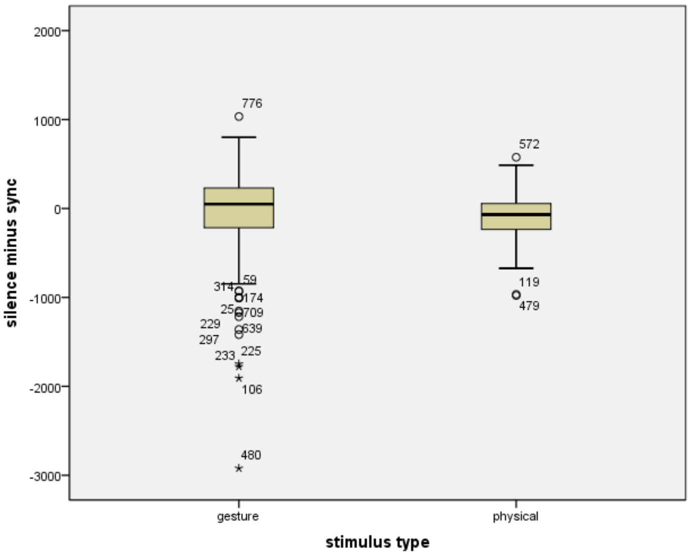


Figure 37: Range of asynchronies set for gestures and physical events in Studies 5 & 6.

Combining the results of Studies 5 and 6, there is a clear variation in synchrony range between gesture types and between physical events (Figure 38).The post hoc Tukey test on the various gesture types and the physical event stimuli revealed a significant impact of the stimulus type on the preferred synchrony only for the emblematic gestures (*p* < .01). The iconic (*p* = .078) and deictic (*p* = .226) gestures with their co-produced speech did not contrast as strikingly with the cause-and-effect stimuli. Taking the iconic gestures as the reference category, emblematic function significantly influences the preferred speech-gesture synchrony (*p* < .01), but deictic gestures are show a marginal significant difference from the iconic ones (*p* = .078).

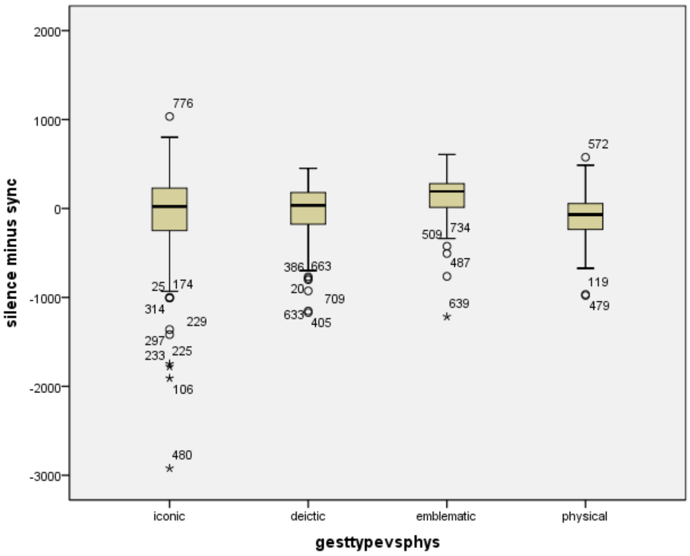


Figure 38: Range of asynchronies set for gesture types and physical events in Studies 5 & 6.

There still is a narrow window for the preferred synchrony of physical events (87 ms VA to 672 ms VA; *SD* = 214.4), and the iconic gestures are synchronized only loosely with their speech (908 ms GS to 778 ms SG; *SD* = 386.4). The resynchronizations of emblematic and deictic gestures show different patterns: Both got resynchronized closer to their original timing than the iconic gestures. The deictics were readjusted more similarly to the physical events (51 ms GS to 1171 ms SG; *SD* = 321.2), with more of a tendency toward an audio advance. The emblematic gestures were also resynchronized more closely with their non-obligatory speech (607 ms GS to 1216 ms SG; *SD* = 284.4) than the iconic ones to their disambiguating speech. It appears there are some conditions for speech-gesture AVI after all.

8.3.2 **General discussion of Studies 5 and 6**

The results for the physical events and speech-gesture utterances show that participants accept delays or advances in both the acoustic and the visual modality. Like the Perceptual Judgment Task, this supports hypothesis (2) of this dissertation that "[l]isteners are able to discriminate variation in the synchrony of spontaneously co-produced speech and gestures and they will prefer a window of AVI encompassing both gestural advance and delay" (p. 85), which has been a major gap in previous research. The Preference Task supports the results of the Perceptual Judgment Task by confirming and even expanding the wide range of accepted offsets: hypothesis (3), that "[l]isteners are able to reproduce the synchronization they prefer between speech and co-produced gestures" (p. 85), could equally be supported by the results. However, while audiovisual stimuli such as physical events and speech-lip signals require a production-like, tight synchrony, the relevance of such a synchrony between speech and gesture is not supported. Deictic and emblematic gestures do seem to entail a closer temporal synchrony to their co-occurring speech than iconic gestures. This may be due to a closer semantic relation between the modalities during the phase of synchronous production.

The audio and video in the physical events stand in a causal relationship while speech and gesture share a semantic, conceptual connection. In multimodal language production, they temporally align to a certain degree. The speech-gesture

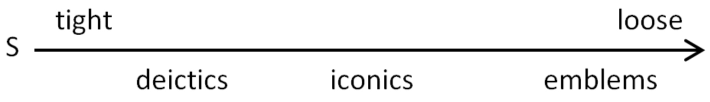


Figure 39: Continuum of semantic synchrony of speech and gesture types.

continua by McNeill (2005, pp. 7ff. based on Kendon, 1988) give a more specific explanation of the varying levels of gesture-speech entrainment. McNeill (2005) classifies gestures along a continuum regarding the obligatoriness of speech: For 'gesticulations', such as iconic and deictic gestures, speech is mandatory for disambiguation and complementation, while for emblems it is optional; for pantomime and sign language speech need not be present. One can modify this continuum to include deictic and iconic gestures in lieu of the encompassing gesticulations (Figure 39):

One can hypothesize that with loosening semantic synchrony the need for temporal synchrony becomes less because of the decreasing disambiguating function of co-occurring speech. Another factor is the theme-rheme frame discussed in Kirchhof (2011; Chapter 6.6), which binds the gesture to a certain sentential and hence temporal frame of an utterance. These frames are present in the stimuli of both the Perceptual Judgment Task and the Preference Task, and the participants accepted larger temporal asynchronies than had been found in production. Hence, I hypothesized that gestures only need to synchronize loosely with their co-occurring speech. The Preference Task disproves this to a certain degree because different windows of AVI are accepted by the participants for different gesture types: Emblems seem to need more synchrony with speech than deictics and deictics than iconics. This information can provide us with a sketch of a temporal continuum (Figure 40) diverging from the semantically governed one:

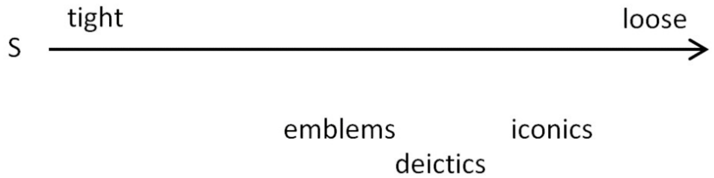


Figure 40: Continuum of temporal speech-gesture synchrony in perception.

The close temporal synchrony between speech and gesture is a well-known production phenomenon, and it seems be more important for AVI than previously thought. Since iconic gestures complement phrases and utterances, the temporal window for their AVI is only bound by the utterance duration and the timing within this boundary is flexible. Deictic gestures correspond to deictic parts of speech (POS), the closest a gesture can be to lexical affiliation with speech. They are semantically and temporally bound and their phases are short, which makes the temporal window for AVI small. Emblematic gestures, then, are a special case. When they occur together with speech they are redundant to certain POS. In the Preference Task, participants synchronized them closely to their temporal production synchrony, which suggests a tight semantic and temporal bound between the two modalities for this gesture category. As with deictic gestures, their phases are short, but, due to their redundancy, the window for AVI is slightly larger.

As de Ruiter (2000) and Kirchhof (2011) already suggested, the relation between gestures and speech is governed by conceptual bounds. For perception, this conceptual package is transmitted by an internal (re)synchronization of the duration of the gphr with the speech it is semantically associated with, by AVI. Within one theme-rheme pair, production-like synchrony is not necessary for the listener. However, it might be restricted to the duration of a full utterance, which might contain more than on theme-rheme pair (Chapter 6.6). I suggest that gesture-speech synchrony within utterance borders is a predominantly production-based phenomenon. This explains why in the Perceptual Judgment Task and the Preference Task there was a wide range of accepted as well as of preferred asynchronies between the speech and co-expressed gesture: Listeners do not require speech-gesture synchrony and hence cannot reproduce it.

As McNeill (2012) speculated on the conceptual transmission of a speech-gesture utterance, “the time limit on growth point asynchrony is probably around 1~2 secs., this being the range of immediate attentional focus” (p. 32). The GP is temporally flexible in perception, with the possibility of either modality preceding the other by up to 1418 ms, depending on the gesture type. One can observe a semiotic connection between the two modalities by analyzing co-produced speech and gestures. What cannot be done so easily is to desynchronize or semantically mismatch speech and gesture during production (cf. Holler et al., 2009). Our results strongly suggest that speech-gesture synchrony is rather a consequence of the production system but, as far as actively set preferences are concerned, seems not to be crucial for comprehension. This finding should allow for a higher tolerance of timing in modeling gestures in virtual agents and robots and could inform and inspire future research into the perception of naturally co-occurring speech and gestures.

As has been briefly discussed in the beginning, this dissertation does not aim to explicitly analyze the relevance of speech-gesture production synchrony for comprehension, but for perception. Transferring the findings of the temporal windows of AVI from the varying sets of studies to the model of the GP-SP transmission cycle will need to take into account the temporal flexibility in perception nevertheless, since any multimodal utterance will have to be integrated by L to facilitate comprehension. The model draft shown in Figure 13 already included alignment mechanisms in the perception module, as well as in the production unit. It is now possible to further specify the temporal tasks and restrictions of the perception module in the model. This as well as other additional factors gained from the results of the Conceptual Affiliation Study (Chapter 6), the Perceptual Judgment Task (Chapter 7), and the Preference Task (Chapter 8) will be discussed in the following and concluding Chapter 9.

9 General Discussion and Conclusion

9.1 On the Relevance of Speech-Gesture Production Synchrony for the Listener

While there has been a growing number of studies on the perception of speech-accompanying gestures over the last years, there has been a lack of investigations on naturally co-produced speech-gesture utterances. Due to methodological restrictions, no considerable data on the ability of listeners to AVI larger speech-gesture asynchronies, or speech proceeding gphr, has been gathered. Such methodological restrictions are, for instance, that EEGs require punctual and not interval targets, that there is often still a focus on lexical affiliates and that the assumption prevails that gestures and speech need to be linked in production-like synchrony, or, if desynchronized, that the gesture may only precede the speech, but not vice versa. Accordingly, no proper assumptions could be made about the tolerable or optimal windows of AVI for speech-gesture utterances, a gap that has long been closed for speech-lip utterances, possibly because of its applicability in the film industry.

To close these gaps in speech-gesture research has been major aim of this dissertation. This was approached by exploring whether listeners can identify larger as well as bidirectional speech-gesture asynchronies and whether they can actively resynchronize such asynchronies. By combining these two approaches of perceptual judgment and preference, the relevance of synchronization between verbal utterances and their accompanying gestures for perception, and from that for comprehension, was to be identified. In the context of investigating the relevance of speech-gesture synchrony, the general understanding of this synchrony was addressed before the background of temporal interval relations. This ensured a consistent usage of the different kinds of synchrony between speech and gesture intervals , each containing a well-formed verbal utterance or a complete gphr, throughout this dissertation (Chapter 2.3).

Prior to investigating the relevance of temporal synchrony, the issue of lexical affiliation had to be addressed, an analytical phenomenon grounded on the temporal linkage of prosodic peaks in speech and the apex of gesture strokes. Particularly in the context of GP externalization, the interplay of holistic gestures with the iterative semiotics of speech had to be analyzed with a focus on temporal and semantic linkage. Based on the GP theory (Chapter 3.2), that is, that within an MU an information package is formed that includes all imagistic and linguistic information necessary to express what the speaker wants to relate to the listener, the SP hypothesis was formed (Chapter 4.5): What is available to the listener through the unpacking of the speaker's GP will be audiovisually integrated by the listener through processes of perception and temporal as well as semantic alignment, and then compressed into the SP. The SP, like the GP, would be an ideational unit maximally resembling the speaker's GP, but modified by the listener's communicative intent, personal background, etc. Throughout this dissertation, a model draft of a GP-SP transmission cycle was being developed and readjusted. In this model draft, findings from previous research, the Conceptual Affiliation Study (Chapter 6), the Perceptual Judgment Task (Chapter 7), and from the Preference Task (Chapter 8) were formalized. To finalize the model of GP-SP transmission, the six hypotheses posed at the beginning of this dissertation will now be readdressed below before the background of these new developments.

Hypothesis (1) implied that “[t]he semiotic-semantic relation between spontaneously co-produced speech and gestures is not restricted to the lexical item(s) of the speech the gesture stroke synchronizes with but encompasses all newsworthy information given in speech” (p. 85). Revisiting the GP theory brought to light that the semiotics of gestures are holistic in nature in that they can communicate various features of an idea at once, in contrast to speech. While all or one feature expressed through a gesture can at times be semantically connected with a word or phrase contained in the co-expressed verbal utterance (lexical affiliation), this is not always the case. The idea of lexical affiliation founded mostly on the fact that parts of the verbal utterance temporally coincide with the apex of the stroke of an accompanying gesture. This factor was excluded in the Conceptual Affiliation Study (Chapter 6) by exposing participants to stimuli created from naturally co-expressed speech and gphrs successively instead of synchronously. Inquiring about which part or parts of the verbal utterance corresponded most with the gphr revealed a variety of answers that excluded the possibility of *the* lexical affiliate for any of the spontaneously produced deictic and iconic gestures used as stimuli in the experiment. The qualitative analysis of the data resulted in the finding that one gphr corresponds to one rheme at a time only, much as one GP relates to one psychological predicate. This was explained through conceptual affiliation, that is, that gphr are co-expressive with the ideational concept that the speaker wants to relate within one GP-SP cycle. Choosing a holistic, conceptual affiliation of speech and gesture allows for cross-utterance semantic relations, while lexical affiliation based on temporal coinciding allows only for atomic, restricted interpretations of meaning. The concept-based understanding of speech-gesture affiliation does not exclude the occasional occurrence of lexical affiliation, however, but rather subsumes it. The function of prosodic emphasis that gestures can also serve (e.g. Wagner, Origlia, Avesani, Christodoulides, D'Imperio, Escudero, Lacheret, et al. 2015) is also preserved within conceptual affiliation.

The disestablishment of temporally bound lexical affiliation[17] as well as findings from the areas of psychophysics and speech-perception research “suggested that participants would prefer audiovisual asynchronies between ±200 ms (SG or GS), while Habets et al. (2011) and Özyürek et al. (2007) found preferred windows of AVI for speech-gesture combinations between -160 ms and -360 ms [(GS)]” (p. 136). These findings provided grounds to investigate hypothesis (2): Listeners are able to discriminate variation in the synchrony of spontaneously co-produced speech and gestures and. . . will prefer a window of AVI encompassing both gestural advance and delay” (p. 85). In are series of studies within the Perceptual Judgment Task (Chapter 7), asynchronies in seven steps of 200 ms were rated for naturalness between 600 ms of S before G and 600 ms G before S in the visibility conditions 'lips visible', 'face blurred', and 'face blocked'. The participants in the 'lips visible' condition rated the stimuli on the basis of former findings on speech-lip synchrony in that they “perceived the original condition without synchrony manipulation, an audio delay of 600 ms (GS), and an audio advance of 400 ms (SG) as

[17] I hope that no new movement of Antidisestablishmentarianism will form.

most natural” (p. 125). This suggests that participants were confused by asynchronies larger than ±200 ms and that a breakdown of AVI occurred shortly after. In future studies, the stimuli used should be analyzed for rhythmic patterns that might explain the participant's ratings for the 600 ms (GS) and 400 ms (SG) stimuli. The participants in the 'face blocked' condition preferred stimuli in their original production synchrony to any asynchrony, but about two thirds of all stimuli were rated as somewhat or fully natural, regardless of the degree of asynchrony. Whether these above-chance ratings speak against or for a breakdown of AVI is debatable. Hypothesis (2), whether listeners can discriminate between different degrees of asynchrony encompassing gestural advances and delays, is supported to a certain degree by the findings of this set of perception studies; the suspected breakdown between 250-500 ms (SG or GS) could not be confirmed due to inconclusive results. However, in the follow-up study with the physical event cause-and-effect stimuli, participants clearly preferred stimuli with and audio advance of 200 ms (AV), which confirms the findings from previous research on the AVI of bimodal signals in psychophysics and speech-lip research. This indicates that participants were indeed able to select what felt most natural to them in the online interface, as did the results for the 'lips visible' condition. In order to further specify the temporal alignment mechanisms of the GP-SP transmission cycle, the point of integrational breakdown as well as the window of AVI for speech-gesture utterances needed to be concretized through the Preference Task.

Participants in the Preference Task (Chapter 8) resynchronized temporally manipulated speech-gesture stimuli as well as physical event cause-and-effect stimuli to what they felt was natural production timing. The results showed an overall main effect of stimulus type (gphr vs. physical) on the degree of synchrony, which was also clearly visible in the temporal ranges the participants set for the different stimuli (p. 148). These ranges, or windows of AVI, selected for the physical stimuli were much smaller than those for speech-gesture utterances in general and showed a clear preference for audio advance (87 ms VA to 672 ms AV). Iconic gestures were allowed about twice as much temporal space by the participants (908 ms GS to 778 ms SG) while deictics were readjusted more similarly to the physical events (51 ms GS to 1171 ms SG); emblematic gestures were also resynchronized more closely with their non-obligatory speech (607 ms GS to 1216 ms SG). With these results, again, hypothesis (2) could not be refuted, in particular not for iconic and emblematic gestures. The preference of audio advance was still strong in the participants regarding the physical and deictic stimuli. These findings can be interpreted to indicate that physical events require production-like synchrony, much like speech-lip signals, both of which are of the cause-and-effect kind to some degree. Deictics and emblems at least seem to require a closer production synchrony for perception, while iconic gestures do not at all - “with loosening semantic synchrony the need for temporal synchrony becomes smaller because of the decreasing disambiguating function of the speech toward the gesture” (p. 111). Emblems in the context of the Perceptual Judgment Task, however, are to be considered with caution because only two incidents were tested. Also, they do not require speech to be disambiguated as iconics do. Maybe the redundancy in an emblematic speech-gesture utterances indicates a tight semantic bound between speech and gesture, akin to semiotic twins[18]. This will definitely have to be explored more extensively in the future.

Returning to the investigative aims proposed initially, the results of the Preference Task partially refuted hypothesis (3), which stated that “”[l]listeners are able to reproduce the synchronization they prefer between speech and co-produced gestures” (p. 85). The synchronies set in the slider interface for the physical events and deictic gestures did in fact reflect the windows of AVI expected based on previous research and the Perceptual Judgment Task. The temporal windows set by the participants for the iconic gestures are wide, but the results for iconics from the Perceptual Judgment Task were inconclusive. This could indicate that listeners do not have any preferred window of AVI for iconic gestures within speech-gesture utterance boarders, which makes the listeners unable to relate any significant synchrony preferences for those gestures in the Preference Task. This lack of preference, along with the preferred windows of AVI for deictic and emblematic gestures, supports hypothesis (4) that “[t]he preferred synchrony of speech and gesture in perception will vary from that produced during spontaneous utterances” (p. 85). While the synchrony between gphr and the co-produced speech can by observa-

[18] @AT: Just bigger!

tion be anchored around the overlap of the prosodic peak and the gesture stroke, the findings from the Perceptual Judgment Task and the Preference Task strongly indicate that production synchrony is not required for perception and AVI. However, particularly for deictic gestures, the acceptable windows of AVI differ from those of iconic gestures in that they are close to physical cause-and effect stimuli. Hypothesis (5), that "[t]he preferred synchrony of speech and gestures will vary for different gesture types as well as for non-speech signals" (p. 85) can hence be supported for deictic and iconic gestures, and, with reservations, for emblems.

The final and central hypothesis (6), whether "[t]here is a perceptual equivalent to the *Growth Point (GP)*, that is, the *Shrink Point (SP)*" (p. 85), can neither be fully refuted nor supported by the findings from previous research or within this dissertation due to its innately hypothetical and metaphorical character. The existence of certain production and processing mechanisms for speech-gesture utterances, however, are undeniable. Presupposing the formation of conceptual packages like GPs and SPs as sub-units of MUs is a convenient means to study multimodal communication on an abstract level. Assuming a GP-SP transmission cycle will help explain how listeners will deal with temporal asynchronies diverging from the usual speech-gesture production synchrony. The Conceptual Affiliation Study (Chapter 6), the Perceptual Judgment Task (Chapter 7), and the Preference Task (Chapter 8) provided a lot of information on how speech-gesture utterances are perceived by listeners in their original production synchrony as well as when temporally shifted between up to 908 ms GS and 778 ms SG (iconics). Before the background of the three sets of studies and the successful testing of the six hypotheses formed at the beginning of this dissertation, the model draft of the GP-SP transmission cycle (Figure 14) can now be optimized further. To do so, the following aspects need to be kept in mind:

- The verbal utterance and the gphr are connected by conceptual bounds stemming from a GP that contains the instructions, so to speak, for both modalities to maximally co-express an ideational unit. Through the processes of perception and integration, the conceptual contents of the GP are reassembled into the SP.

- The production as well as the perception of speech is iterative in that meaningful unit is succeeded by meaningful unit in the form of phonemes, words, clauses, or utterances, while gestures convey their whole meaning at once during the stroke phase.

- These meaningful speech and gphrs can be regarded as temporal intervals that overlap during the production of speech-gesture utterances. Within this dissertation, complete verbal utterances containing one rheme as well as gphr containing one stroke are regarded as full intervals. That way these intervals overlap, or synchronize, has been described by Thies (2003) based on Allen's (1983) temporal logic, that is, either G contains S, S contains G, S overlaps G, G overlaps S, S occurs before G, or G occurs before S.

- A common observation in production is that the verbal pitch accent is minimally preceded temporally by the onset of the gesture stroke, that is, G overlaps S, or that the apex of the gesture stroke hits the verbal pitch accent within a word or phrase, that is, S contains G.

- The empirical investigations within the scope of this dissertation resulted in the findings that such precise synchrony as in production is not required for the perception, or AVI of speech gesture utterances. Still, for certain gesture types, for example deictics, a certain degree of synchrony is required by the listener (51 ms GS to 1171 ms SG), while for other types, such as iconics, the acceptable window of AVI is much wider (908 ms GS to 778 ms SG) or, quite possibly, not there at all.

A model of the GP-SP transmission cycle based on a Leveltian model of speech-gesture production belonging to de Ruiter's (2007) category of Window Architectures should be able to integrate these conceptual as well as temporal constraints on alignment processes in production as well as in perception. In addition, it needs to be able to explain errors within the transmission cycle, for example caused by interruptions or impairments. I hence propose the following “Model of GP-SP transmission” (Figure 41):

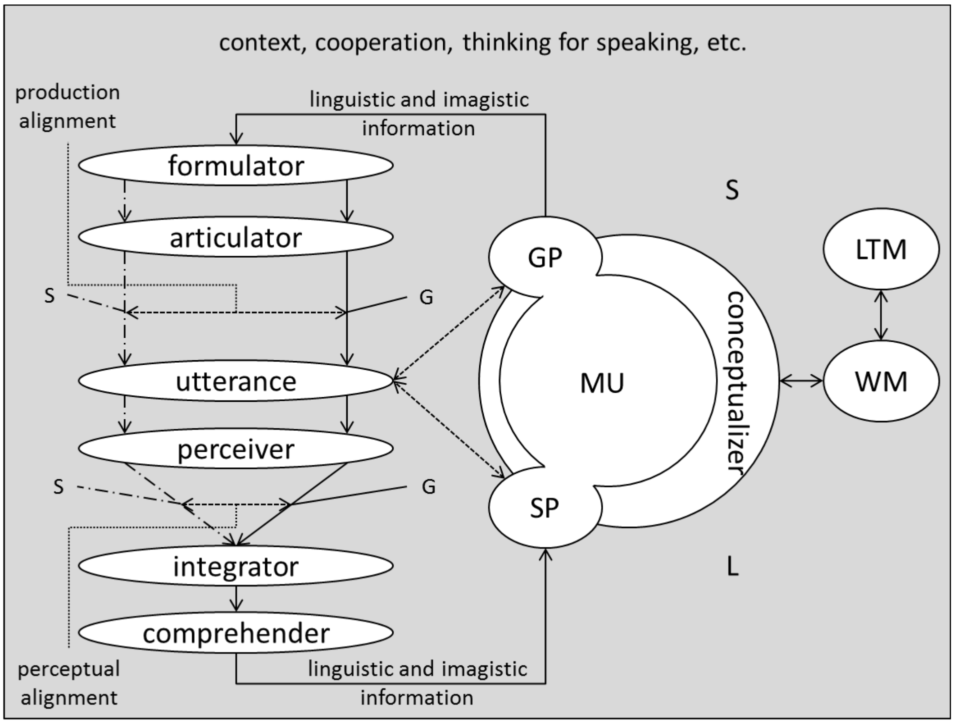


Figure 41: Model of GP-SP transmission based on the Window Architecture by de Ruiter (2007).

Going back to Figure 5, in which S' said, “dann <ähm> kommt die omma aber an /”[19] while making a hitting motion with their right hand, which grabs an invisible, stick-like object, the workings of the model can be demonstrated. S' had watched the Canary Row series, which contains various scenes in which Sylvester the cat is chased by Tweetybird's owner, the granny. She regularly beats Sylvester up with an umbrella or motions as if she would. From watching these scenes, and particularly the one that is currently the topic of conversation, S' will have an imagistic as well as linguistic resemblance in mind of what she wants to narrate to L.

Utterance production: While the memory of Canary Row as well as the experimental instructions reside in S's *WM*, with a general knowledge of cartoon logic possibly in her *LTM*, the scene to be currently described would be an *MU* assembled within the *conceptualizer*. Taking into account factors such as the communicative context, background knowledge, intentions, instructions, etc., the conceptualizer forms a GP containing all imagistic and linguistic information to be related to L. At this stage, the GP is an unsorted heap of information, so to speak, that needs to be channeled through the *formulator* to be processed into a comprehensible utterance adhering to the laws of physics as well as to certain linguistic constraints.

[19] “<ehm> but then granny comes along /”

The formulator takes into account the requirements for the successful transmission of the GP, for example that one gesture should only be co-expressed within one theme-rheme set, as well as syntactical rules and other factors. The *articulator* then takes care of the externalization of the conceptually and syntactically aligned speech-gesture *utterance*, during the process of which the GP is unpacked while speech and gphr temporally overlap. Through the gesture, the additional information that the granny is either hitting Sylvester with the umbrella, or pretending to do so, is added to the verbal utterance. Which modality initiates the overlap as well as within which temporal window is irrelevant for the successful integration of the speech-gesture utterance, but due to the iterativeness of speech, S will possibly contain or overlap G most of the time.

Utterance reception: Since L shares general cognitive and linguistic competence with S, as well as knowledge about the current context and task at hand, L will hopefully attend to utterances by S'. The *perceiver* will take note of the multimodal utterance produced by S' as well of the surroundings of the current communicative setting. Through mechanisms of prediction, for instance that what S' relates is relevant, and pattern matching, for example regarding content or speech-gesture co-expressivity, the perceiver will send signals to the *integrator*, indicating that information relevant to the communicative situation is being uttered by S'. The integrator will then combine bits and pieces of information as potentially referring to the same message, for example information co-expressed by speech and gesture across the utterance. With the help of the communicative intent of L, which in our experimental setting is that L will have to retell what S' narrates, as well as with background knowledge and other factors, the *comprehender* then will re-assemble selected information from the speech-gesture utterance into the *SP*. The currently discussed SP should contain some imagistic version of a granny chasing a cat with an umbrella at hand. The SP will then merge into a new *MU*, from which, much like in cell division, a new *GP* will emerge, after which the cycle re-initiates.

In a perfect communicative setting, with direct visibility between S' and L and full cooperation between the two, the GP-SP transmission cycle as described above would run until their conversation was over, which is rarely the case. Later in the same conversation, for instance, S' encounters a ToT state when describing the granny sitting in the cage. S' says, “<ähm> ja sie [sitzt eigentlich k* <ä] [hm> / % j] [a % sitzt] [in dem käfig <im]mer>,” during which four gphrs (gphr 1326-1329; indicated by square brackets) are performed. The gphrs, or their strokes, re-initiate with every attempt S' makes until she finally resolves the ToT state and the gesture can retract. The articulator ensures that speech and gesture are properly co-expressed, and will re-initiate co-expression until it succeeds.

In case of a semantic speech-gesture mismatch, should it naturally occur, the error would have already been processed through the formulator, and the articulator would operate on the assumption that the instructions given were correct. The same would be true for the perception process. While listeners should often contemplate whether the information they have received is true and sound, this does not happen at the perception or integration stage, but only later on in interactions between the conceptualizer and WM, which is simplifying thought processes within the GP-SP transmission cycle to a bare minimum here.

Should temporal asynchronies occur within the speech-gesture utterance, the perceiver will take all signals in regardless. The integrator, however, is only able to process linguistics or imagistic information within certain temporal constraints, for example between 51 ms GS up to 1171 ms SG for deictic gestures. Multimodal information perceived outside these constraints will not be considered for further processing: An AVI breakdown occurs. In case of such a breakdown, possibly either only the verbal or the gestural information will be passed on, or L will initiate actions to accommodate the situation. With video streaming issues, for instance, one would check the internet connection or the settings of the video player. In face-to-face communication, L would probably ask S' to repeat their utterance or take other actions to reach their communicative goal.

All these potential ways in which the model of the GP-SP transmission cycle would deal with errors, as well as with well-formed speech-gesture utterances, are hypothetical at the moment. Combining the model with knowledge and tools from computer science and previously designed computational models of speech and/or gesture processing will hopefully contribute to a deeper understanding of speech-gesture interaction in humans as well as in other communicating agents.

9.2 Theoretical Implications

The fact that listeners do not require production-like synchrony for a successful integration of speech-gesture utterances will hopefully broaden the area of focus for those still assuming the temporal coincidence of prosodic peak and gesture stroke as essential to the communicativeness of gestures. Along these lines, lexical affiliation should be expanded to conceptual affiliation in future studies on multimodal utterance interpretation, recognizing that the GP is temporally flexible in perception. That gesture-speech synchrony might be a consequence of the production system, but is not be essential for comprehension, at least to a certain degree, could also influence research on speech-gesture comprehension, particularly regarding the area of semantic, or so-called temporal mismatches. Possibly, mismatches as determined or constructed by gesture experts will not be noticed as such by listeners, which would entail a rethinking of this field of research. Personally, I would like to combine the findings on the informational integration from semantically mismatching speech and gestures and the methodology of video editing used for the experiments in this dissertation to investigate, for example, how neurolinguistic programming or semantic framing and gestures can be used to underline the urgency of medical checkups or the workings of complex processes in promotional or instructional videos.

An additional point of intersection would be to connect the integrational patterns for gestures in correlation with prosody and rhythm, as has already been discussed in relation to the results of Study 3 of the Perceptual Judgment Task. This should be telling as to what other mechanisms are involved in the AVI of speech-gesture utterances. Wheatland et al. (2015), for instance, in the context of modeling synthesized hand motions, have noted that “models based purely on prosody recognize the important correlation between gesture timing and audio changes. . ., but cannot account for deep semantics. Newer work seeks to address both. . .” (p.

17). The GP-SP transmission cycle involves both temporal and semantic factors, so an extension towards prosody and computational modeling is promising.

The findings of this dissertation, specifically regarding the temporal dimensions, should also lead to testing whether higher tolerances in the programming of the synchrony of speech and modeled gestures in virtual agents and robots would be feasible. For now, at least in the video gaming industry, actors are often recorded and digitized to achieve the highest naturalness in character modeling. Previous presentations of parts of this dissertation have already inspired investigations into the necessity of precise speech-gesture timing in social robots, for example by Srinivasan et al. (2014) regarding the programming of their Survivor Buddy Robot (p. 776). Further investigations into this area would also be possible in Bielefeld, for instance with the NAO or iCub robotics projects of the applied informatics group.

The truth is out there ...

10 References

Adobe Systems Inc. (2010). Adobe Premiere Pro CS5 (Scully) [video editing software]. Available at http://www.adobe.com/premiere/

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468-523.

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44*(2), 169-188.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of ACM, 26*(11), 832-843.

Argyle, M. (1975). *Bodily Communication*. London: Methuen.

Bavelas, J. B., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58,* 495-520.

Beattie, G., & Aboudan, R. (1994). Gestures, pauses, and speech: An experimental investigation of the effects of changing social context on their precise temporal relationships. *Semiotica, 99*(3-4), 1-40.

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology, 90,* 35-56.

Becker, A. (Producer). (2012, February 2). *Test & Testbericht SCORPION EXO-500 deutsch* [Video file]. In *Nanokultur.de*. Retrieved June 8, 2016, from https://www.youtube.com/watch?v=rF65hyaRxNQ&usg=AFQjCNE2N42soAZ9wA4vn6awa_X9lvtXXQ

Bergmann, K., Aksu, V., & Kopp, S. (2011, Sept. 5-7). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In P. Wagner, C. Kirchhof, & Malisz, Z. (Eds.), *Proceedings of GESPIN – Gesture and Speech in Interaction* [CD]. Paper presented at GESPIN 2011: Gesture and Speech in Interaction, Bielefeld, Germany (n.p.).

Bergmann, K., Kahl, S., & Kopp, S. (2014). How is information distributed across speech and gesture? A cognitive modeling approach. *Cognitive processing, Special Issue: Proceedings of KogWis 2014*, 84-87.

Bühler, K. (1990). *Theory of Language: The representational function of language*. Amsterdam: Benjamins.

Butterworth, B. L., & Beattie, G. W. (1978). Gesture and silence as indicators of planning in speech. In P. T. Smith & R. Campbell (Eds.), *Recent Advances in the psychology of language: Formal and experimental approaches* (pp. 347-360). New York: Plenum.

Butterworth, B. L., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review,* 96(1), 168-174.

Butterworth, B. L., Hine, R. R., & Brady, K. D. (1977). Speech and interaction in sound-only communication channels. *Semiotica, 20*(1-2), 81-100.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16*(5), 805-816.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 32*, 85-89.

Cassell, J., & McNeill, D. (1991). Non-verbal imagery and the poetics of prose. *Poetics Today, 12*(3), 375-404.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*, 1-34.

Chafe, W. L. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chen, M. (2002, Dec. 1-6). Achieving effective floor control with a low-bandwidth gesture-sensitive video-conferencing system. In L. A. Rowe, B. Mérialdo, M. Mühlhäuser, K. W. Ross, & N. Dimitrova (Eds.), *Proceedings of the 10th ACM International Conference on Multimedia 2002*. Paper presented at he 10th ACM International Conference on Multimedia, Juan les Pins, France (476-483). New York, NY: ACM.

Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In C. Cieri & M. Liberman (Eds.), Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation [CD]. Presented at LREC 2008, Sixth International Conference on Language Resources and Evaluation, Marrakesh, Morocco (n.p.). Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands (Dev.). ELAN [multimodal annotation software]. Available at http://tla.mpi.nl/tools/tla-tools/elan/

Deco Bliss (Producer). (2011, April 12). *How to hammer nail* [video file]. In *Deco Bliss*. Retrieved June 8, 2016, from https://www.youtube.com/watch?v=GDE8R51FqJw

de Ruiter, J. P. (1998). *Gesture and speech production*. MPI series in psycholinguistics. Wageningen: Ponsen & Looijen.

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture* (pp. 284-311). Cambridge, UK: Cambridge University Press.

de Ruiter, J. P. (2003). The function of hand gesture in spoken conversation. In M. Bickenback, A. Klappert, & H. Pompe (Eds.), *Manus Loquens* (pp. 338-347). Cologne: DuMont.

de Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *International Journal of Speech-Language Pathology, 8*(2), 124-127.

de Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture, 7*(1), 21-38.

de Ruiter, J. P. (2012, August 23). Trial recording at Nat.CoMM/HD lab [Digital image]. Retrieved April 4, 2016, from http://www.uni-bielefeld.de/lili/personen/jruiter/NatCoMMHD.html

de Ruiter, J. P., & de Beer, C. (2013). A critical evaluation of models of gesture and speech production for understanding gesture in aphasia, *Aphasiology, 42*(2), 257-270.

de Ruiter, J. P., & Wilkins, D. (1998). The synchronization of gesture and speech in Dutch and Arrernte (an Australian Aboriginal language): A cross-cultural comparison. In S. Santi (Ed.), *Oralité et Gestualité* (pp. 603-607). Paris: L'Harmattan.

de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expression: Investigating the tradeoff hypothesis. *Topics in Cognitive Science, 4*(2), 232-248.

de Saussure, F. (1972/1983). *Course in General Linguistics (3rd ed.).* (R. Harris, Trans.). Chicago: Open Court Publishing Company.

Duncan, S. (2001; 2006). Co-expressivity of speech and gesture: Manner of motion in Spanish, English, and Chinese. In A. Simspon (Ed.), *Proceedings of the 27th Berkeley Linguistics Society Annual Meeting*. Plenary held at the 27th annual meeting of the Berkeley Linguistics Society, University of California at Berkeley, CA (pp. 353-370). Berkeley, CA: Berkeley University Press.

Duncan, S. (2008). Gestural imagery and cohesion in normal and impaired discourse. In I. Wachsmuth, M. Lenzen, & G. Knoblich (Eds.), *Embodied communication in humans and machines* (pp. 305-328). Oxford: Oxford University Press.

Duncan, S. (2009, Sept. 24-26). Gesture and speech prosody in relation to structural and affective dimensions of natural discourse. In E. Jarmołowicz-Nowikow, K. Juszczyk, Z. Malisz, & M. Szczyszek, (Eds.), *Proceedings of GESPIN – Gesture and Speech in Interaction* [CD]. Paper presented at GESPIN 2009: Gesture and Speech in Interaction, Poznan, Poland (n.p.).

Duncan, S., Parrill, F., & Loehr, D. (2005, June 15-18). *Discourse factors in gesture and speech prosody*. Paper presented at the 2nd Conference of the International Society for Gesture Studies (ISGS), Lyon, France.

Duncan, S., Galati, A., Goodrich, L., Ramig, L., & Brandabur, M. (2004, March 18-21). Impairments in complex language and coverbal gestures: Idiopathic Parkinson Disease. Paper presented at the 12th biennial conference on motor speech: Motor speech disorders & speech motor control, Albuquerque, NM.

Efron, D. (1941/1972). *Gesture, race, and culture*. The Hague: Mouton.

Einstein, A. (1905/2005). Über die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. In F. A. C. Gren, L. W. Gilbert, J. C. Poggendorf, G. Wiedemann, & E. Wiedemann (Eds.), *Annalen der Physik* (Vol. 17) (pp. 164-181). Leipzig: Paul Drude.

Ekman, P. (1992). Facial expression of emotion: New findings, new questions. *Psychological Science, 3(1),* 34-38.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. New York: Times Books.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding, *Semiotica, 1*, 49-98.

Elicitation Protocol. (n.d.). Retrieved March 23, 2016, from mcneilllab.uchicago.edu/analyzing-gesture/elicitation-protocol.html

Fast, J. (1971). *Body Language*. London: Pan Books.

Feyerabend, P. (1998). *Widerstreit und Harmonie. Trentiner Vorlesungen*. In P. Engelmann (Ed.), Wien: Passagen.

Feyereisen, P. (1983). Manual activity during speaking in aphasic subjects. International Journal of Psychology*, 18*, 545-556.

Feyereisen, P. (2007). How do gesture and speech production synchronise? *Current Psychology Letters, 22*(2), 1-12.

Feyereisen, P., & Seron, X. (1982). Nonverbal communication and aphasia: A review. II. Expression, *Brain and Language,16*, 213-236.

Freleng, F. (Director). (1950). Canary Row [Motion picture]. USA: Warner Bros.

Fricke, E. (2008). *Grundlagen einer multimodalen Grammatik des Deutschen: Syntaktische Strukturen und Funktionen*. Habilitation thesis, Europa-Universität Viadrina, Frankfurt (Oder), Germany.

Fricke, E. (2012). *Grammatik multimodal: Wie Wörter und Gesten zusammenwirken*. Berlin/Boston: De Gruyter.

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166*(3), 455-464.

Fujisaki, W., & Nishida, S. (2007). Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. V*ision Research, 47*, 1075-1093.

Fujisaki, W., & Nishida, S. (2008). Top-down featurebased selection of matching features for audio-visualsynchrony discrimination. *Neuroscience Letters, 433*, 225-230.

Gahl, S., Garnsey, S. M., Fisher, C., & Matzen, L. (2006, July 26-29). “That sounds unlikely”: Phonetic cues to garden paths. In R. Sun, & n. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* [CD]. Paper presented at the 28th Annual Conference of the Cognitive Science Society (CogSci/ICCS), Vancouver, BC (n.p.). New York, NY: Psychology Press.

Gawne, L., Kelly, B. F., & Unger, A. (2010). Gesture categorisation and understanding speaker attention to gesture. In Y. Treis & R. de Busser (Eds.), *Selected Papers from the 2009 Conference of the Australian Linguistic Society*. Paper presented at the 2009 Conference of the Australian Linguistic Society, La Trobe University, Melbourne, VA (pp. 216-233). Retrieved May 30, 2016, from http://www.als.asn.au

Gebre, B. G., Wittenburg, P., & Lenkiewicz, P. (2012). Towards automatic gesture stroke detection. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, . . . S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* Paper presented at the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey (pp. 231-235). Paris, France: ELRA.

Gibbon, D. (2009). Gesture theory is linguistics: On modelling multimodality as prosody. In O. Kwong (Ed.), *PACLIC 2009.* Paper presented at the 23rd Pacific Asia Conference on Language, Information and Computation (pp. 9-18). Hong Kong: City University of Hong Kong Press.

Gibbon, D., Hell, B., Looks, K., & Trippel, T. (2003). Formal syntax of gesture: CoGesT 1.1. *Technical Report 2*, Bielefeld University, Bielefeld.

Gibbon, D., Winski, R., & Moore, R. (Eds.). (1997). *Handbook of standards and resources for spoken language systems*. Berlin/Boston: De Gruyter Mouton.

Goldin-Meadow, S. (2003). *Hearing gesture: how our hands help us think*. Cambridge: Harvard University Press.

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology, 64*, 257-84.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100*, 279-297.

Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology, 91*(4), 720-730.

Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and Speech, 1*, 226-231.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Grice, H. P. (1975). Logic and conversation. In Cole, P., & J. Morgan (Eds.), *Syntax and semantics, Vol. 3, Speech acts* (pp. 41-58). New York: Academic Press.

Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics & Cognition, 7*, 35-63.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition, 14*, 53-82.

Gullberg, M. & Kita, S. (2009). Modulating addressees' attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior, 33*(4), 251-277.

Gunter, T. C., & Bach, P. (2004). Communicating hands: ERPs elicited by meaningful symbolic hand postures. *Neuroscience Letters*, 372, 52-56

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience, 23*(8), 1845-54.

Hadar, U. & Butterworth, B. (1997). Iconic gestures, imagery and word retrieval in speech. *Semiotica, 115*, 147-172.

Harland, R. (1987/2007). *Superstructuralism. New accents*. London: Methuen.

Hassemer, J., Joue, G., Willmes, K., & Mittelberg, I. (2011, Sept. 5-7). Dimensions and mechanisms of form constitution: Towards a formal description of gestures. In P. Wagner, C. Kirchhof, & Z. Malisz (Eds.), *Proceedings of GESPIN – Gesture and Speech in Interaction* [CD]. Paper presented at GESPIN 2011: Gesture and Speech in Interaction, Bielefeld, Germany (n.p.).

Herrmann, T. (1985). *Allgemeine Sprachpsychologie. Grundlagen und Probleme*. Munich/Vienna/ Baltimore: Urban & Schwarzenberg.

Hielscher-Fastabend, M. (1996). *Emotion und Textverstehen: Eine Untersuchung zum Stimmungs-kongruenzeffekt. Psycholinguistische Studien*. Opladen: Westdeutscher Verlag.

Hogrefe, K., Ziegler, W., Wiesmayer, S., Weidinger, N., & Goldenberg, G. (2013). The actual and potential use of gestures for communication in aphasia. *Aphasiology: Special Issue: Aphasia and Gesture*, *27*, 1070-1089.

Hoiting, N., & Slobin, D. I. (2007). From gestures to signs in the acquisition of sign language. In S. Duncan, J. Cassell, & E. T. Levy (Eds.), *Gesture and the dynamic dimension of language: Essays in honor of David McNeill* (pp. 51-65). Amsterdam/Philadelphia: John Benjamins.

Holler, J., Shovelton, H. K., & Beattie, G. W. (2009). Do iconic hand gestures really contribute to the communication semantic information in a face-to-face context? *Journal of Nonverbal Behavior, 33*, 73-88.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2*, 1-16.

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

Kendon, A. (1972). Some relationships between body motion and speech. An analysis of an example. In A. Siegman & B. Pope (Eds.), *Studies in Dyadic Communication* (pp. 177-210). Elmsford, NY: Pergamon.

Kendon, A. (1979) Some theoretical and methodological aspects of the use of film in the study of social interaction. In G. P. Ginsburg (Ed.), *Emerging Strategies in Social Psychological Research* (pp. 67-91). London/New York: John Wiley.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207-227). The Hague: Mouton.

Kendon, A. (1988). How gestures can become like words. In F. Poyatos (Ed.), *Crosscultural perspectives in nonverbal communication* (pp. 131-141). Toronto: C. J. Hogrefe.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics*, *23*(3), 247-279.

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

Kipp, M., & Martin, J.-C. (2009, Sept. 10-12). Gesture and emotion: Can basic gestural form features discriminate emotions? In J. Cohn, A. Nijholt, & M. Pantic (Eds.), *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII).* Paper presented at ACII 2009: International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands (1-8).

Kirchhof, C. (2010). *The Truth about Mid-Life Singles in the USA: A Corpus-Based Analysis of Printed Personal Advertisements*. München: GRIN.

Kirchhof, C. (2011, Sept. 5-7). So What's Your Affiliation With Gesture? In P. Wagner, C. Kirchhof, & Z. Malisz (Eds.), *Proceedings of GESPIN – Gesture and Speech in Interaction* [CD]. Paper presented at GESPIN 2011: Gesture and Speech in Interaction, Bielefeld, Germany (n.p.).

Kirchhof, C., & de Ruiter, J. P. (2012, July 24-27). On the audiovisual integration of speech and gesture. Paper presented at the 5th Conference of the International Society for Gesture Studies (ISGS 5), Lund, Sweden.

Kita, S., & Özyürek, A. (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*(1), 16-32.

Kok, K., Bergmann, K., Cienki, A., & Kopp, S. (2016). Mapping out the multifunctionality of speakers' gestures. *Gesture. 15*(1), 37-59.

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261-283). New York: Cambridge University Press.

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000a). Lexical gestures and lexical access: A process model. Draft.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61*, 743.

Latiolais-Hargrave, J. (2008). *Strictly business: Body language – using nonverbal communication for power and success* (2nd ed.). Dubuque, IA: Kendall/Hunt.

Lee, A. (2010). VirtualDub (1.9.11) [video editing software]. Available at http://virtualdub.sourceforge.net/

Leiner, D. J. (2014). SoSci Survey (Version 2.5.00-i) [survey design interface]. Available at http://www.soscisurvey.com

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: M.I.T. Press/Bradford Books.

Lewandowski, T. (1985). *Linguistisches Wörterbuch (4th ed.).* Heidelberg: Quelle & Meyer.

Loehr, D. (2004). *Gesture and intonation*. Unpublished PhD thesis, Georgetown University, Washington, DC.

Luigi1872 (Producer). (2014, August 11). *Einen Luftballon platzen lassen* [Video file]. In *Menschen & Blogs*. Retrieved June 8, 2016, from https://www.youtube.com/watch?v=5CEqPqPmAJ8

Lutzeier, P. R. (2006). Lexical fields. In K. Brown (Ed.), *Encyclopedia of language and linguistics (2nd ed.)* (pp. 79-87). Oxford: Elsevier.

Marstaller, L., & Burianová, H. (2014). The multisensory perception of co-speech gestures – a review and meta-analysis of neuroimaging studies, *Journal of Neurolinguistics, 30*, 69-77.

Massaro, D. W. (1989). Speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences, 12*, 741-794.

Massaro, D. W., & Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13*, 127-134.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. J*ournal of the Acoustical Society of America, 100*, 1777-1786.

Mayberry, M., Crocker, M., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science, 33*(1), 449-496.

McClave, E. (1994). Gestural beats: The rhythm hypothesis, *Journal of Psycholinguistic Research, 23*, 45-66.

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

McNeill, D. (1975). Semiotic Extension. In L. Solso (Ed.), *Information processing and cognition* (pp. 351-380). Hillsdale, NJ: Erlbaum.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review; 92*(3), 350-371.

McNeill, D. (1989). A straight path – to where? Reply to Butterworth and Hadar. *Psychological Review; 96*(1), 175-179.

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

McNeill, D. (2000). *Language and Gesture*. Cambridge, UK: Cambridge University Press.

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

McNeill, D. (2011). *How language began - gesture and speech in human evolution*. Draft.

McNeill, D. (2012). *How language began - gesture and speech in human evolution*. Cambridge, UK: Cambridge University Press.

McNeill, D. (2015). Why we gesture – The surprising role of hand movements in communication. Cambridge, UK: Cambridge University Press.

McNeill, D. (n.d.). *Growth points and modeling*. Unpublished presentation, University of Chicago, Chicago. Retrieved March 08, 2016, from http://mcneilllab.uchicago.edu/writing/essays.html

McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative effects of speech mismatched gestures. *Research on Language and Social Interaction, 27*(3)*,* 223-237.

McNeill, D., & Duncan, S. (2000). Growth points in thinking for speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141-161). Cambridge, UK: Cambridge University Press.

McNeill, D., & Levy, E. (1982). Conceptual representations in language activity and gesture. In R. Jarvella & W. Klein (Eds.), *Speech, place, and action: Studies in deixis and related topics* (pp. 271-295). Chichester, England: Wiley.

McNeill, D., & Levy, E. (1993). Cohesion and gesture. *Discourse Processes, 16*(4), 363-386.

McNeill, D., Quaeghebeur, L., & Duncan, S. (2008). IW – “the man who lost his body”. In S. Gallagher & D. Schmickin (Eds.), *Handbook of phenomenology and cognitive sciences* (pp. 519-546). Dordrecht: Springer.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4*(2), 119-141.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 615-623.

Morris, D. (1967). *The naked ape.* London: Cape.

Morris, D. (1982). *Manwatching: A field guide to human behaviour*. London: Granada.

Morris, D. (2002). *People watching: The Desmond Morris guide to body language*. London: Vintage.

Navarro, J., & Karlins, M. (2008). *What every body is saying: An ex-FBI agent’s guide to speed-reading people*. New York: Harper Collins.

Neff, M., Kipp, M., Albrecht, I., & Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics, 27*(1), 1-24.

Nishida, S. (2006, July 9-12). Interactions and integrations of multiple sensory channels in human brain. In L. Guang & H.-J. Zhang (Eds.), *Proceedings of ICME 2006*. Paper presented at the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON (pp. 509-512).

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

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (2000). Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension. *Cognitive Studies. Bulletin of the Japanese Cognitive Science Society, 7*, 86-92.

NorPix, Inc. (2008) StreamPix 4 (Release 4.14.0) [video recording software]. Available at https://www.norpix.com/products/streampix/streampix.php

Oertel, C. (2010). Identification of cues for the automatic detection of hotspots. Unpublished master's thesis, Bielefeld University, Bielefeld.

Olshausen, B. A. (2000, Oct. 10). *Aliasing* (PSC 129 - Sensory Processes). Lecture presented at Redwood Center for Theoretical Neuroscience, Berkeley, CA.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*(4), 605-616.

Peirce, C. S. (1894/1998), What Is a Sign? In The Peirce Edition Project (Eds.), *The essential Peirce – selected philosophical writings. Volume 2 (1893-1913)* (pp. 4-10). Bloomington, IN: Indiana University Press.

Petrini, K., Holt, S. P., & Pollick, F. (2010). Expertise with multisensory events eliminates the effect of biological motion rotation on audiovisual synchrony perception. *Journal of Vision, 10*(5), 2-14.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 1-57.

Pika, S., Liebal, K., Call, J., & Tomasello, M. (2005). The gestural communication of apes. *Gesture, 5*, 41-56.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226-231.

Rickheit, G., Sichelschmidt, L., & H. Strohner (2002). Gedanken ausdrücken und Sprachen verstehen. In Müller, H. (Ed.), *Arbeitsbuch Linguistik* (pp. 382-405). Paderborn: Schöningh.

Rossini, N., & Gibbon, D. (2011, Sept. 5-7). Why gesture without speech but not talk without gesture? Paper presented at GESPIN 2011: Gesture and Speech in Interaction, Bielefeld, Germany (n.p.).

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

Sekine, K., Stam, G., Yoshioka, K., Tellier, M., & Capirci, O. (2015). Cross-linguistic views of gesture usage. *Vigo International Journal of Applied Linguistics (VIAL), 12,* 91-105.

Seyfeddinipur, M., & Kita, S. (2003). Gestures and speech disfluencies. In A. Simspon (Ed.), *Proceedings of the 27th Berkeley Linguistics Society Annual Meeting*. Paper presented at the 27th annual meeting of the Berkeley Linguistics Society, University of California at Berkeley, CA (pp. 457-464). Berkeley, CA: Berkeley University Press.

Slobin, D. I. (1987). Thinking for speaking. In *Proceedings of the 13th annual meeting of the Berkeley Linguistics Society*, 435-444.

Sowa, T., Kopp, S., Duncan, S., McNeill, D., & Wachsmuth, I. (2008). Implementing a non-modular theory of language production in an embodied conversational agent. In I. Wachsmuth, M. Lenzen, & G. Knoblich (Eds.), *Embodied Communication in Humans and Machines* (pp. 425-449). Oxford University Press.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition, *24*(6), 1521-1543.

Srinivasan, V., Bethel, C. L., & Murphy, R. R. (2014). Evaluation of head gaze loosely synchronized with real-time synthetic speech for social robots. *IEEE Transactions on Human-Machine Systems, 44* (6), 767-778.

Stanek, J. (Producer). (2010, October 05). *Great sound from this outlaw kick drum* [Video file]. In *Dixon Drums*. Retrieved June 8, 2016, from https://www.youtube.com/watch?v=77cRPvcvBps

Stokoe, W. C. (1960/2000). *Sign language structure: An outline of the visual communication systems of the American Deaf*. Studies in linguistics: Occasional papers (No. 8). Buffalo: Dept. of Anthropology and Linguistics, University of Buffalo.

Sugiura, R. (Producer). (2011, January 17). *How to snap your fingers Tutorial (my method)* [Video file]. In *PileSpiral*. Retrieved June 8, 2016, from https://www.youtube.com/watch?v=hqet9qN8dIA

The Audacity Team (2011). Audacity Sound Editor 1.3.13 [sound editing software]. Available at http://audacity.sourceforge.net/download

Thies, A. (2003). *First the hand, then the word: On gestural displacement in non-native English speech*. Unpublished SEII thesis, Bielefeld University, Bielefeld.

Trautmann-Voigt, S. & Voigt, B. (Eds.). (2009). *Grammatik der Körpersprache. Ein integratives Lehre- und Arbeitsbuch zum Embodiment*. (2nd Edition). Schattauer: Stuttgart.

Trofatter, C., Kontra, C., Beilock, S., & Goldin-Meadow, S. (2015). Gesturing has a larger impact on problem-solving than action, even when action is accompanied by words. *Language, Cognition and Neuroscience, 30* (3), 251-260.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*, 598-607.

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

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., D'Imperio, M., Escudero Mancebo, D., Lacheret, A., et al. (2015). *Disentagling and Connecting Different Perspectives on Prosodic Prominence*. Presented at the ICPL 2015, Cologne.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: Temporal order versus simultaneity judgments. *Experimental Brain Research, 185*(3), 521-529.

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L. S. Vygotsky, Vol. 1: Problems of general psychology* (pp. 39-285). New York/London: Plenum Press.

Wachsmuth, I., de Ruiter, J., Jaecks, P., & Kopp, S. (Eds.). (2013). *Alignment in communication: Towards a new theory of communication*. Advances in Interaction Studies (Vol. 6). Amsterdam: Benjamins.

Watling, J. (1950). The causal theory of perception. *Mind, 59,* 539-540.

Watzlawick, P., Helmick Beavin, J., & Jackson, D. (1967). Some tentative axioms of communication. In P. Watzlawick, J. Beavin-Bavelas, & D. Jackson (Eds.). *Pragmatics of human communication - A study of interactional patterns, pathologies and paradoxes* (pp. 275-288). New York: W. W. Norton & Company, Inc.

Wheatland, N., Wang, Y., Song, H., M. Neff, Zordan, V., & Jörg, S. (2015). State of the art in hand and finger modeling and animation. *Computer Graphics Forum, 34*, 735-760.

Winter, V., & Müller, H. M. (2010). ERP analysis of audiovisual integration during language comprehension. In J. Haack, H. Wiese, A. Abraham, & C. Chiarcos (Eds.), *Proceedings of the 10th Biannual Meeting of the German Society for Cognitive Science (KogWis 2010).* Poster presented at the 10th Biannual Meeting of the German Society for Cognitive Science, Potsdam, Germany (p. 236).

11 Appendix

11.1 Corpus

11.1.1 Meta data for narration recordings

recording	dur (mm:ss.f)	S_sex	S_age	S_hand	S_gphr	comment
10.06.57.995	04:03,000	-	-	-	-	insufficient video quality
10.17.48.959	05:48,598	f	18	r	211	-
10.54.29.104	06:12,303	f	22	r	44	audio off sync after 51 sec
11.00.31.621	06:08,093	f	26	l	214	-
11.17.45.463	05:26,299	f	24	r	39	-
12.05.31.682	06:05,415	m	25	l	112	audio off sync after 01:07,000
12.37.35.766	01:26,681	f	23	r	33	video damaged after 01:19,417
12.50.52.223	04:34,961	f	21	r	30	audio broken after 01:08,110
13.09.12.480	06:44,459	m	34	r	4	audio broken after 01:20,381
13.14.02.898	06:52,030	f	25	l	50	audio broken after 01:18,340
14.01.46.033	05:22,458	-	-	-	-	video fully damaged
14.15.34.268	03:43,000	-	-	-	-	excluded due to pen in hand
14.27.42.306	08:16,702	m	32	r	95	pilot
14.40.09.299	07:12,073	f	30	l	25	audio broken after 05:16,814
15.04.57.785	04:37,085	m	32	r	51	-
15.27.51.757	04:06,665	f	24	r	7	audio broken after 01:43,941
15.41.04.113	05:00,978	f	22	r	71	-
16.11.09.878	06:18,962	m	24	l	28	audio broken after 01:36,892
16.26.56.109	04:48,923	m	24	r	4	audio broken after 03:11,299
16.36.00.692	05:43,655	f	24	l	140	-
16.43.50.013	05:31,484	m	26	r	36	audio broken after 01:35,423
16.51.31.649	07:08,566	f	22	r	73	audio broken after 06:24,543
17.14.30.283	06:55,000	m	21	r	0	no gestures

11.1.2 Meta data for physical stimuli

1	clap	clap	3060
2	glass	glass	1080
3	keyboard	keyboard	2070
4	knock	knock	3110
5	sekt	sekt	3020
6	hammer	How To Hammer A Nail (Deco Bliss, 2011)	1410
7	snap	How to snap your fingers Tutorial (Sugiura,2011)	5070
a	a	Great sound from this Outlaw Kick Drum (Stanek, 2010)	940
b	b	Einen Luftballon platzen lassen (Luigi1872, 2014)	5240

11.1.3 **Flyer for gathering participants**

!!! 12. - 24.10.2010 ganztägig !!!

Wollt ihr gemeinsam über Sylvester und Tweety lachen?

http://tinyurl.com/35ahyen

Dann kommt **zu zweit** im Interaktionslabor vorbei!
Ihr schaut Euch einen Cartoon an und sprecht danach darüber. Das ganze dauert ca. 30 Minuten.
Neben ein Paar Keksen könnt ihr auch einen VP-Schein ergattern!

Meldet euch einfach unter
[redacted]@uni-bielefeld.de oder 0176-[redacted]
mit einem Terminvorschlag zwischen
08:00 und 20:00h!

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

Sylvester & Tweety
[redacted]@uni-bielefeld.de
0176-[redacted]

11.1.4 **Video recording consent form**

Seite **1** von **2**

Information und Einwilligungserklärung zum Forschungsprojekt

Bitte lesen Sie sich dieses Formular sorgfältig durch und fragen Sie nach, wenn Sie etwas nicht verstehen oder weitere Erklärungen wünschen.

I. Persönliche Angaben

Studie: Sylvester & Tweety	
Name, Vorname:	
E-Mail:	
Alter:	
Muttersprache (n):	
Weitere Sprachen:	
In welcher Region aufgewachsen:	
Heutiger Wohnort:	
Sehhilfe: □ ja □ nein	**Wenn ja, welche:**
Beruf:	
Bei Studierenden **Studiengang:**	**Semester:**
Höchster Bildungsabschluss:	
Händigkeit:	**Geschlecht: □ männlich □ weiblich**

II. Allgemeine Informationen zum Forschungsprojekt

Die Arbeitsgruppe „Psycholinguistik" erstellt einen Korpus, der für Kommunikations- und Interaktionsstudien genutzt werden soll. In den verschiedenen Forschungsprojekten der Arbeitsgruppe stehen die strukturellen Aspekte in Konversationen im Vordergrund. Die Gesprächsinhalte an sich spielen dabei keine Rolle und werden deshalb nicht analysiert.

III. Sobald es der Forschungszweck zulässt, werden Ihre personenbezogenen Daten und die Audio- und Videoaufnahmen vernichtet bzw. gelöscht.

IV. Ihre Einwilligung ist freiwillig. Durch eine Verweigerung der Einwilligung entstehen Ihnen keine Nachteile. Sie können Ihre Einwilligung jederzeit mit Wirkung für die Zukunft widerrufen und die Löschung bzw. Vernichtung ihrer Daten und der Audio- und Videoaufnahmen verlangen.

Seite **2** von **2**

V. – Unzutreffendes bitte streichen -

1. Sie erklären sich mit Ihrer Unterschrift damit einverstanden, dass die Arbeitsgruppe „Psycholinguistik" der Fakultät für Linguistik und Literaturwissenschaften der Universität Bielefeld die im Rahmen der Studie erhobenen Daten in Form von Audio und Videoaufzeichnungen und die Angaben unter I.) speichert und für Forschungszwecke nutzt.

2. Sie erklären sich mit Ihrer Unterschrift damit einverstanden, dass die Arbeitsgruppe „Psycholinguistik" der Fakultät für Linguistik und Literaturwissenschaften der Universität Bielefeld einzelne Video- bzw. Tonsequenzen im Rahmen der Präsentation von Forschungsergebnissen Dritten vorführt.

3. Sie erklären sich mit Ihrer Unterschrift damit einverstanden, dass die Arbeitsgruppe „Psycholinguistik" der Fakultät für Linguistik und Literaturwissenschaften der Universität Bielefeld Ihre Angaben unter I.) in eine Versuchspersonenliste aufnimmt. Daten aus der Versuchspersonenliste und die im Rahmen der Studie erhobenen Daten in Form der Audio- und Videoaufzeichnungen werden folgenden Forschungsprojekten zur Verfügung gestellt:

a. Dem EU-Forschungsprojekt JAMES (Joint Action for Multimodal Embodied Social Systems),Projektnummer FP7-ICT-2009.2.1a

b. Der Forschungsgruppe des Research Area C des Citec (Situated Communication)

c. Dem Sonderforschungsbereich SFB 673 - Alignment in Communication

4. Sie erklären sich außerdem mit Ihrer Unterschrift damit einverstanden, dass Sie für eine mögliche Teilnahme an anderen Studien von der Arbeitsgruppe „Psycholinguistik" der Fakultät für Linguistik und Literaturwissenschaften der Universität Bielefeld und den unter V. 3.a.-c) genannten Forschungsstellen per E-Mail kontaktiert werden dürfen.

VI. Ich bin mit der vorgesehenen Verarbeitung meiner Daten einverstanden.

Bielefeld, den __.___.2010________________________

Ort, Datum Unterschrift

11.1.5 ELAN annotation scheme

```
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION_DOCUMENT AUTHOR="" DATE="2016-07-13T15:32:56+01:00"
FORMAT="2.8" VERSION="2.8" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance"
xsi:noNamespaceSchemaLocation="http://www.mpi.nl/tools/elan/EAFv2.8.xsd">
 <HEADER MEDIA_FILE="" TIME_UNITS="milliseconds"/>
 <TIME_ORDER/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="default-lt" TIER_ID="S"/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="default-lt" TIER_ID="L"/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="default-lt" TIER_ID="for
      studies"/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="gphr" TIER_ID="S_gphr"/>
 <TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="default-lt" TIER_ID="for
      desync"/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="stories"
TIER_ID="scenes"/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="default-lt"
TIER_ID="notes"/>
 <TIER DEFAULT_LOCALE="de" LINGUISTIC_TYPE_REF="BG" TIER_ID="BG"/>
 <TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="default-lt"
      TIER_ID="S_EN_word"/>
 <LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="default-
lt"   TIME_ALIGNABLE="true"/>
 <LINGUISTIC_TYPE CONTROLLED_VOCABULARY_REF="gphr"
GRAPHIC_REFERENCES="false"   LINGUISTIC_TYPE_ID="gphr"
TIME_ALIGNABLE="true"/>
 <LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="desync"
      TIME_ALIGNABLE="true"/>
 <LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="BG"
      TIME_ALIGNABLE="true"/>
 <LINGUISTIC_TYPE CONTROLLED_VOCABULARY_REF="stories"
      GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="stories"
      TIME_ALIGNABLE="true"/>
 <LOCALE COUNTRY_CODE="DE" LANGUAGE_CODE="de"/>
 <LOCALE LANGUAGE_CODE="en" VARIANT="ASCII"/>
 <LANGUAGE LANG_DEF="http://cdb.iso.org/lg/CDB-00130975-001"
LANG_ID="und"     LANG_LABEL="undetermined (und)"/>
 <CONSTRAINT DESCRIPTION="Time subdivision of parent annotation's time
      interval, no time gaps allowed within this interval"
      STEREOTYPE="Time_Subdivision"/>
 <CONSTRAINT DESCRIPTION="Symbolic subdivision of a parent annotation.
      Annotations refering to the same parent are ordered"
      STEREOTYPE="Symbolic_Subdivision"/>
 <CONSTRAINT DESCRIPTION="1-1 association with a parent annotation"
      STEREOTYPE="Symbolic_Association"/>
 <CONSTRAINT DESCRIPTION="Time alignable annotations within the parent
      annotation's time interval, gaps are allowed"
STEREOTYPE="Included_In"/>
 <CONTROLLED_VOCABULARY CV_ID="stories">
 <DESCRIPTION LANG_REF="und"/>
 <CV_ENTRY_ML CVE_ID="cveid0">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">intro</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid1">
 <CVE_VALUE DESCRIPTION=""
```

```
            LANG_REF="und">bird_watchers_society</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid2">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">outside-pipe</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid3">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">bowling_ball</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid4">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">monkey</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid5">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">hotel</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid6">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">weight</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid7">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">rope_swing</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid8">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">streetcar</CVE_VALUE>
 </CV_ENTRY_ML>
 <CV_ENTRY_ML CVE_ID="cveid9">
 <CVE_VALUE DESCRIPTION="" LANG_REF="und">credits</CVE_VALUE>
 </CV_ENTRY_ML>
 </CONTROLLED_VOCABULARY>
 <CONTROLLED_VOCABULARY CV_ID="gphr">
 <DESCRIPTION LANG_REF="und"/>
 <CV_ENTRY_ML CVE_ID="cveid0">
 <CVE_VALUE DESCRIPTION="gphr" LANG_REF="und">gphr</CVE_VALUE>
 </CV_ENTRY_ML>
 </CONTROLLED_VOCABULARY>
</ANNOTATION_DOCUMENT>
```

11.1.6 **List of annotated gphr**

gphr	recording	gphr_start (mm:ss.f)	gphr_end (mm:ss.f)	gphr_dur (s.f)
1	10.17.48.959.eaf	00:05.212	00:05.829	0.617
2	10.17.48.959.eaf	00:05.944	00:06.764	0.82
3	10.17.48.959.eaf	00:07.714	00:08.197	0.483
4	10.17.48.959.eaf	00:11.156	00:11.855	0.699
5	10.17.48.959.eaf	00:12.270	00:12.991	0.721
6	10.17.48.959.eaf	00:14.206	00:16.274	2.068
7	10.17.48.959.eaf	00:16.274	00:18.380	2.106
8	10.17.48.959.eaf	00:19.162	00:20.093	0.931
9	10.17.48.959.eaf	00:20.093	00:21.459	1.366
10	10.17.48.959.eaf	00:21.459	00:22.097	0.638
11	10.17.48.959.eaf	00:22.097	00:23.123	1.026
12	10.17.48.959.eaf	00:23.527	00:24.195	0.668
13	10.17.48.959.eaf	00:24.195	00:25.170	0.975
14	10.17.48.959.eaf	00:26.380	00:26.839	0.459
15	10.17.48.959.eaf	00:26.839	00:27.238	0.399
16	10.17.48.959.eaf	00:27.356	00:28.058	0.702
17	10.17.48.959.eaf	00:28.058	00:28.920	0.862
18	10.17.48.959.eaf	00:31.083	00:32.166	1.083
19	10.17.48.959.eaf	00:32.166	00:34.544	2.378
20	10.17.48.959.eaf	00:37.080	00:38.700	1.62
21	10.17.48.959.eaf	00:39.080	00:39.998	0.918
22	10.17.48.959.eaf	00:39.998	00:40.936	0.938
23	10.17.48.959.eaf	00:40.936	00:42.319	1.383
24	10.17.48.959.eaf	00:42.321	00:42.978	0.657
25	10.17.48.959.eaf	00:42.978	00:43.965	0.987
26	10.17.48.959.eaf	00:44.975	00:46.318	1.343
27	10.17.48.959.eaf	00:46.318	00:47.298	0.98
28	10.17.48.959.eaf	00:47.298	00:47.992	0.694
29	10.17.48.959.eaf	00:47.992	00:49.384	1.392
30	10.17.48.959.eaf	00:49.385	00:50.648	1.263
31	10.17.48.959.eaf	00:50.648	00:51.938	1.29
32	10.17.48.959.eaf	00:52.218	00:53.678	1.46
33	10.17.48.959.eaf	00:53.678	00:54.536	0.858
34	10.17.48.959.eaf	00:59.970	01:00.878	0.908
35	10.17.48.959.eaf	01:01.334	01:03.950	2.616
36	10.17.48.959.eaf	01:04.720	01:05.542	0.822
37	10.17.48.959.eaf	01:07.175	01:08.155	0.98
38	10.17.48.959.eaf	01:08.155	01:09.232	1.077
39	10.17.48.959.eaf	01:13.559	01:15.076	1.517
40	10.17.48.959.eaf	01:15.077	01:17.224	2.147

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
41	10.17.48.959.eaf	01:17.224	01:17.794	0.57
42	10.17.48.959.eaf	01:20.160	01:21.806	1.646
43	10.17.48.959.eaf	01:21.965	01:22.926	0.961
44	10.17.48.959.eaf	01:22.926	01:24.454	1.528
45	10.17.48.959.eaf	01:24.658	01:25.121	0.463
46	10.17.48.959.eaf	01:25.121	01:25.872	0.751
47	10.17.48.959.eaf	01:25.872	01:27.024	1.152
48	10.17.48.959.eaf	01:35.116	01:35.790	0.674
49	10.17.48.959.eaf	01:35.790	01:36.481	0.691
50	10.17.48.959.eaf	01:39.506	01:40.345	0.839
51	10.17.48.959.eaf	01:41.243	01:42.794	1.551
52	10.17.48.959.eaf	01:43.150	01:43.305	0.155
53	10.17.48.959.eaf	01:44.950	01:46.744	1.794
54	10.17.48.959.eaf	01:53.744	01:55.452	1.708
55	10.17.48.959.eaf	01:55.452	01:56.862	1.41
56	10.17.48.959.eaf	01:56.862	01:57.857	0.995
57	10.17.48.959.eaf	01:57.857	01:58.735	0.878
58	10.17.48.959.eaf	01:58.735	01:59.812	1.077
59	10.17.48.959.eaf	02:02.311	02:03.529	1.218
60	10.17.48.959.eaf	02:05.458	02:06.704	1.246
61	10.17.48.959.eaf	02:06.704	02:08.142	1.438
62	10.17.48.959.eaf	02:08.144	02:08.748	0.604
63	10.17.48.959.eaf	02:08.748	02:09.634	0.886
64	10.17.48.959.eaf	02:09.634	02:10.537	0.903
65	10.17.48.959.eaf	02:10.537	02:12.296	1.759
66	10.17.48.959.eaf	02:12.296	02:15.070	2.774
67	10.17.48.959.eaf	02:15.070	02:15.504	0.434
68	10.17.48.959.eaf	02:15.504	02:16.914	1.41
69	10.17.48.959.eaf	02:17.186	02:18.483	1.297
70	10.17.48.959.eaf	02:18.483	02:19.154	0.671
71	10.17.48.959.eaf	02:21.971	02:23.279	1.308
72	10.17.48.959.eaf	02:23.279	02:24.471	1.192
73	10.17.48.959.eaf	02:24.471	02:25.430	0.959
74	10.17.48.959.eaf	02:25.791	02:26.313	0.522
75	10.17.48.959.eaf	02:26.488	02:28.455	1.967
76	10.17.48.959.eaf	02:28.455	02:29.438	0.983
77	10.17.48.959.eaf	02:29.438	02:31.231	1.793
78	10.17.48.959.eaf	02:35.892	02:36.949	1.057
79	10.17.48.959.eaf	02:37.345	02:38.795	1.45
80	10.17.48.959.eaf	02:38.965	02:39.506	0.541
81	10.17.48.959.eaf	02:39.506	02:40.065	0.559
82	10.17.48.959.eaf	02:40.065	02:41.550	1.485
83	10.17.48.959.eaf	02:41.550	02:42.798	1.248

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
84	10.17.48.959.eaf	02:42.798	02:43.717	0.919
85	10.17.48.959.eaf	02:43.717	02:45.613	1.896
86	10.17.48.959.eaf	02:47.228	02:48.713	1.485
87	10.17.48.959.eaf	02:51.159	02:51.676	0.517
88	10.17.48.959.eaf	02:51.676	02:52.564	0.888
89	10.17.48.959.eaf	02:52.564	02:54.188	1.624
90	10.17.48.959.eaf	02:55.852	02:56.979	1.127
91	10.17.48.959.eaf	02:56.979	02:57.664	0.685
92	10.17.48.959.eaf	02:59.369	03:00.106	0.737
93	10.17.48.959.eaf	03:00.106	03:00.657	0.551
94	10.17.48.959.eaf	03:00.657	03:01.691	1.034
95	10.17.48.959.eaf	03:03.759	03:04.774	1.015
96	10.17.48.959.eaf	03:04.774	03:05.554	0.78
97	10.17.48.959.eaf	03:11.184	03:12.120	0.936
98	10.17.48.959.eaf	03:12.120	03:12.964	0.844
99	10.17.48.959.eaf	03:12.964	03:13.866	0.902
100	10.17.48.959.eaf	03:13.866	03:15.252	1.386
101	10.17.48.959.eaf	03:15.473	03:16.490	1.017
102	10.17.48.959.eaf	03:16.490	03:17.291	0.801
103	10.17.48.959.eaf	03:17.291	03:18.037	0.746
104	10.17.48.959.eaf	03:18.037	03:18.188	0.151
105	10.17.48.959.eaf	03:18.188	03:19.442	1.254
106	10.17.48.959.eaf	03:21.345	03:21.876	0.531
107	10.17.48.959.eaf	03:21.877	03:22.788	0.911
108	10.17.48.959.eaf	03:22.788	03:23.700	0.912
109	10.17.48.959.eaf	03:23.700	03:24.125	0.425
110	10.17.48.959.eaf	03:24.125	03:25.642	1.517
111	10.17.48.959.eaf	03:25.642	03:26.252	0.61
112	10.17.48.959.eaf	03:26.252	03:26.617	0.365
113	10.17.48.959.eaf	03:26.617	03:27.842	1.225
114	10.17.48.959.eaf	03:27.842	03:28.545	0.703
115	10.17.48.959.eaf	03:30.042	03:30.388	0.346
116	10.17.48.959.eaf	03:30.476	03:30.715	0.239
117	10.17.48.959.eaf	03:30.715	03:31.398	0.683
118	10.17.48.959.eaf	03:31.769	03:32.393	0.624
119	10.17.48.959.eaf	03:32.393	03:32.930	0.537
120	10.17.48.959.eaf	03:33.068	03:33.965	0.897
121	10.17.48.959.eaf	03:34.676	03:35.144	0.468
122	10.17.48.959.eaf	03:39.486	03:40.691	1.205
123	10.17.48.959.eaf	03:40.691	03:41.564	0.873
124	10.17.48.959.eaf	03:41.564	03:42.681	1.117
125	10.17.48.959.eaf	03:42.681	03:43.778	1.097
126	10.17.48.959.eaf	03:43.778	03:44.798	1.02
127	10.17.48.959.eaf	03:44.798	03:45.934	1.136
128	10.17.48.959.eaf	03:45.934	03:47.895	1.961
129	10.17.48.959.eaf	03:49.104	03:50.475	1.371
130	10.17.48.959.eaf	03:51.554	03:52.315	0.761
131	10.17.48.959.eaf	03:52.315	03:52.876	0.561
132	10.17.48.959.eaf	03:52.876	03:53.446	0.57
133	10.17.48.959.eaf	03:53.446	03:54.056	0.61
134	10.17.48.959.eaf	03:54.637	03:55.500	0.863
135	10.17.48.959.eaf	03:55.500	03:55.507	0.007
136	10.17.48.959.eaf	03:55.507	03:56.739	1.232
137	10.17.48.959.eaf	03:58.700	03:59.100	0.4
138	10.17.48.959.eaf	04:00.398	04:01.134	0.736
139	10.17.48.959.eaf	04:04.351	04:04.980	0.629
140	10.17.48.959.eaf	04:15.393	04:16.812	1.419
141	10.17.48.959.eaf	04:16.812	04:17.524	0.712
142	10.17.48.959.eaf	04:17.524	04:18.261	0.737
143	10.17.48.959.eaf	04:18.261	04:18.544	0.283
144	10.17.48.959.eaf	04:18.544	04:19.085	0.541
145	10.17.48.959.eaf	04:19.617	04:20.436	0.819
146	10.17.48.959.eaf	04:20.436	04:20.929	0.493
147	10.17.48.959.eaf	04:22.207	04:22.558	0.351
148	10.17.48.959.eaf	04:22.558	04:23.553	0.995
149	10.17.48.959.eaf	04:23.553	04:23.978	0.425
150	10.17.48.959.eaf	04:23.978	04:24.890	0.912
151	10.17.48.959.eaf	04:28.100	04:28.837	0.737
152	10.17.48.959.eaf	04:29.686	04:30.110	0.424
153	10.17.48.959.eaf	04:33.690	04:34.280	0.59
154	10.17.48.959.eaf	04:34.280	04:35.373	1.093
155	10.17.48.959.eaf	04:38.082	04:38.548	0.466
156	10.17.48.959.eaf	04:38.548	04:39.144	0.596
157	10.17.48.959.eaf	04:39.792	04:40.768	0.976
158	10.17.48.959.eaf	04:46.080	04:47.490	1.41
159	10.17.48.959.eaf	04:53.787	04:55.363	1.576
160	10.17.48.959.eaf	04:55.612	04:56.525	0.913
161	10.17.48.959.eaf	04:57.002	04:58.734	1.732
162	10.17.48.959.eaf	04:58.734	04:59.265	0.531
163	10.17.48.959.eaf	04:59.265	04:59.787	0.522
164	10.17.48.959.eaf	04:59.787	05:00.485	0.698
165	10.17.48.959.eaf	05:01.128	05:01.895	0.767
166	10.17.48.959.eaf	05:01.895	05:02.451	0.556
167	10.17.48.959.eaf	05:02.451	05:03.236	0.785
168	10.17.48.959.eaf	05:03.236	05:04.210	0.974
169	10.17.48.959.eaf	05:04.210	05:06.643	2.433

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
170	10.17.48.959.eaf	05:06.643	05:07.563	0.92
171	10.17.48.959.eaf	05:07.563	05:08.148	0.585
172	10.17.48.959.eaf	05:08.148	05:09.651	1.503
173	10.17.48.959.eaf	05:09.651	05:12.124	2.473
174	10.17.48.959.eaf	05:12.124	05:12.719	0.595
175	10.17.48.959.eaf	05:12.719	05:13.514	0.795
176	10.17.48.959.eaf	05:13.514	05:14.841	1.327
177	10.17.48.959.eaf	05:14.841	05:16.104	1.263
178	10.17.48.959.eaf	05:16.104	05:17.128	1.024
179	10.17.48.959.eaf	05:17.128	05:18.047	0.919
180	10.17.48.959.eaf	05:18.047	05:18.314	0.267
181	10.17.48.959.eaf	05:18.314	05:18.558	0.244
182	10.17.48.959.eaf	05:18.558	05:18.870	0.312
183	10.17.48.959.eaf	05:18.870	05:19.158	0.288
184	10.17.48.959.eaf	05:19.158	05:19.831	0.673
185	10.17.48.959.eaf	05:20.421	05:21.036	0.615
186	10.17.48.959.eaf	05:21.894	05:22.250	0.356
187	10.17.48.959.eaf	05:22.250	05:22.948	0.698
188	10.17.48.959.eaf	05:22.948	05:23.621	0.673
189	10.17.48.959.eaf	05:23.621	05:24.685	1.064
190	10.17.48.959.eaf	05:24.685	05:25.572	0.887
191	10.17.48.959.eaf	05:25.572	05:27.006	1.434
192	10.17.48.959.eaf	05:27.006	05:28.616	1.61
193	10.17.48.959.eaf	05:28.616	05:28.619	0.003
194	10.17.48.959.eaf	05:28.619	05:29.553	0.934
195	10.17.48.959.eaf	05:29.553	05:30.109	0.556
196	10.17.48.959.eaf	05:30.109	05:30.768	0.659
197	10.17.48.959.eaf	05:35.641	05:36.206	0.565
198	10.17.48.959.eaf	05:36.206	05:37.372	1.166
199	10.17.48.959.eaf	05:38.398	05:39.476	1.078
200	10.17.48.959.eaf	05:39.793	05:40.811	1.018
201	10.17.48.959.eaf	05:40.811	05:41.236	0.425
202	10.17.48.959.eaf	05:41.236	05:41.239	0.003
203	10.17.48.959.eaf	05:41.239	05:42.382	1.143
204	10.17.48.959.eaf	05:42.382	05:42.928	0.546
205	10.17.48.959.eaf	05:42.928	05:43.313	0.385
206	10.17.48.959.eaf	05:43.313	05:44.675	1.362
207	10.17.48.959.eaf	05:44.870	05:45.431	0.561
208	10.17.48.959.eaf	05:46.733	05:47.114	0.381
209	10.17.48.959.eaf	05:47.114	05:47.757	0.643
210	10.17.48.959.eaf	05:47.757	05:48.216	0.459
211	10.17.48.959.eaf	05:48.216	05:49.157	0.941
212	10.54.29.104.eaf	00:03.536	00:04.136	0.6
213	10.54.29.104.eaf	00:11.576	00:13.456	1.88
214	10.54.29.104.eaf	00:14.687	00:15.091	0.404
215	10.54.29.104.eaf	00:16.184	00:17.326	1.142
216	10.54.29.104.eaf	00:17.326	00:17.960	0.634
217	10.54.29.104.eaf	00:17.960	00:18.468	0.508
218	10.54.29.104.eaf	00:18.468	00:20.189	1.721
219	10.54.29.104.eaf	00:20.189	00:20.560	0.371
220	10.54.29.104.eaf	00:20.560	00:21.719	1.159
221	10.54.29.104.eaf	00:21.719	00:22.537	0.818
222	10.54.29.104.eaf	00:22.537	00:23.015	0.478
223	10.54.29.104.eaf	00:23.015	00:23.811	0.796
224	10.54.29.104.eaf	00:23.811	00:24.648	0.837
225	10.54.29.104.eaf	00:24.648	00:25.684	1.036
226	10.54.29.104.eaf	00:25.684	00:26.977	1.293
227	10.54.29.104.eaf	00:26.977	00:27.682	0.705
228	10.54.29.104.eaf	00:27.682	00:28.322	0.64
229	10.54.29.104.eaf	00:28.322	00:29.408	1.086
230	10.54.29.104.eaf	00:29.408	00:30.165	0.757
231	10.54.29.104.eaf	00:30.165	00:31.266	1.101
232	10.54.29.104.eaf	00:31.266	00:32.186	0.92
233	10.54.29.104.eaf	00:32.186	00:33.211	1.025
234	10.54.29.104.eaf	00:33.211	00:33.894	0.683
235	10.54.29.104.eaf	00:33.894	00:34.866	0.972
236	10.54.29.104.eaf	00:35.146	00:35.832	0.686
237	10.54.29.104.eaf	00:35.832	00:37.112	1.28
238	10.54.29.104.eaf	00:41.832	00:43.546	1.714
239	10.54.29.104.eaf	00:43.546	00:44.748	1.202
240	10.54.29.104.eaf	00:44.748	00:47.307	2.559
241	10.54.29.104.eaf	00:47.307	00:47.864	0.557
242	10.54.29.104.eaf	00:47.864	00:48.659	0.795
243	10.54.29.104.eaf	00:48.659	00:50.859	2.2
244	10.54.29.104.eaf	00:50.859	00:51.812	0.953
245	10.54.29.104.eaf	00:51.812	00:52.852	1.04
246	10.54.29.104.eaf	00:52.852	00:54.612	1.76
247	10.54.29.104.eaf	00:54.612	00:57.652	3.04
248	10.54.29.104.eaf	00:57.652	00:59.292	1.64
249	10.54.29.104.eaf	00:59.292	01:01.712	2.42
250	10.54.29.104.eaf	01:01.712	01:03.392	1.68
251	10.54.29.104.eaf	01:03.392	01:04.478	1.086
252	10.54.29.104.eaf	01:04.478	01:04.958	0.48
253	10.54.29.104.eaf	01:04.958	01:07.099	2.141
254	10.54.29.104.eaf	01:07.099	01:10.824	3.725
255	10.54.29.104.eaf	01:11.144	01:13.544	2.4

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
256	11.00.31.621.eaf	00:05.373	00:06.173	0.8
257	11.00.31.621.eaf	00:16.377	00:17.240	0.863
258	11.00.31.621.eaf	00:17.240	00:18.365	1.125
259	11.00.31.621.eaf	00:18.365	00:19.565	1.2
260	11.00.31.621.eaf	00:20.725	00:22.437	1.712
261	11.00.31.621.eaf	00:22.649	00:23.401	0.752
262	11.00.31.621.eaf	00:25.001	00:26.284	1.283
263	11.00.31.621.eaf	00:26.284	00:27.097	0.813
264	11.00.31.621.eaf	00:30.697	00:31.739	1.042
265	11.00.31.621.eaf	00:31.739	00:34.564	2.825
266	11.00.31.621.eaf	00:34.564	00:34.887	0.323
267	11.00.31.621.eaf	00:34.887	00:35.647	0.76
268	11.00.31.621.eaf	00:36.567	00:37.027	0.46
269	11.00.31.621.eaf	00:37.130	00:37.221	0.091
270	11.00.31.621.eaf	00:38.237	00:38.493	0.256
271	11.00.31.621.eaf	00:38.493	00:39.547	1.054
272	11.00.31.621.eaf	00:40.328	00:41.351	1.023
273	11.00.31.621.eaf	00:41.351	00:42.340	0.989
274	11.00.31.621.eaf	00:42.340	00:43.081	0.741
275	11.00.31.621.eaf	00:43.081	00:45.121	2.04
276	11.00.31.621.eaf	00:46.522	00:47.066	0.544
277	11.00.31.621.eaf	00:48.756	00:49.712	0.956
278	11.00.31.621.eaf	00:49.712	00:50.967	1.255
279	11.00.31.621.eaf	00:50.967	00:51.488	0.521
280	11.00.31.621.eaf	00:51.488	00:52.627	1.139
281	11.00.31.621.eaf	00:52.627	00:54.244	1.617
282	11.00.31.621.eaf	01:00.093	01:00.402	0.309
283	11.00.31.621.eaf	01:00.402	01:01.843	1.441
284	11.00.31.621.eaf	01:01.963	01:02.525	0.562
285	11.00.31.621.eaf	01:02.637	01:03.597	0.96
286	11.00.31.621.eaf	01:03.597	01:04.602	1.005
287	11.00.31.621.eaf	01:04.602	01:04.802	0.2
288	11.00.31.621.eaf	01:07.126	01:07.987	0.861
289	11.00.31.621.eaf	01:08.208	01:08.631	0.423
290	11.00.31.621.eaf	01:08.631	01:10.031	1.4
291	11.00.31.621.eaf	01:10.249	01:11.211	0.962
292	11.00.31.621.eaf	01:11.211	01:12.011	0.8
293	11.00.31.621.eaf	01:12.159	01:13.114	0.955
294	11.00.31.621.eaf	01:15.912	01:16.966	1.054
295	11.00.31.621.eaf	01:18.126	01:19.503	1.377
296	11.00.31.621.eaf	01:19.503	01:20.103	0.6
297	11.00.31.621.eaf	01:20.103	01:23.759	3.656
298	11.00.31.621.eaf	01:29.406	01:29.741	0.335
299	11.00.31.621.eaf	01:32.589	01:33.469	0.88
300	11.00.31.621.eaf	01:33.691	01:33.931	0.24
301	11.00.31.621.eaf	01:35.382	01:35.739	0.357
302	11.00.31.621.eaf	01:35.739	01:36.635	0.896
303	11.00.31.621.eaf	01:36.635	01:38.315	1.68
304	11.00.31.621.eaf	01:38.447	01:38.747	0.3
305	11.00.31.621.eaf	01:39.167	01:39.412	0.245
306	11.00.31.621.eaf	01:40.864	01:41.957	1.093
307	11.00.31.621.eaf	01:41.957	01:42.277	0.32
308	11.00.31.621.eaf	01:42.288	01:42.803	0.515
309	11.00.31.621.eaf	01:43.451	01:43.814	0.363
310	11.00.31.621.eaf	01:43.814	01:44.574	0.76
311	11.00.31.621.eaf	01:44.574	01:45.074	0.5
312	11.00.31.621.eaf	01:45.074	01:45.401	0.327
313	11.00.31.621.eaf	01:45.401	01:46.384	0.983
314	11.00.31.621.eaf	01:46.384	01:46.792	0.408
315	11.00.31.621.eaf	01:46.792	01:47.410	0.618
316	11.00.31.621.eaf	01:47.410	01:48.161	0.751
317	11.00.31.621.eaf	01:48.161	01:48.668	0.507
318	11.00.31.621.eaf	01:48.668	01:50.529	1.861
319	11.00.31.621.eaf	01:54.741	01:55.312	0.571
320	11.00.31.621.eaf	01:57.272	01:59.006	1.734
321	11.00.31.621.eaf	02:02.036	02:04.167	2.131
322	11.00.31.621.eaf	02:04.167	02:04.868	0.701
323	11.00.31.621.eaf	02:04.868	02:05.383	0.515
324	11.00.31.621.eaf	02:06.543	02:07.280	0.737
325	11.00.31.621.eaf	02:07.280	02:09.094	1.814
326	11.00.31.621.eaf	02:09.094	02:09.562	0.468
327	11.00.31.621.eaf	02:10.722	02:12.122	1.4
328	11.00.31.621.eaf	02:12.543	02:12.738	0.195
329	11.00.31.621.eaf	02:12.815	02:13.789	0.974
330	11.00.31.621.eaf	02:16.374	02:16.598	0.224
331	11.00.31.621.eaf	02:18.111	02:19.124	1.013
332	11.00.31.621.eaf	02:19.124	02:20.637	1.513
333	11.00.31.621.eaf	02:20.637	02:21.344	0.707
334	11.00.31.621.eaf	02:22.241	02:22.537	0.296
335	11.00.31.621.eaf	02:22.537	02:23.011	0.474
336	11.00.31.621.eaf	02:23.011	02:23.266	0.255
337	11.00.31.621.eaf	02:23.266	02:23.475	0.209
338	11.00.31.621.eaf	02:25.775	02:25.886	0.111
339	11.00.31.621.eaf	02:29.118	02:29.927	0.809
340	11.00.31.621.eaf	02:31.613	02:31.802	0.189
341	11.00.31.621.eaf	02:31.802	02:32.649	0.847

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
342	11.00.31.621.eaf	02:32.761	02:33.989	1.228
343	11.00.31.621.eaf	02:33.989	02:34.589	0.6
344	11.00.31.621.eaf	02:34.589	02:36.167	1.578
345	11.00.31.621.eaf	02:36.167	02:36.441	0.274
346	11.00.31.621.eaf	02:36.441	02:37.881	1.44
347	11.00.31.621.eaf	02:38.081	02:38.648	0.567
348	11.00.31.621.eaf	02:39.476	02:39.961	0.485
349	11.00.31.621.eaf	02:40.112	02:40.556	0.444
350	11.00.31.621.eaf	02:40.556	02:41.796	1.24
351	11.00.31.621.eaf	02:41.796	02:42.478	0.682
352	11.00.31.621.eaf	02:44.813	02:45.042	0.229
353	11.00.31.621.eaf	02:47.919	02:48.439	0.52
354	11.00.31.621.eaf	02:51.959	02:52.439	0.48
355	11.00.31.621.eaf	02:54.319	02:55.039	0.72
356	11.00.31.621.eaf	02:56.033	02:58.141	2.108
357	11.00.31.621.eaf	02:58.261	02:59.927	1.666
358	11.00.31.621.eaf	03:00.741	03:01.021	0.28
359	11.00.31.621.eaf	03:01.021	03:02.381	1.36
360	11.00.31.621.eaf	03:04.623	03:05.707	1.084
361	11.00.31.621.eaf	03:07.278	03:08.632	1.354
362	11.00.31.621.eaf	03:08.688	03:08.839	0.151
363	11.00.31.621.eaf	03:10.545	03:12.524	1.979
364	11.00.31.621.eaf	03:25.823	03:26.518	0.695
365	11.00.31.621.eaf	03:26.518	03:27.083	0.565
366	11.00.31.621.eaf	03:27.161	03:28.299	1.138
367	11.00.31.621.eaf	03:28.299	03:29.247	0.948
368	11.00.31.621.eaf	03:29.247	03:29.930	0.683
369	11.00.31.621.eaf	03:29.930	03:30.727	0.797
370	11.00.31.621.eaf	03:30.727	03:30.991	0.264
371	11.00.31.621.eaf	03:30.991	03:31.623	0.632
372	11.00.31.621.eaf	03:31.623	03:33.035	1.412
373	11.00.31.621.eaf	03:33.035	03:33.622	0.587
374	11.00.31.621.eaf	03:33.622	03:34.224	0.602
375	11.00.31.621.eaf	03:34.224	03:34.722	0.498
376	11.00.31.621.eaf	03:34.722	03:34.949	0.227
377	11.00.31.621.eaf	03:34.949	03:35.277	0.328
378	11.00.31.621.eaf	03:35.277	03:36.647	1.37
379	11.00.31.621.eaf	03:36.647	03:37.032	0.385
380	11.00.31.621.eaf	03:37.032	03:37.331	0.299
381	11.00.31.621.eaf	03:37.331	03:39.164	1.833
382	11.00.31.621.eaf	03:39.164	03:39.717	0.553
383	11.00.31.621.eaf	03:39.717	03:40.643	0.926
384	11.00.31.621.eaf	03:41.714	03:42.908	1.194
385	11.00.31.621.eaf	03:42.908	03:43.393	0.485
386	11.00.31.621.eaf	03:43.393	03:47.291	3.898
387	11.00.31.621.eaf	03:47.291	03:48.967	1.676
388	11.00.31.621.eaf	03:53.198	03:54.648	1.45
389	11.00.31.621.eaf	03:54.648	03:55.345	0.697
390	11.00.31.621.eaf	03:55.345	03:55.569	0.224
391	11.00.31.621.eaf	03:55.569	03:56.122	0.553
392	11.00.31.621.eaf	03:56.122	03:56.686	0.564
393	11.00.31.621.eaf	03:56.686	03:57.841	1.155
394	11.00.31.621.eaf	03:57.841	03:58.497	0.656
395	11.00.31.621.eaf	03:58.497	03:58.968	0.471
396	11.00.31.621.eaf	03:58.968	03:59.667	0.699
397	11.00.31.621.eaf	03:59.667	04:01.235	1.568
398	11.00.31.621.eaf	04:01.235	04:02.980	1.745
399	11.00.31.621.eaf	04:02.980	04:04.051	1.071
400	11.00.31.621.eaf	04:04.051	04:05.013	0.962
401	11.00.31.621.eaf	04:05.013	04:05.915	0.902
402	11.00.31.621.eaf	04:05.915	04:06.532	0.617
403	11.00.31.621.eaf	04:07.454	04:07.810	0.356
404	11.00.31.621.eaf	04:08.438	04:09.496	1.058
405	11.00.31.621.eaf	04:09.496	04:10.216	0.72
406	11.00.31.621.eaf	04:11.884	04:12.583	0.699
407	11.00.31.621.eaf	04:12.583	04:13.008	0.425
408	11.00.31.621.eaf	04:13.008	04:13.423	0.415
409	11.00.31.621.eaf	04:13.423	04:13.989	0.566
410	11.00.31.621.eaf	04:13.989	04:14.258	0.269
411	11.00.31.621.eaf	04:15.698	04:15.965	0.267
412	11.00.31.621.eaf	04:15.965	04:16.295	0.33
413	11.00.31.621.eaf	04:16.295	04:17.255	0.96
414	11.00.31.621.eaf	04:27.390	04:27.701	0.311
415	11.00.31.621.eaf	04:27.701	04:28.101	0.4
416	11.00.31.621.eaf	04:28.500	04:30.177	1.677
417	11.00.31.621.eaf	04:33.377	04:34.337	0.96
418	11.00.31.621.eaf	04:34.337	04:34.737	0.4
419	11.00.31.621.eaf	04:35.657	04:36.273	0.616
420	11.00.31.621.eaf	04:36.273	04:36.686	0.413
421	11.00.31.621.eaf	04:36.686	04:37.494	0.808
422	11.00.31.621.eaf	04:37.494	04:39.073	1.579
423	11.00.31.621.eaf	04:41.219	04:42.256	1.037
424	11.00.31.621.eaf	04:46.520	04:46.632	0.112
425	11.00.31.621.eaf	04:46.632	04:47.042	0.41
426	11.00.31.621.eaf	04:47.042	04:47.362	0.32
427	11.00.31.621.eaf	04:48.370	04:49.234	0.864

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
428	11.00.31.621.eaf	04:49.234	04:49.583	0.349
429	11.00.31.621.eaf	04:49.583	04:50.783	1.2
430	11.00.31.621.eaf	04:51.566	04:51.966	0.4
431	11.00.31.621.eaf	04:52.225	04:52.616	0.391
432	11.00.31.621.eaf	04:53.353	04:55.285	1.932
433	11.00.31.621.eaf	04:55.285	04:55.605	0.32
434	11.00.31.621.eaf	04:57.896	04:58.883	0.987
435	11.00.31.621.eaf	05:03.026	05:03.301	0.275
436	11.00.31.621.eaf	05:03.301	05:03.736	0.435
437	11.00.31.621.eaf	05:08.093	05:09.088	0.995
438	11.00.31.621.eaf	05:09.088	05:10.709	1.621
439	11.00.31.621.eaf	05:11.703	05:11.931	0.228
440	11.00.31.621.eaf	05:11.931	05:13.558	1.627
441	11.00.31.621.eaf	05:13.558	05:13.868	0.31
442	11.00.31.621.eaf	05:13.868	05:15.438	1.57
443	11.00.31.621.eaf	05:15.438	05:16.817	1.379
444	11.00.31.621.eaf	05:16.817	05:18.342	1.525
445	11.00.31.621.eaf	05:19.222	05:20.542	1.32
446	11.00.31.621.eaf	05:20.542	05:20.978	0.436
447	11.00.31.621.eaf	05:21.419	05:21.756	0.337
448	11.00.31.621.eaf	05:21.756	05:22.196	0.44
449	11.00.31.621.eaf	05:22.196	05:23.897	1.701
450	11.00.31.621.eaf	05:23.897	05:24.073	0.176
451	11.00.31.621.eaf	05:24.073	05:25.771	1.698
452	11.00.31.621.eaf	05:25.771	05:27.349	1.578
453	11.00.31.621.eaf	05:29.996	05:30.124	0.128
454	11.00.31.621.eaf	05:30.404	05:31.116	0.712
455	11.00.31.621.eaf	05:31.116	05:31.436	0.32
456	11.00.31.621.eaf	05:31.667	05:33.827	2.16
457	11.00.31.621.eaf	05:33.827	05:35.089	1.262
458	11.00.31.621.eaf	05:35.089	05:35.809	0.72
459	11.00.31.621.eaf	05:35.809	05:36.284	0.475
460	11.00.31.621.eaf	05:36.757	05:37.197	0.44
461	11.00.31.621.eaf	05:37.197	05:39.211	2.014
462	11.00.31.621.eaf	05:39.211	05:40.249	1.038
463	11.00.31.621.eaf	05:40.249	05:40.939	0.69
464	11.00.31.621.eaf	05:40.939	05:42.299	1.36
465	11.00.31.621.eaf	05:42.299	05:43.539	1.24
466	11.00.31.621.eaf	05:47.299	05:47.806	0.507
467	11.00.31.621.eaf	05:51.316	05:52.053	0.737
468	11.00.31.621.eaf	05:55.013	05:55.216	0.203
469	11.00.31.621.eaf	05:56.302	05:57.739	1.437
470	11.17.45.463.eaf	00:10.601	00:10.980	0.379
471	11.17.45.463.eaf	00:11.337	00:12.053	0.716
472	11.17.45.463.eaf	00:12.264	00:13.541	1.277
473	11.17.45.463.eaf	00:19.787	00:21.350	1.563
474	11.17.45.463.eaf	00:21.656	00:22.656	1.0
475	11.17.45.463.eaf	00:27.717	00:28.966	1.249
476	11.17.45.463.eaf	00:36.961	00:38.381	1.42
477	11.17.45.463.eaf	00:45.321	00:46.273	0.952
478	11.17.45.463.eaf	01:06.297	01:07.434	1.137
479	11.17.45.463.eaf	01:08.307	01:10.300	1.993
480	11.17.45.463.eaf	01:31.385	01:32.384	0.999
481	11.17.45.463.eaf	01:32.850	01:33.630	0.78
482	11.17.45.463.eaf	01:34.628	01:36.096	1.468
483	11.17.45.463.eaf	01:37.538	01:38.536	0.998
484	11.17.45.463.eaf	01:40.235	01:40.844	0.609
485	11.17.45.463.eaf	01:43.365	01:44.175	0.81
486	11.17.45.463.eaf	02:04.368	02:04.978	0.61
487	11.17.45.463.eaf	02:05.755	02:06.601	0.846
488	11.17.45.463.eaf	02:09.790	02:10.790	1.0
489	11.17.45.463.eaf	02:11.974	02:12.358	0.384
490	11.17.45.463.eaf	02:29.343	02:30.501	1.158
491	11.17.45.463.eaf	02:31.953	02:32.343	0.39
492	11.17.45.463.eaf	02:50.881	02:51.876	0.995
493	11.17.45.463.eaf	02:52.167	02:53.568	1.401
494	11.17.45.463.eaf	02:55.689	02:56.232	0.543
495	11.17.45.463.eaf	02:56.232	02:56.808	0.576
496	11.17.45.463.eaf	03:08.155	03:08.711	0.556
497	11.17.45.463.eaf	03:35.828	03:36.408	0.58
498	11.17.45.463.eaf	03:40.530	03:41.539	1.009
499	11.17.45.463.eaf	03:45.483	03:46.139	0.656
500	11.17.45.463.eaf	03:48.910	03:49.733	0.823
501	11.17.45.463.eaf	03:54.253	03:54.953	0.7
502	11.17.45.463.eaf	04:02.090	04:02.787	0.697
503	11.17.45.463.eaf	04:05.613	04:06.272	0.659
504	11.17.45.463.eaf	04:09.301	04:10.012	0.711
505	11.17.45.463.eaf	04:14.044	04:14.916	0.872
506	11.17.45.463.eaf	04:32.260	04:33.658	1.398
507	11.17.45.463.eaf	04:55.755	04:56.346	0.591
508	11.17.45.463.eaf	05:02.113	05:02.574	0.461
509	12.05.31.682.eaf	00:01.662	00:02.734	1.072
510	12.05.31.682.eaf	00:03.613	00:04.453	0.84
511	12.05.31.682.eaf	00:06.430	00:06.651	0.221
512	12.05.31.682.eaf	00:10.691	00:11.214	0.523
513	12.05.31.682.eaf	00:11.214	00:11.682	0.468

gphr	recording	gphr_start (mm:ss.f)	gphr_end (mm:ss.f)	gphr_dur (s.f)
514	12.05.31.682.eaf	00:11.682	00:12.827	1.145
515	12.05.31.682.eaf	00:16.599	00:17.812	1.213
516	12.05.31.682.eaf	00:18.780	00:19.500	0.72
517	12.05.31.682.eaf	00:23.944	00:24.194	0.25
518	12.05.31.682.eaf	00:24.194	00:25.471	1.277
519	12.05.31.682.eaf	00:30.223	00:30.962	0.739
520	12.05.31.682.eaf	00:42.809	00:43.728	0.919
521	12.05.31.682.eaf	00:45.862	00:46.177	0.315
522	12.05.31.682.eaf	00:46.497	00:46.988	0.491
523	12.05.31.682.eaf	00:47.192	00:47.625	0.433
524	12.05.31.682.eaf	00:47.850	00:48.164	0.314
525	12.05.31.682.eaf	00:48.164	00:48.609	0.445
526	12.05.31.682.eaf	00:48.609	00:49.197	0.588
527	12.05.31.682.eaf	00:49.197	00:49.700	0.503
528	12.05.31.682.eaf	00:50.735	00:51.614	0.879
529	12.05.31.682.eaf	00:55.984	00:56.395	0.411
530	12.05.31.682.eaf	00:57.854	00:59.077	1.223
531	12.05.31.682.eaf	00:59.857	01:01.148	1.291
532	12.05.31.682.eaf	01:01.148	01:02.333	1.185
533	12.05.31.682.eaf	01:05.279	01:06.016	0.737
534	12.05.31.682.eaf	01:13.365	01:13.802	0.437
535	12.05.31.682.eaf	01:18.044	01:18.756	0.712
536	12.05.31.682.eaf	01:21.828	01:23.411	1.583
537	12.05.31.682.eaf	01:27.013	01:27.755	0.742
538	12.05.31.682.eaf	01:29.328	01:29.528	0.2
539	12.05.31.682.eaf	01:30.887	01:32.540	1.653
540	12.05.31.682.eaf	01:33.879	01:34.407	0.528
541	12.05.31.682.eaf	01:34.638	01:35.671	1.033
542	12.05.31.682.eaf	01:35.671	01:37.138	1.467
543	12.05.31.682.eaf	01:37.361	01:38.551	1.19
544	12.05.31.682.eaf	01:39.303	01:39.924	0.621
545	12.05.31.682.eaf	01:40.514	01:41.180	0.666
546	12.05.31.682.eaf	01:41.989	01:42.279	0.29
547	12.05.31.682.eaf	01:47.641	01:47.925	0.284
548	12.05.31.682.eaf	01:48.385	01:48.983	0.598
549	12.05.31.682.eaf	01:55.333	01:55.886	0.553
550	12.05.31.682.eaf	02:02.266	02:03.106	0.84
551	12.05.31.682.eaf	02:08.308	02:09.308	1.0
552	12.05.31.682.eaf	02:15.296	02:16.340	1.044
553	12.05.31.682.eaf	02:18.145	02:19.082	0.937
554	12.05.31.682.eaf	02:19.757	02:20.548	0.791
555	12.05.31.682.eaf	02:21.916	02:22.621	0.705
556	12.05.31.682.eaf	02:22.810	02:23.344	0.534
557	12.05.31.682.eaf	02:24.952	02:25.471	0.519
558	12.05.31.682.eaf	02:25.471	02:26.045	0.574
559	12.05.31.682.eaf	02:27.040	02:27.923	0.883
560	12.05.31.682.eaf	02:29.868	02:30.438	0.57
561	12.05.31.682.eaf	02:32.684	02:33.443	0.759
562	12.05.31.682.eaf	02:34.162	02:35.376	1.214
563	12.05.31.682.eaf	02:35.376	02:36.194	0.818
564	12.05.31.682.eaf	02:36.863	02:38.169	1.306
565	12.05.31.682.eaf	02:56.342	02:57.186	0.844
566	12.05.31.682.eaf	02:57.908	02:58.556	0.648
567	12.05.31.682.eaf	03:02.002	03:02.701	0.699
568	12.05.31.682.eaf	03:03.183	03:04.035	0.852
569	12.05.31.682.eaf	03:05.752	03:06.749	0.997
570	12.05.31.682.eaf	03:07.537	03:07.910	0.373
571	12.05.31.682.eaf	03:08.620	03:09.070	0.45
572	12.05.31.682.eaf	03:10.383	03:11.673	1.29
573	12.05.31.682.eaf	03:12.397	03:13.924	1.527
574	12.05.31.682.eaf	03:18.384	03:19.048	0.664
575	12.05.31.682.eaf	03:19.694	03:19.961	0.267
576	12.05.31.682.eaf	03:20.883	03:21.802	0.919
577	12.05.31.682.eaf	03:23.229	03:23.699	0.47
578	12.05.31.682.eaf	03:23.940	03:24.445	0.505
579	12.05.31.682.eaf	03:27.381	03:28.542	1.161
580	12.05.31.682.eaf	03:29.259	03:30.256	0.997
581	12.05.31.682.eaf	03:32.046	03:32.968	0.922
582	12.05.31.682.eaf	03:33.379	03:34.358	0.979
583	12.05.31.682.eaf	03:42.314	03:42.909	0.595
584	12.05.31.682.eaf	03:51.515	03:53.288	1.773
585	12.05.31.682.eaf	03:53.845	03:54.169	0.324
586	12.05.31.682.eaf	04:04.843	04:05.621	0.778
587	12.05.31.682.eaf	04:06.382	04:07.058	0.676
588	12.05.31.682.eaf	04:11.884	04:12.690	0.806
589	12.05.31.682.eaf	04:12.690	04:13.109	0.419
590	12.05.31.682.eaf	04:13.109	04:14.081	0.972
591	12.05.31.682.eaf	04:14.453	04:15.031	0.578
592	12.05.31.682.eaf	04:16.643	04:17.621	0.978
593	12.05.31.682.eaf	04:18.028	04:18.465	0.437
594	12.05.31.682.eaf	04:19.922	04:20.678	0.756
595	12.05.31.682.eaf	04:21.182	04:22.048	0.866
596	12.05.31.682.eaf	04:26.723	04:27.403	0.68
597	12.05.31.682.eaf	04:29.473	04:30.064	0.591
598	12.05.31.682.eaf	04:31.122	04:31.641	0.519
599	12.05.31.682.eaf	04:32.086	04:33.489	1.403

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
600	12.05.31.682.eaf	04:33.608	04:34.287	0.679
601	12.05.31.682.eaf	04:34.482	04:35.013	0.531
602	12.05.31.682.eaf	04:35.566	04:36.137	0.571
603	12.05.31.682.eaf	04:36.263	04:36.997	0.734
604	12.05.31.682.eaf	04:36.997	04:38.508	1.511
605	12.05.31.682.eaf	04:39.479	04:39.837	0.358
606	12.05.31.682.eaf	04:43.044	04:43.433	0.389
607	12.05.31.682.eaf	04:43.856	04:45.428	1.572
608	12.05.31.682.eaf	04:46.942	04:48.393	1.451
609	12.05.31.682.eaf	04:48.394	04:49.820	1.426
610	12.05.31.682.eaf	04:50.117	04:51.477	1.36
611	12.05.31.682.eaf	04:53.144	04:53.914	0.77
612	12.05.31.682.eaf	05:04.664	05:05.323	0.659
613	12.05.31.682.eaf	05:05.947	05:06.851	0.904
614	12.05.31.682.eaf	05:09.028	05:10.105	1.077
615	12.05.31.682.eaf	05:11.183	05:11.778	0.595
616	12.05.31.682.eaf	05:18.337	05:18.699	0.362
617	12.05.31.682.eaf	05:18.793	05:19.661	0.868
618	12.05.31.682.eaf	05:21.838	05:22.338	0.5
619	12.05.31.682.eaf	05:24.088	05:25.833	1.745
620	12.05.31.682.eaf	05:27.329	05:27.841	0.512
621	12.37.35.766.eaf	00:03.673	00:04.793	1.12
622	12.37.35.766.eaf	00:08.783	00:09.262	0.479
623	12.37.35.766.eaf	00:09.262	00:10.142	0.88
624	12.37.35.766.eaf	00:10.142	00:10.880	0.738
625	12.37.35.766.eaf	00:10.880	00:11.742	0.862
626	12.37.35.766.eaf	00:11.742	00:12.342	0.6
627	12.37.35.766.eaf	00:12.342	00:12.982	0.64
628	12.37.35.766.eaf	00:13.742	00:16.222	2.48
629	12.37.35.766.eaf	00:16.222	00:16.282	0.06
630	12.37.35.766.eaf	00:17.202	00:17.742	0.54
631	12.37.35.766.eaf	00:21.302	00:22.222	0.92
632	12.37.35.766.eaf	00:22.222	00:23.362	1.14
633	12.37.35.766.eaf	00:23.362	00:24.274	0.912
634	12.37.35.766.eaf	00:24.274	00:24.954	0.68
635	12.37.35.766.eaf	00:24.954	00:25.411	0.457
636	12.37.35.766.eaf	00:25.411	00:26.971	1.56
637	12.37.35.766.eaf	00:26.971	00:27.741	0.77
638	12.37.35.766.eaf	00:27.741	00:28.527	0.786
639	12.37.35.766.eaf	00:28.527	00:29.327	0.8
640	12.37.35.766.eaf	00:29.327	00:30.287	0.96
641	12.37.35.766.eaf	00:30.287	00:31.160	0.873
642	12.37.35.766.eaf	00:31.160	00:32.380	1.22
643	12.37.35.766.eaf	00:32.380	00:33.050	0.67
644	12.37.35.766.eaf	00:33.050	00:33.590	0.54
645	12.37.35.766.eaf	00:40.030	00:41.551	1.521
646	12.37.35.766.eaf	00:41.551	00:42.311	0.76
647	12.37.35.766.eaf	00:42.311	00:44.231	1.92
648	12.37.35.766.eaf	00:44.231	00:45.111	0.88
649	12.37.35.766.eaf	00:45.340	00:46.500	1.16
650	12.37.35.766.eaf	00:51.182	00:51.879	0.697
651	12.37.35.766.eaf	01:06.676	01:07.592	0.916
652	12.37.35.766.eaf	01:07.592	01:08.032	0.44
653	12.37.35.766.eaf	01:08.032	01:09.316	1.284
654	12.50.52.223.eaf	00:06.895	00:07.833	0.938
655	12.50.52.223.eaf	00:09.050	00:10.080	1.03
656	12.50.52.223.eaf	00:10.080	00:10.793	0.713
657	12.50.52.223.eaf	00:11.275	00:11.897	0.622
658	12.50.52.223.eaf	00:12.361	00:13.740	1.379
659	12.50.52.223.eaf	00:14.535	00:14.788	0.253
660	12.50.52.223.eaf	00:17.076	00:17.888	0.812
661	12.50.52.223.eaf	00:18.337	00:18.892	0.555
662	12.50.52.223.eaf	00:19.430	00:20.371	0.941
663	12.50.52.223.eaf	00:21.826	00:22.629	0.803
664	12.50.52.223.eaf	00:26.884	00:27.710	0.826
665	12.50.52.223.eaf	00:27.710	00:28.533	0.823
666	12.50.52.223.eaf	00:29.424	00:29.892	0.468
667	12.50.52.223.eaf	00:30.233	00:31.201	0.968
668	12.50.52.223.eaf	00:31.521	00:32.409	0.888
669	12.50.52.223.eaf	00:34.985	00:35.910	0.925
670	12.50.52.223.eaf	00:41.293	00:42.121	0.828
671	12.50.52.223.eaf	00:42.121	00:44.339	2.218
672	12.50.52.223.eaf	00:45.680	00:46.450	0.77
673	12.50.52.223.eaf	00:46.450	00:47.451	1.001
674	12.50.52.223.eaf	00:48.689	00:50.418	1.729
675	12.50.52.223.eaf	00:51.191	00:52.264	1.073
676	12.50.52.223.eaf	00:52.264	00:52.532	0.268
677	12.50.52.223.eaf	00:52.532	00:53.884	1.352
678	12.50.52.223.eaf	00:53.884	00:55.440	1.556
679	12.50.52.223.eaf	00:56.605	00:57.816	1.211
680	12.50.52.223.eaf	01:01.780	01:02.850	1.07
681	12.50.52.223.eaf	01:03.085	01:04.085	1.0
682	12.50.52.223.eaf	01:04.924	01:05.851	0.927
683	12.50.52.223.eaf	01:06.610	01:07.610	1.0
684	13.09.12.480.eaf	00:04.888	00:05.862	0.974
685	13.09.12.480.eaf	00:36.972	00:38.037	1.065

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
686	13.09.12.480.eaf	00:50.189	00:51.184	0.995
687	13.09.12.480.eaf	00:59.625	01:00.540	0.915
688	13.14.02.898.eaf	00:03.926	00:04.868	0.942
689	13.14.02.898.eaf	00:05.609	00:06.748	1.139
690	13.14.02.898.eaf	00:06.813	00:07.465	0.652
691	13.14.02.898.eaf	00:07.712	00:08.451	0.739
692	13.14.02.898.eaf	00:08.743	00:09.400	0.657
693	13.14.02.898.eaf	00:09.406	00:10.413	1.007
694	13.14.02.898.eaf	00:10.413	00:17.268	6.855
695	13.14.02.898.eaf	00:17.268	00:18.363	1.095
696	13.14.02.898.eaf	00:18.363	00:19.171	0.808
697	13.14.02.898.eaf	00:19.171	00:19.956	0.785
698	13.14.02.898.eaf	00:20.785	00:21.575	0.79
699	13.14.02.898.eaf	00:21.575	00:23.252	1.677
700	13.14.02.898.eaf	00:23.252	00:24.543	1.291
701	13.14.02.898.eaf	00:24.543	00:25.570	1.027
702	13.14.02.898.eaf	00:25.570	00:28.410	2.84
703	13.14.02.898.eaf	00:28.410	00:30.149	1.739
704	13.14.02.898.eaf	00:30.149	00:30.954	0.805
705	13.14.02.898.eaf	00:30.954	00:33.111	2.157
706	13.14.02.898.eaf	00:33.111	00:34.797	1.686
707	13.14.02.898.eaf	00:35.793	00:36.118	0.325
708	13.14.02.898.eaf	00:36.118	00:36.988	0.87
709	13.14.02.898.eaf	00:36.988	00:38.106	1.118
710	13.14.02.898.eaf	00:38.106	00:39.263	1.157
711	13.14.02.898.eaf	00:39.263	00:39.991	0.728
712	13.14.02.898.eaf	00:39.991	00:42.418	2.427
713	13.14.02.898.eaf	00:42.418	00:45.178	2.76
714	13.14.02.898.eaf	00:45.781	00:46.780	0.999
715	13.14.02.898.eaf	00:46.780	00:47.282	0.502
716	13.14.02.898.eaf	00:47.282	00:48.526	1.244
717	13.14.02.898.eaf	00:48.526	00:48.961	0.435
718	13.14.02.898.eaf	00:48.961	00:50.894	1.933
719	13.14.02.898.eaf	00:50.894	00:51.842	0.948
720	13.14.02.898.eaf	00:51.842	00:53.548	1.706
721	13.14.02.898.eaf	00:53.548	00:54.396	0.848
722	13.14.02.898.eaf	00:54.893	00:56.891	1.998
723	13.14.02.898.eaf	00:56.891	00:57.999	1.108
724	13.14.02.898.eaf	00:57.999	00:58.773	0.774
725	13.14.02.898.eaf	00:58.773	00:59.921	1.148
726	13.14.02.898.eaf	00:59.921	01:01.830	1.909
727	13.14.02.898.eaf	01:01.830	01:02.551	0.721
728	13.14.02.898.eaf	01:03.672	01:04.157	0.485
729	13.14.02.898.eaf	01:04.157	01:05.582	1.425
730	13.14.02.898.eaf	01:06.329	01:07.703	1.374
731	13.14.02.898.eaf	01:07.703	01:08.731	1.028
732	13.14.02.898.eaf	01:08.731	01:09.796	1.065
733	13.14.02.898.eaf	01:09.796	01:10.584	0.788
734	13.14.02.898.eaf	01:10.584	01:11.772	1.188
735	13.14.02.898.eaf	01:11.772	01:13.376	1.604
736	13.14.02.898.eaf	01:13.376	01:14.259	0.883
737	13.14.02.898.eaf	01:14.259	01:17.962	3.703
738	14.40.09.299.eaf	00:46.302	00:47.653	1.351
739	14.40.09.299.eaf	00:54.327	00:54.716	0.389
740	14.40.09.299.eaf	01:23.404	01:24.900	1.496
741	14.40.09.299.eaf	01:47.417	01:49.931	2.514
742	14.40.09.299.eaf	02:42.884	02:43.297	0.413
743	14.40.09.299.eaf	02:48.190	02:49.198	1.008
744	14.40.09.299.eaf	03:01.531	03:01.737	0.206
745	14.40.09.299.eaf	03:01.737	03:02.566	0.829
746	14.40.09.299.eaf	03:02.566	03:03.872	1.306
747	14.40.09.299.eaf	03:09.817	03:10.900	1.083
748	14.40.09.299.eaf	03:37.645	03:38.254	0.609
749	14.40.09.299.eaf	03:39.002	03:39.449	0.447
750	14.40.09.299.eaf	03:41.733	03:42.273	0.54
751	14.40.09.299.eaf	03:42.577	03:43.235	0.658
752	14.40.09.299.eaf	03:43.967	03:45.187	1.22
753	14.40.09.299.eaf	03:52.223	03:52.762	0.539
754	14.40.09.299.eaf	04:07.583	04:09.109	1.526
755	14.40.09.299.eaf	04:22.811	04:24.847	2.036
756	14.40.09.299.eaf	04:39.396	04:41.340	1.944
757	14.40.09.299.eaf	04:43.165	04:43.821	0.656
758	14.40.09.299.eaf	04:46.238	04:47.464	1.226
759	14.40.09.299.eaf	04:59.313	04:59.762	0.449
760	14.40.09.299.eaf	05:00.701	05:01.172	0.471
761	14.40.09.299.eaf	05:13.407	05:13.813	0.406
762	14.40.09.299.eaf	05:15.629	05:15.944	0.315
763	15.04.57.785.eaf	00:09.898	00:10.628	0.73
764	15.04.57.785.eaf	00:22.999	00:23.741	0.742
765	15.04.57.785.eaf	00:24.886	00:26.103	1.217
766	15.04.57.785.eaf	00:26.931	00:27.709	0.778
767	15.04.57.785.eaf	00:29.604	00:30.694	1.09
768	15.04.57.785.eaf	00:39.156	00:39.979	0.823
769	15.04.57.785.eaf	01:00.373	01:02.065	1.692
770	15.04.57.785.eaf	01:02.511	01:05.060	2.549
771	15.04.57.785.eaf	01:15.495	01:16.079	0.584

gphr	recording	gphr_start (mm:ss.f)	gphr_end (mm:ss.f)	gphr_dur (s.f)
772	15.04.57.785.eaf	01:20.781	01:21.587	0.806
773	15.04.57.785.eaf	01:21.587	01:23.410	1.823
774	15.04.57.785.eaf	01:38.188	01:40.220	2.032
775	15.04.57.785.eaf	01:42.611	01:44.324	1.713
776	15.04.57.785.eaf	01:52.649	01:54.311	1.662
777	15.04.57.785.eaf	02:03.150	02:04.062	0.912
778	15.04.57.785.eaf	02:04.277	02:04.687	0.41
779	15.04.57.785.eaf	02:05.406	02:07.100	1.694
780	15.04.57.785.eaf	02:14.393	02:14.988	0.595
781	15.04.57.785.eaf	02:16.046	02:16.580	0.534
782	15.04.57.785.eaf	02:24.272	02:24.925	0.653
783	15.04.57.785.eaf	02:25.534	02:28.472	2.938
784	15.04.57.785.eaf	02:29.234	02:30.215	0.981
785	15.04.57.785.eaf	02:30.215	02:31.208	0.993
786	15.04.57.785.eaf	02:31.208	02:32.147	0.939
787	15.04.57.785.eaf	02:32.729	02:33.243	0.514
788	15.04.57.785.eaf	02:33.243	02:33.849	0.606
789	15.04.57.785.eaf	02:33.849	02:35.569	1.72
790	15.04.57.785.eaf	02:35.569	02:37.822	2.253
791	15.04.57.785.eaf	02:37.822	02:39.312	1.49
792	15.04.57.785.eaf	02:39.542	02:39.876	0.334
793	15.04.57.785.eaf	02:39.876	02:40.652	0.776
794	15.04.57.785.eaf	02:40.652	02:41.930	1.278
795	15.04.57.785.eaf	02:49.975	02:50.540	0.565
796	15.04.57.785.eaf	03:01.384	03:02.315	0.931
797	15.04.57.785.eaf	03:02.315	03:02.721	0.406
798	15.04.57.785.eaf	03:07.402	03:07.999	0.597
799	15.04.57.785.eaf	03:08.324	03:09.238	0.914
800	15.04.57.785.eaf	03:09.703	03:10.123	0.42
801	15.04.57.785.eaf	03:11.942	03:12.544	0.602
802	15.04.57.785.eaf	03:12.544	03:12.988	0.444
803	15.04.57.785.eaf	03:12.988	03:13.581	0.593
804	15.04.57.785.eaf	03:26.047	03:26.898	0.851
805	15.04.57.785.eaf	03:33.034	03:34.090	1.056
806	15.04.57.785.eaf	03:35.613	03:36.690	1.077
807	15.04.57.785.eaf	03:36.690	03:38.540	1.85
808	15.04.57.785.eaf	03:38.540	03:40.001	1.461
809	15.04.57.785.eaf	04:00.921	04:02.354	1.433
810	15.04.57.785.eaf	04:02.354	04:04.464	2.11
811	15.04.57.785.eaf	04:12.881	04:14.284	1.403
812	15.04.57.785.eaf	04:27.022	04:28.158	1.136
813	15.04.57.785.eaf	04:30.090	04:32.122	2.032
814	15.27.51.757.eaf	01:10.498	01:11.349	0.851
815	15.27.51.757.eaf	01:13.412	01:13.835	0.423
816	15.27.51.757.eaf	01:20.599	01:21.474	0.875
817	15.27.51.757.eaf	01:24.829	01:25.801	0.972
818	15.27.51.757.eaf	01:25.801	01:27.379	1.578
819	15.27.51.757.eaf	01:30.725	01:32.184	1.459
820	15.27.51.757.eaf	01:42.367	01:45.146	2.779
821	15.41.04.113.eaf	00:29.614	00:30.191	0.577
822	15.41.04.113.eaf	00:54.008	00:54.893	0.885
823	15.41.04.113.eaf	00:56.787	00:58.324	1.537
824	15.41.04.113.eaf	01:00.083	01:01.549	1.466
825	15.41.04.113.eaf	01:03.665	01:05.404	1.739
826	15.41.04.113.eaf	01:12.494	01:13.394	0.9
827	15.41.04.113.eaf	01:13.394	01:14.534	1.14
828	15.41.04.113.eaf	01:14.876	01:16.187	1.311
829	15.41.04.113.eaf	01:17.827	01:18.720	0.893
830	15.41.04.113.eaf	01:20.376	01:21.679	1.303
831	15.41.04.113.eaf	01:22.572	01:23.563	0.991
832	15.41.04.113.eaf	01:25.159	01:26.591	1.432
833	15.41.04.113.eaf	01:26.591	01:27.698	1.107
834	15.41.04.113.eaf	01:45.423	01:47.145	1.722
835	15.41.04.113.eaf	01:47.145	01:48.231	1.086
836	15.41.04.113.eaf	01:48.231	01:49.283	1.052
837	15.41.04.113.eaf	01:49.283	01:49.974	0.691
838	15.41.04.113.eaf	01:50.178	01:50.842	0.664
839	15.41.04.113.eaf	01:50.842	01:52.223	1.381
840	15.41.04.113.eaf	01:52.223	01:54.478	2.255
841	15.41.04.113.eaf	01:55.419	01:57.385	1.966
842	15.41.04.113.eaf	01:59.530	02:00.445	0.915
843	15.41.04.113.eaf	02:02.398	02:03.768	1.37
844	15.41.04.113.eaf	02:03.768	02:04.821	1.053
845	15.41.04.113.eaf	02:06.777	02:08.291	1.514
846	15.41.04.113.eaf	02:21.533	02:23.680	2.147
847	15.41.04.113.eaf	02:25.019	02:25.524	0.505
848	15.41.04.113.eaf	02:25.524	02:26.587	1.063
849	15.41.04.113.eaf	02:28.734	02:29.551	0.817
850	15.41.04.113.eaf	02:47.814	02:49.313	1.499
851	15.41.04.113.eaf	02:49.313	02:50.109	0.796
852	15.41.04.113.eaf	02:50.109	02:51.900	1.791
853	15.41.04.113.eaf	02:51.900	02:53.090	1.19
854	15.41.04.113.eaf	02:54.080	02:55.361	1.281
855	15.41.04.113.eaf	03:04.416	03:05.324	0.908
856	15.41.04.113.eaf	03:05.324	03:06.723	1.399
857	15.41.04.113.eaf	03:25.903	03:27.232	1.329

gphr	recording	gphr_start (mm:ss.f)	gphr_end (mm:ss.f)	gphr_dur (s.f)
858	15.41.04.113.eaf	03:27.232	03:28.489	1.257
859	15.41.04.113.eaf	03:29.240	03:30.308	1.068
860	15.41.04.113.eaf	03:30.308	03:31.097	0.789
861	15.41.04.113.eaf	03:31.097	03:31.532	0.435
862	15.41.04.113.eaf	03:31.532	03:32.332	0.8
863	15.41.04.113.eaf	03:33.885	03:34.931	1.046
864	15.41.04.113.eaf	03:42.175	03:44.843	2.668
865	15.41.04.113.eaf	03:46.373	03:47.020	0.647
866	15.41.04.113.eaf	03:50.770	03:51.923	1.153
867	15.41.04.113.eaf	03:58.569	03:59.192	0.623
868	15.41.04.113.eaf	04:02.518	04:03.543	1.025
869	15.41.04.113.eaf	04:03.821	04:05.025	1.204
870	15.41.04.113.eaf	04:05.025	04:05.714	0.689
871	15.41.04.113.eaf	04:05.714	04:06.328	0.614
872	15.41.04.113.eaf	04:07.575	04:08.920	1.345
873	15.41.04.113.eaf	04:09.797	04:11.201	1.404
874	15.41.04.113.eaf	04:11.201	04:14.226	3.025
875	15.41.04.113.eaf	04:14.226	04:14.947	0.721
876	15.41.04.113.eaf	04:14.947	04:16.191	1.244
877	15.41.04.113.eaf	04:18.717	04:20.286	1.569
878	15.41.04.113.eaf	04:21.365	04:21.963	0.598
879	15.41.04.113.eaf	04:21.963	04:22.838	0.875
880	15.41.04.113.eaf	04:31.740	04:32.471	0.731
881	15.41.04.113.eaf	04:32.471	04:33.307	0.836
882	15.41.04.113.eaf	04:33.307	04:34.098	0.791
883	15.41.04.113.eaf	04:34.775	04:35.124	0.349
884	15.41.04.113.eaf	04:35.124	04:37.469	2.345
885	15.41.04.113.eaf	04:37.760	04:38.546	0.786
886	15.41.04.113.eaf	04:39.590	04:40.683	1.093
887	15.41.04.113.eaf	04:45.136	04:45.973	0.837
888	15.41.04.113.eaf	04:46.534	04:47.732	1.198
889	15.41.04.113.eaf	04:47.732	04:48.753	1.021
890	15.41.04.113.eaf	04:50.100	04:51.454	1.354
891	15.41.04.113.eaf	04:52.221	04:53.775	1.554
892	16.11.09.878.eaf	00:10.037	00:10.173	0.136
893	16.11.09.878.eaf	00:10.173	00:11.546	1.373
894	16.11.09.878.eaf	00:11.546	00:13.056	1.51
895	16.11.09.878.eaf	00:16.261	00:17.481	1.22
896	16.11.09.878.eaf	00:17.864	00:18.651	0.787
897	16.11.09.878.eaf	00:19.094	00:19.746	0.652
898	16.11.09.878.eaf	00:20.425	00:21.311	0.886
899	16.11.09.878.eaf	00:21.311	00:22.777	1.466
900	16.11.09.878.eaf	00:23.403	00:24.613	1.21
901	16.11.09.878.eaf	00:27.948	00:29.649	1.701
902	16.11.09.878.eaf	00:30.827	00:31.668	0.841
903	16.11.09.878.eaf	00:34.206	00:35.314	1.108
904	16.11.09.878.eaf	00:37.078	00:39.119	2.041
905	16.11.09.878.eaf	00:40.784	00:41.750	0.966
906	16.11.09.878.eaf	00:45.954	00:47.119	1.165
907	16.11.09.878.eaf	00:57.623	00:58.655	1.032
908	16.11.09.878.eaf	00:58.655	00:59.664	1.009
909	16.11.09.878.eaf	00:59.664	01:01.381	1.717
910	16.11.09.878.eaf	01:05.517	01:06.366	0.849
911	16.11.09.878.eaf	01:08.628	01:10.518	1.89
912	16.11.09.878.eaf	01:11.858	01:12.560	0.702
913	16.11.09.878.eaf	01:12.560	01:13.629	1.069
914	16.11.09.878.eaf	01:13.853	01:14.596	0.743
915	16.11.09.878.eaf	01:14.596	01:16.023	1.427
916	16.11.09.878.eaf	01:16.750	01:17.785	1.035
917	16.11.09.878.eaf	01:18.982	01:19.700	0.718
918	16.11.09.878.eaf	01:19.700	01:21.053	1.353
919	16.11.09.878.eaf	01:25.742	01:26.768	1.026
920	16.26.56.109.eaf	00:33.826	00:34.341	0.515
921	16.26.56.109.eaf	00:35.293	00:36.084	0.791
922	16.26.56.109.eaf	00:37.440	00:37.902	0.462
923	16.26.56.109.eaf	00:50.741	00:54.343	3.602
924	16.36.00.692.eaf	00:04.692	00:05.341	0.649
925	16.36.00.692.eaf	00:05.833	00:07.049	1.216
926	16.36.00.692.eaf	00:09.688	00:10.404	0.716
927	16.36.00.692.eaf	00:10.609	00:12.255	1.646
928	16.36.00.692.eaf	00:14.827	00:16.830	2.003
929	16.36.00.692.eaf	00:22.009	00:23.712	1.703
930	16.36.00.692.eaf	00:26.324	00:27.513	1.189
931	16.36.00.692.eaf	00:28.013	00:28.812	0.799
932	16.36.00.692.eaf	00:30.031	00:30.605	0.574
933	16.36.00.692.eaf	00:30.605	00:31.239	0.634
934	16.36.00.692.eaf	00:31.239	00:32.868	1.629
935	16.36.00.692.eaf	00:33.721	00:35.102	1.381
936	16.36.00.692.eaf	00:36.854	00:38.036	1.182
937	16.36.00.692.eaf	00:38.036	00:38.941	0.905
938	16.36.00.692.eaf	00:38.941	00:39.951	1.01
939	16.36.00.692.eaf	00:44.635	00:45.139	0.504
940	16.36.00.692.eaf	00:45.139	00:45.664	0.525
941	16.36.00.692.eaf	00:45.664	00:46.868	1.204
942	16.36.00.692.eaf	00:48.880	00:49.818	0.938
943	16.36.00.692.eaf	00:50.348	00:50.722	0.374

gphr	recording	gphr_start (mm:ss.f)	gphr_end (mm:ss.f)	gphr_dur (s.f)
944	16.36.00.692.eaf	00:50.722	00:51.394	0.672
945	16.36.00.692.eaf	00:52.297	00:53.721	1.424
946	16.36.00.692.eaf	00:55.648	00:56.836	1.188
947	16.36.00.692.eaf	01:04.823	01:06.174	1.351
948	16.36.00.692.eaf	01:09.042	01:10.825	1.783
949	16.36.00.692.eaf	01:16.166	01:17.153	0.987
950	16.36.00.692.eaf	01:18.172	01:19.399	1.227
951	16.36.00.692.eaf	01:23.998	01:25.118	1.12
952	16.36.00.692.eaf	01:27.278	01:27.647	0.369
953	16.36.00.692.eaf	01:27.647	01:28.956	1.309
954	16.36.00.692.eaf	01:30.120	01:31.435	1.315
955	16.36.00.692.eaf	01:34.390	01:35.761	1.371
956	16.36.00.692.eaf	01:35.761	01:36.883	1.122
957	16.36.00.692.eaf	01:38.720	01:39.404	0.684
958	16.36.00.692.eaf	01:39.404	01:40.116	0.712
959	16.36.00.692.eaf	01:40.116	01:41.812	1.696
960	16.36.00.692.eaf	01:44.161	01:45.020	0.859
961	16.36.00.692.eaf	01:45.157	01:45.999	0.842
962	16.36.00.692.eaf	01:46.921	01:47.609	0.688
963	16.36.00.692.eaf	01:47.609	01:48.997	1.388
964	16.36.00.692.eaf	01:49.982	01:51.021	1.039
965	16.36.00.692.eaf	01:51.021	01:52.084	1.063
966	16.36.00.692.eaf	01:52.745	01:53.826	1.081
967	16.36.00.692.eaf	01:54.695	01:55.538	0.843
968	16.36.00.692.eaf	01:56.107	01:57.310	1.203
969	16.36.00.692.eaf	02:00.039	02:00.640	0.601
970	16.36.00.692.eaf	02:13.345	02:13.747	0.402
971	16.36.00.692.eaf	02:13.747	02:15.352	1.605
972	16.36.00.692.eaf	02:15.352	02:16.236	0.884
973	16.36.00.692.eaf	02:16.236	02:17.927	1.691
974	16.36.00.692.eaf	02:18.522	02:19.491	0.969
975	16.36.00.692.eaf	02:19.491	02:22.078	2.587
976	16.36.00.692.eaf	02:22.716	02:25.026	2.31
977	16.36.00.692.eaf	02:26.204	02:27.231	1.027
978	16.36.00.692.eaf	02:27.231	02:27.938	0.707
979	16.36.00.692.eaf	02:28.737	02:29.562	0.825
980	16.36.00.692.eaf	02:29.788	02:30.894	1.106
981	16.36.00.692.eaf	02:32.158	02:33.707	1.549
982	16.36.00.692.eaf	02:33.909	02:34.680	0.771
983	16.36.00.692.eaf	02:34.680	02:35.712	1.032
984	16.36.00.692.eaf	02:36.361	02:37.663	1.302
985	16.36.00.692.eaf	02:37.663	02:38.742	1.079
986	16.36.00.692.eaf	02:40.670	02:40.995	0.325
987	16.36.00.692.eaf	02:40.995	02:42.300	1.305
988	16.36.00.692.eaf	02:45.312	02:45.927	0.615
989	16.36.00.692.eaf	02:45.927	02:46.976	1.049
990	16.36.00.692.eaf	02:47.522	02:48.009	0.487
991	16.36.00.692.eaf	02:48.204	02:49.592	1.388
992	16.36.00.692.eaf	02:49.592	02:50.124	0.532
993	16.36.00.692.eaf	02:50.124	02:51.047	0.923
994	16.36.00.692.eaf	02:54.153	02:55.047	0.894
995	16.36.00.692.eaf	02:56.825	02:57.518	0.693
996	16.36.00.692.eaf	02:59.204	02:59.892	0.688
997	16.36.00.692.eaf	03:01.094	03:03.231	2.137
998	16.36.00.692.eaf	03:03.821	03:04.583	0.762
999	16.36.00.692.eaf	03:09.078	03:10.325	1.247
1000	16.36.00.692.eaf	03:11.155	03:12.078	0.923
1001	16.36.00.692.eaf	03:12.533	03:18.017	5.484
1002	16.36.00.692.eaf	03:18.143	03:18.852	0.709
1003	16.36.00.692.eaf	03:21.351	03:22.144	0.793
1004	16.36.00.692.eaf	03:32.706	03:33.085	0.379
1005	16.36.00.692.eaf	03:33.085	03:34.647	1.562
1006	16.36.00.692.eaf	03:35.454	03:36.289	0.835
1007	16.36.00.692.eaf	03:36.289	03:37.564	1.275
1008	16.36.00.692.eaf	03:38.974	03:39.578	0.604
1009	16.36.00.692.eaf	03:39.578	03:40.045	0.467
1010	16.36.00.692.eaf	03:40.045	03:40.780	0.735
1011	16.36.00.692.eaf	03:40.883	03:41.496	0.613
1012	16.36.00.692.eaf	03:41.496	03:42.275	0.779
1013	16.36.00.692.eaf	03:44.808	03:45.515	0.707
1014	16.36.00.692.eaf	03:46.194	03:48.461	2.267
1015	16.36.00.692.eaf	03:49.975	03:50.473	0.498
1016	16.36.00.692.eaf	03:50.984	03:51.657	0.673
1017	16.36.00.692.eaf	03:52.333	03:54.004	1.671
1018	16.36.00.692.eaf	03:57.098	03:57.917	0.819
1019	16.36.00.692.eaf	04:00.720	04:01.720	1.0
1020	16.36.00.692.eaf	04:01.720	04:02.634	0.914
1021	16.36.00.692.eaf	04:03.104	04:04.374	1.27
1022	16.36.00.692.eaf	04:04.824	04:05.663	0.839
1023	16.36.00.692.eaf	04:08.365	04:09.403	1.038
1024	16.36.00.692.eaf	04:10.593	04:11.715	1.122
1025	16.36.00.692.eaf	04:12.881	04:14.866	1.985
1026	16.36.00.692.eaf	04:14.866	04:15.598	0.732
1027	16.36.00.692.eaf	04:15.598	04:17.018	1.42
1028	16.36.00.692.eaf	04:17.773	04:18.849	1.076
1029	16.36.00.692.eaf	04:21.230	04:21.754	0.524

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
1030	16.36.00.692.eaf	04:26.364	04:27.010	0.646
1031	16.36.00.692.eaf	04:27.010	04:28.795	1.785
1032	16.36.00.692.eaf	04:30.688	04:31.457	0.769
1033	16.36.00.692.eaf	04:31.457	04:32.165	0.708
1034	16.36.00.692.eaf	04:32.165	04:32.913	0.748
1035	16.36.00.692.eaf	04:33.089	04:34.218	1.129
1036	16.36.00.692.eaf	04:35.551	04:36.384	0.833
1037	16.36.00.692.eaf	04:36.384	04:38.322	1.938
1038	16.36.00.692.eaf	04:38.954	04:39.929	0.975
1039	16.36.00.692.eaf	04:40.795	04:42.117	1.322
1040	16.36.00.692.eaf	04:43.059	04:45.298	2.239
1041	16.36.00.692.eaf	04:45.298	04:47.171	1.873
1042	16.36.00.692.eaf	04:47.477	04:48.361	0.884
1043	16.36.00.692.eaf	04:48.361	04:49.717	1.356
1044	16.36.00.692.eaf	04:49.717	04:51.098	1.381
1045	16.36.00.692.eaf	04:52.538	04:53.850	1.312
1046	16.36.00.692.eaf	05:00.145	05:02.448	2.303
1047	16.36.00.692.eaf	05:02.448	05:05.212	2.764
1048	16.36.00.692.eaf	05:06.596	05:07.832	1.236
1049	16.36.00.692.eaf	05:09.843	05:11.222	1.379
1050	16.36.00.692.eaf	05:11.534	05:12.963	1.429
1051	16.36.00.692.eaf	05:13.474	05:14.582	1.108
1052	16.36.00.692.eaf	05:14.696	05:15.384	0.688
1053	16.36.00.692.eaf	05:15.723	05:16.445	0.722
1054	16.36.00.692.eaf	05:18.318	05:21.149	2.831
1055	16.36.00.692.eaf	05:21.149	05:21.818	0.669
1056	16.36.00.692.eaf	05:22.955	05:23.432	0.477
1057	16.36.00.692.eaf	05:25.863	05:26.975	1.112
1058	16.36.00.692.eaf	05:28.445	05:29.104	0.659
1059	16.36.00.692.eaf	05:29.104	05:31.611	2.507
1060	16.36.00.692.eaf	05:33.288	05:34.031	0.743
1061	16.36.00.692.eaf	05:34.909	05:35.767	0.858
1062	16.36.00.692.eaf	05:36.041	05:36.843	0.802
1063	16.36.00.692.eaf	05:38.401	05:39.384	0.983
1064	16.43.50.013.eaf	00:10.329	00:11.428	1.099
1065	16.43.50.013.eaf	00:11.428	00:12.201	0.773
1066	16.43.50.013.eaf	00:12.201	00:13.185	0.984
1067	16.43.50.013.eaf	00:13.185	00:13.767	0.582
1068	16.43.50.013.eaf	00:13.767	00:14.236	0.469
1069	16.43.50.013.eaf	00:14.824	00:15.421	0.597
1070	16.43.50.013.eaf	00:15.421	00:16.032	0.611
1071	16.43.50.013.eaf	00:16.252	00:17.197	0.945
1072	16.43.50.013.eaf	00:17.197	00:18.010	0.813

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
1073	16.43.50.013.eaf	00:18.010	00:18.940	0.93
1074	16.43.50.013.eaf	00:19.066	00:20.354	1.288
1075	16.43.50.013.eaf	00:21.081	00:21.677	0.596
1076	16.43.50.013.eaf	00:26.893	00:27.568	0.675
1077	16.43.50.013.eaf	00:31.977	00:33.310	1.333
1078	16.43.50.013.eaf	00:35.457	00:35.891	0.434
1079	16.43.50.013.eaf	00:37.397	00:37.967	0.57
1080	16.43.50.013.eaf	00:42.749	00:43.045	0.296
1081	16.43.50.013.eaf	00:43.045	00:43.589	0.544
1082	16.43.50.013.eaf	00:46.888	00:47.515	0.627
1083	16.43.50.013.eaf	00:50.927	00:51.927	1.0
1084	16.43.50.013.eaf	00:54.083	00:54.672	0.589
1085	16.43.50.013.eaf	00:56.672	00:57.302	0.63
1086	16.43.50.013.eaf	01:00.104	01:00.938	0.834
1087	16.43.50.013.eaf	01:01.522	01:01.947	0.425
1088	16.43.50.013.eaf	01:05.355	01:06.237	0.882
1089	16.43.50.013.eaf	01:17.195	01:18.011	0.816
1090	16.43.50.013.eaf	01:18.698	01:19.428	0.73
1091	16.43.50.013.eaf	01:19.676	01:20.053	0.377
1092	16.43.50.013.eaf	01:20.492	01:21.107	0.615
1093	16.43.50.013.eaf	01:21.441	01:22.062	0.621
1094	16.43.50.013.eaf	01:26.200	01:26.839	0.639
1095	16.43.50.013.eaf	01:26.839	01:28.785	1.946
1096	16.43.50.013.eaf	01:30.391	01:30.955	0.564
1097	16.43.50.013.eaf	01:30.955	01:31.740	0.785
1098	16.43.50.013.eaf	01:32.674	01:33.360	0.686
1099	16.43.50.013.eaf	01:34.076	01:35.158	1.082
1100	16.51.31.649.eaf	00:12.963	00:14.091	1.128
1101	16.51.31.649.eaf	00:16.284	00:16.943	0.659
1102	16.51.31.649.eaf	00:18.363	00:19.214	0.851
1103	16.51.31.649.eaf	00:22.107	00:22.636	0.529
1104	16.51.31.649.eaf	00:31.391	00:31.984	0.593
1105	16.51.31.649.eaf	00:33.469	00:34.256	0.787
1106	16.51.31.649.eaf	00:41.988	00:43.139	1.151
1107	16.51.31.649.eaf	01:00.517	01:01.496	0.979
1108	16.51.31.649.eaf	01:03.901	01:04.548	0.647
1109	16.51.31.649.eaf	01:37.287	01:39.165	1.878
1110	16.51.31.649.eaf	01:51.765	01:53.215	1.45
1111	16.51.31.649.eaf	01:57.932	01:58.886	0.954
1112	16.51.31.649.eaf	01:58.886	01:59.779	0.893
1113	16.51.31.649.eaf	02:02.749	02:03.458	0.709
1114	16.51.31.649.eaf	02:08.734	02:09.484	0.75
1115	16.51.31.649.eaf	02:16.586	02:17.365	0.779

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
1116	16.51.31.649.eaf	02:29.902	02:33.020	3.118
1117	16.51.31.649.eaf	02:43.698	02:44.777	1.079
1118	16.51.31.649.eaf	02:46.384	02:46.969	0.585
1119	16.51.31.649.eaf	02:48.227	02:48.732	0.505
1120	16.51.31.649.eaf	02:48.893	02:50.227	1.334
1121	16.51.31.649.eaf	02:50.966	02:51.857	0.891
1122	16.51.31.649.eaf	02:51.857	02:53.322	1.465
1123	16.51.31.649.eaf	03:05.833	03:06.183	0.35
1124	16.51.31.649.eaf	03:06.183	03:07.107	0.924
1125	16.51.31.649.eaf	03:20.071	03:20.996	0.925
1126	16.51.31.649.eaf	03:25.134	03:25.853	0.719
1127	16.51.31.649.eaf	03:30.621	03:31.938	1.317
1128	16.51.31.649.eaf	03:33.073	03:34.260	1.187
1129	16.51.31.649.eaf	03:36.662	03:37.321	0.659
1130	16.51.31.649.eaf	03:40.299	03:40.954	0.655
1131	16.51.31.649.eaf	03:40.954	03:42.059	1.105
1132	16.51.31.649.eaf	03:43.872	03:44.775	0.903
1133	16.51.31.649.eaf	04:01.586	04:02.876	1.29
1134	16.51.31.649.eaf	04:05.580	04:06.669	1.089
1135	16.51.31.649.eaf	04:06.669	04:07.784	1.115
1136	16.51.31.649.eaf	04:07.784	04:08.682	0.898
1137	16.51.31.649.eaf	04:08.682	04:09.381	0.699
1138	16.51.31.649.eaf	04:09.381	04:10.639	1.258
1139	16.51.31.649.eaf	04:25.634	04:26.161	0.527
1140	16.51.31.649.eaf	04:27.659	04:29.169	1.51
1141	16.51.31.649.eaf	04:40.841	04:41.975	1.134
1142	16.51.31.649.eaf	04:41.975	04:42.674	0.699
1143	16.51.31.649.eaf	04:44.717	04:45.650	0.933
1144	16.51.31.649.eaf	04:46.053	04:46.971	0.918
1145	16.51.31.649.eaf	04:49.336	04:49.858	0.522
1146	16.51.31.649.eaf	04:50.350	04:51.220	0.87
1147	16.51.31.649.eaf	04:57.124	04:58.143	1.019
1148	16.51.31.649.eaf	04:58.143	04:58.758	0.615
1149	16.51.31.649.eaf	04:59.811	05:00.482	0.671
1150	16.51.31.649.eaf	05:01.344	05:01.965	0.621
1151	16.51.31.649.eaf	05:02.417	05:03.136	0.719
1152	16.51.31.649.eaf	05:08.388	05:09.128	0.74
1153	16.51.31.649.eaf	05:11.399	05:12.409	1.01
1154	16.51.31.649.eaf	05:17.303	05:18.309	1.006
1155	16.51.31.649.eaf	05:26.399	05:27.444	1.045
1156	16.51.31.649.eaf	05:28.864	05:30.044	1.18
1157	16.51.31.649.eaf	05:31.125	05:32.074	0.949
1158	16.51.31.649.eaf	05:32.074	05:33.597	1.523
1159	16.51.31.649.eaf	05:36.662	05:37.185	0.523
1160	16.51.31.649.eaf	05:37.185	05:38.055	0.87
1161	16.51.31.649.eaf	05:42.616	05:43.256	0.64
1162	16.51.31.649.eaf	05:43.397	05:44.116	0.719
1163	16.51.31.649.eaf	05:47.795	05:48.586	0.791
1164	16.51.31.649.eaf	05:50.627	05:51.487	0.86
1165	16.51.31.649.eaf	05:58.019	05:59.602	1.583
1166	16.51.31.649.eaf	06:00.775	06:02.123	1.348
1167	16.51.31.649.eaf	06:07.418	06:08.909	1.491
1168	16.51.31.649.eaf	06:08.909	06:10.274	1.365
1169	16.51.31.649.eaf	06:11.536	06:12.325	0.789
1170	16.51.31.649.eaf	06:12.873	06:13.724	0.851
1171	16.51.31.649.eaf	06:13.971	06:14.673	0.702
1172	16.51.31.649.eaf	06:16.738	06:17.511	0.773
1173	17.29.17.354.eaf	00:18.652	00:19.222	0.57
1174	17.29.17.354.eaf	00:25.750	00:26.473	0.723
1175	17.29.17.354.eaf	00:26.473	00:26.952	0.479
1176	17.29.17.354.eaf	00:30.108	00:31.081	0.973
1177	17.29.17.354.eaf	00:58.868	01:01.089	2.221
1178	17.29.17.354.eaf	01:05.075	01:07.320	2.245
1179	17.29.17.354.eaf	01:10.839	01:12.009	1.17
1180	17.29.17.354.eaf	01:14.334	01:15.074	0.74
1181	17.29.17.354.eaf	01:16.301	01:18.160	1.859
1182	17.29.17.354.eaf	01:21.929	01:22.923	0.994
1183	17.29.17.354.eaf	01:29.899	01:30.289	0.39
1184	17.29.17.354.eaf	01:35.589	01:37.021	1.432
1185	17.29.17.354.eaf	01:49.004	01:49.989	0.985
1186	17.29.17.354.eaf	01:57.189	01:57.956	0.767
1187	17.29.17.354.eaf	01:59.263	02:00.215	0.952
1188	17.29.17.354.eaf	02:01.208	02:01.907	0.699
1189	17.29.17.354.eaf	02:02.714	02:04.177	1.463
1190	17.29.17.354.eaf	02:07.729	02:08.448	0.719
1191	17.29.17.354.eaf	02:08.934	02:09.679	0.745
1192	17.29.17.354.eaf	02:12.665	02:13.767	1.102
1193	17.29.17.354.eaf	02:17.188	02:17.440	0.252
1194	17.29.17.354.eaf	02:20.202	02:21.008	0.806
1195	17.29.17.354.eaf	02:23.081	02:23.786	0.705
1196	17.29.17.354.eaf	02:24.244	02:25.097	0.853
1197	17.29.17.354.eaf	02:25.328	02:26.093	0.765
1198	17.29.17.354.eaf	02:26.296	02:26.851	0.555
1199	17.29.17.354.eaf	02:26.851	02:27.750	0.899
1200	17.29.17.354.eaf	02:28.102	02:29.041	0.939
1201	17.29.17.354.eaf	02:43.050	02:43.814	0.764

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
1202	17.29.17.354.eaf	02:56.312	02:57.232	0.92
1203	17.29.17.354.eaf	02:58.054	02:59.193	1.139
1204	17.29.17.354.eaf	03:00.656	03:01.451	0.795
1205	17.29.17.354.eaf	03:08.502	03:09.323	0.821
1206	17.29.17.354.eaf	03:11.533	03:12.514	0.981
1207	17.29.17.354.eaf	03:13.523	03:14.688	1.165
1208	17.29.17.354.eaf	03:15.785	03:16.542	0.757
1209	17.29.17.354.eaf	03:17.703	03:19.276	1.573
1210	17.29.17.354.eaf	03:23.504	03:24.514	1.01
1211	17.29.17.354.eaf	03:25.520	03:27.991	2.471
1212	17.29.17.354.eaf	03:28.492	03:29.314	0.822
1213	17.29.17.354.eaf	03:29.669	03:30.962	1.293
1214	17.29.17.354.eaf	03:32.732	03:33.825	1.093
1215	17.29.17.354.eaf	03:41.860	03:43.495	1.635
1216	17.29.17.354.eaf	03:43.495	03:44.361	0.866
1217	17.29.17.354.eaf	03:46.282	03:47.145	0.863
1218	17.29.17.354.eaf	03:47.145	03:47.999	0.854
1219	17.29.17.354.eaf	03:47.999	03:48.565	0.566
1220	17.29.17.354.eaf	03:49.763	03:50.764	1.001
1221	17.29.17.354.eaf	03:52.192	03:52.813	0.621
1222	17.29.17.354.eaf	03:52.813	03:53.613	0.8
1223	17.29.17.354.eaf	03:56.063	03:57.579	1.516
1224	17.29.17.354.eaf	04:09.860	04:10.658	0.798
1225	17.29.17.354.eaf	04:10.658	04:11.911	1.253
1226	17.29.17.354.eaf	04:18.038	04:18.734	0.696
1227	17.29.17.354.eaf	04:24.132	04:25.278	1.146
1228	17.29.17.354.eaf	04:25.278	04:26.080	0.802
1229	17.29.17.354.eaf	04:28.230	04:28.523	0.293
1230	17.29.17.354.eaf	04:29.091	04:29.782	0.691
1231	17.29.17.354.eaf	04:32.290	04:33.003	0.713
1232	17.29.17.354.eaf	04:36.572	04:37.843	1.271
1233	17.29.17.354.eaf	04:42.080	04:42.989	0.909
1234	14.27.42.306.eaf	07:22.558	07:23.065	0.507
1235	14.27.42.306.eaf	00:00.000	00:01.386	1.386
1236	14.27.42.306.eaf	00:09.670	00:10.746	1.076
1237	14.27.42.306.eaf	00:15.854	00:16.537	0.683
1238	14.27.42.306.eaf	00:24.639	00:25.073	0.434
1239	14.27.42.306.eaf	00:31.605	00:32.220	0.615
1240	14.27.42.306.eaf	00:35.244	00:36.507	1.263
1241	14.27.42.306.eaf	00:45.054	00:46.454	1.4
1242	14.27.42.306.eaf	00:46.971	00:48.008	1.037
1243	14.27.42.306.eaf	00:52.817	00:53.739	0.922
1244	14.27.42.306.eaf	00:57.217	00:58.760	1.543
1245	14.27.42.306.eaf	00:59.867	01:00.990	1.123
1246	14.27.42.306.eaf	01:08.565	01:09.330	0.765
1247	14.27.42.306.eaf	01:12.027	01:12.963	0.936
1248	14.27.42.306.eaf	01:35.180	01:36.360	1.18
1249	14.27.42.306.eaf	01:51.550	01:54.560	3.01
1250	14.27.42.306.eaf	02:05.775	02:06.775	1.0
1251	14.27.42.306.eaf	02:09.823	02:10.604	0.781
1252	14.27.42.306.eaf	02:12.360	02:13.228	0.868
1253	14.27.42.306.eaf	02:35.453	02:36.492	1.039
1254	14.27.42.306.eaf	02:55.959	02:57.610	1.651
1255	14.27.42.306.eaf	02:57.618	02:59.037	1.419
1256	14.27.42.306.eaf	02:59.193	02:59.705	0.512
1257	14.27.42.306.eaf	02:59.706	03:00.408	0.702
1258	14.27.42.306.eaf	03:00.408	03:01.227	0.819
1259	14.27.42.306.eaf	03:01.227	03:02.856	1.629
1260	14.27.42.306.eaf	03:09.242	03:11.066	1.824
1261	14.27.42.306.eaf	03:11.066	03:11.608	0.542
1262	14.27.42.306.eaf	03:11.608	03:12.203	0.595
1263	14.27.42.306.eaf	03:13.880	03:15.630	1.75
1264	14.27.42.306.eaf	03:17.331	03:18.486	1.155
1265	14.27.42.306.eaf	03:22.481	03:23.476	0.995
1266	14.27.42.306.eaf	03:25.457	03:25.803	0.346
1267	14.27.42.306.eaf	03:28.180	03:28.922	0.742
1268	14.27.42.306.eaf	03:31.464	03:32.093	0.629
1269	14.27.42.306.eaf	03:37.491	03:38.225	0.734
1270	14.27.42.306.eaf	03:39.601	03:40.679	1.078
1271	14.27.42.306.eaf	03:57.826	03:59.055	1.229
1272	14.27.42.306.eaf	04:15.118	04:15.996	0.878
1273	14.27.42.306.eaf	04:18.494	04:19.554	1.06
1274	14.27.42.306.eaf	04:21.422	04:22.490	1.068
1275	14.27.42.306.eaf	04:22.778	04:23.768	0.99
1276	14.27.42.306.eaf	04:24.363	04:24.910	0.547
1277	14.27.42.306.eaf	04:33.675	04:34.266	0.591
1278	14.27.42.306.eaf	04:38.188	04:39.040	0.852
1279	14.27.42.306.eaf	04:39.041	04:39.422	0.381
1280	14.27.42.306.eaf	04:39.422	04:40.217	0.795
1281	14.27.42.306.eaf	04:40.217	04:40.776	0.559
1282	14.27.42.306.eaf	04:42.630	04:44.620	1.99
1283	14.27.42.306.eaf	04:44.626	04:45.324	0.698
1284	14.27.42.306.eaf	04:52.548	04:53.300	0.752
1285	14.27.42.306.eaf	04:58.612	04:59.934	1.322
1286	14.27.42.306.eaf	05:00.383	05:01.183	0.8
1287	14.27.42.306.eaf	05:01.188	05:01.773	0.585

gphr	recording	gphr_start	gphr_end	gphr_dur
		(mm:ss.f)	(mm:ss.f)	(s.f)
1288	14.27.42.306.eaf	05:15.080	05:16.422	1.342
1289	14.27.42.306.eaf	05:16.422	05:18.358	1.936
1290	14.27.42.306.eaf	05:18.358	05:18.734	0.376
1291	14.27.42.306.eaf	05:23.851	05:24.397	0.546
1292	14.27.42.306.eaf	05:24.397	05:25.017	0.62
1293	14.27.42.306.eaf	05:48.125	05:48.671	0.546
1294	14.27.42.306.eaf	05:49.559	05:50.056	0.497
1295	14.27.42.306.eaf	05:53.251	05:54.036	0.785
1296	14.27.42.306.eaf	05:54.164	05:55.193	1.029
1297	14.27.42.306.eaf	05:55.193	05:55.534	0.341
1298	14.27.42.306.eaf	05:56.460	05:56.870	0.41
1299	14.27.42.306.eaf	05:56.870	05:57.378	0.508
1300	14.27.42.306.eaf	05:59.059	05:59.821	0.762
1301	14.27.42.306.eaf	06:00.382	06:04.109	3.727
1302	14.27.42.306.eaf	06:04.109	06:05.201	1.092
1303	14.27.42.306.eaf	06:07.332	06:09.533	2.201
1304	14.27.42.306.eaf	06:11.123	06:12.299	1.176
1305	14.27.42.306.eaf	06:15.196	06:16.025	0.829
1306	14.27.42.306.eaf	06:16.025	06:16.934	0.909
1307	14.27.42.306.eaf	06:16.942	06:17.415	0.473
1308	14.27.42.306.eaf	06:18.991	06:19.815	0.824
1309	14.27.42.306.eaf	06:20.400	06:22.330	1.93
1310	14.27.42.306.eaf	06:23.504	06:24.450	0.946
1311	14.27.42.306.eaf	06:24.450	06:24.963	0.513
1312	14.27.42.306.eaf	06:25.099	06:25.333	0.234
1313	14.27.42.306.eaf	06:45.738	06:47.050	1.312
1314	14.27.42.306.eaf	06:52.284	06:53.445	1.161
1315	14.27.42.306.eaf	07:05.347	07:06.869	1.522
1316	14.27.42.306.eaf	07:06.869	07:07.561	0.692
1317	14.27.42.306.eaf	07:08.415	07:08.873	0.458
1318	14.27.42.306.eaf	07:15.229	07:15.746	0.517
1319	14.27.42.306.eaf	07:15.746	07:16.063	0.317
1320	14.27.42.306.eaf	07:20.351	07:20.795	0.444
1321	14.27.42.306.eaf	07:21.330	07:22.381	1.051
1322	14.27.42.306.eaf	07:22.558	07:23.058	0.5
1323	14.27.42.306.eaf	07:23.058	07:23.873	0.815
1324	14.27.42.306.eaf	07:23.873	07:24.425	0.552
1325	14.27.42.306.eaf	07:25.981	07:26.826	0.845
1326	14.27.42.306.eaf	07:27.615	07:28.741	1.126
1327	14.27.42.306.eaf	07:30.390	07:33.150	2.76
1328	14.27.42.306.eaf	07:55.975	07:57.653	1.678
1329	14.27.42.306.eaf	08:12.780	08:13.698	0.918

11.2 Conceptual Affiliation Study

11.2.1 Participant form

Studie:

VP#

Alter: ________

Geschlecht: m □ w □

Muttersprache: Deutsch □ andere: □ (_________)

Sehhilfe: ja □ nein □

Sprachw. Hintergrund: ja □ nein □

11.2.2 Conceptual Affiliation form

Hörensehen?

Im Folgenden werden Sie kurze Audio- und Videoclips abspielen von denen jeweils die beiden gleich nummerierten zusammengehören. Während Sie in den Audioclips Aussagen in voller Länge hören, so enthalten die Videos je nur eine Geste.

Die Aufgabe:

1. Spielen Sie je die zusammengehörigen Clips so oft Sie mögen ab, jedoch niemals beide gleichzeitig.
2. Schreiben Sie den Teil der Aussage der von seiner Bedeutung her mit der Geste zusammenpasst in die unten stehende Tabelle. Bitte unterstreichen Sie diesen Teil zur Sicherheit auch noch.

Vielen Dank und viel Spaß!

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	

11.3 Perceptual Judgment Task

11.3.1 Instructions for participants

Anleitung (following a form on sociodemographic data)

Box 1:

Sie werden 24 kurze Clips aus Nacherzählungen eines Sylvester und Tweety Cartoons sehen. In den Clips sind die Gesichter verpixelt um die Anonymität der Gefilmten zu gewährleisten. Des Weiteren sind Ton und Bild der Videos zum Teil manipuliert. Ihre Aufgabe wird es sein, die Natürlichkeit der Clips in folgenden Abstufungen zu bewerten:

Box 2:

Völlig natürlich:

Sie nehmen den Clip als völlig normal wahr; Sie können keine Manipulation feststellen.

Ziemlich natürlich:

Sie nehmen den Clip als normal wahr, haben aber das Gefühl, dass etwas nicht ganz richtig ist.

Eher unnatürlich:

Sie nehmen den Clip als ein wenig unnormal wahr, aber es ist noch “okay”.

Völlig unnatürlich:

Sie nehmen den Clip als unnatürlich wahr; das Video wirkt für Sie unecht.

Box 3:

Spielen Sie die Clips ab, indem Sie „Play” klicken (Sie dürfen jedes Video mehrfach anschauen). Dann wählen Sie die Einstufung, die Ihrer Meinung nach den Clip am Besten beschreibt. Nachdem Sie eine Whal getroffen haben, klicken Sie „Next“ um zum nächsten Clip zu gelangen.

Wenn möglich benutzen Sie Kopfhörer und stellen Sie den Ton auf Zimmerlautstärke.

Zunächst kommt ein kurzer Testdurchlauf.

Button: Zum Testdurchlauf

11.3.2 Disclaimer after study had been completed by a participant

Vielen Dank für Ihre Teilnahme

Für Kommentare bin ich Ihnen sehr dankbar! Schreiben Sie einfach Ihre Meinung, Tipps und Wünsche in das untenstehende Feld und klicken Sie auf „Abschicken“. Vielen Dank.

[KOMMENTARFELD]

Button: Abschicken

Für Fragen und Ratschläge kontaktieren Sie bitteckirchhof AT uni-bielefeld.de.

11.3.3 Participant form for Lab Replication study

Studie:

VP#

Alter: ________

Geschlecht: m□ w□

Muttersprache: Deutsch □ andere: □ (_________)

Sehhilfe: ja □ nein □

Sprachw. Hintergrund: ja □ nein □

11.3.4 Selection sheet used in Lab Replication study

Bitte kreuzen Sie innerhalb der 3er-Gruppen jeweils den Clip an, der Ihnen am natürlichsten vorkommt.

Original:

1. Clip
   a) □
   b) □
   c) □

2. Clip
   a) □
   b) □
   c) □

3. Clip
   a) □
   b) □
   c) □

4. Clip
   a) □
   b) □
   c) □

5. Clip
   a) □
   b) □
   c) □

Gesicht verpixelt:

6. Clip
   a) □
   b) □
   c) □

7. Clip
   a) □
   b) □
   c) □

8. Clip
   a) □
   b) □
   c) □

9. Clip
   a) □
   b) □
   c) □

10. Clip
   a) □
   b) □
   c) □

Gesicht geblockt:

11. Clip
   a) □
   b) □
   c) □

12. Clip
   a) □
   b) □
   c) □

13. Clip
   a) □
   b) □
   c) □

14. Clip
   a) □
   b) □
   c) □

15. Clip
   a) □
   b) □
   c) □

11.4 Preference Task

11.4.1 Verbose transcript sheet for Study 5

Testphase

Stimulus	Inhalt
a	Hammer
b	Fingerschnippen

Experiment

Stimulus	Text
1	Und dann lockt er den Affen mit ner Banane.
2	Also, er macht den Käfig auf.
3	Und dann haste den Käfig da stehen.
4	Der rennt mit der Geldspendebüchse rum.
5	Draufhin wird er in die Luft katapultiert.
6	Klettert er erst rauf.
7	Will dann die Decke runter machen.
8	Und dann zieht er den Hut so hoch und dann erkennste, dass das ne Katze ist.
9	# Und haut ihm mit dem Regenschirm wieder fleißig übern Detz (Kopf).
10	Er ist in dem Regenrohr.
11	Sylvester fliegt sofort wieder raus.
12	Wo er dann als Roomboy verkleidet ist und anklopft. #
13	Ja, hier ist 'n Penny.
14	Er # klettert das Abwasserrohr hoch.
15	# Er schluckt die Kugel. #

# = Atmen oder sonstiges Geräusch.

11.4.2 **Verbose transcript sheet for Study 6**

Stimulus	**Inhalt**
1	Ein Buch wird zugeklappt.
2	Jemand klatscht.
3	Eine Gabel schlägt gegen ein Glas.
4	Eine Taste wird gedrückt.
5	Jemand klopft auf den Tisch.
6	Eine Flasche Sekt wird geöffnet.
PAUSE	
7	Wo er dann als Roomboy verkleidet ist und anklopft.
8	Und in dem Film geht es darum, dass Sylvester scharf auf (äh) den Vogel ist.
10	Klettert er erst rauf.
11	Und halt mit'm Fernglas durch die Gegend guckt.
12	Also, das war'n so zwei Hochhäuser au* an so ner Straße.
13	Klasse.
14	Und dann klingelt's.
PAUSE	
15	Draußen auf dem Schild steht (äh) „Hunde und Katzen verboten".
16	Will dann die Decke runter machen.
17	Also, das Telefon klingelt.
18	Warum auch immer #
19	Da drüben is ja endlich der leckere Vogel.
20	Dort geht es links zum Elevator.