Gesture, Prosody and Information Structure Synchronisation in Turkish

by

Olcay Türk

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Linguistics.

Victoria University of Wellington
2020

Abstract

This thesis investigates the synchronisation of gesture with prosody and information structure in Turkish. Speech and gesture have a close relationship in human communication, and they are tightly coordinated in production. Research has shown that gestural units are synchronised with prosodic units on a prominence-related micro level (i.e., pitch accents and gesture apexes), however these studies have largely been on a small number of languages of a similar prosodic type, not including Turkish, which has prominence-less prosodic words. It is known that both gesture and speech, through prosody, are hierarchically structured with nested phrasal constituents, but little is known about gesture-prosody synchronisation at this macro level. Even less is known about the timing relationships of gesture with information structure, which is also closely related to prosody. This thesis links gesture to information structure as a part of a three-way synchronisation relationship of gesture, prosody, and information structure.

Four participants were filmed in a narrative task, resulting in three hours of Turkish natural speech and gesture data. Selected sections were annotated for prosody using an adapted scheme for Turkish in the Autosegmental-Metrical framework, for information structure and for gesture. In total, there were over 20,000 annotations.

The synchronisation of gesture and speech units was systematically investigated at (1) the micro level, and (2) the macro level. At the micro level, this thesis asked which tones apexes are synchronised with, and whether this synchronisation depends on other prosodic and gestural features. It was found

that gesture apexes were synchronised with pitch accents if there were pitch accents in the relevant prosodic phrases; if not, they were synchronised with low tones that marked the onsets of prosodic words. This synchronisation pattern was largely consistent across different prosodic and gestural contexts, although it was tighter in the nuclear area. These findings confirm prominence as a constraint on synchronisation with evidence of pitch accent-apex synchronisation. The findings also extend our knowledge of the typology of micro-level synchronisation to cases where prominence is locally absent showing that micro-level synchronisation also obeys the prosodic hierarchy.

At the macro level, the aim was to find the prosodic anchor for single gesture phrases while testing for the possible effects of prosodic, gestural and information structural contexts. The findings showed that there was no one-to-one synchronisation of single gesture phrases with single intermediate or intonational phrases. However, it was found that gesture phrases often spanned over multiple consecutive intermediate phrases, and the synchronisation of gesture phrase boundaries was with the boundaries of these intermediate phrase groupings. In addition, these groupings tended to be combinations of pre-nuclear and nuclear intermediate phrases constituting the default focus position in Turkish. This synchronisation behaviour over the focal domain implied that there might be another speech element governing the speech-gesture synchronisation which also informs prosody, i.e. information structure.

Based on this finding and a few other associations in the earlier studies, it was hypothesised that gesture is also informed by and synchronised with information structure. In order to test this hypothesis, it was investigated whether gesture phrases were synchronised with information structural units, i.e., topics, foci and background. The findings showed that gesture phrases tended to accompany discursively prominent foci over topics and background. However, gesture phrases did not show perfect synchronisation with any of

these information structure units, although there was a systematic overlap in which foci and topics were contained within the duration of complete gesture phrases. Further investigations revealed that gesture phrase parts that bear apex related meaning provided a much better anchor for the synchronisation of information structure units. The preference for accompanying and synchronisation with the parts of gesture bearing gesturally prominent apical meaning also highlighted that prominence is a driving factor of synchronisation at the macro level as well as at the micro level.

This thesis has revealed pivotal links between gesture, prosody and information structure through a systematic investigation of synchronisation of these structures. The implications of these links have also been discussed within the thesis, and a model of speech and gesture production integrating synchronisation has been proposed. Overall, the thesis contributes to a deeper understanding of speech and gesture production, explaining how these interact during natural speech.

Contents

List of Figures

List of Tables

1

Introduction

People from all linguistic and cultural backgrounds spontaneously move their hands and other body parts when they speak (Feyereisen & De Lannoy, 1991). This behaviour surfaces very early in children even before the emergence of language (Acredolo & Goodwyn, 1988; Goldin-Meadow & Butcher, 2003; Özçalışkan & Goldin-Meadow, 2005; Esteve-Gibert & Prieto, 2014). It is innate and not solely learned by watching others since it is also observed for congenitally blind interlocutors (Iverson & Goldin-Meadow, 1997; Iverson, Tencer, Lany, & Goldin-Meadow, 2000; Iverson & Goldin-Meadow, 2001). These spontaneous bodily movements that accompany speech, i.e., gestures, are not randomly produced movements - they have been shown to serve various functions such as complementing or supplementing information delivered in speech (Goldin-Meadow, Alibali, & Church, 1993; Goldin-Meadow, 1999; Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001), regulating conversation (Bavelas, Chovil, Lawrie, & Wade, 1992; Bavelas, Gerwing, Sutton, & Prevost, 2008; Bavelas, Gerwing, & Healing, 2014) as well as assisting speech by helping word retrieval (Butterworth & Beattie, 1978; Morrel-Samuels &

Krauss, 1992; Aboudan & Beattie, 1996).

All these observed interactions between speech and gesture have led to the claim that speech and gesture production processes must be tightly linked (McNeill, 1992, 2005; Kendon, 2004). Consequently, researchers have tried to explain these empirical findings of gesture-speech interactions with theories of gesture production integrated into speech production. As a result, several unified psycholinguistic models of speech-gesture production have been proposed (McNeill & Duncan, 2000; Kita, 2000; Krauss, Chen, & Gottesman, 2000; De Ruiter, 2000; Kita & Özyürek, 2003; Hostetter & Alibali, 2008). These models have explained where in their productions speech and gesture interact, largely focusing on their interaction at the earliest stage, i.e., a common origin (or common source) (McNeill, 1992). These explanations have relied on observed interactions between gesture and speech to pinpoint their common origin. However, any potential links between speech and gesture productions apart from their common origin, i.e., interactions at the later stages of productions such as interactions with linguistic structure, have been largely ignored.

The reliance on such interactions for linking speech and gesture productions has imposed some limitations on these models. As we will see in Section 2.1.1, a great deal is known about *how* people gesture (i.e., form). Research has also revealed much on *why* people gesture (i.e., function). However, not much is known about *when* people gesture in relation to speech (i.e., synchronisation) - there are still a lot of open questions about the exact nature of synchronisation of gesture and its co-speech. Existing studies on synchronisation have also not addressed what their findings imply for the integration of speech and gesture productions in general. Accordingly, synchronisation has not been accounted for in these production models sufficiently. These models either have not explained synchronisation at all or only commented on how basic observations of co-occurrence can be explained based on already existing assumptions within the models. Systematic synchronisation of speech and gesture parts encoded separately at the later stages of their respective productions implies links between speech and gesture throughout the production processes, not just at the earlier stages. A working theory of gesture production should account for all types of interactions, including temporal interactions (not just functional ones) between these modalities. Therefore, synchronisation with all its aspects, needs to be taken into account by unified production models.

We shall see in Sections 2.5 and 2.6 that there have been a number of studies that have explored synchronisation although none have interpreted their findings within the frame of earlier production models. Moreover, early proposals about speech-gesture synchronisation adopted a more general understanding of the term. Synchronisation was used to explain why speech and co-occurring gesture carry the same semantic and pragmatic content without detailing the concept of co-occurrence itself (McNeill, 1992). Therefore, synchronisation was seen more as the sharing of certain features rather than temporal co-occurrence. Other studies have focused more on defining actual timewise synchronisation and investigated timing relationships between gestures and their lexical affiliates (Butterworth & Beattie, 1978; Morrel-Samuels & Krauss, 1992; Chui, 2005; Ferré, 2010). These studies have reported a common observation that gesture precedes its lexical affiliate. This precedence of gesture has also been interpreted as synchronisation, which has found some explanation in the proposed production models (McNeill & Duncan, 2000; Krauss et al., 2000; De Ruiter, 2000).

1.1 Synchronisation

Despite various observations of some form of synchronisation, what exactly constitutes synchronisation, and what units within speech and gesture are synchronised has remained largely undefined. The main approach in earlier studies have been that synchronisation was treated as a qualitatively determined overlapping phenomenon, describing a more general understanding of synchronisation where certain speech gesture parts are more likely to cooccur rather than not (Sections 2.5.1 and 2.5.2). For example, if two units overlapped within the same domain such as a syllable, word or whole clause, then these units were considered synchronised regardless of the actual time distance between them. The present study's view on synchronisation differs greatly from this approach. Within this study, synchronisation is defined as the systematic co-occurrence of two units in time. Two synchronised units can take place at the same time or at fixed distances from each other as long as there is consistency. The present study uses actual time measurements to reveal such consistent synchronisation behaviour while also testing these measurements against a phonologically meaningful and statistically tested synchronisation criterion which is explained in detail in Chapter 3.

In terms of which component of speech is synchronised with gesture, there has been some consensus in the literature. Expressive, structural and temporal connections between prosody and gesture were noticed quite early on. Bloomfield (1933) associated gesture with intonation on the basis of their expressive similarity, highlighting that gestural movements are used in parallel with pitch features, for instance when humans talk harshly, sneeringly, cheerfully and so on. Similarly, Bolinger (1983) famously stated that "intonation belongs more with gesture than with grammar" (p.157) and that pitch features and gestural movements contribute to a discourse in the same manner - pitch and body parts move in harmony, responding to emotional tension and relaxation. Early gesture and prosody interaction was also observed in the regulation of interaction (Starkey & Fiske, 1977), and in studies concerning rhythmicity (Condon, 1976).

On the structural and temporal coordination of prosody and gesture, one of the earliest comments came from Birdwhistell (1952) who observed that intonational contours and events aligned with general gestural movements. Later, Kendon (1972, 1980) introduced hierarchical gestural units and described how they were coordinated with hierarchical intonational units. Since these early studies, researchers have come to understand more about how gesture and prosody are interrelated, and many now consider their connection as an inherent facet of the synchronisation of speech and gesture, although there are many open questions about the nature of this connection. In line with the literature, the present study also links gesture to prosody in its investigation of synchronisation.

1.1.1 Synchronisation with Prosody at the Micro level

McNeillean synchronisation rules (McNeill, 1992) and indications of synchronisation in the form of precedence at the lexical level have led researchers to look for finer temporal relationships between speech and gesture. Most studies conducted for this purpose have linked prosody to gesture as the main speech component that regulates speech-gesture synchronisation (Section 2.5.1). These studies have mostly focused on the synchronisation of atomic landmarks (i.e., the smallest possible anchors) within gesture and prosody (i.e., micro level). In general, they have reported that prominent points in time in prosody and gesture are synchronised. However, these studies employed different methodologies which had a number of shortcomings. In particular, the definition of synchronisation, determining what these prominent units are in gesture and prosody, and accounting for the effects of prosodic and gestural contexts (e.g., prosodic structure, and the semantic function of gesture) have been problematic in these studies.

The present study interrogates the claims of synchronisation between prominent markers in gesture and prosody in natural multimodal data in Turkish. The prosodic structure of Turkish is described using the most widely-used formal phonological framework, i.e., the Autosegmental-Metrical model of intonational phonology (Ladd, 2008; Kamali, 2011; Ipek & Jun, 2013; Güneş, 2015). As we will see in Section 2.5.1.2, Turkish prosodic structure presents a challenge for the claims of synchronisation between prominent units because some prosodic constituents in the prosodic structure of Turkish do not bear any prominence. The synchronisation of gesture and prosody in such prominence-less cases is unclear. Moreover, thanks to the adoption of a formal phonological framework, both the identification of synchronisation anchors in prosody, and the investigation of synchronisation itself are informed by the relevant prosodic research. The study accounts for the effects of different prosodic contexts on synchronisation through a series of analyses testing whether the synchronisation behaviour shows variation depending on the prosodic structure of utterances accompanied by gesture. Unlike earlier studies, the present study investigates synchronisation for different types of gesture to determine whether the synchronisation of units shows variation depending on the type of gesture involved (Section 3.4.1.1).

The analysis of synchronisation in the present study is important for three main reasons. First, it addresses the shortcomings of previous studies and offers a methodologically sounder account of synchronisation. Secondly, by analysing synchronisation using a phonological framework, it adds a linguistic perspective to synchronisation which has been overlooked by most previous studies. Comprehensive details about the effects of prosodic structure in the analyses are offered to encourage future research to conduct linguistically informed synchronisation analyses. Thirdly, previous studies have been mainly in English or in other prosodically similar languages. Turkish is prosodically very different from these languages, therefore its prosodic features, such as the lack of prominence on certain prosodic constituents, are bound to extend our understanding of gesture-prosody synchronisation. Finally, the results of synchronisation (Chapter 4) have implications on speech and gesture production models as the systematic synchronisation of micro units encoded at the later stages of production reveal that speech and gesture production is more interactive at later stages than previously assumed.

1.1.2 Synchronisation with Prosody at the Phrasal level

The claims of gesture-prosody synchronisation in the literature have been mainly based on the synchronisation between the smallest anchors (i.e., points in time). However, gesture and prosody are also alike structurally in that they both have phrasal constituents nested within each other, forming a structural hierarchy. Very few studies have investigated whether the synchronisation of gesture and prosody also persists at the phrasal level (i.e., macro level). It therefore remains an open question as to whether gestural phrases are synchronised with prosodic phrases. We will see in Section 2.5.2 that some of the shortcomings in the studies that investigated synchronisation at the micro level have also been observed at the macro level. In terms of the definition of synchronisation, these studies also have not defined how close these phrases should be in order to be considered as synchronised. Most of these studies have not reported a synchronisation between gestural and prosodic phrases (cf. Loehr, 2004). Instead, either precedence (i.e., gestural phrase precede prosodic phrases) or overlap (i.e., one phrase occurs within the duration of the other) cases were observed. All these studies concluded that gesture and prosody synchronisation at the phrasal level is disputable and evidently less strong than the synchronisation at the micro level.

The present study extends the investigation of synchronisation to the phrasal level in order to define the scope of synchronisation between gesture and prosody. That is, it must be systematically investigated whether gesture and prosody are only synchronised at the micro level or the synchronisation of these also persists between gestural and prosodic phrases. Consequently, the present study tests whether gestural phrases are synchronised with prosodic phrases defined in Turkish (Section 2.5.1.2). The analyses are conducted on the same data, using same definition of synchronisation and statistical tests as in the analyses at the micro level. These tests also account for aspects that were overlooked in the earlier studies such as the inclusion of different gesture types and the effects of the organisation of prosodic phrases within the prosodic structure.

These analyses into synchronisation at the macro level are important because the analyses are complementary to the analyses of synchronisation at the micro level. If we want to establish links between gesture and prosody, we need to consider their interactions at every level within their hierarchical organisation. Moreover, the use of a definition of synchronisation that is statistically tested makes the present study one of the first to systematically investigate synchronisation between gestural and prosodic phrases. The present study offers insight into synchronisation using data from a language that has never been investigated for this purpose before. The results can be compared to the results of earlier studies, which helps define the extent of synchronisation between gesture and prosody cross-linguistically. Finally, as is the case for synchronisation at the micro level, the results of synchronisation at the macro level (Chapter 5) have implications for unified models of speech and gesture production, which also have not been considered in the models proposed so far (Section 2.3).

1.1.3 Synchronisation with Information Structure

It can be said that prosody has been seen as the sole driver behind the synchronisation of speech and gesture in the literature. No other linguistic interface has been considered to have a similar role. Based on some associations made in earlier works (Section 2.6.2), the present study hypothesises that gesture is also synchronised with information structural units, i.e., topic and focus. There have been many studies showing that gestures serve communicative functions within discourse (Section 2.2.1). Therefore, it is possible that gesture can be sensitive to the information structure units that organise information in relation to a discourse in line with the communicative intention of the interlocutor. Gesture and information structure have also been implicitly associated with each other in gesture production theories (Section 2.3). In these models, both speech and gesture are claimed to originate from the same source as departures of thought from the presupposed background. This explanation of the origins of gesture and speech overlaps a great deal with information structural notions defined in Section 2.6.1. A number of studies have also associated gesture and information structural notions from formal semantic and developmental perspectives (Section 2.6.2). Furthermore, a link between gesture and information structure can also be established through gesture's close relationship with prosody. Prosody is a one of the principal cues to information structure for many languages - the features of prosodic structure are synchronised with information structural units in order to mark them in the speech signal (Section 2.6.2). This means that gesture, which has already been shown to be synchronised with prosody to some extent, can also be synchronised with information structure through the medium of prosody.

Overall, the present study postulates that prosody and information structure both interact with gesture and govern the synchronisation of gesture with speech as an ensemble. To test this hypothesis, the present study investigates whether gestural phrases are synchronised with information structural units. The investigation uses the same methods as the investigation of synchronisation with prosodic phrases, using statistical approaches to define synchronisation as well as to test potential gestural, prosodic, and information structural effects on it.

The synchronisation of gesture and information structure has not been systematically tested before (see Ebert, Evert, & Wilmes, 2011) - there has been no empirical evidence showing whether gesture is synchronised information structure. In addition, the present study is the first study to consider a three-way synchronisation of gesture, prosody and information structure. The synchronisation behaviours observed in the present study can show evidence for the fact that gesture production is informed by information structural and prosodic processes in speech production (Chapter 6). Such temporal sensitivities have implications for speech and gesture production, therefore they must also be accommodated by psycholinguistic models that integrate speech and gesture production.

1.2 Organisation

This thesis has seven chapters. Chapter 2 presents reviews of previous works that sets the background to the investigation in this thesis. It starts by defining gesture and its form and function based classifications as used in the thesis. It moves onto establishing links between speech and gesture by discussing early works that revealed various temporal and non-temporal interactions which led to conflicting views about gesture's function in communication. It then explains how these interactions have been represented in four influential psycholinguistic models of speech and gesture production. As the main interest of the thesis, the extent in which these models have incorporated synchronisation is further discussed. Later, the chapter introduces prosody as a speech component that enables gesture synchronisation.

It gives a brief historical account of how prosody has been linked to gesture in different ways before turning to more recent studies on gesture-prosody synchronisation.

Studies on synchronisation are covered at the micro and macro levels separately. At both levels, selected studies are reviewed in detail focusing on the methodologies employed, identifying open questions in the literature which have led to the research questions of this thesis. These research questions are framed within the prosodic structure of Turkish in order to highlight the multiplicity of potential synchronisation scenarios and to emphasise the importance of prosodic analysis in investigation of synchronisation.

Chapter 3 includes descriptions of the participant profile, and data collection design and procedures. In addition, it presents guidelines that were used for the annotation of gesture, prosody, and information structure. These guidelines define every unit annotated for each modality, and give careful details about the annotation procedure. Finally, the chapter ends with the descriptions of statistical tests used throughout the thesis.

Chapter 4 presents the results of the analyses of synchronisation at the micro level. The synchronisation between gestural anchors (i.e., apexes) and prosodic anchors (i.e., tonal events) was tested through a series of analyses which observed synchronisation behaviour in different prosodic and gestural contexts. The results showed significant effects of both contexts. Prominent events in gesture and prosody were confirmed to be synchronised. However, in contexts where prosodic prominence was absent, gesture apexes showed sensitivity to prosodic hierarchy by systematically synchronising with the tonal event that marked the boundary of the smallest prosodic phrase in the hierarchy (the prosodic word) while avoiding synchronisation with tonal events marking the prosodic phrases higher in the hierarchy. Synchronisation behaviour was also found to be depended on the type of gesture. Rhythmic behaviour commonly observed in some gesture types predicted tighter synchronisation between anchors.

Synchronisation behaviours at the macro level are presented in Chapter 5. This chapter first finds the best phrasal prosodic anchor (i.e., intermediate phrase or intonational phrase) for gesture phrases. The results showed that there was no synchronisation of single gesture phrases with single intermediate or intonational phrases. However, as reported by some previous studies, an overlap between phrases was found. A single gesture phrase was observed to span over multiple consecutive intermediate phrases, and gesture phrases were synchronised with these intermediate phrase groupings rather than synchronising with only one intermediate phrase. These intermediate phrase groupings tended to occur over the default focus position in Turkish, which signalled a possible synchronisation with information structural categories (i.e., topic and focus).

Following the hypothesis that gesture is informed by information structure, which was also supported by the results presented in Chapters 5, 6 tests whether gesture phrases were synchronised with topics and foci. The results showed a delayed synchronisation where topics and foci were contained within gesture phrases occurring at fixed distances away from the boundaries of gesture phrases. Further synchronisation tests revealed that topics and foci were synchronised with the meaning bearing segments of gesture phrases containing apex related information (i.e., apical area).

Chapter 7 first gives a summary of results presented in Chapters 4, 5, and 6. It then compares these results with the results of relevant previous works while discussing the significance and implications of these results. In general, this thesis shows that gesture is informed by the encodings of prosody and information structure, which is manifested in the synchronisation of units through the prosodic phrasing and prominence. The chapter then goes on to discuss the results in relation to psycholinguistic models of speech-gesture production. It proposes a new model as an extension to one of the earlier models (i.e., the Interface Hypothesis, Kita & Özyürek, 2003) showing how the observed synchronisations can be accommodated in a unified model of speech and gesture production. The chapter finally concludes the thesis with a summary that includes future directions.

2
Review

The present study is interested in the synchronisation of gesture with prosody and information structure and the implications of this synchronisation for models of speech and gesture production. The synchronisation of three different verbal and non-verbal components of communication is a complex process, and in order to be contextualised, it requires knowledge about not only the mechanisms and structures of each, but also the basic interactions between them. Moreover, existing speech and gesture production theories must also be introduced in order to be able to discuss the implications of synchronisation for the models.

The objective of this chapter is to elaborate on each of these points progressively, leading to the research questions of the present study. Since gesture is the main modality of interest, a definition of gesture and its classifications of its forms and functions (i.e., gestural structure) are presented first (Section 2.1). Gesture has been observed to have many interactions with speech, and these interactions have been used to create models of speech and gesture production (Sections 2.2 and 2.3). This chapter presents these next in order to give an overview of the extent of the relationship between gesture and speech. We will see that although synchronisation is one of the key interactions between gesture and speech, it has found limited representation in these models or not been represented at all.

As stated in Chapter 1, gesture has been linked to prosody as the main driver of speech-gesture synchronisation. In Section 2.5, a review of what is known to date about the synchronisation of gesture and prosody is provided, which will give in-depth details about the connection between prosody and gesture. Parallel to the aims of the present study, earlier studies on the synchronisation of gesture and prosody at the micro and macro level are reviewed separately, and then the implications of their findings are discussed within the frame of the prosodic structure of Turkish, leading to the research questions of the present study.

Finally, information structure is introduced as another modality that may be governing speech-gesture synchronisation. Earlier studies that have shown the close relationship between prosody and information structure, and various associations between gesture and information structure are reviewed in Section 2.6 in order to show why information structure is relevant for the synchronisation of gesture and speech.

2.1 Gesture

This section introduces the working definition of gesture as used in the present study along with characteristics of its forms and functions since these are essential in understanding the analysis of gesture in the study. Since the main interest here is to examine timing relationships between speech and gesture units, the most relevant aspects of gesture are gesture segmentation, which reveals the temporal structure of gesture, and gesture identification, which puts any timing relationship into a semantic context. Accordingly, these are the main aspects covered in this section.

Broadly, a gesture is any kind of bodily movement with a communicative purpose. Within the frame of human communication, in his seminal work McNeill (1992) distinguished four types of gestures in his “Kendon’s continuum” (see Figure 2.1).[1] *Gesticulations* are the most common gesture type observed in daily communication. They carry speech-related information, and have been shown to coincide with their co-expressive speech. *Emblems* are conventionalized gestures such as the popular “thumbs-up”. They are culture specific (i.e., they can have different meanings in different cultures), and are optionally accompanied by speech as they are meaningful by themselves without any accompaniment. *Pantomimes* are sequences of gestures with a narrative, and are obligatorily performed without speech. *Signs* are words of sign languages, such as New Zealand Sign Language, which have their own linguistic structure. As would be implied by their evolution, sign languages naturally do not require any accompanying speech.

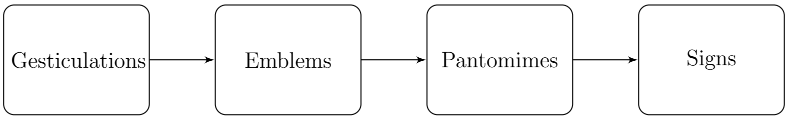


Figure 2.1: Kendon’s continuum

In line with these definitions, two changes can be observed moving along the continuum from left to right: (1) obligatory speech accompaniment decreases, and (2) gesture shows more language-like behaviour. Gesticulations or co-speech gestures are at the left end of the continuum, meaning that they require de facto speech accompaniment, bear the least amount of linguistic properties, are not conventionalized, and are most closely tied semantically

[1] Additional continua were introduced later in McNeill (2005).

to units of speech. Therefore, it is gesticulations that must be examined in order to comprehensively understand the relationship between speech and gesture. In sum, within the frame of this study, a gesture is any kind of bodily movement that is spontaneously produced in relation to and accompanied by speech, [2] i.e., co-speech gestures (Kendon, 2004). In the following chapters, unless otherwise stated, the word "gesture" is used to refer to gesticulations/co-speech gestures for the sake of simplicity.

2.1.1 Gesture Classifications

Gestures can be further classified into various categories according to their functions (Gut & Milde, 2003), forms (McNeill, 1992; Martell, 2002), and roles in discourse (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). Many such gesture classifications focus on manual gestures, i.e., the gestural movements of the hands and arms. This is because their typical kinematic specifications are less challenging to determine because of their dynamic affordability compared to, for instance, head gestures (Altorfer et al., 2000). Amongst the different classifications of gestural movements, this review focuses on the well-known classifications of manual gestures based on their form (i.e., segmentation) and on their semantic function, as these are the most relevant aspects for the analysis of synchronisation in the present study.

In form-oriented classifications, gestures are often categorised according to features such as multidimensional hand positions including distance (e.g. far/close relative to the body), height (e.g., above head), radial orientation (e.g., inward) (Kipp, Neff, & Albrecht, 2007), and shape (the ASL handshape inventory is often adopted as in McNeill, 1992). Along with these fundamental form features, the gestural excursions of the hand can be divided into a sequence of dynamically discrete segments (Kendon, 1980; McNeill, 1992).

[2] The semantic relation to speech naturally excludes self-touching, e.g., itching and hair combing.

Such a linear segmentation of gestural movements from beginning to end allows us to capture the temporal structure of a gesture, which is key for the present study given its aim of capturing the synchronisation of gesture and speech. These gestural segments also exist in a hierarchy, in that the smallest segments come together to form a larger segment higher in the hierarchy (see Figure 2.2). These are summarised below, moving from the largest to smallest.

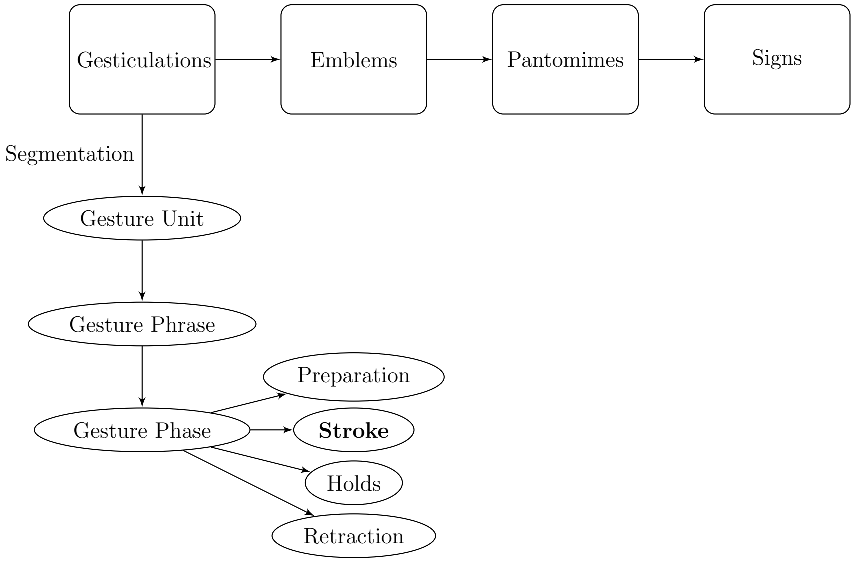
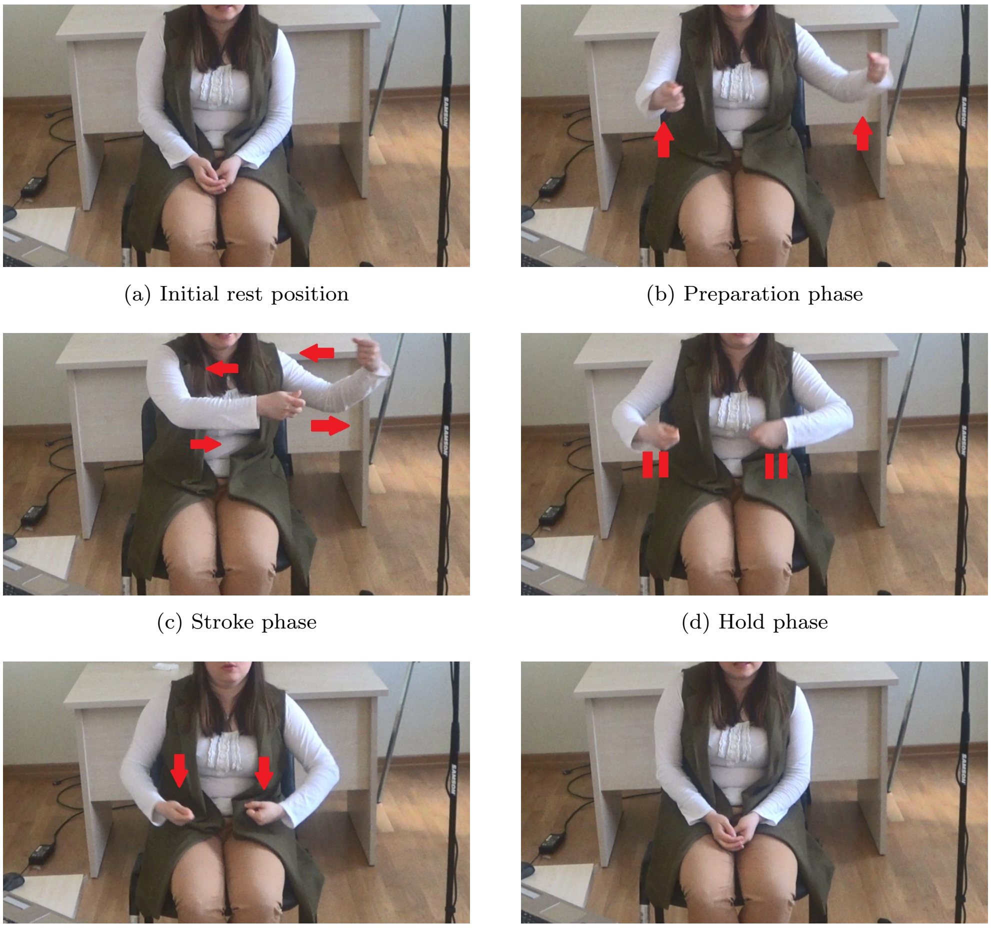


Figure 2.2: Segmentation of co-speech gestures

The initialisation and termination of a gestural movement occurs at rest positions. A *rest position* is a stable state of the hand where it is supported by an object or a part of the body (see Figures 2.3a and 2.3f). Any kind of gestural hand movement between two rest positions, from the initialisation of movement from one rest position to its termination at another (or the same) rest position, constitutes the largest segment, the *gesture unit.* Figure 2.3 shows an example of a complete gesture unit between two rest positions where the gestural movement describes an act of swapping. First, starting from a rest position, both hands are raised to chest level where they perform a semi-circular movement going away and returning to her body (describes swapping). Once the hands are back at chest level, they hang in the air briefly before being retracted to a rest position.

Figure 2.3 shows only one meaningful gesture within the gesture unit. However, interlocutors can perform multiple distinctly meaningful gestures within a gesture unit, and each of these meaningful gestures is called a *gesture phrase*. In other words, the hand does not need to return to a rest position every time in order to start a new gesture, instead the interlocutor can chain multiple gesture phrases between rest positions, forming one gesture unit. In the context of Figure 2.3, this means that instead of the hands being retracted to a rest position, another (or multiple) meaningful gesture is performed and it is only after the second gesture that the hands are retracted to a rest position.

The definition of gesture phrase, in this sense, depends on the expressiveness of movement. However, not all movements are expressive within a gesture phrase - gesture phrases can be segmented into dynamically discrete and linearly ordered phases called, *gesture phases*. The first phase in this organization is the *preparation phase* where the hand departs from a rest position (e.g., raising of the hands in Figure 2.3b) to enable the execution of the *gesture stroke*. A gesture stroke carries the meaning of the gesture phrase and is executed with maximum effort (e.g., the semi-circular movement in Figure 2.3c). Because of its expressive content, the gesture stroke is the only obligatory phase within a gesture phrase. The stroke can be preceded and followed by *holds* (if it comes before the stroke, it is called a *pre-hold*, otherwise it is a *post-hold*), in which the hands are frozen in location (e.g., Figure 2.3d). Finally, in a *retraction phase*, the hand returns to a rest position (e.g., hands are retracting in Figure 2.3e to the final rest position in Figure 2.3f). Altogether, Figure 2.3 shows a representation of the segmentation of a gesture unit as described thus far.



(a) Initial rest position

(b) Preparation phase

(c) Stroke phase

(d) Hold phase

(e) Retraction phase

(f) Final rest position

Figure 2.3: Gesture phases forming a single gesture phrase describing an act of swapping

There have been several additions to these segments (e.g., “recoil” in Kipp, 2005). One of these additions, “the apex” (Loehr, 2004) or “hit” (Yasinnik, Renwick, & Shattuck-Hufnagel, 2004), is central to the present study as it is used in the analyses of synchronisation at the micro level (see Section 2.5.1). The apex is the dynamically most prominent point in time within the stroke. The dynamic prominence of this unit has been attributed to seemingly different but overlapping qualities such as the point of maximum extension (Leonard & Cummins, 2011), the peak of the stroke or gesture target (Loehr, 2004; Jannedy & Mendoza-Denton, 2005), and abrupt stops (Yasinnik et al., 2004; Shattuck-Hufnagel, Yasinnik, Veilleux, & Renwick, 2007). These definitions of the apex are discussed in Section 2.5.1. All gestural units/segments introduced thus far are further explained with examples in Chapter 3.

Different annotation schemata have treated gesture segmentation in different ways. However, many follow or build on the heuristic segmentation suggestions of McNeill (1992) (Caldognetto, Poggi, Cosi, Cavicchio, & Merola, 2004; Trippel et al., 2004; Kipp et al., 2007; Lausberg & Sloetjes, 2009; Lücking, Bergman, Hahn, Kopp, & Rieser, 2013). These schemata were designed to capture and detail specific aspects of gesture, generally focusing on either function or form depending on research interests. Therefore each comes with its own advantages and disadvantages. One aspect in which these schemata vary considerably is the level of detail available for the actual annotation practice. Some of these include broad descriptions and leave the practical application of these descriptions to the annotator (e.g., Caldognetto et al., 2004), whereas others offer complex but clearer guidelines for segmentation practice (e.g., Martell, 2002 and Kita, Van Gijn, & Van der Hulst, 1998). Kita et al. (1998) designed a form-oriented and syntagmatic rule-based scheme that can be used for annotating both gesture and signs. It involves segmenting body movements into “movement units, phrases, and phases”, a segmentation that is based solely on movement dynamics without any assignment of meaning and function. The resulting segments are fundamentally the same concepts as in Kendon (1980) and McNeill (1992) as summarised in Figure 2.2. However, Kita et al. (1998) include clear descriptive criteria for the identification of the boundaries of each segment and what the annotator can observe during these segments. The annotation scheme used in the present study is based on their guidelines with some adaptations (e.g., the annotation of the apex) and with more detailed explanation and solutions for potential issues that an annotator can face. However, the gesture segments will be referred to using McNeill’s (1992) widely-known terminology (e.g., gesture phrase and gesture phase) for the sake of easier comparability with previous studies. For details of how these units were defined and annotated in the present study, the reader is referred to Section 3.4.1.

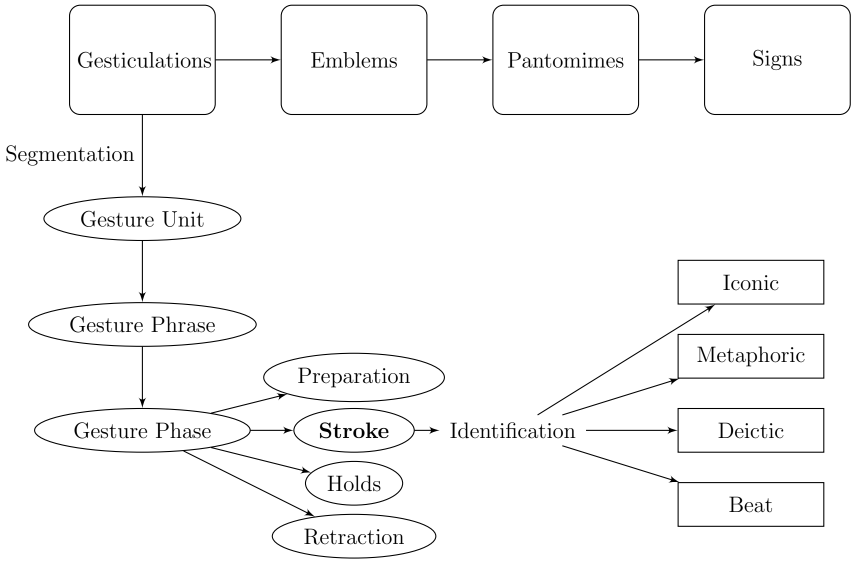
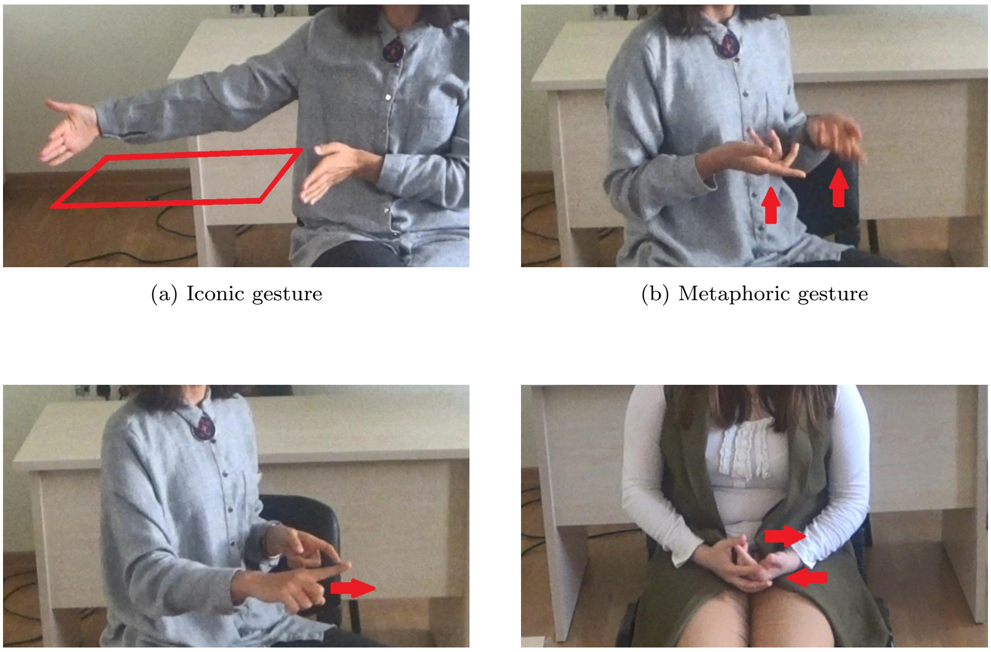


Figure 2.4: Segmentation and identification of co-speech gestures

Manual gestures consisting of units with basic form characteristics as described in Figure 2.2 can assume numerous functions in human communication. One of the most studied functions of these gestures is their semantic function. McNeill (1992) categorises manual gestures according to the semantic content expressed within the stroke phase, the meaning-bearing phase (see Figure 2.4).

Iconic gestures (or iconics) have a close semantic relationship with co-expressive speech in that they represent the physical aspects of the information encoded in the speech (e.g., gesturing to describe the shape of a table as in Figure 2.5a). *Metaphoric gestures (metaphorics)* function in the same way as iconics, except they represent abstract contents (e.g., gesturing with open palms facing up to show "empty hands" which indicates uncertainty in Figure 2.5b). *Deictic gestures (deictics)* are pointing gestures indicating the locations of entities in space (e.g., pointing to side with the index finger Figure 2.5c). *Beat gestures (beats)* are flicks of the hand. They do not bear any semantic content themselves, but have instead been shown to be coordinated with prosodic events (e.g., Loehr, 2004; Leonard & Cummins, 2011; Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016; Shattuck-Hufnagel et al., 2016; Shattuck-Hufnagel & Ren, 2018), functioning as a visual highlighter (e.g., a quick sideways flick as in Figure 2.5d). The definition and annotation of these semantic functions is further detailed with examples in Section 3.4.1.



(a) Iconic gesture

(b) Metaphoric gesture

(c) Deictic gesture

(d) Beat gesture

Figure 2.5: Categorisation of gestures according to their semantic function

McNeill (2005) emphasised that this semantic typology of gestures should not be viewed as a typology with mutually exclusive categories but one with dimensions. He refined the original categories as dimensions which are iconicity, metaphoricity, deixis and temporal highlighting (beat related).[3] The aim of the introduction of a multidimensional classification was to recognize that a single gesture can portray characteristics of multiple dimensions. For instance, a pointing gesture that points at a location in space to represent an abstract concept such as past or future while also containing superimposed beats over its segments would be an example of the multidimensionality of gesture. While it is easy to define these dimensions, it can be challenging to fully capture these on real visual data. The semantic function identification process within the present study is discussed in detail in Section 3.4.1.

This section has focused on a general introduction of gesture, its formal characteristics and semantic functions in the interest of the main objective of the present study - to investigate the synchronisation of gesture with prosody and information structure. These basic concepts will be encountered in the review of previous studies below, but more importantly, they will be used to define gestural units to test synchronisation in this study. The following Section 2.2 deals with how co-speech gestures (as well as their formal and functional characteristics) interact with speech in communication and introduces speech-gesture production models that have arisen to account for these interactions.

2.2 Interaction of Speech and Gesture

The act of speaking leads universally to gesturing (Kita, 2009). The combination of speech and gesture happens spontaneously and in a systematic way. This systematic combination of speech and gesture has been observed as early as the one-word production stage in infants (Iverson & Goldin-Meadow, 2005), and during the speech production of congenitally blind chil-

[3] McNeill (2005) also defined social interactivity as a new dimension which arose from the gesture's role in communication organization.

dren (Iverson & Goldin-Meadow, 2001), which shows how resilient and deeply entrenched speech-gesture coupling is. However, gesture's role in human communication has been a subject of debate for gesture researchers, and created a dichotomy in the literature. One view is that gestures are generated for the benefit of the speaker, and the meaning in gesture is redundant to the meaning conveyed via speech, implying that gestures are derived from speech. On the other hand, gestures are seen as communicative tools which are generated for the sake of an addressee, and they convey a message that is complementary to that of speech. This type of interaction however, does not mean that gestures are derived from speech. Instead, speech and gesture are generated in parallel and share the same computational stage.

These issues have been at the centre of speech and gesture production models which were formulated to account for the observed interactions of these two modalities. The production models resulting from these studies can mainly be distinguished by the degree of integration of speech and gesture (i.e., unified versus separate), in which synchronisation of modalities plays an important role. However, the models vary in the extent to which they account for various aspects of synchronisation. This section gives a summary of various interactions observed by previous studies, followed by a brief introduction of how different speech and gesture production models accounted for them. Finally, this section highlights the importance of synchronisation in understanding speech and gesture interaction and comments on how each model incorporates synchronisation patterns observed in relevant research.

Section 2.2.1 gives a summary of empirical findings on the interaction between speech and gesture, and reviews how existing psycholinguistic models of speech and gesture production account for these findings. Numerous investigations have revealed various interactions between speech and gesture, and some of these have shown contradictory findings, resulting in considerable disagreement in the field. The disagreements have revolved around overlapping issues such as the role of gesture in communication (i.e., is gesture communicative or only for the speaker's benefit?), its co-expressivity (i.e., is gestural content redundant or complementary to speech?), and its co-occurrence (i.e., does gesture occur with relevant speech or during pauses?).

2.2.1 Gesture is a Communicative Tool

McNeill (1985, 1992) is perhaps the most influential proponent of gesture's communicative role. His main proposition was that gesture and speech share an early computational stage where semantic and pragmatic functions are decided and performed in parallel (McNeill, 1985, p. 354). The idea that gesture and speech share a common cognitive origin can be traced back to Kendon (1980). However, McNeill is the first researcher to provide a theoretical framework for this origin. He defines three synchronisation rules that govern speech and gesture interaction.[4] *Phonological synchronisation* claims that gesture strokes either come slightly before or end at the same time as "phonological peaks" (referring to stressed syllables) in speech. This claim sits at the heart of studies that link phonological units to gestural units, including the present study; therefore it will be covered in detail in the following sections. *Semantic synchronisation* states that speech and accompanying gestures must carry the same meaning. This conceptual linkage of the two modalities is observed in the form of a gesture anticipating its speech counterpart. That is, the preparation of a gesture anticipates the semantically linked speech counterpart by a short duration (Kendon, 2004). The semantic synchronisation claim has bearing on the determination of the communicativeness of gestures in that if gestural content is the same as speech content then gesture is communicatively redundant because it does not contribute

[4] The term synchronisation is used loosely to refer to the linkage between speech and gesture - it does not necessarily mean actual temporal synchronisation. This is discussed in Section 2.4.

to communication. However, if gesture complements speech by encoding messages that are not in speech, then gesture can be said to contribute to communication by complementing intended messages in a different modality. Finally, *pragmatic synchronisation* suggests that gesture and speech converge at a pragmatic level. A gesture can indicate an interlocutor's stance, parallel a speech act, emphasize speech chunks that are thought to be important, and organize turn-taking (Kendon, 2004). All of these show that both interlocutors and communicative interaction itself can benefit from gesture-speech couplings.

Many researchers have tested these synchronisation types as well as other aspects of communication in order to capture the communicative nature of gestures. Over the years, one line of investigation with this goal has been the examination of gesture production in terms of the frequency and quality of gestures under different visibility conditions between interlocutors (Cohen & Harrison, 1973; Cohen, 1977; Rimé, 1982; Bavelas et al., 1992; Krauss, Dushay, Chen, & Rauscher, 1995; Alibali, Heath, & Myers, 2001; Emmorey & Casey, 2001; De Ruiter, 2000). In the experimental design of these studies, speakers with varying tasks (e.g., giving directions, telling stories) talked to an addressee in conditions where they can see the addressee (e.g., face-to-face), and when they cannot (e.g., via an intercom or through a partition) (see Bavelas et al., 2008 for an extended overview and discussion). These varying visibility conditions offer insight into the communicative function of gesture because if there is no gesturing or a decrease in gesture under the no-visibility condition, this can be interpreted as showing that gestures are intended for an addressee who can see and understand the message they convey. In all these studies, it was indeed the case that speakers gestured more while they were talking to a visible addressee. Note that gesture production was never completely gone in these studies - they were only produced at a lower rate. It is challenging to explain the persistence of gestures from a communicative perspective. However, Cohen and Harrison (1973) commented that these may be habitual productions spilling over from the common use of gesture in which an addressee is visually present. The fact that there was gesture production in the no-visibility condition can also be an argument for the view that gesture production is not for a visible addressee but for the speaker, which will be covered in Section 2.2.2.

In addition to visibility, researchers also investigated whether the mode of communication, i.e., dialogue or monologue, has an effect on gestural behaviour. It was found that compared to monologues, when interlocutors engaged in dialogues, they produced larger gestures and also gestured at a higher rate (Bavelas et al., 2008). Moreover, some gestures were shown to function particularly to regulate the conversation with an addressee, e.g., managing turn-taking (Bavelas et al., 1992, 2008, 2014) or to repair communication (Holler & Beattie, 2003). Findings such as these suggest that gesture is deemed to be an effective tool that is produced in response to an addressee's immediate communicative needs, showing that gestures are designed for a recipient, and are therefore communicative. De Ruiter, Bangerter, and Dings (2012) showed further evidence for the communicative function of gesture in a study where they investigated the redundancy of speech and gesture combinations. They found that interlocutors' gesture and speech combinations were often redundant at the start of dialogues, which they interpreted as serving to decrease the chance of misunderstanding early in the discourse. In this sense, gesture was an effective way of quickly addressing this demand while also keeping the "joint effort in communication" at a minimum. Similarly, addressee-oriented gesture design in dialogues was also observed in studies that investigated whether interlocutors' opinions of mutual understanding can influence speech and gesture. In their review of the social functions of gesture, Holler and Bavelas (2017) showed that interlocutors used fewer gestures as well as fewer words when the level of shared knowledge between interlocutors was high. The finding shows that gesture and speech productions parallel each other depending on what will be and not be for the benefit of the addressee, as determined by the degree of already established knowledge.

Another way in which the communicative function of gesture has been tested has involved looking at gestural content. Relative to the speech content, a gesture can convey supplementary information (i.e., information that is not encoded in speech), complementary information (i.e., information that is encoded in speech but gesture tokenizes an additional aspect) or redundant information (i.e., information already encoded in speech) (Goldin-Meadow et al., 1993; Goldin-Meadow, 1999; Goldin-Meadow et al., 2001). For instance, a child pointing at a football while saying "I want to play" would be an example of a complementary gesture, whereas a gesture indicating a certain length along with the utterance "the snake was this long" would be a supplementary gesture. A redundant gesture, for example, would indicate the shape and size of a tennis ball in an utterance where it is mentioned. The information that an addressee has access to during such parallel uses of gesture can be used to predict communicativeness, in that if a gesture encodes extra information (i.e., it conveys complementary or supplementary information), then this implies that gesture has a communicative purpose for the listener instead of only aiding the speaker in their production of speech.

Beattie and Shovelton (1999, 2002) used interviews to find out the amount of information participants take in when they were introduced to short excerpts of cartoon narrations describing single events. The interviews were structured so that participants could only listen to excerpts, or listen to and watch them at the same time. They found that compared to when they just listened to the speech, the participants absorbed significantly more precise information about the events depicted when they both listened to and watched the excerpt. Although the extra information in gestures was found to be limited to the position of objects in narrative space and to object size, they concluded that "the beneficial effect of gestural communication was significant" (Beattie & Shovelton, 1999, p. 458).

Taken together, investigations into visibility, mode of conversion, belief of mutual understanding, and redundancy of gestural information show that although gestures are fundamentally illustrators produced alongside speech, their use as illustrators depend on these communicative factors. That gesture production is sensitive to the presence of an addressee and the mode of interaction suggests that gestures are planned for an intended communicative role. However, this view was not without challenge - there been several other studies revealing empirical data that conflicted with the communicative gesture view, claiming that gesture is mainly produced for speakers, functioning as a facilitator for speech production.

2.2.2 Gesture is a Facilitator

The communicative theory of gesture has been immensely influential in gesture research, however there have been other views. The main opposing view argued that gesture is a facilitator for speech, and that gestures are generated for the benefit of the speaker, primarily functioning to assist speech production processes.

Some proponents of this view made use of Goldman-Eisler's (1967) "cognitive rhythm" theory to challenge the communicative theory of gesture. Cognitive rhythm is a rhythmic property in speech that is a "manifestation of a cycle of acts of planning and verbal production" (p. 127). In speech, planning takes place in hesitant phases containing a high number of pauses and shorter verbal expressions, whereas main verbal production takes place in fluent phases having a smaller number of hesitations along with relatively fluent speech. Aboudan and Beattie (1996) tested the effects of this pause/speech ratio on gesture. They posited that if speakers have shorter hesitant phases it will be more difficult for them to retrieve lexical items in fluent phases (due to having less time for the planning process itself), resulting in more gestures to assist speech production in order to compensate. Indeed, in their experiment, when hesitant phases were shorter than normal, more gestures were observed in fluent phases, which provided evidence that gesture assists speech production. Butterworth and Beattie (1978) also found that in these fluent phases, the beginnings of gestures were more likely to occur during pauses, and therefore precede their semantically related speech by a long margin (cf. Kendon, 2004). Their interpretation was that some of the pauses in fluent speech were caused by difficulty in accessing a desired lexical item, which triggered gesture production to assist speech. They provided evidence for their interpretation with an analysis of word frequency. Low frequency words hypothesized to be more difficult to retrieve were more often accompanied by gestures. Similarly, Morrel-Samuels and Krauss (1992) investigated the duration of the delay between a gesture and its lexical affiliate and found that if a lexical item was less familiar to the speaker, the gesture was more likely to precede its affiliate and the lag between the two was also increased. They attributed this delay to the difficulty in lexical retrieval and suggested that the timing relationship between gesture and speech is essentially managed by the ease of the retrieval process.

Schegloff's (1984) analysis extended Butterworth and Beattie's (1978)to gesture strokes (i.e., meaning bearing parts which he called "acmes" or "thrusts"). He found that in addition to gesture onsets (i.e., the preparation phase), strokes also preceded their lexical affiliates, and he claimed that gesture and speech are therefore not affiliated by synchronisation. He postulated that gestures are secondary to speech and they become meaningful "only when the bit of talk they are built to accompany arrives" (p. 291). Krauss, Morrel-Samuels, and Colasante (1991) arrived at similar conclusions when they analysed whether gestures conveyed additional meaning to the speech. Their study claimed that "[j]udgements of a gesture's semantic category were determined principally by the accompanying speech rather than gestural form" (p. 743). Although they agreed that gestures can convey information to a limited extent, they claimed that the information contained in gesture is mostly redundant.

Interesting findings have been reported in studies that investigated the effect of visibility on speech and gesture that a communicative perspective could not fully explain, as previously mentioned. Bavelas and Chovil (2000) and Bavelas, Kenwood, Johnson, and Phillips (2002) investigated gesturing rate through different addressee designs such as face-to-face, telephone and tape-recorder, and found evidence supporting the communicativeness of gesture. However, they also reported that speakers still gestured on the phone when there was no visible addressee, and that suggests gestures are for the benefit of the speaker only, which, in fact, was also recognised by the authors: "When their gestures would not be seen, speakers are much likely to make them redundant, and therefore not essential to their recipients" (Bavelas et al., 2002, p. 15). It is also important to note that in telephone and tape-recorder designs speakers pointed at and followed the shapes and lines in pictures they described with their hands. The authors classified these gestures that were clearly not aimed at an addressee as "self-prompting gestures" and these can be considered as evidence for gesture's speech facilitator function as the speaker is the sole beneficiary of these self-prompting gestures.

Looking at the role of gesture from a more biological point of view, Iverson and Goldin-Meadow (1997) and Iverson et al. (2000) reported that congenitally blind children still gesture even though their gestures are relatively different from sighted children in terms of shape, content, and frequency. It can be claimed that because congenitally blind children were never able to witness the informative value of gestures and they still gestured emphasizes speech facilitator function of gestures.

Overall, the findings and their interpretations reported here prescribe different motives behind the interaction of speech and gesture. It is likely that these differences are partly a result of different methodologies adopted - studies on both sides of the dichotomy defined, annotated, and elicited gesture through varying methods. On the whole, a working theory of gesture production must account for both its communicative and facilitation roles. In Section 2.3, psycholinguistic speech and gesture production models that were formulated to account for the findings discussed in this section are introduced.

2.3 Speech-Gesture Production Models

The previous section showed that gesture and speech interact. These interactions have led to claims that speech and gesture must be linked in production and perception. Consequently, researchers aimed to generate production theories for gesture by associating it with already existing psycholinguistic models of speech production. The observed interactions between speech and gesture have been used to formulate theories about how these productions are integrated.

Several models have been put forward to offer explanations for the observed interactions between speech and gesture. The discrepancies between these models stem from different points of view regarding gesture communicativeness, redundancy, and a few other aspects. In this section, the basic principles of four influential models are outlined, and brief accounts of how these models have integrated speech and gesture at different stages of production are given. Then, what these models predict about synchronisation is discussed. As we will see, synchronisation has not received much attention in these models. The summaries provided here are intended to prepare the ground for the discussion of the implications of the synchronisation results in the present study (Chapter 7). The reader is referred to the original studies for more details about the models.

2.3.1 Growth Point Theory

In Growth Point (GP) theory (McNeill, 1992; McNeill & Duncan, 2000; McNeill, 2005), speech-gesture production involves a blend of imagistic and linguistic thinking. The imagistic mode of thinking leads to gesture (the model is concerned only with iconic gesture production), and the linguistic mode leads to speech. It is assumed that these productions unfold from a pre-linguistic common origin. The productions take place in parallel - the interplay between them is preserved throughout. The GP theory refers to these initial ideational units (i.e., the common origin) from which gesture and speech are produced as “growth points”. They are minimal psychological units which aim to convey the most “noteworthy” information in a given context as a result of being born as a ”novel departure of thought from the presupposed background” (McNeill, 1992, p. 220). They are therefore not just non-redundant but actually rich in terms of content.

Placing the common origin of speech and gesture at such a conceptual level (i.e., the conceptualiser level in Levelt’s (1989) model, see Figure 2.6) suggests that gesture is integrated into production at the earliest stages of communication. Any kind of communicative intention, therefore, is externalized via both verbal and gestural means as potentially equal partners, which accounts for gesture’s communicative role evidenced by previously mentioned studies. The GP theory is based largely on gestural behaviour during speech disruptions such as delayed auditory feedback experiments, aphasia (McNeill, 1992), and clinical stuttering (Mayberry & Jaques, 2000), where gesture and speech coordination was unimpaired even when speech production was greatly disrupted. Since this model is not a computational model that shows clear interactions of speech and gesture at distinct stages, studies that make a general argument for the communicative role of gesture through visibility, non-redundancy, and mode of conversation can be accommodated in the GP theory without much difficulty (see Section 2.2.1).

Overall, GP theory can account for the relevant findings of studies that advocate for the communicativeness of gesture. It essentially makes the same argument for any interaction there may be between speech and gesture - they come from the same representation. However, the theory itself is not very clear about how conceptual gestural imageries are translated into sequences of hand movements beyond the common origin. There is not sufficient elaboration on what kind of mechanism handles this translation during production, and at precisely which stages of production there are interactions. Yet, it offers a reasonable explanation for the generation of iconic gestures representing physical entities with concrete features. However, it seems to fall short on explaining how gestures that represent abstract concepts (i.e., metaphoric gestures) are generated. For example, it is not fully clear how the theory would handle the generation of metaphorics whose physical features are not shared by the speech they accompany (Krauss et al., 2000). As will be shown, other production models that consider gesture as communicative draw from the GP theory. Therefore it has been very influential in the development of psycholinguistic models.

2.3.2 Lexical Retrieval Theory

The Lexical Retrieval (LR) theory (Krauss & Hadar, 1999; Krauss et al., 2000) has the facilitation function of gesture at its core, rather than communication. Its main claim is that iconic/metaphoric gestures (referred to as lexical gestures) do not predominantly function to deliver imagistic messages to an addressee. Instead, they operate speaker-internally to facilitate lexical retrieval for the speaker's own benefit. The model assumes that gesture and speech are two separate production systems that interact only at certain points. These interactions with speech are explained using Levelt's

(1989) speech production model in which three main stages are described: *conceptualiser, formulator,* and *articulator*. The right branch of Figure 2.6 shows a simplified schematic for this production model. In this figure (and in the schematics of other models in this section), the rectangles represent processors within the model; the arrows represent the direction of output between the processors; and the ellipses represent external information storages connected to the processors with dashed lines.

The conceptualiser is the level that constructs communicative intentions, and generates a *pre-verbal message* which is an ideational container for semantic specifications of concepts to be coded in speech. The formulator, through a *grammatical encoder*, converts the specifications in the pre-verbal message into a syntactic surface structure by mapping lexical and syntactic information stored in the lexicon onto these specifications. Then, a *phonological encoder* generates appropriate prosodic and phonetic plans (i.e., *internal speech*) for this surface structure. The articulator generates overt speech which is monitored for repair purposes (see Levelt, 1989 for a detailed account).

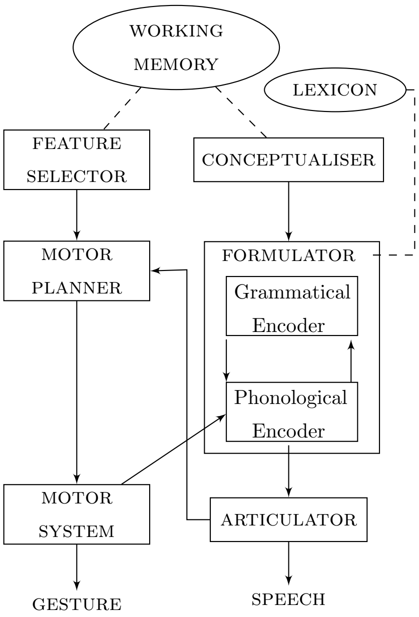


Figure 2.6: A simplified schematic of speech and gesture production in the Lexical Retrieval Theory

In the LR theory, gesture production starts from representations in the working memory independently from speech production. First, the representation that is going to be expressed is selected from the working memory. At this stage, the *feature selector* picks certain features to be encoded in the gesture from that representation, since not all features get to surface, much like speech. The selector translates the selected dynamic features into abstract properties of movements which are then translated into a set of instructions by the *motor planner*. The *motor system* executes these instructions and outputs gestural movements (see Figure 2.6).

As previously mentioned, in the GP theory's account, gesture begins with the imagistic information contained in the conceptualiser level, which is also where communicative intent resides. However, Krauss et al. (2000) argue that gestures primarily function to facilitate speech, not to deliver any communicative intent (see Krauss et al., 1995 for their review of relevant literature). Their model indicates the working memory as the source of the representations encoded in gesture. This implies that speech and gesture production systems diverge at a very early stage before any communicative intention can be constructed. In this view, encoded gestural features may or may not have an overlap with the communication intention of interlocutors; however, the potential impact of gestural contribution to communication is seen as "on average, negligible" (Krauss et al., 2000, p. 6).

The predominant function of speech facilitation in LR theory is represented as the motor system feeding into the phonological encoder (see Figure 2.6) where, Krauss et al. (2000) claim, the actual facilitation takes place thanks to the phonological encoder having access to word forms stored in the lexicon. They state that the features stored in the motoric form of gesture "facilitate retrieval of the word form by a process of cross modal priming" (p.13). Moreover, the model also links the speech output of the articulator to the motor program. The model suggests that hearing the articulation of the lexical affiliate is the cue for gesture termination following findings by Morrel-Samuels and Krauss's (1992) mentioned earlier. These late interactions between the two production flows are important for the present study because they may enable making assumptions about the synchronisation of gesture and speech. The findings of the present study will be discussed in relation to the interactions in this model in Chapter 7.

The LR theory accommodates the results of research on fluency and semantic asynchrony, which were discussed in Section 2.2.2. The findings that disfluency causes more gesturing, and hindrance of gesturing causes disfluency fit well within the theory's description of gesture as an assistant to speech. The reported asynchrony of gesture and its lexical affiliate can also be justified in the theory as it predicts no synchronisation due to the assumption of two separate production chains for speech and gesture. However, the LR theory is not without drawbacks. An important one comes from the strong association it makes between gesture and lexical items. Kita and Özyürek (2003) criticize this in a well-known example "roll down the hill". When the meaning of an accompanying gesture can be associated with the entire clause (e.g., encoding both rolling and downward movement and possibly the ground), the theory cannot explain which word is actually facilitated and what led to such a complex gesture. Plenty of similar cross-linguistic evidence, along with the findings that advocate a communicative function for gesture, strengthened the view that the source of gesture production may actually be conceptual, rather than lexical, which in turn motivated the introduction of different speech-gesture models.

2.3.3 Sketch Model

The Sketch Model (SM) proposed by De Ruiter (2000) is one such model that links speech and gesture at a conceptual level. It is also based on Levelt's (1989) model. However, unlike the GP theory or the LR theory, it not only explains the production of iconic/metaphoric gesture, but it also gives detailed accounts of the production of other gesture types and nongesticulations (i.e., emblems and pantomimes).[5]

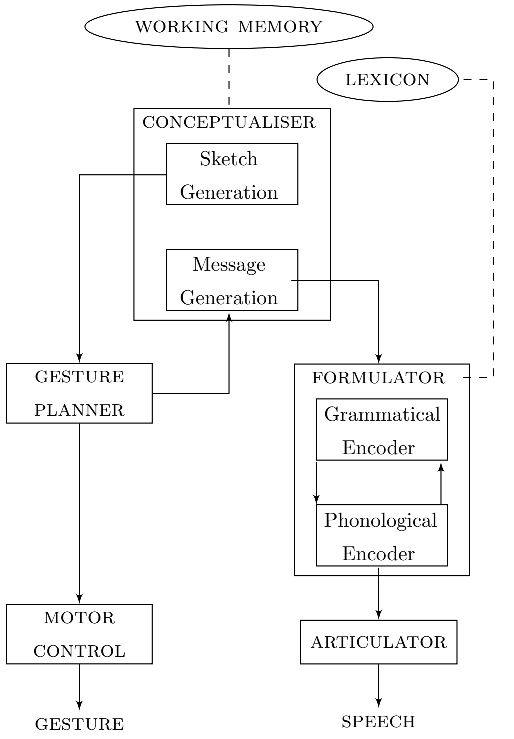


Figure 2.7: A simplified schematic of speech and gesture production in the Sketch Model

The SM is a computational model influenced by the GP theory, and therefore there is some overlap between their assumptions. The central similarity is that gesture and speech both have a communicative function, and are borne out of the same communicative intention in production. The signature assumption of the SM is that the conceptualiser, where gesture and speech originate from, performs a distribution of "communicative load" over speech and gesture channels. The conceptualiser generates a pre-verbal message via a *message generator* while additionally generating a *sketch* via the *sketch generator* (see Figure 2.7). The sketch is essentially the gestural counterpart of the pre-verbal message, and it contains an "abstract spatio-temporal trajectory that is as yet underspecified with respect to concrete motoric parameters (such as size, speed and location)" (de Ruiter & de Beer, 2013, p. 1022). In case the conceptualiser has difficulty or is restricted in the production of either the pre-verbal message or the sketch, it will allocate a higher load to the other channel in order to compensate (De Ruiter, 2006; De Ruiter et al., 2012). The feedback from the *gesture planner* and the formulator to the conceptualiser ensures that the communicative intention is preserved (this monitoring is not represented

[5] Beats are not accounted for in the model.

in Figure 2.7, see Levelt, 1989), and the load between channels is managed in case of any trouble in the later stages of production.

The most relevant assumption in the SM for the present study is about the gesture planner and formulator stages. The model assumes that these two act independently - there is no linkage between them. Therefore, any association between the channels related to the processes taking place in the gesture planner and the formulator, i.e. the temporal coordination of gestures with linguistic properties, has to be somehow arranged at the conceptualiser before any conventional linguistic planning. This issue will be further discussed in Chapter 7.

The SM accommodates the arguments for the communicative intention of gesture well by setting the origins of speech and gesture at the conceptual level where this intention is constructed. It can also accommodate arguments for the facilitator function related to speech disfluency and gesture hindrance, and word finding problems. The hardships in either channel would be handled by the conceptualiser's load distribution mechanism triggered by the feedback loops. Overall, the SM adopts the original hypothesis of the GP theory and organizes it according to a computational framework while also accounting for gesture's potential for speech facilitation. However, it cannot explain temporal coordination relating to processors below the conceptualiser.

2.3.4 Interface Hypothesis

In the Interface Hypothesis (IH) (Kita & Özyürek, 2003), speech and gesture are seen as outputs of separate but interactive production processes. It is also based on Levelt's (1989) speech model like the SM and LR theory. The model sets out to explain how iconic/metaphoric gesture content is coded. The authors specifically state that speech-gesture synchronisation is not within the scope of their hypothesis. However, it may be possible to establish temporal relations between the channels thanks to its highly interactive nature.

The model has overlaps with the previously discussed models in its assumptions, but it diverges in its architecture, which is shaped by the intriguing cross-linguistic investigations of Kita and Özyürek (2003). First, they showed that the availability of words for a particular concept in a language affects the gesture accompanying that concept communicated in speech. If there is no word for a concept in a language, the speakers of that language use fewer gestures in encoding that concept (compared to languages that have a word), which implies that gesture adapts itself to speech content. Second, they observed a relationship between syntactic packaging of motion event components (i.e., manner and path) and gesture types. Within the classic example "roll down", English interlocutors conflate components for manner (i.e., roll) and path (i.e., down) in one clause (e.g., it rolled down the hill), whereas Turkish interlocutors tend to express manner and path in separate clauses (e.g., it descended the hill by rolling). They found that this pattern was also paralleled in gesture. That is, English interlocutors also gesture in a way that encodes both manner and path at the same time. Conversely, in Turkish, interlocutors do not conflate these components in their gesture, and only gesture for one of the two. Both of these observations indicate that syntactic and lexical properties of speech have an effect on gesture production. The IH was mainly designed to accommodate such linguistic effects on gesture.

A simplified flow of production within this model is shown in Figure 2.8. Fundamentally, the model is not too different from the SM. One of the differences lies in the organization of the original conceptualiser components in Levelt's (1989) model. The conceptualiser in the IH is divided into two separate parts: (1) the *communication planner* which generates communicative intention, and distributes information load to each channel as in De Ruiter

(2000), and the *message generator* which prepares the pre-verbal message. One interpretation based on this split is that the *action generator* is not considered to be within Levelt's conceptualiser, but is instead a separate general mechanism. Note that De Ruiter's (2000) model places the origin of gesture production in the sketch generator (equivalent to the action generator here) within the conceptualiser. In the IH, the action generator is tasked with planning the contents of gesture not by itself, but by jointly using input from the communication planner, selected features of the representation from working memory, and feedback from the formulator through the message generator (the formulator operates as per (Levelt, 1989)). The message generator and action generator process the representations accessed through the working memory simultaneously and in coordination. The definitive feature that separates the IH from the previous models is the direct connection of speech and gesture planning components (i.e., the message and action generators) rather than occasional feedback mechanisms (cf. De Ruiter, 2000). That is, unlike the SM, the coordination of gesture and speech is maintained internally and is not entrusted to external auditory feedback. More importantly, there is also a bidirectional link between the formulator and the message generator, which is how the model allows linguistic formulation (i.e., grammatical and lexical) taking place in the formulator to configure gestural content in line with the findings in Kita and Özyürek (2003).

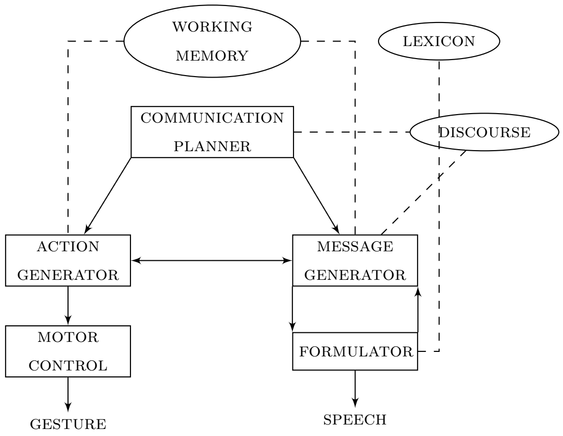


Figure 2.8: A simplified schematic of speech and gesture production in the Interface Hypothesis

The linkage that enables linguistic properties to shape gesture can also be used to explain temporal coordination phenomena such as the phonological synchronisation rule (McNeill, 1992). The phonological encoder within the formulator, which is tasked with creating a prosodic plan, has a mechanism for informing the action generator through the message generator (see Figure 2.8). This type of on-line information management can sustain the temporal relationships between phonological and gestural units. I will discuss this further in below and in Chapter 7.

Another strength of the model is that it describes the workings of the communication planner and its coordination with external information storages such as the *discourse model* and the *environment module* in good detail.[6] In the model, the discourse model tracks what has been communicated or not communicated in the conversation (amongst other things). The IH claims that gesture production is sensitive to this record, based on studies that showed that gesture can be used to refer to parts of the previous discourse (Bavelas et al., 1992) as well as to set a general frame (i.e., a preview) for the following discourse (Melinger & Levelt, 2004). These descriptions show that gesture is informed by the organization of information in the utterances, which is managed by information structure in speech. Therefore, although not explicitly stated, the model links gesture production to the information structure of speech, without commenting on a possible temporal coordination. I will discuss information structure as a modality that can be synchronised with gesture in Section 2.6.

The high interactivity of the IH makes it easier to accommodate the observed interactions related to the functions of gesture discussed in Section 2.2 as well as Kita and Özyürek's (2003) own findings. All models described so far focus on the origin of gesture and speech - where their co-production is initiated. Any linkage at the later stages of production seems to focus on

[6] The environment module is not described here. See Kita, 2014 for a description.

accommodating the specific findings presented by researchers. As previously noted, synchronisation has received relatively less attention in these models. I will describe the extent they were concerned with temporal coordination in Section 2.4.

2.4 Synchronisation

As we have seen in the previous section, the conceptual link in production derived from the association of these channels has been extensively represented in the speech-gesture production models. However, the synchronisation of speech and gesture has found very little coverage in the models.

To begin with, the GP theory is compatible with synchronisation. It suggests that the coordination of speech and gesture is not broken even if there is a disruption in one of the channels. This means that they are synchronised to the extent that they inform one another in case one runs into problems during production. However, there is not sufficient information on how this is managed in the theory - any type of synchronisation there may be between speech and gesture can only be attributed to the fact that they come from the same representation (i.e., common source). The LR theory briefly comments on gestural synchronisation from the perspective of gesture termination. Since it focuses on the lexical retrieval of words, the model needs a means to let the motor planner know that lexical retrieval has been completed in order to terminate gesturing. This is achieved via auditory feedback (i.e., the actual sound of the word) from the articulator to the motor planner (Figure 2.6). The model does not make any other predictions related to synchronisation. In the SM, the production paths of speech and gesture diverge after the conceptualiser stage (see Figure 2.7). If there is any synchronisation between speech and gesture, it has to be planned by the conceptualiser. The model is not able to account for the synchronisation of gesture with linguistic properties since these are encoded by the formulator after the conceptualiser stage. The IH does not make any prediction regarding synchronisation. However, bidirectional communication between the action generator and the formulator (through the message generator; see Figure 2.8) potentially enables a dynamic management of gesture timing both at early and late stages of production. As can be understood from these summaries, fundamental questions such as “what is synchronisation?”, “which units in speech and gesture are synchronised?” and “what kind of mechanisms are in charge of the synchronisation of gesture and speech?” remain unanswered.

One of the earliest definitions of synchronisation came from McNeill (1985, 1992). In his discussions of the degree of connection between the production of speech and gesture, he proposed three rules of synchronisation in support of the interconnection of these two channels (see Section 2.2.1). To recap, these rules are the following:

a. *Phonological Synchronisation:* Gesture stroke occurs before or at the same time as the stressed (most prominent) syllable.

b. *Semantic Synchronisation:* Co-occurring gestures and speech bear the same semantic content.

c. *Pragmatic Synchronisation:* Co-occurring gestures and speech have the same pragmatic function.

First of all, the term “synchronisation” is used loosely in these rules. It does not imply precise simultaneous co-occurrence at fixed times but rather proximity. This is especially clear for the pragmatic synchronisation rule in which no real temporal coordination is implied (i.e., it does not predict synchronisation). Instead, it describes the shared function of gesture and speech, assuming that they co-occur. Evidence for shared pragmatic functions was previously presented in Section 2.2 while discussing the communicative and facilitator roles of gesture; therefore, it will not be covered here again.

Similarly, the semantic synchronisation rule affiliates gesture and speech based on their semantic content without prescribing a temporal coordination pattern for their co-occurrence. However, some studies investigated this affiliation from a temporal synchronisation point of view. As previously described in Section 2.2.2, many studies reported that gestures generally precede their lexical affiliates (Butterworth & Beattie, 1978; Schegloff, 1984; Morrel-Samuels & Krauss, 1992; De Ruiter, 2000). More recent studies also supported this finding for different languages such as German, Mandarin, and French (Chui, 2005; Ferré, 2010; Bergmann, Aksu, & Kopp, 2011). The temporal precedence of gesture relative to its speech affiliate has come to be considered as evidence for synchronisation. However, other studies showed that the degree of synchronisation between gesture and its semantic affiliate was also not thought be very strong or stable since this synchronisation was reported to be affected by lexical familiarity (Morrel-Samuels & Krauss, 1992). This semantic coordination was also challenged by studies which manipulated gesture time. In a perception study, Kirchhof and Ruiter (2012) shifted the natural position of a gesture by 600ms (towards either way) and reported that even such long delays do not prevent the gesture's semantic integration with spoken material by interlocutors (also cf. Leonard & Cummins, 2011). This implies that the semantic relations between speech and gestural elements do not necessarily induce a strict synchronisation.

The precedence phenomenon has found some explanation in speech-gesture production models. The GP theory assumes that the precedence arises from gesture's being free from rigorous linguistic processing. In the LR theory, semantic features of gesture and speech are processed separately by their respective production systems without any coordination. The theory assumes that gesture precedes its lexical affiliate because accessing representations in the working memory for gesture planning takes a shorter time than for speech planning (Morrel-Samuels & Krauss, 1992). A precedence of gesture can also be observed in the phonological synchronisation rule in that the stroke can end before the stressed syllable but not after. McClave (1991) showed evidence for this rule in a study where she found that when multiple gestures were performed in quick succession, gesture phases within these gestures were compressed and “fronted” so that the gesture can end before a stressed syllable. Wlodarczak, Buschmeier, Malisz, Kopp, and Wagner (2012) showed that head gestures used for feedback purposes tended to precede their co-speech affiliate by 200ms on average. On the perception side, Leonard and Cummins (2011) reported that listeners identified synthesised beat gestures as mismatches only when they occurred 200ms later than their natural anchors, whereas beats occurring before their natural anchors were often not recognised as mismatches.

All the studies mentioned so far point out some form of synchronisation between speech and gesture channels at different linguistic levels, i.e., pragmatic, lexical, and phonological. However, the precise nature of speech-gesture synchronisation as well as the mechanisms behind synchronisation are still far from clear. In order to investigate these, there is a need for disambiguation about what the definition of synchronisation is and which segments or landmarks within speech and gesture streams are synchronised. Gesture itself has its own form and function related typologies each of which can show synchronisation with speech. On the linguistic side, there are different potential anchors with their own specific features (e.g., tones, syllables, words, phrases, constituents). It is unclear which of these segments are synchronised with gestural segments and whether other gestural and linguistic properties have any effect on synchronisation. Modelling a computational relationship between gesture and speech relies on a full-scale understanding of synchronisation between these since synchronisation behaviour can be used to validate, extend and establish speech-gesture production models.

To address these issues, the present study first has defined what it considers synchronisation is in Section 1.1. To reiterate, any systematic co-occurrence of two units is considered as synchronisation. This means that two units occurring simultaneously or with a fixed delay is accepted as synchronised so long as these occurrences are consistent. In determining these, the present study does not rely on vaguely defined concepts such as precedence or overlaps. Instead, the actual time distances between the units under consideration are calculated, and whether the calculated time distances indicate consistency is statistically tested. The tests used for this purpose are detailed in Sections 2.5.1.2 and 3.5.1. Secondly, the present study has also defined which components within gesture and speech are expected to be synchronised (Section 1.1). Gesture has been linked to prosody and information structure, and it has been hypothesised that atomic landmarks as well as phrasal constituents in gesture, prosody, and information structure can show synchronisation and other forms sensitivity to each other. Sections 2.5 to 2.6.2 discuss why prosody and information structure are relevant for the investigation of synchronisation with gesture while also specifying which constituents in these are expected to be synchronised. Relevant earlier studies are also reviewed in those sections.

2.5 Prosody as a Synchronisation Anchor

Chapter 1 established that prosody has been considered to have a very close relationship to gesture in many respects including synchronisation. In addition to earlier studies that dealt with precedence and McNeillean synchronisation, in the last two decades, there have also been studies that dealt with synchronisations between better defined gestural and prosodic anchors. The present study shares this general interest in a more precise understanding of synchronisation. Accordingly, Section 2.5 reviews these earlier studies critically inspecting their methodologies while also setting the ground for and introducing the research questions of the present study.

The section starts with this review, which provides an overall idea about what is known to date about the synchronisation of gesture and prosody while also making the connection between prosody and gesture clearer. As we will see, the majority of studies that investigated the synchronisation of gesture and prosody focused on the synchronisation of atomic landmarks within gesture and prosody (i.e., micro level), whereas the synchronisation of larger phrasal constituents in these modalities (i.e., macro level) has not received the same amount of attention. Accordingly, this review will cover the synchronisation of units at the micro and macro levels separately (Section 2.5.1 and 2.5.2). After the reviews of relevant studies, the research questions of the present study are introduced within the frame of Turkish prosodic structure.

2.5.1 Micro Level

Synchronisation at the micro level involves the investigation of temporal relations between the smallest units in gesture and prosody. Gesture and prosody are both analogue in the sense that information in these signals is represented by continuously varying features - they are continuous streams. In order to be able to determine whether these streams show synchronisation, certain anchor points within the flow of these streams must be identified. The measurements of time distance between these anchors then enable making predictions about synchronisation. Research on gesture and prosody had independently identified various points or chunks within their structure. However, as we will see, studies dealing with synchronisation have not made use of these points very consistently.

2.5.1.1 Previous studies

In earlier studies, the synchronisation of gesture and prosody was investigated between units associated with some form of peak effort in production (Schegloff, 1984). For example, Schegloff's (1984) qualitative investigation showed a tendency in which the maximum point of dynamic effort in beats (i.e., downbeats) aligned with the stressed syllable. As it will be presented in this section, later studies often tested the synchronisation of gesture with stressed syllables because stressed syllables are often considered to be produced with maximum articulatory effort (Barry & Andreeva, 2001). Note that such kinematic and prosodic peaks of effort manifest themselves as expressions of prominence (i.e., locally highlighted entities) in both modalities. Therefore, the main hypothesis behind these studies has been that prominent units in gesture and prosody are synchronised.

Defining prominence and selecting measurable cues of it have been a fundamental problem for studies on synchronisation. This has been especially true for prosodic prominence. Entire syllables (e.g., McNeill, 1992; Yasinnik et al., 2004) and various phonetic points measured over syllables have been used to make prominence related synchronisation claims (see Table 2.1). However, prominence, as invoked in these studies, is a phonological category which occupies a position in the prosodic structure (more on this in Section 2.5.1.2). The smallest unit that can bear any prominence is the entire syllable (not individual parts of it), which is an interval, not a point in time. Despite this, previous studies were interested in a more refined relationship between micro-level points in time in gesture and prosody, rather than a relationship between intervals. This led to the introduction of anchor points which are points in time measured to serve as synchronisation targets. However, we will see that the identification and selection of such points has been fundamentally problematic.

Table 2.1: A table summarising some of the earlier studies on synchronisation at the micro level

Study and Language	**Gesture Type**	**Anchor Definition**	**Prosodic/Phonetic Anchor**	**Finding**
Natural Speech				
Loehr (2004): English	All gesture types	Apex: the kinetic goal or target of a stroke	Pitch accents	Apexes synchronise with pitch accents
Yasinnik et al. (2004); Shattuck-Hufnagel et al. (2007): English	All gesture types containing abrupt stops, bouncy-jerky movements - "Discrete gestures"	Hit: abrupt stops of movement followed by direction change	Stressed syllables with pitch accents	Hits occur on pitch accents
Jannedy and Mendoza-Denton (2005): English	All gesture types	Apex: the kinetic goal or target of a stroke	Pitch accents	Apexes co-occur with pitch accents
Experimental				
Rochet-Capellan, Laboissière, Galván, and Schwartz (2008): Portuguese Brazilian	Deictic gesture	Apex: Finger-target alignment onset	Jaw openings on stressed syllables	Apexes synchronise with jaw openings
Roustan and Dohen (2010): French	Deictic and Beat gesture	Apex: index finger extension, beat endpoint	F0, intensity and articulatory movements	Deictic apexes synchronise with articulatory vocalic targets, beat apexes synchronise with prosodic peaks
Leonard and Cummins (2011): English	Beat gestures	Apex: point of maximum extension	Vowel onsets, pitch peaks on stressed syllables	Apexes synchronise with pitch peaks
Rusiewicz (2010) and Rusiewicz, Shaiman, Iverson, and Szuminsky (2013): English	Deictic gesture	Apex: endpoint of gesture stroke	Vowel midpoint	No synchronisation

Most studies represented in Table 2.1 used pitch accented syllables in their investigation because pitch accents have long been considered as the main correlates of prominence (although not necessarily so for every language) in line with definitions such as:

> A pitch accent may be defined as a local feature of a pitch contour - usually but not invariably a pitch change, and often involving a local minimum and maximum - which signals that the syllable with which it is associated is prominent in the utterance.[7]
>
> Ladd, 2008, p. 48

Other studies assumed very phonetic definitions of what prominence is by characterising it with single acoustic parameters such as articulatory vocalic targets or jaw openings without providing insight into why those parameters were chosen (see Table 2.1). This view is in conflict with prosodic research which points out a complex relationship of acoustic cues to prominence where multiple cues are shown to contribute to the perception of prominence (see Breen, Fedorenko, Wagner, & Gibson, 2010; Cole, 2015; Arnhold & Kyröläinen, 2017; Kügler & Calhoun, in press; Baumann & Winter, 2018. The cues to prominence depend on multiple factors such as language in question and position in prosodic structure (see Section 2.5.1.2). Accordingly, if particular points in time are needed to be measured for synchronisation tests, this measure should not ignore prosodic structure and how the acoustic measure is used in the language in question. In fact, some recent work on synchronisation showed that the prosodic structure shapes the patterns of synchronisation as the positions of gestural anchor points and prosodic anchor points were found to change in tandem depending on the prosodic phrasing (Esteve-Gibert & Prieto, 2013; Esteve-Gibert, Borràs-Comes, Asor, Swerts, & Prieto, 2017) (these studies are detailed later in this section). Also, in terms of the usage of acoustic cues, Rohrer, Prieto, and Delais-Roussarie

[7] This should not be confused with "lexical pitch accents" which refer to lexically specified pitch features in languages such as Norwegian.

(2019) showed that in French gesture anchors are synchronised with pitch accents at much lower rates compared to the rates reported in studies in English (Shattuck-Hufnagel & Ren, 2018). They reported that when pitch accents are absent, gesture anchors may be synchronising with the onsets of prosodic phrases acting independently from the encoding of prominence. Although their analysis is inconclusive, the study shows that the prosodic structure and its constituents may influence gesture synchronisation. These highlight why the selection of prosodically informed acoustic measures is important for synchronisation. From this perspective, the present study determined its prosodic anchor points as F0 turning points (i.e., tonal events, see Section 2.5.1.2). These are seen as the most reliable cues to prominence because they are sensitive to both the prosodic structure of language and how acoustic cues are used per language, and they have been consistently shown to align within syllables in different languages (Ladd, 2008). Consequently, the present study adopts tonal events as anchor points against which gesture synchronisation is tested. What these events are in Turkish prosodic structure (and Turkish prosodic structure itself) is explained further in Section 2.5.1.2.

In contrast with prosody, the units of measurement have been more consistent for gesture. The studies agreed on what constitutes a prominent dynamic point in gesture (i.e., an apex) although the way this term is defined was not uniform (see Table 2.1). Loehr (2004) defined a gesture apex as "the kinetic goal of the stroke" (p.89), which he identified differently for different gesture types depending on their directionality. In uni-directional gestures, the apex is identified as the endpoint of the stroke; and in multi-directional gestures, apexes are points where directional changes occur in the stroke. Similarly, in other studies in Table 2.1, the apex definitions revolved around *endpoints* or *direction changes* in parallel to the gesture types investigated. Deictics and beats have simpler kinematics and are often uni-directional. Therefore, the endpoints, where the hand completes its extension in these gestures, can easily be identified as the point that the gesture wants to reach, i.e., the prominent target (Rochet-Capellan et al., 2008; Roustan & Dohen, 2010; Leonard & Cummins, 2011; Rusiewicz et al., 2013). A broader account of kinetic prominence was needed to identify apexes in iconics and metaphorics. The representational nature of iconics and metaphorics can affect the kinematic complexity of gesture in that multiple directional changes can occur over the stroke in order to represent the desired features of an entity, making each direction change prominent (Loehr, 2004; Yasinnik et al., 2004; Jannedy & Mendoza-Denton, 2005). For instance, direction changes would be prominent in a gesture where the index finger traces four corners in the air to represent the rectangle shape of a table.

Despite the differences in the anchor points used across these studies, most studies reported a synchronisation between their chosen anchors. Loehr (2004) investigated the synchronisation of pitch accents and apexes of all gesture types produced by interlocutors engaged in natural conversations in English. The study employed a synchronisation criterion, ±275ms, i.e., apexes were only considered to be synchronised with the nearest pitch accent only if they occurred within 275ms of each other. This criterion was decided based on the average distance between the nearest gestural markings and prosodic markings (regardless of type). The study reported "a significant tendency" of apexes synchronising with pitch accents over other prosodic events 75% of the time with an average time distance of 17ms (sd=341ms). The study did not report such time measurements for the cases where the apexes were synchronised with "non-pitch accents" or describe what those tonal events were. In addition, there was no report on whether the synchronisation behaviour showed any difference depending on gesture type.

Using samples from academic lectures in English, Yasinnik et al. (2004) investigated the synchronisation of pitch accents and *hits* in *discrete gestures*. The presence of a hit defines discrete gestures (i.e, a gesture is discrete if it has a hit). A hit is "an abrupt stop or pause in movement, which breaks the flow of the gesture" appearing as "bouncing, jerky movements, changes in the direction of movement, or as complete stops in movement" (p.98). The study reported different findings depending on the method they used for the annotation of hits. When hits were annotated while also listening to speech, it was found that 90% of hits occurred on a pitch accented syllable. When the annotator did not listen to speech whilst annotating, only about 50% of the annotated hits occurred on pitch accented syllables. The study also looked at such overlaps at the word level in the latter annotation condition and reported that 90% of polysyllabic words that overlapped with a hit also contained a pitch accent (i.e., hits did not overlap with pitch accents but were within the same word). This percentage was 65% for monosyllabic words. These findings imply a methodological influence where the perception of pitch accents affected the annotation of hits. More importantly, the study showed evidence for some form of synchronisation. Hits did not necessarily take place during the pitch accented syllables themselves but tended to take place somewhere during words that had a pitch accent. These findings highlight the importance of defining a synchronisation criterion. The gestural and prosodic anchors do not have to strictly overlap in order to show synchronisation. Systematic occurrences of the two regardless of their distance can also indicate synchronisation. It is difficult to make further assumptions about the nature of synchronisation in Yasinnik et al.'s study (2004) because it did not report actual time measurements tested against an indicative but not absolute synchronisation criterion. It also seems that the overlapping of hits with pitch accents was sensitive to the number of syllables in a word. What the cause of such relationship may be was not covered in the study. The study also did not discuss whether the semantic functions of gestures (i.e., gesture types) had an effect on the patterns they observed, or what other prosodic events were available around the hits when they did not overlap with the pitch accents (see Section 2.5.1.2 for possible tonal events).

Jannedy and Mendoza-Denton (2005) analysed different aspects of the gesture-prosody relationship in English multimodal data excised from a town hall meeting. One of these aspects was the synchronisation of pitch accents and apexes (adopting Loehr's 2004 definition). Their analysis was qualitative and only focused on four brief segments of speech. They did not present any time measurements or establish any clear definition of synchronisation. They only counted the number of times where the interlocutor produced pitch accents as well as apexes in their utterances. Their finding showed that 95.7% of apexes "co-occurred" with pitch accents. It is unclear what the co-occurrence claim was based on as there were no measurements of time distances or a synchronisation frame such as overlapping over a syllable or word duration as in Yasinnik et al. (2004). The authors also highlighted the importance of investigating natural speech and gesture data compared to data elicited in experimental settings. However, their data came from a public speaking performance that may or may not be heavily rehearsed, which can influence the naturalness of speech and gesture production, much like gestures elicited in an experimental setting.

In experimental studies, various prosodic and gestural anchors were used. For example, Rochet-Capellan et al. (2008) investigated the synchronisation of deictic gesture apexes and jaw openings (i.e., the maximum height of jaw opening on vowels) in a task where speakers of Brazilian Portuguese had to point at a target they saw on a screen while saying the corresponding disyllabic words which only differed in their stress position (i.e., either first or second syllable was stressed). The participants received prior training about reading stress (as marked on words) and were instructed to produce gesture and words simultaneously. The study measured the onset and offset timings of *pointing plateau* (i.e., the duration in which the finger stayed on the target) as well as the timing when the jaw reached its maximum height. The pointing plateau corresponds to the post-hold phase described in Section 2.1.1 in which the hand remains frozen after the completion of the stroke phase.

In turn, this means that the onset of the pointing plateau corresponds to the apex in Loehr's (2004) terms because the endpoint of the stroke is the apex in deictic gestures. Their findings showed that when the stress was on the first syllable, apexes were synchronised with the jaw openings of stressed syllables. On the other hand, when the stress was on the second syllable, apexes occurred midpoint between two jaw openings, but significantly closer to the jaw opening of the second syllable compared to the first-syllable stress condition. This resulted in a different synchronisation pattern - apexes were synchronised with the onset of the jaw movement (not the maximum height) in the second syllable stress condition. Note that the study did not define a synchronisation criterion (e.g., anchors must be within x distance to be considered synchronised). The synchronisation was assumed based on the significant difference between the measurements in the two stress conditions. Regardless of what the jaw movement anchor was (i.e., the maximum opening or onset of jaw), the apex was found to be sensitive to the stress position of words under the experimental conditions, which confirmed the hypothesis that prominences are synchronised.

One striking finding was that the duration of the post-holds was significantly longer in the second-syllable stress condition than in the first-syllable stress condition. The hand waited to retract to a rest position until the realisation of the jaw opening on the stressed syllable was completed. That is, the post-hold phase was maintained until the execution of the segment that was bearing prominence. This lengthening implied that gesture adapted to the changes in stress position. Moreover, the lengthening constitutes the first hint in the literature suggesting that apical prominence stored in the post-hold (the hand is frozen at the apex) is sensitive to prominence in the speech signal, and that the post-hold phase is used to ensure that the production of prominences synchronise. The apical prominence in post-hold phases is a key point in the findings of the present study and is discussed extensively in Section 6.4 and Chapter 7.

Rusiewicz (2010) and Rusiewicz et al. (2013) had very similar designs to Rochet-Capellan et al. (2008) (pointing + naming task but in English) in which they analysed the effects of prosodic stress and its position (first or second syllable) on the synchronisation of deictic apexes and vowel midpoints as well as on the synchronisation of gesture and word onsets. However, their analysis resulted in different findings. They reported that gesture apexes were not synchronised with the midpoints of stressed vowels which they considered to be "the most consistent acoustic correlate of prosodic stress" (Rusiewicz et al., 2013, p. 462). In fact, for example, in the first-syllable stress condition, the distance between apexes and the midpoints of stressed vowels were significantly longer, contrary to the expectation. In contrast, gesture and word onsets showed sensitivity to stress - the distance between these anchors was significantly shorter under the stress condition. In other words, they found that gesture apex and vowel midpoints were not synchronised. Instead, the onsets of gestures and the onsets of their affiliated words were observed to show better synchronisation. Their findings contradict the findings of most studies on synchronisation at the micro level.

One potential similarity between these findings and those of Rochet-Capellan et al. (2008) was that in the second-syllable stress condition, gesture and word onset time distances were the shortest, and more importantly, total gesture time was the longest. Although the study did not report whether gestures ended before or after the realisation of the stressed syllable, the increase in gesture time could be related to the maintenance of the post-hold phase until the stress in the second syllable is realised as in Rochet-Capellan et al. (2008). Such a finding would confirm the post-hold phase's role of managing the synchronisation of prominences in a different language.

Esteve-Gibert and Prieto (2013) used a similar design to Rusiewicz (2010) and Rochet-Capellan et al. (2008) (i.e., manual pointing + naming) where gesture production is heavily controlled, and they tested whether apexes were synchronised with pitch accents and whether this synchronisation is sensitive to the prosodic structure. As predicted, they found that apexes co-occurred with the intonational peaks in accented syllables. They also found that the time distance between apexes and intonational peaks were affected by the presence of an upcoming prosodic phrase offset (i.e., phrase boundary). It was reported that the intonational peaks within accented syllables are temporally retracted when there is a phrase boundary upcoming immediately after the peak (Prieto, Van Santen, & Hirschberg, 1995). Their results showed that the anchoring of apexes behaves in the same way. If the intonational peaks were retracted because of an upcoming phrase boundary, the anchoring position of apexes was also retracted. These results were also found to be consistent for head gestures (Esteve-Gibert et al., 2017). Overall, similar to Rusiewicz (2010) and Rochet-Capellan et al. (2008), Esteve-Gibert and Prieto (2013) did not define what constituted synchronisation precisely (although they seem to assume that apexes would occur before the end of accented syllables as per the phonological synchrony rule in McNeill, 1992). Rather, they tested whether the positioning of selected anchors is affected by prosodic structural constraints.

Roustan and Dohen (2010) observed different synchronisation patterns in an experimental task in French. Participants listened to audio files where two speakers produced SVO sentences differing on one element only such as: (S1) “Baba holds the baby” and (S2) “S1 said Mumu holds the baby?”. The participants were instructed to verbally correct the statement in S2 while also pointing at the correct pictorial description of the statement appearing on a screen. In a separate condition, they were instructed to perform a beat gesture (instead of pointing) as they produced the correct utterance. This design aimed to elicit sentence level prosodic prominence on subject and object separately in order to test whether sentence level prosodic prominence attracted gestures, and whether there was synchronisation between apexes (maximum extension and downbeats, i.e., endpoints) and various prosodic anchors measured over the prominent constituent such as F0 and intensity peaks, duration, and peaks of amplitude of lip opening and protrusion (articulatory targets). The study considered apexes synchronised with the prosodic anchor that was the closest.

The study claimed that sentence level prominence attracts gesture in the sense that gestures and sentence level prominence occur “close” to each other. No comparisons with other possible attractors were made to justify whether this closeness was meaningful or consistent, however. They observed that apexes were synchronised with different anchors depending on gesture type. Deictic apexes were synchronised with articulatory vocalic targets, whereas beat apexes were synchronised with F0 and intensity peaks on prominent constituents. They did not comment on why synchronisation showed such variation between gesture types. They also reported inconsistent synchronisation cases where apexes were synchronised with different prosodic anchors depending on the position of sentence level prominence (i.e., subject or object). Yet, the study, at no point, commented on why there were inconsistencies between subject and object prominence conditions although these were controlled as a part of their experimental design. The inconsistency they observed could be to do with the differences in realisations of subject and object prominence in French, which was not considered.

Leonard and Cummins (2011) set out to find the most precise anchor for the synchronisation of beat gestures in a highly controlled experiment in English. Participants read texts while standing up and holding their arm against their chest. They were instructed to perform beat gestures on preselected stressed syllables as they read. These gestures shared the same form because the arm was expected to extend outwards from the same initial position and return to its original position, which caused rather large and durationally long beats (1 second on average). Such beats are possible but not typical (McNeill, 1992). The study designated three prosodic/phonetic anchors which were the vowel onset, perceptual center (estimated based on another study of one of the authors), and pitch peak of a stressed syllable; and five gestural anchors which were the onset and offset of movement, point of maximum extension, and peak velocity of hand extension and retraction. The distances between these anchors were calculated and tested against each other. Their findings indicated that the alignment with these prosodic/phonetic anchors were most consistent for the maximum point of extension. This endpoint of the stroke is considered to be the apex for beat gestures as per Loehr (2004) and other studies mentioned so far. In addition, it was observed that these apexes were performed approximately 200ms after the vowel onsets, synchronising with the pitch peaks most tightly.

As can be seen in the studies reviewed so far, there is variation in the methodologies employed. Each methodology has its own advantages and disadvantages. Common methodological shortcomings in these studies can be summarised as follows:

i. *There was no clear definition of what constitutes synchronisation.* This issue was more evident in studies that investigated natural speech with the exception of Loehr (2004) whom I credit for giving me insight about how synchronisation should be framed. Qualitative observations of co-occurrences and overlaps are useful in general but they lack precision. In experimental studies, the definition of synchronisation relied on time distance measurements under different conditions. That is, if the time distances between gesture apexes and prosodic anchors were significantly less in one condition and more in another, then the reduction in measured distance was considered as synchronisation. How much actual time difference there was between these conditions was often left out. A significant difference between two conditions might be caused by, for example, 20ms of actual time difference between anchors under consideration in those two conditions. In such cases, would the significant difference between conditions be indicative of a synchronisation relationship when the actual average difference between anchors is so little? How far away should two anchors be so as not to be considered synchronised (and vice versa)? These questions so far remained unanswered in the literature. Studies on synchronisation need a standard timeframe to base their synchronisation interpretations on. This frame must also be tolerant of this type of minor timing differences but also indicative enough to represent a meaningful timing relationship between anchors.

As previously described, Loehr (2004) proposed 275ms, the average distance between any gestural and prosodic markings in his data, as a short enough distance between anchors to assume synchronisation. Namely, he considered two anchors synchronised if the distance between them was less than the average distance between any prosodic and gestural markings, regardless of type. However, assuming synchronisation based on such a criterion is fundamentally circular as the criterion itself is not independent of annotations. The present study uses average syllable duration in order to establish synchronisation between anchors. This duration was selected first because it is phonologically meaningful as it is the smallest phonological unit that carry any prominence (i.e., the main concept that synchronisation has been built on). Secondly, this duration is estimated to be long enough to tolerate effects that cause slight asynchronies between anchors; and at the same time short enough to describe fine synchronisation patterns (see Chapter 3 for how this duration was used in statistical tests).

ii. *The conditions used in some of these studies were highly constrained in order to meet the experimental demands.* The elicitation of gesture from participants in particular can be considered as a major point of argument against the relevance of findings of these studies. Participants in such studies in Table 2.1 were instructed to gesture as well as to align their gestures with their verbal productions which were only single sentences in isolation. For instance, Leonard and Cummins (2011) heavily controlled gesture productions - on which syllables the participants should perform their gestures, and the form of their gestures were pre-determined. In contrast, gestures that humans perform and observe everyday are typically produced unselfconsciously with clear intentions. On the subject of intentions, experimental gestures also lack communicative intent which was claimed to be an integral element of gesture production (see Sections 2.2.1 and 2.3). In the studies summarised above, participants gestured at a target on a screen because they were told to do so without the presence of an addressee, highlighting that these movements were produced only for the sake of movement in a ritualised manner. Novack and Goldin-Meadow (2017) distinguish gestures from other bodily actions based on the goal of movement, stating that gestures are produced to accomplish representation of information and communication, whereas actions are produced to achieve communication-external goals (e.g., hair combing). Within these definitions, experimentally elicited gestures seem to be closer to bodily actions than they are to gesticulations.

Another important point to note is that in these experimental designs, participants were highly aware of the fact that their gesture timing was under investigation due to the instructions they received. For instance in Rochet-Capellan et al. (2008), the participants even received prior training about stress and how to read stress markings on words. Therefore, they could establish the focus of that particular study to be about stress and gesture, and adapt their productions according to what they think was expected in the experiments. Leonard and Cummins (2011) reported a participant's comments on the nature of the task saying: "... (beat gesture placement) resembled the task of placing stress, focus or accent in a required point within a sentence" (p.1469). Within their study, this type of awareness would not make a difference as the participant was told when to gesture in the first place. However, this shows evidence for the fact that participants were able to establish that such tasks were about some form of prominence relationship. Therefore, any synchronisation between prominent markers could only show that participants could consciously detect prominences in both channels, and adapt their productions so that these are close to each other in a way they conceive to be appropriate. Overall, the findings on experimental and natural gestures may have reached the same conclusions about the synchronisation, but possibly for different reasons. For all these reasons mentioned, experimentally elicited gestures make poor alternates for naturally occurring ones and lack generalisability because of this. As also stated as a limitation in most of these experimental studies in Table 2.1, if the aim is to capture actual synchronisation relationships between gesture and prosody, natural speech data should be investigated.

iii. *These studies investigated different gesture types.* The experimental studies concentrated on deictics and beats, whereas the studies that used natural speech data did not make any distinctions between gesture types and included iconics and metaphorics in their analyses as well. However, these studies using natural speech data did not report any observations or statistical analyses showing whether and how synchronisation patterns differed depending on these gesture types. For the experimental studies, the decision to analyse only deictics and/or beats is likely to be motivated by the relative ease of the identification of apexes since they mainly used motion capture technologies for annotation purposes. Regardless, only Roustan and Dohen (2010) analysed both of these and reported different synchronisation anchors for deictic and beat apexes. The studies that investigated beat apex synchronisation agreed on the synchronisation anchor, i.e., pitch peaks. However, the findings on deictic apex synchronisation were less consistent (see Table 2.1). Based on different within/across groups findings on deictic and beat apexes, it is possible that iconic and metaphoric apexes may also exhibit distinct synchronisation behaviours with different anchors. This seems even more likely when we consider the different definitions of the apex for iconics and metaphorics compared to deictic and beats (i.e., direction change versus endpoint). Overall, there is no systematic analysis showing how the apexes of each gesture type are synchronised with prosodic anchors within the same data set. Therefore, it is currently unclear whether the semantic function of gesture has an effect on gesture-prosody synchronisation.

iv. *These studies did not describe the prosodic structure of the languages that they investigated.* That is, there are various descriptions of prosodic/phonetic anchors, but where these anchors existed in the larger prosodic context remained unexplained. For instance, some of these studies focused on pitch accents on stressed syllables. Pitch accents are just one type of tonal events taking place in the prosodic stream - there are others serving different functions. Based on the general assumption that prominences are synchronised, only pitch accents were examined, and the other tonal events were not even considered as potential synchronisation anchors. Moreover, pitch accents and other tonal events exist as members of prosodic phrases and are affected by the feature of these phrases. For example, it is possible to categorise prosodic phrases depending on their relative position to sentence level prosodic prominence. Depending on their position, prosodic phrases exhibit different pitch trends affecting the realisation of pitch accents as well as of other tonal events. Potentially, the synchronisation of atomic units can be affected by the prosodic context these units inhabit. Such an effect has not been investigated in synchronisation studies before. Tonal events, prosodic phrases, and other prosodic context related concepts are explained within Turkish prosodic structure in Section 2.5.1.2 and Section 3.4.2.

Broadly speaking, these studies produced similar findings confirming the prominence-based synchronisation hypothesis. However, they used different methods which had their own shortcomings. These common problems were summarised in the list above. The present study addresses all these concerns in its design. First, as mentioned above, the study adopts a well-defined synchronisation criterion which is statistically tested (see Section 3.5.1 for details). Second, to address potential issues that may arise from the artificial nature of experimentally elicited gestures, the study uses natural speech and gesture data elicited in a meaningful task in the presence of an addressee. The data elicitation and other relevant information regarding the design of the study is detailed in Sections 3.2 and Section 3.3. Third, all gesture types (iconics, metaphorics, deictics and beats) are included in this study. Whether synchronisation is sensitive to gesture types that apexes are in is statistically tested (see Section 3.5). Finally, the study also tests whether prosodic context has any effect on synchronisation. The prosodic structure of languages varies meaning that prosodic contexts that can affect synchronisation are language specific. Since this study investigates synchronisation in Turkish, the next section describes the prosodic structure of Turkish while also placing the research question of this study within the frame of Turkish prosody.

2.5.1.2 Investigation of micro level synchronisation in the present study

Section 2.1.1 presented a hierarchical phrasal structure for gesture organised around a prominent unit while also describing kinematic features that distinguish these phrases from each other. We will see in this section that prosody is similar to gesture as it also consists of a hierarchy of phrases. This structural similarity between gesture and prosody makes it possible that synchronisation could be observed between corresponding events and phrases at every level within these hierarchies. This section focuses on synchronisation at the micro level. It defines which anchors in gesture and prosody are tested for synchronisation and frames the research questions within Turkish prosodic structure.

As seen in Section 2.5.1.1, the way in which anchors are defined is crucial for the establishment of meaningful synchronisation relationships between gesture and prosody. The present study diverges from previous studies in how it determines these anchors. It does not pre-select two anchors in gesture and prosody and test synchronisation just between these two anchors. Instead, it defines a prominent gestural anchor and aims to find the best synchronisation anchor amongst a set of prosodic events that have well-defined prosodic functions (i.e., including a prominence-lending function) within Turkish prosody.

First of all, the gesture apex is selected as the gesture anchor to be tested for synchronisation. The present study adopts Loehr's (2004) definition which is applicable to all gesture types. A gesture apex is "the kinetic goal of the stroke" (p.89). It constitutes the target that gesture aims to reach, and is therefore considered to be dynamically prominent. Apexes are typically identified as the endpoint of the stroke (common in deictics and beats) or as points in time at which directional changes occur in the stroke (common in iconics and metaphorics) (see Section 3.4.1.3 for a detail description and annotation guidelines). It is likely that differences in the definitions of apex can affect the patterns of synchronisation found between apexes and prosodic anchors. This possible effect is accounted for within the present study where synchronisation of apexes is compared across all gesture types (i.e., iconics, deictics, metaphorics, and beats) through regression modelling (see Section 3.5).

In Section 2.5.1.1, it was shown that the selection of a prosodic anchor has been problematic in previous studies. Various acoustic cues during prominent syllables have been measured, with the assumption that they cue prominence. However, relevant research on prosody has shown that multiple acoustic cues contribute to the perception of prominence. It has also been shown that these cues cannot be expected to be the same for every syllable, regardless of factors (1) position in prosodic structure, (2)how those acoustic cues are used in that language (Breen et al., 2010; Arnhold & Kyröläinen, 2017; Baumann & Winter, 2018; Kügler & Calhoun, in press). Therefore, the prosodic anchors that are tested for synchronisation should be sensitive to these factors. The present study uses F0 minima and maxima (i.e., F0 turning points or tonal events) as prosodic anchors. These have been shown to be sensitive to both of these factors, and to be consistently aligned within syllables across different languages (Prieto et al., 1995; Arvaniti, Ladd, & Mennen, 1998; Xu, 1998; Atterer & Ladd, 2004; Prieto & Torreira, 2007; Ladd, 2008).

F0 turning points have different functions within the prosodic hierarchy. The prosodic hierarchy is constructed by syllables coming together to form prosodic words which in turn group into larger prosodic phrases (similar to the gestural hierarchy of phrases in Section 2.1.1). The inventory of prosodic phrases and tonal events is language specific. Prosodic phrases are related to but not isomorphic with morphosyntactic phrases, i.e., prosodic structure reflects morphosyntactic structure imperfectly. Within the Autosegmental-Metrical framework (AM) of intonational phonology, these phrases in prosodic hierarchies are defined based on intonational properties (although not purely) (Pierrehumbert, 1980; Ladd, 2008). In this framework, intonation can be considered as the linguistically controlled and pragmatically meaningful use of pitch. Pitch is employed as a sequence of high (H) and low (L) tonal targets, i.e., F0 turning points (see Figure 2.10. These tonal events either mark prominence (i.e., pitch accents) or mark the boundaries of prosodic phrases (i.e., phrase accent or boundary tone).

Turkish

In Turkish, there are three prosodic phrases, which are: (1) prosodic word (PW) (≈ morphological word), (2) intermediate phrase (ip) (≈ syntactic phrase), and (3) intonational phrase (IP) (≈ syntactic clause) (Kamali, 2011; Ipek & Jun, 2013; Güneş, 2015). Each of these is marked by specific tonal

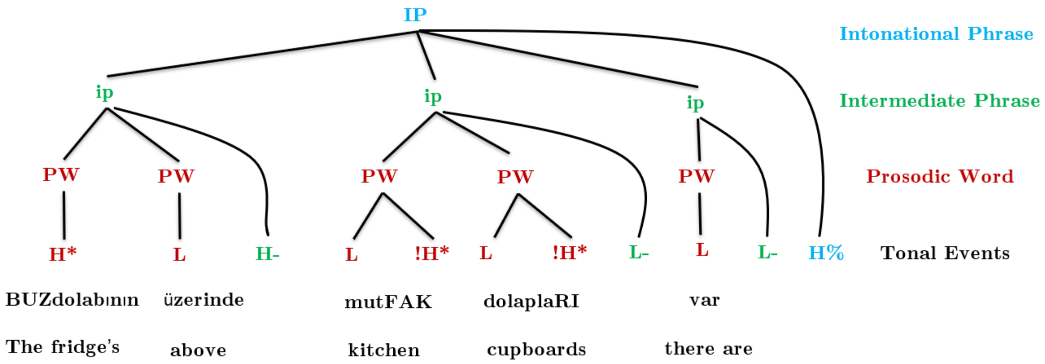


Figure 2.9: A schematic for the prosodic structure of Turkish shown on the utterance "There are kitchen cupboards above the fridge". Capital letters indicate stressed syllables. Prosodic phrases and tonal events associated with these phrases have the same colour.

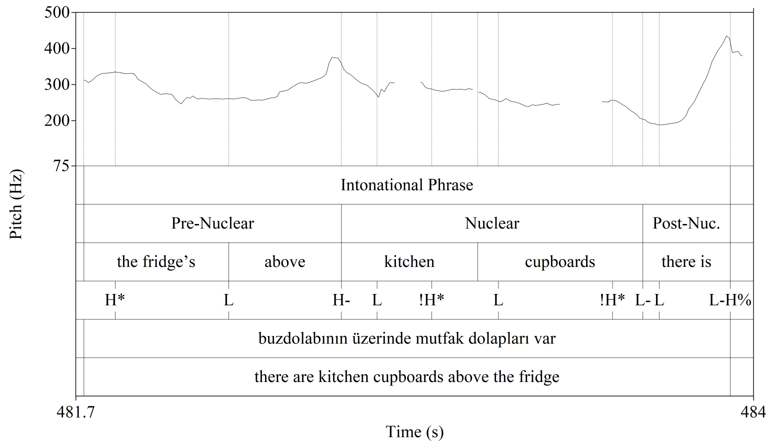


Figure 2.10: The pitch track of the utterance in Figure 2.9 showing example tonal events and prosodic phrases. The first four annotation tiers show IP, ips, PWs and tonal events respectively.

events. The IP is the largest phrase in the hierarchy, and it is marked by either high or low boundary tone at the right edge, indicated by %, i.e., L% or H%. An IP contains at least one ip. Similar to the IP, the ip is marked at the right edge with a phrase accent, indicated by -, i.e., L- or H-. PWs, on the other hand, are marked with a L tone at their left edge. The only other tonal event within the PW is the pitch accent (H*, !H*, or L*). Pitch accents are associated with stressed syllables of PWs and mark prominence. A basic schematic for the prosodic structure of Turkish is shown in Figure 2.9. A more detailed account is presented in Section 3.4.2.

It can be seen in Figures 2.9 and 2.10 that the phrase accents and pitch accents found within certain ips can be realised differently. For example, the first ip in the utterance (*buzdolabinin* 'the fridge's above') is marked at the right edge with a H-, whereas the following ip is marked by a L-. This difference in tonal marking is due to the position of ips relative to the ip that contains sentence level prosodic prominence, i.e., the nuclear ip ('kitchen cupboards' in Figure 2.10). Therefore, nuclear ips are perceived to be central prominent unit. In Turkish, ips occurring before the nuclear ip, i.e., pre-nuclear ips, are usually marked with a H- phrase accent, whereas nuclear and post-nuclear ips are usually marked with a L- (Kamali, 2011; Ipek & Jun, 2013). This difference in tonal marking shows how these tonal events are sensitive to the organisation of prosodic structure. This sensitivity has not been reflected in studies on gesture-prosody synchronisation - it is unclear whether apex synchronisation with prosodic anchors would show any difference depending on the relative position of phrases to the nuclear constituent.

Figure 2.10 shows that in Turkish, there can be multiple prominence-lending and boundary marking tonal events occurring close to each other. For the reasons explained in Section 2.5.1, studies on gesture-prosody synchronisation have generally concentrated on prominence-based synchronisation. This means that these studies isolated stressed syllables in the continuous event-rich prosodic signal and only checked whether apexes are synchronised with acoustic measures taken from these syllables when isolated from their prosodic context (cf. Loehr, 2004). Because of this, other possible synchronisations with anchors such as the tonal events that mark boundaries have been ignored, which creates a bias in the analysis of synchronisation. It is unclear whether gesture apexes will stay synchronised with prominent pitch accents when other types of tonal events are included in the analyses.

Turkish presents an ideal prosodic context to interrogate prominence-based synchronisation for two main reasons. Firstly, prosodic phrases in Turkish are tonally crowded in that there can be multiple tonal events occurring close to each other within short durations of PWs (PW initial L tone, pitch accent, and a phrase tone; see Figure 2.10). Systematic synchronisation with one of these tonal events over others, especially when there is little time difference between them, would indicate a systematic selection of particular tonal events for apex synchronisation.

Secondly, there can be prosodic phrases without a pitch accent in Turkish, and in such cases there are therefore no tonal events that encode prominence. For example, post-nuclear ips, by definition, do not contain pitch accents (see 'there is' in Figure 2.10). In addition, words with regular stress are also claimed by some researchers to be accentless in Turkish (Kamali, 2011; Güneş, 2015). Regularly-stressed words in Turkish are where the lexical stress is on the PW final syllable, whereas in irregularly-stressed words the lexical stress is on a non-final syllable (Sezer, 1981; Kabak & Vogel, 2001).

In pre-nuclear ips, stressed syllables are typically marked with a H* (Kamali, 2011; Ipek & Jun, 2013). If the final word in a pre-nuclear ip has regular stress, the word-final H* and the H- which marks the right edge of the ip coincide (see 'then' in Figure 2.11). In such cases, it is unclear what status is of the rise in pitch (to the H tone), i.e. whether it is a part of the pitch accent or the phrase accent.

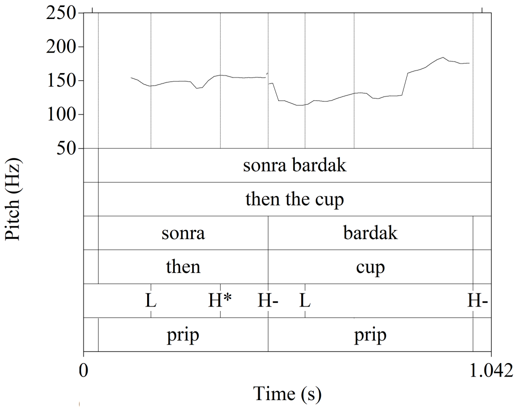
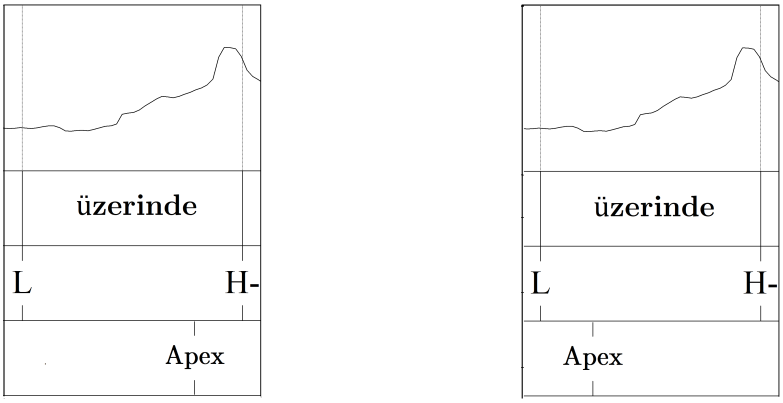


Figure 2.11: The pitch track of two pre-nuclear ips (i.e., prips) first of which has a pitch accent and the second does not

There are two views regarding this issue. Ipek and Jun (2013) suggests that the H tone functions both as a part of the pitch accent and as part of the phrase accent. On the other hand, Kamali (2011) and Güneş (2015) claim that the H tone is a property of the ip only, which in turn proposes that words with regular stress are accentless in Turkish. Research on this issue seems to be inconclusive so far (see Altmann, 2006; Domahs, Genc, Knaus, Wiese, & Kabak, 2013). The present study allows both accented (as in 'then' in Figure 2.11) and accentless phrases (as in 'cup' in Figure 2.11 and above in Figure 2.10). The reader is referred to Section 3.4.2.2 for details about how the distinction between these phrases was made in the present study.

Prominence-based synchronisation of gesture and prosody cannot predict synchronisation in cases where there are no pitch accents since the prosodic structure has no prescribed prominent anchor. For instance, in the case of a gesture accompanying the word *uzerinde* 'above' in Figure 2.10, there is no prominent pitch accent with which the apex can be synchronised - it is therefore unclear which tonal event the apex will show consistent synchronisation with. The synchronisation options in such cases are limited to the PW initial L and the H- phrase accent. Synchronisation with either of these tonal events has implications for both gesture-prosody synchronisation and can shed light on the accentlessness of words with regular stress. Assuming that prominence-based synchronisation also applies to Turkish in general, and apexes are synchronised with pitch accents when they are present in prosodic phrases, there are two possibilities:



(a) H- phrase accent

(b) L

Figure 2.12: Two hypothetical gesture apex synchronisation cases with H- or L the word *uzerinde* ‘above’ in Figure 2.10

1. If apexes are synchronised with H- (see Figure 2.12a), this implies that the H tone may actually be a part of the pitch accent as well as the phrase accent, supporting Ipek and Jun’s (2013) double function claim. The presence of a prominent pitch accent at the location of the rise attracts apexes.

2. If apexes are synchronised with PW initial L tones instead (see Figure 2.12b), this means that the H tone does not function as a part of a pitch accent, but only marks the end of the ip as a part of the phrase accent as claimed by Kamali (2011) and Güneş (2015). In other words, if there was a pitch accent at the location of the H tone, then the apex would be synchronised with that tonal event as dictated by the prominence-based synchronisation theory. The observation that apexes are systematically synchronised with PW-initial Ls would imply that synchronisation is not managed by prominence only. In the absence of pitch accents that are associated with PWs, apexes are synchronised with PW-initial L tones which are also associated with PWs. This implies that apex synchronisation is sensitive to the prosodic hierarchy in

that synchronisation at the micro level occurs only between the members of the micro level (i.e., the PW) - apexes avoid synchronising with tonal events that are markers of higher level constituents such as phrase accents and boundary tones (see Figure 2.9 for a visual representation of which tonal events are associated with which phrases).

To summarise, descriptions of the hierarchical structures of gesture and prosody in Turkish reveal similarities between these modalities. It is possible to make one-to-one mapping of phrases and events based on their position in their hierarchies (see Figure 2.13). In terms of the temporal relationships between gesture and prosody, synchronisation has been mainly investigated between apexes and various prosodic anchors at the micro level in a limited number of languages. The common hypothesis of these studies has been that prominences in gesture and prosody are synchronised. The present study puts this hypothesis to the test and checks which tonal events within Turkish prosodic structure gesture apexes are synchronised with.

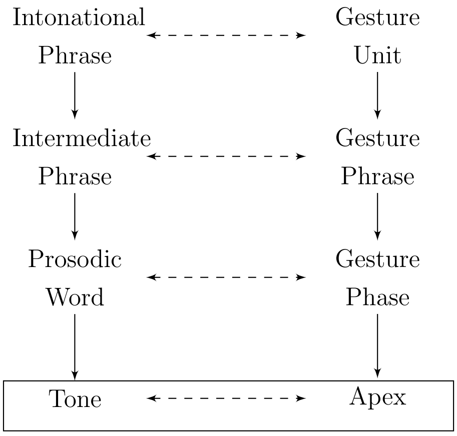


Figure 2.13: Pairing of structural hierarchies

Present Study

Within this study, gesture apex synchronisation is considered with any tonal event (compare Table 2.1). An apex is considered to be synchronised with the nearest tonal event only if the time distance between them is less than the average syllable duration (160ms), which is tested statistically. Accordingly, the analysis of synchronisation is twofold. First, the nearest tonal event to each apex is identified. These nearest apex and tonal events are called *pair-*

ings (e.g., an iconic apex and H* pairing). Note that at this point a pairing of an apex with a tonal event does not mean synchronisation but only indicates proximity. After the identification of pairings, the study reports if there are any consistent patterns in the pairing of apexes and tonal events. The consistency of these pairings is checked in several different prosodic and gestural contexts to reveal whether apexes tend to be near certain tonal events in different contexts (e.g., whether iconic apexes are usually synchronised with nuclear H*). Then, synchronisation between these pairings is statistically tested given the explained criterion. The calculation of time distance and statistical methods used are explained in Sections 4.1 and 3.5 respectively. This statistical testing also factors in prosodic context (i.e., prosodic phrasing relative to sentence prominence) and gestural context (i.e., gesture types) since these may have an effect on how apexes and tonal events are synchronised. Altogether, these lead to the following research questions:

1. Which tonal events are apexes synchronised with in Turkish?

    1a. Is the synchronisation of apexes and tonal events affected by prosodic, gestural, and information structural contexts?

Note that this question 1a also allows for the possible effects of information structure on apex-tonal event synchronisation. As previously mentioned in Chapter 1, the present study also proposes information structure as a speech component that may govern the synchronisation of speech and gesture. The study tests whether gestural units show any synchronisation with information structural categories such topic, focus, and contrast (see Section 2.6.2). In line with this proposal, the study needs to account for the possibility that interactions of information structural categories with gestural and prosodic units can affect synchronisation patterns at every level. For example, it may be the case that tone-apex pairings occurring within foci may show tighter synchronisation compared to the same pairings occurring within topics. The inclusion of information structure in the analysis intends to capture such effects. Details on information structure categories and the proposal concerning the synchronisation of information structure and gesture is explained further in Section 2.6.

This section and Section 2.5.1 summarised the previous studies on micro level synchronisation and set the background leading to the first of the research questions investigated in the present study. As previously mentioned, this study aims to capture synchronisation relationships between gesture and prosody between larger phrasal units in addition to micro units introduced in this section. Accordingly, the following Section 2.5.2 reviews relevant studies that have investigated the synchronisation of prosodic and gestural phrases (i.e., the macro level) before formulating the relevant research questions.

2.5.2 Macro Level

In Section 2.5.1, it was highlighted that gesture-prosody synchronisation has been mainly tested between the smallest units at the bottom of these hierarchies (see Figure 2.13), and that those investigations reported observations of consistent synchronisation in general. Any claims of synchronisation of gesture and prosody stem only from observations of synchronisation at this micro level. However, as portrayed in Section 2.5.1.2, there is a structural similarity between gesture and prosody in that both are hierarchically constructed with nested phrasal constituents organised around central prominent phrase (i.e., nuclear ip and stroke). If we want to establish that gesture and prosody are synchronised in production, we have to investigate synchronisation not only between points in time at the micro level but also between phrases that are higher in the hierarchies, i.e., macro level. However, much less attention has been paid to whether or not gesture and prosody show synchronisation between phrasal constituents. Section 2.5.2.1 reviews the small number of studies that investigated synchronisation at this level.

2.5.2.1 Previous studies

It was pointed out in Section 2.5.1 that determining stable anchors within the continuous streams of gesture and prosody for synchronisation tests has been problematic. This has not been the case for phrasal synchronisation because phrases are intervals with relatively easily identifiable onsets and offsets that demarcate them. Testing the synchronisation of onsets and offsets of corresponding phrases in gesture and prosody can explain whether these phrases are synchronised or not. For gestures, gestural phrases that may be tested for synchronisation were described in Section 2.1.1. Amongst these, the studies in Table 2.2 used in their analysis either gesture phrases (i.e., combinations of gesture phases around a single stroke) or gesture units (i.e., single or multiple gesture phrase combinations that occur between rest positions) (see Section 3.4.1.1 for more details on segmentation of gestures).

Definitions of prosodic phrases to be tested for synchronisation followed well-established linguistic frameworks. Most of the studies in Table 2.2 used the AM theory of intonation (Ladd, 2008). Even when prosodic units were defined according to earlier studies that were part of the emerging AM framework (Selkirk, 1980, 1984), it is possible to assume equivalence between defined prosodic phrases since they share a number of underlying assumptions (e.g., tonal events in AM are linked to prosodic edges and prosodic heads Selkirk, 1980, 1984). The correspondence between phrases is noted in the reviews of studies in this section.

Table 2.2: A table summarising some of the earlier studies on synchronisation at the macro level

Study and Language	**Gestural Phrase**	**Prosodic Phrase**	**Finding**
Loehr (2004); English	Gesture phrases	Intermediate phrase (Ladd, 2008)	A single gesture phrase synchronises with a single intermediate phrase - there can be multiple gesture phrases within an intermediate phrase
Yasinnik et al. (2004); English	Gesture units	Intonational phrase groupings (Ladd, 2008)	Gesture units span over intonational phrase groupings
Karpiński, Jarmołowicz-Nowikow, and Malisz (2009): Polish	Gesture phrases	Major intonational phrases (A. Wagner, 2008)	No synchronisation but varying degrees of overlaps were observed
Ferré (2010): French	Iconic gesture phrases	Intonation phrase (Selkirk, 1980)	No synchronisation but a single gesture phrase overlaps with a single intonation phrase
Shattuck-Hufnagel, Ren, and Tauscher (2010): English	Torso leans	Intermediate phrases (Ladd, 2008)	The majority of leans overlap with intermediate phrases
House, Alexanderson, and Beskow (2015): Swedish	Gesture units	Talk spurt: intervals of continuous speech	The onsets of talk spurts slightly precede the onsets of gesture units

In general, it can be said that the findings of studies in Table 2.2 did not indicate a strong synchronisation between selected gestural and prosodic phrases compared to the studies on micro-level synchronisation. However, these findings revealed some form of containment of multiple gestural phrases within one prosodic phrase (or vice versa) across languages.

As a follow-up of his investigation on apex-pitch accent synchronisation, Loehr (2004) also investigated whether gesture phrases and intermediate phrases (ips) ($\approx$ syntactic phrase) were synchronised in English natural speech data. The study made use of the same synchronisation criterion as the one used for apex-pitch accent synchronisation in the study in which two anchors were considered as synchronised only if they occurred within 275ms of each other. The synchronisation of phrases was assumed to be the case if there was synchronisation of boundaries. That is, synchronisation was tested between onsets and offsets of phrases separately. If gesture phrases were synchronised with ips both at the onsets as well as the offsets, then the synchronisation of these phrases was achieved.

The findings of Loehr (2004) showed that approximately two-thirds of 115 gesture phrases were synchronised with an ip. The means of time difference between onsets (m=37ms) and offsets (-14ms) were quite close to zero. However, the standard deviation of time differences was greater with 553ms (onsets) and 557ms (offsets) compared to the one observed for apex synchronisation (sd=341ms). This was interpreted to show a less tight synchronisation behaviour between these phrases in English.

Checking onset and offset synchronisation separately meant that these onsets and offsets may belong to different ips. For example, there may be multiple gesture phrases occurring within the duration of a single ip, and the boundaries of such an intermediate phrase can still be synchronised with the onset and offset of the grouping of gesture phrases contained within this ip.

Loehr (2004) did not present a detailed analysis of such phenomena but only stated that indeed there were "often" multiple gesture phrases taking place within a single ip. Although the typical pattern was stated to be one-to-one synchronisation of a single gesture phrase with a single ip, it is unclear at what rate ips spanned over gesture phrases, and whether or not this had any effect on the synchronisation behaviour. The high standard deviation observed for the time difference between onsets and offsets may be a consequence of such an effect. In addition, the finding of ips spanning over multiple phrases indicates a potential durational mismatch between gesture phrases and intermediate phrases. It may be the case that gesture phrases were synchronised with ip internal components such as PWs. The analysis also overlooked whether prosodic context (e.g., pre-nuclear, nuclear ips) and gesture type (e.g., iconic, deictic) had any effect on synchronisation. In brief, the study concluded that gesture phrases showed synchronisation with ips - whenever there was a gesture phrase boundary, there was an ip boundary within 275ms. However, it also noted that the temporal correlation between gesture phrases and ips in their data was looser than the correlation between apexes and pitch accents which was indicated by the lack of uniformity in the distribution as expressed by a high standard deviation.

Yasinnik et al. (2004) had a different approach to investigating synchronisation. The study tested the synchronisation of gesture units with intonational phrases (IPs) (≈ syntactic clause) in English natural speech data. In the study, adjacent IPs were grouped together if they were accompanied by the same gesture unit, which created "IP groupings". Pause durations between IPs within the groupings (i.e., IP-grouping medial pause) were measured and then compared to the pause at the end of the groupings. The hypothesis was that if within-IP grouping pauses were shorter than the grouping-final pause durations (i.e., if the longest pause was grouping-final and not grouping-internal), this would indicate that gesture units create a formation of intonational phrase groupings which may be corresponding to "higher level constituents" or "larger organizational elements of discourse" (p.100). It is not entirely clear what such larger units might be, but they seem to be either entire utterances or information structural categories such as topic and focus, which are related but different speech components. Moreover, the grouping/spanning phenomenon was loosely defined. An IP was included in a grouping regardless of exactly how much overlap there was between the gesture unit and the IP - a gesture unit might have just started at the very end of an IP but this IP would still be a part of the grouping as per their annotation guidelines (p.99). Therefore, it would be difficult to make any meaningful associations between such IPs and gesture units, especially if they were linked to some discursive function as suggested.

The study reported that the expected prediction of the hypothesis was observed for 12 out of 14 multiple IP groupings where grouping-medial pauses were shorter compared to the grouping-final pauses. Although there was no general report of actual time differences or statistical tests, there was only one example showing an IP grouping containing 4 IPs where the final pause was about 120ms longer than the second and third pauses within the grouping. In their discussion, Yasinnik et al. (2004) did not specify exactly which higher level structures these gesture units marked in speech or what functions these IP groupings served (e.g., did the groupings correspond to information structural categories or to discursive movements?). They commented only that gesture units tended to span the boundaries of adjacent IPs.

Overall, Yasinnik et al. (2004) only showed compression of pauses between IPs when they were accompanied by gestures. The most relevant finding for the purposes of the present study was that single gesture units were not synchronised with single IPs. Only half of the observations showed possible synchronisations of these two phrases. In the rest of the observations, gesture units spanned multiple IPs. As implied by Yasinnik et al. (2004), it is possible that another modality such as information structure may be governing speech and gesture synchronisation at the macro level. The present study explores further the possibility of such a relationship between gesture and information structure in Section 2.6.

Karpiński et al. (2009) investigated the synchronisation of gesture phrases and major intonational phrases (as defined in A. Wagner, 2008) in Polish natural speech data. These major intonational phrases equate to intonational phrases (IP) (≈ syntactic clause) within the AM framework. The study did not present any actual time distances between these phrases, nor did it define what was considered synchronisation of phrases. The authors only presented a probabilistic model that indicated how single IPs were positioned in relation to single gesture phrases. First, IPs outnumbered gesture phrases approximately four to one, which the authors explained to be due to the overlap of gesture phrases with more than one IP. This shows that, similar to Yasinnik et al. (2004), there was a durational mismatch between these phrases. Their probabilistic model demonstrated that IPs could occur at various positions relative to their semantically related gesture phrases. IPs could be contained within a single gesture phrase, or they could stretch across gesture phrase boundaries, with parts of the IPs outside the related gesture phrase (see Figure 2.14). These observations rule out a one-to-one synchronisation between single gesture phrases and IPs in Polish. Durational mismatch and gesture phrases overlapping multiple IPs portray a similar synchronisation behaviour to what was observed in Yasinnik et al. (2004) although the synchronisation (or rather co-occurrence) was tested with different gestural phrases (i.e., gesture phrase versus gesture units). There was no comment on what type of gestures were used in the study.

Ferré (2010) checked the synchronisation of iconic gesture phrases and intonation phrases (as defined in Selkirk, 1980) in French natural speech data. The author states that intonation phrases equate to intermediate phrases (ip) (≈ syntactic phrase) in the AM model. Accordingly, this study parallels Loehr's (2004) in terms of the phrases used. The study did not define a synchronisation rule such as Loehr's (2004) 275ms rule. Instead, it reported whether gesture phrases started/ended before or after the ips they accompanied. Synchronisation was statistically tested, in order to determine whether the mean time difference of onsets/offsets were significantly greater than zero. The study did not explicitly state what significant differences would mean in terms of synchronisation. Presumably, if the mean time difference between onsets and offsets was significantly greater than zero, i.e., the perfect synchronisation condition, then these boundaries were not synchronised. The findings showed that 70% of gesture phrases started before their semantically related ips by a mean of 190ms. The standard deviation was also high with 790ms. It was also shown that the mean time difference was significantly greater than zero. The synchronisation between offsets was even weaker. 61% of gesture phrases ended after their ips with higher values for the mean (220ms) and the standard deviation (860ms). The mean time difference with zero was also statistically significant. Based on their way of testing synchronisation, gestures phrases and intermediate phrases in French were not found to be synchronised. It must be noted that statistical testing of synchronisation in this way is more likely to indicate asynchrony than synchrony. Unless boundaries of phrases occur almost at the same time, which should not be expected during spontaneous speech, measured time differences are likely to be significantly different from zero. With such a strict constraint, analyses of synchronisation can miss out important consistent synchronisations that are not centered on zero (e.g., consistent precedence cases reported in Section 2.2). Compared to the findings of Loehr (2004), these findings show a weaker temporal co-occurrence of the two phrases. Another difference between the findings of these studies was that Ferré (2010) did not report ips spanning over multiple gesture phrases. Instead, they reported an overlap. Single ips tended to be contained within a single gesture phrase with the corresponding boundaries of the units at these two levels approximately 200ms away from each other (see Figure 2.14). This implies that the definition of prosodic phrases in the prosodic structure of languages may have an effect on synchronisation of phrases.

Shattuck-Hufnagel et al. (2010) is different from the studies covered so far because it made use of sideways torso swings instead of hand gestures to establish timing relationships with intermediate phrases in English. There were no reports of time measurements or a definition of synchronisation for these units. The findings of the study showed that the majority of torso leans show 90-100% overlap with intermediate phrases. The study also reported cases where torso movements continued across multiple intermediate phrases, which was interpreted to be a cue to groupings of ips into larger prosodic constituents much like Yasinnik et al. (2004). Overall, the findings of this study were only observational and inconclusive. However, they pointed at a possibility that torso movements, which are not necessarily speech related (see gesticulations in Section 2.1.1), may also show some kind of coordination with prosodic structure.

In Swedish natural speech data, House et al. (2015) investigated the synchronisation of gesture units and “talk spurts” which were defined as “intervals of continuous speech” (p.64). Note that this definition is not a phonological definition, therefore it is difficult to equate them to a phrase in the AM model. They used an automatic speech activity detection algorithm which segmented continuous speech into talk spurts (minimum 500ms) according to durations of silence - if there was a silence longer than 200ms, this constituted the offset of a talk spurt. Accordingly, the resulting segments can be considered to be intonational phrase groupings (IP groupings) similar to Yasinnik et al. (2004). IPs are the largest phrases defined in the AM model, and such long speech streams should contain at least one IP.

The study only tested the synchronisation of talk spurts and gesture units at the onsets. It reported considerable variation across the two participants and two dialogue conditions included in the study, but there was a general tendency showing that the onsets of talk spurts slightly preceded (by about 100ms) the onsets of gesture units. Note that this observation of precedence of talk spurts contradicted the general precedence expectation where gesture is expected to precede its speech counterpart (see Section 2.2). The study interpreted this precedence as a synchronisation of onsets. However, gestures studied in House et al. (2015) only occur when there is speech by default (see gesticulations in Section 2.1.1). Therefore, it is not very surprising that such large segments of speech and gesture start around the same time. In order to establish a more meaningful relationship between these, discursive functions of talk spurts should also be considered. It is unclear which aspect of speech these talk spurts (and by proxy gesture units) were associated with in the study.

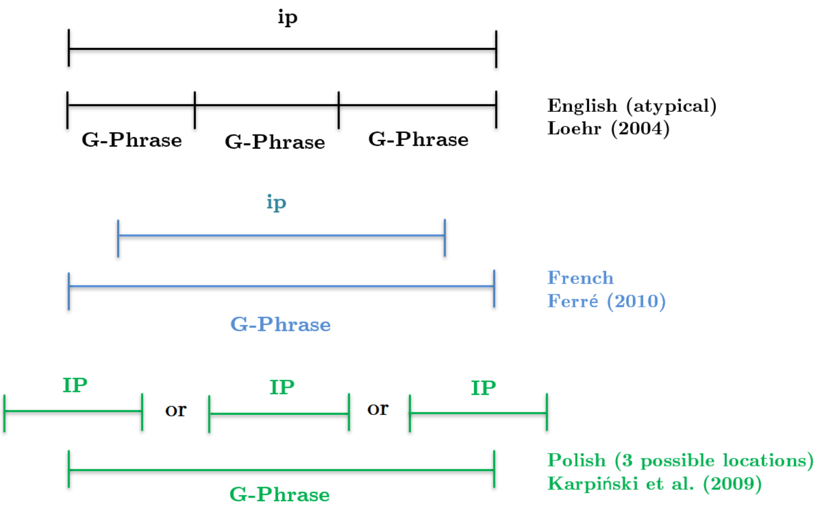


Figure 2.14: Simple representations of gesture phrase (G-phrase) and intermediate phrase (ip) synchronisations in English, French, and Polish

The studies in Table 2.2 tested synchronisation between different gestural phrases, but most of them do not report data from which synchronisation can be judged as it is defined in the present study (apart from Loehr, 2004). Interestingly, there were cases of overlap observed as either one phrase overlapping with a smaller chunk of the other phrase or one phrase fully containing the other within its duration. For instance, Loehr (2004) reported atypical cases where multiple gesture phrases were contained within a single ip in English, whereas Ferré (2010) reported that typically a single gesture phrase contained a single ip in French (see Figure 2.14). On the other hand, Karpiński et al. (2009) tested the synchronisation of gesture phrases and intonational phrases (IP), and they reported a very different synchronisation behaviour to Loehr (2004) and Ferré (2010). IPs are larger constituents than ips within the prosodic hierarchy. Based on the observations of Loehr (2004) and Ferré (2010), these larger prosodic phrases would be expected to contain gesture phrases. However, contrary to this expectation, Karpiński et al. (2009) showed that in Polish, gesture phrases were much larger than IPs, as seen in Figure 2.14. Note that the representation for Polish in the figure shows how a single IP can be positioned at three different locations in relation to the gesture phrase to which it is semantically bound.

These findings reveal varying patterns of synchronisation across these languages. It is possible that synchronisation is affected by differences in the prosodic phrasing across these languages. Durational mismatches between gestural phrases and prosodic phrases can make one-to-one synchronisation of gestural and prosodic phrases unlikely. However, this does not necessarily mean that prosodic phrases and gestural phrases do not show any synchronisation. Synchronisation can be manifested in terms of a sensitivity of boundaries of units under consideration, regardless of how many phrases are contained within the other. For example, in the representation of English in Figure 2.14, assume that the ip onset is synchronised with the first gesture phrase onset, and the ip offset is synchronised with the final gesture phrase offset. Such a synchronisation case would imply that there is no one-to-one synchronisation of ips and gesture phrases due to their durational mismatch, but onsets and offsets at each level are sensitive to each other as events (e.g., whenever there is a gesture phrase starting, there is an ip starting nearby). Only Loehr (2004) showed evidence for such a synchronisation due to the study's implementation of a synchronisation rule. In the other studies, this possibility was overlooked.

Another explanation of these mismatches might be that the selected phrases were not the best candidates for synchronisation. These studies set out to test synchronisation between two pre-selected phrases only. This means that if there was another phrase defined within the hierarchy that might show a tighter synchronisation with the selected phrase, these studies would not be able to capture that. Consequently, selecting a phrase in one hierarchy (e.g., gesture phrase) and testing its synchronisation with all possible phrases defined in the other hierarchy (e.g., IP, ip, and PW) could produce better synchronisation results.

In general, the shortcomings in the analyses of synchronisation at the macro level paralleled those at the micro level. First, conditions of synchronisation were not set in most of the studies in Table 2.2. In particular, analyses involving constituents higher up in the hierarchy remained highly observational. Only Loehr (2004) and Ferré (2010) could make generalisations about synchronisation, as only these studies statistically defined what constituted synchronisation.

Potential differences between gesture type were not a problem for the studies investigating gesture units since these are intervals that can contain multiple gestures with distinct semantic functions. In the investigation of gesture phrases, only Ferré (2010) controlled for such an effect of gesture type by focusing only on iconic gestures. There were no reports of gesture types investigated in the other studies. Therefore, it is unknown whether the synchronisation of gesture phrases and prosodic phrases shows any variation depending on the semantic function of gestures. Similarly, the effect of the position of ips relative to sentence prominence on synchronisation has not been investigated. At the micro level, synchronisation has shown to be between prominent anchors. It might be the case that ips carrying the maximum prosodic prominence (i.e., nuclear ips) show a different synchronisation pattern (possibly tighter) with gesture phrases compared to pre-nuclear and post-nuclear ips.

The present study aims to test synchronisation between phrasal constituents in gestural and prosodic hierarchies while also addressing concerns described above. Section 2.5.2.2 revisits Turkish prosodic structure and frames the research questions for the investigation of phrasal synchronisation between gesture and prosody.

2.5.2.2 Investigation of macro level synchronisation in the present study

The present study previously noted a structural similarity between gesture and prosody, showing that they both consist of hierarchical phrasal constituents. Section 2.5.1 highlighted that previous studies on synchronisation of gesture and prosody reported consistent synchronisation of prominent atomic points defined within these phrases. Using the same multimodal data used in the investigations of micro-level synchronisation, the present study investigates whether there is synchronisation between gesture and prosody beyond these atomic points by testing whether gestural phrases and prosodic phrases are synchronised with each other in Turkish.

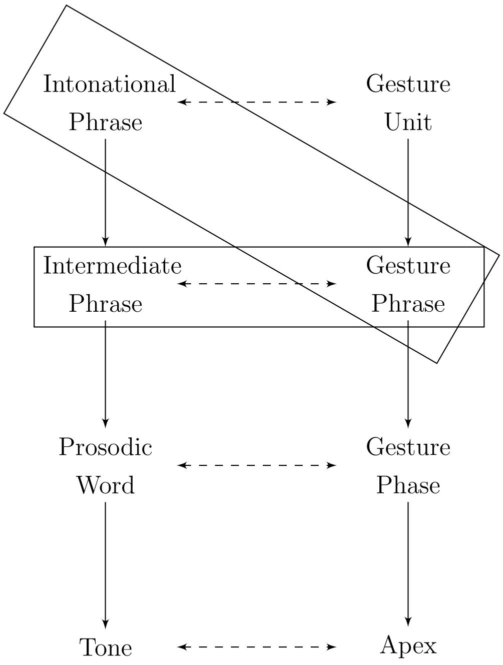


Figure 2.13: Pairing of structural hierarchies (repeated from page 75)

The investigation of synchronisation in the present study focuses on gesture phrases. Gesture phrases have been investigated more carefully in previous studies working on different languages (Loehr, 2004; Ferré, 2010); therefore, comparing findings with those studies can give insight into the universals of synchronisation. Moreover, investigating synchronisation at the gesture phrase level also enables the testing of variation in synchronisation patterns depending on gesture types. The present study includes all gesture types in its analysis (iconics, deictics, and metaphorics) except for beats. Beats are rhythmic units often with very short durations due to being formed by one or two gesture phases (see Section 2.1.1). Therefore, they are not suitable candidates for synchronisation with larger prosodic phrases. How gestures are segmented and how their types are identified within this study are detailed in Section 3.4.1.

The aim of the analyses here is to find the prosodic phrase that is best synchronised with gesture phrases. A qualitative look at the data showed that ips and IPs are the most likely candidates for synchronisation with gesture phrases within Turkish prosodic structure (see Section 5.1 and Figure 2.13). Therefore, synchronisation of gesture phrases is tested with these prosodic phrases. The present study also investigates whether gesture phrase and ip synchronisation is affected by the position of ips relative to sentence level prominence (i.e., ip types: pre-nuclear, nuclear, and post-nuclear; e.g., are nuclear ips better synchronised with gesture phrases than post-nuclear ips?). Moreover, there may also be interactions between these ip types and gesture types that can affect the synchronisation of phrases. For example, deictic gestures may be more likely to accompany and be synchronised with pre-nuclear ips. The present study also integrates such analyses into its analysis of synchronisation. How these prosodic phrases and their types are identified within the AM model of intonational phonology is described in detail in Section 3.4.2.

The analysis of synchronisation is done in the same manner as the analysis at the micro level as explained in Section 2.5.1.2. The analyses are conducted separately for onsets and offsets. The synchronisation of phrases is assumed to be based on the synchronisation of boundaries of corresponding phrases. First, time distances between prosodic phrase boundaries and accompanying gesture phrase boundaries are calculated (see Section 5.1), and the nearest boundaries to each other are identified and paired (i.e., a pairing as in Section 2.5.1.2). At this point, a series of analyses is carried out to reveal whether there are any pairing patterns between these phrases according to ip types and gestures types (e.g., deictic gesture onsets may often be nearest to pre-nuclear onsets). Then, given the same synchronisation criterion (i.e., average syllable duration), it was tested whether these pairings achieve synchronisation. This procedure is repeated both for the onsets and offsets of both ips and IPs. The resulting research questions relating to macro-level synchronisation between gesture and prosody are the following:

2. Are the onsets/offset of gesture phrases synchronised with the onset/offsets of intermediate phrases?

   2.a Is the synchronisation of gesture phrases and intermediate phrases affected by prosodic, gestural, and information structural contexts?

3. Are the onsets/offsets of gesture phrases synchronised with the onsets/offsets of intonational phrases?

   3.a Is the synchronisation of gesture phrases and intonational phrases affected by prosodic, gestural, and information structural contexts?

Note that as with the research questions in Section 2.5.1.2, these questions also account for the possibility that information structure can have an effect on the synchronisation of gesture phrases and ips/IPs in line with the proposal that gesture is sensitive to information structural categories. The present study proposes that at the macro level, gestures may in fact be synchronised with information structural categories. Section 2.6 introduces what these categories are, and establishes the details of the association of information structure with prosody and gesture.

2.6 Information Structure as a Synchronisation Anchor

The main goal of the present study is to reveal temporal relationships between speech and gesture. As can be understood from the reviews of previous studies above, these relationships are complex, drawing from various aspects of speech and gesture. A general impression from these earlier studies is that the relevant research on synchronisation has not always fully integrated or even considered various areas of linguistic enquiry. As discussed in Sections 2.5.1 and 2.5.2, research so far has mainly focused on prosody as the main linguistic component that gesture is synchronised with while also not fully recognising its richness as a phonological system. Investigations of prosody as an anchor for gesture has shown consistent and meaningful synchronisation behaviours at the micro level as covered in Section 2.5.1. However, we have seen in Section 2.5.2 that larger phrasal units in gesture and prosody either have shown less tight and consistent synchronisation or have not shown synchronisation at all. This opens up further discussions about what other suitable linguistic anchors there might be in speech for gesture synchronisation. The present study proposes information structure as such an anchor and tests whether gesture is synchronised with categories that are defined by information structure. Information structure can be considered relevant because it is linked to prosody, and it has been implicitly associated with gesture although not in terms of synchronisation (covered in Section 2.6.2). This section first defines information structure and its categories as used in the present study, and then establishes its relevance to prosody and gesture by reviewing relevant studies.

2.6.1 Information Structure

Information structure describes the salience and organisation of information in relation to a discourse (Calhoun, 2007). Discourse can be taken to be any coherent spoken expression involving multiple utterances. It is constructed from a series of propositions that interlocutors accept to be understood and known by the other, contributing to a model of the relevant world of discourse. Information structure functions to map out how new utterances that are added to the discourse relate to or update this discourse model. Linguistic theories of information structure assume that the construction of a discourse is affected by interlocutors' general world knowledge, personal attitudes (Kruijff-Korbayová & Steedman, 2003). We will see how gesture plays a part in discourse building in Section 2.6.2.

The present study defines information structure on three dimensions which are *information status*, *topic/focus/background*, and *contrast*. These dimensions and their annotations were based on Skopeteas et al. (2006) and Götze et al. (2007), but minor amendments were made in accordance with the purposes of the present study. These dimensions are briefly introduced in this section in order to help interpret the reviews of relevant studies in Section 2.6.2. The reader is referred to Section 3.4.3 for further details on how these dimensions are annotated in the data.

The information status dimension categorises discourse referents (i.e., constituents that refer to entities in a discourse) according to their retrievability from discourse. A referent that is explicitly mentioned in previous discourse is *given*. If a referent was not mentioned explicitly but can be inferred from the context or through common sense, it is *accessible*. If a referent was not mentioned in previous discourse and cannot be inferred from the context, then this referent is *new*. Example 1 shows an example utterance for these referents.

(1)	The pub	is very crowded. There is	a bouncer	at its entrance.
	given		**new**	**accessible**

In the example, 'the pub' must have been given in the previous discourse, and it is now being re-introduced. 'At its entrance' is an accessible referent because the knowledge that a place such as a pub would have an entrance is available in the assumed world knowledge of the interlocutor. 'A bouncer' is newly introduced to the discourse.

The Topic/Focus/Background dimension serves to distinguish the part of an utterance that moves the discourse forward (i.e., focus) and the part that relates it to a previous discourse (i.e., topic). In other words, focus constitutes the main predication of an utterance, whereas topic constitutes the entity that the main predication is about. Background is a negative category that is not defined in Skopeteas et al. (2006) and Götze et al. (2007). It consists of any post-focal backgrounded information in an utterance that is not a topic or a focus (see "tails" in Vallduví & Engdahl, 1996). Examples 2 and 3 show example utterances containing topic, focus, and background.

(2) The pub is very crowded.
topic **focus**

(3) Who works in the pub?

George works in the pub.
focus **background**

In the example, 'the pub' relates the following part of the utterance to an entity that was previously established in the discourse - the interlocutor knows the pub in question. The main predication is the rest of the utterance that updates the status of the pub, moving the discourse forward. Note that information status and topic/focus dimensions do not necessarily overlap. For instance, a focus updating a discourse can contain referents that are given in the discourse, and topics can likewise contain discourse new referents (see Example 4).

(4) A good friend of mine worked *in that pub.*
topic - new **focus** *given*

The final dimension is contrast. Contrasted parts of utterances evoke a notion of contrast with a previous element in the discourse, distinguishing it from a set of alternatives. Contrast can overlap with other dimensions, meaning that any element regardless of whether it is topic or focus (i.e., contrastive topic/foci) and given, new or accessible can be contrasted (except for background). Example 5 shows an example utterance that contains contrasted elements.

(5) I don't like white wine, so I will have a stout.
Contrast **Contrast**

In the example, ‘white wine’ and ‘a stout’ are contrasted elements because they are presented as two possible alternatives of beverages. Within the utterance one available alternative is exhausted (i.e., white wine) and the other is selected (i.e., a stout).

Overall, these categories structure information within utterances according to a discourse in order to augment information transfer between interlocutors. They are mental representations of discursive entities which make up the interface between cognitive systems that guide communication and grammar of languages (Zimmermann & Féry, 2010). Languages express these categories via different linguistic modules such as syntax, morphology, and prosody. In light of the interests of the present study, Section 2.6.2 outlines information structure’s relationship with prosody as a linguistic module which has already been shown to be linked to gesture. The section further explains why information structure may have a temporal relationship with gesture through a review of studies which have highlighted connections between information structure and gesture in aspects other than synchronisation.

2.6.2 Why is Information Structure Relevant?

The present study treats the synchronisation of speech and gesture as a possible three-way synchronisation of gesture, prosody, and information structure. The relevance of information structure in terms of gesture synchronisation can be assumed indirectly via its association with prosody, and directly via its association with gesture in terms of discourse management (explained in Section 2.6.2.2).

First of all, Sections 2.2, 2.5.1, and 2.5.2 showed that gesture and prosody have been linked in many aspects including synchronisation - units within these modalities have been shown to be synchronised. As we will see in

Chapters 4 and 5, the findings of the present study also consolidate the observations of synchronisation between gestural units and prosodic units, and the view that gesture production is informed by prosody. This link between gesture and prosody is represented in Figure 2.16 which demonstrates a simple schematic showing the association of gesture, prosody, and information structure.

2.6.2.1 Information structure and prosody

It can be argued that information structure is indirectly linked to gesture through its close relationship with prosody (see Figure 2.16). Prosody is one of the principal cues to information structure for many languages (Kügler & Calhoun, in press). Amongst the types of information structure categories, prosodic marking of focus has received the most attention. Prosodic marking of focus means that prosody highlights the focus by distinguishing it from any surrounding elements, making use of pitch accent placement and type, prosodic phrasing and/or pitch register (Kügler, 2011).

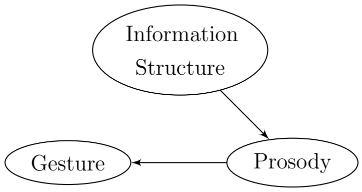


Figure 2.16: Three-way association of gesture, prosody, and information structure

In languages which rely on pitch accents in the marking of information structure (e.g., English, German, Dutch, Swedish Ladd, 2008; Féry & Kügler, 2008; Gussenhoven, 2004; Myrberg & Riad, 2016), the typical observation is that the focused word bears the maximum prosodic prominence. From a phonological point of view, phonetic cues (e.g., F0) mark nuclear pitch accents which in turn mark focus - nuclear accent placement is governed by focus (Gussenhoven, 1984; Ladd, 2008; Calhoun, 2010a). In addition, some languages make use of pitch accent and boundary tone types in order to distinguish between information structure categories.

For example, in English, it has been claimed that contrastive foci are marked with L+H* whereas neutral (i.e., non-contrastive) foci are marked with H* (Watson, Tanenhaus, & Gunlogson, 2008). Similarly, topics and foci can be signalled with different accent types (Steedman, 2014) (also see Frota & Prieto, 2015).

As for information status, degrees of givenness have been shown to correlate with degrees of prosodic prominence (Baumann, Grice, & Steindamm, 2006; Röhr & Baumann, 2011; Baumann, 2012; Baumann & Kügler, 2015; Genzel, Ishihara, & Surányi, 2015). Given elements are generally deaccented and/or associated with low prominence, regardless of whether given elements are in focus or not (see Example 4). New elements, on the other hand, tend to be accented.

Another group of languages makes use of prosodic phrasing to signal focus. What separates these languages from languages that employ pitch accents is that they usually do not have lexical prominence; therefore, cueing of focus via pitch accenting is not an option. In the absence of such prominence, prosodic phrasing plays a more active role. Prosodic marking of focus requires the addition of phrase boundaries which separate focus from other elements within an utterance. For example, in Korean, the onset of focused elements is marked with the insertion of a prosodic phrase boundary (Jun & Kim, 2007; Jeon & Nolan, 2017). Similarly, a phrase boundary is inserted at the offset of focused elements in Chichewa (Downing & Pompino-Marschall, 2013) (see Beckman & Pierrehumbert, 1986; Féry, 2001 for other languages).

A final group of languages makes use of pitch register manipulation to cue information structure. Pitch register deals with the relative pitch scaling of tonal events. In this type of languages, focused elements are given prominence through amplified realisations of tonal events (e.g., the maximum F0 for an H tone under focus is increased). In addition, the prominence of post-focal elements can be reduced through the compression of pitch register (e.g., the maximum F0 for an H tone within a post-focal element is decreased). Mandarin Chinese (Xu, 1999; Wang & Xu, 2011) and West Greenlandic (Arnhold, 2014) are examples of languages that use register to cue information structure.

What kind of prosodic marking system Turkish uses has not been agreed on in the literature. Kamali (2011) and Ipek (2011) claim that focus is not marked with pitch accents or a particular pitch expansion. Instead, focus is aligned with prosodic phrases that show a pitch plateau that is higher than the following phrases. On the other hand, Göksel and Özsoy (2000) and Özge and Bozsahin (2010) argue that focus is marked by pitch accents in Turkish. As we will see in Section 3.4.2, the observations in the present study seem to be in line with Kamali (2011) and Ipek (2011) - focus is aligned with nuclear prosodic phrases exhibiting a flat pitch contour and existing in the immediately pre-verbal area. Based on this, Turkish seems to have a register-based system since tonal events in phrases that align with focus are reduced (i.e., pitch plateau) and post-focal phrases have a lower pitch profile than pre-focal and focal phrases (explained further in Section 3.4.2.2).

How does Prosody-Information Structure Interface Relate to Gesture?

Typologically varying prosodic means of expressing information structure categories were summarised above. These studies provide clear cross-linguistic evidence for the strong link between information structure and prosody - information structure informs prosodic structure. What is important in this association for the present study is that marking/cueing behaviour inherently posits a form of synchronisation between these speech components. That is, the positions of information structure categories within utterances are marked by prosodic phenomena systematically co-occurring with these categories. The present study considers such systematic co-occurrences as synchronisation. As was mentioned in Sections 2.5.1 and 2.5.2, some of these prosodic phenomena have also been shown to be synchronised with gestural units. Therefore, it is reasonable to propose that both prosody and gesture productions are informed by information structure where communication of information structure and gesture is indirect and mediated by prosody (as represented in Figure 2.16). For instance, in a language that uses pitch accenting to cue focus, focus is signalled by the nuclear pitch accent which an apex can be synchronised with. Such a scenario, in turn, would imply a synchronisation of focused word and apex by proxy.

In fact, information structure categories, especially focus and contrast have been used in many studies that were interested in gesture and prosody synchronisation without invoking a potential synchronisation of information structure and gesture per se (in the sense of the present study). For example, the synchronisation of gestural and prosodic prominence has been checked in contrastive focus conditions in order to take advantage of intonational expressiveness of contrastive constructs (Roustan & Dohen, 2010; Rusiewicz et al., 2013; Esteve-Gibert & Prieto, 2013; Fung & Mok, 2018). These studies investigated manual deictic gestures in similar highly experimental designs and mostly confirmed the prominence-based synchronisation of gesture and prosody. Other studies that have dealt with focus concentrated on head gestures, eye blinks, and eyebrows. In general, the findings of these studies showed that combinations of these non-manual gestures can cooccur with focused constituents (Ambrazaitis, Svensson Lundmark, & House, 2015; Ambrazaitis & House, 2017), and interlocutors are more likely to use such non-manual gestures with focused constituents than with non-focused ones (Alexanderson, House, & Beskow, 2013; Ferré, 2014; Dimitrova et al., 2016; Esteve-Gibert, Loevenbruck, Dohen, & D'imperio, 2019), and also with greater form related variation (Beskow, Granström, & House, 2006). These findings have also been reinforced by perception studies which claimed that non-manual gestures facilitate the recognition of prominent events such as contrastive focus (Dohen & Loevenbruck, 2004; Krahmer & Swerts, 2007; Prieto, Puglesi, Borràs-Comes, Arroyo, & Blat, 2015). Overall, it can be said that these studies have established focus as the domain where the apexes of non-manual gestures are synchronised with the prosodic prominence (Esteve-Gibert et al., 2017).

To reiterate, although information structure categories have been used in these studies, they have not assumed a direct synchronisation of information structure and non-manual gestures. Instead, they have either reported a frequency-based co-occurrence of focus/contrast and gesture or assigned focus/contrast as the domain or the attractor of gesture and prosody synchronisation. Whether there is a direct synchronisation of entire topical/focal areas (with or without contrast) with larger gestural phrases that can match their duration remains untested. To the author's knowledge, synchronisation of information structure and gesture at the macro level has not been systematically investigated before (see Ebert et al., 2011). The present study tests this possibility in Turkish as a part of its investigation of macro level synchronisation of speech and gesture. In addition to the studies mentioned in this section, there have also been studies that established direct links between information structural notions and gesture although not from a synchronisation perspective. The next section reviews these studies in order to further explain the relevance of information structure to gesture, and then states the final research questions of the present study.

2.6.2.2 Information structure and gesture

In addition to the indirect associations of gesture and information structure through prosody, there have been studies that investigated gestural behaviour under the effect of information structure directly. These studies are reviewed in this section.

Perhaps the simplest form of association between information structure and gesture can be made through their underlying communicative orientation. From an information structural perspective, the organisation of information in utterances is a strategy to shape the form of the message so that it is well understood by an addressee, meaning that this organisation is in place to meet the immediate communicative needs of interlocutors. In this sense, information structural categories introduced in Section 2.6.1 emerge as a response to these communicative needs (Chafe, 1976; Féry & Krifka, 2008).

Similarly, Section 2.2.1 already highlighted that gesture also responds to the communicative needs of interlocutors. In that section it was explained how interlocutors coordinate their gestural behaviour in a way that is sensitive to the presence/visibility of an addressee and the mode of conversation (i.e., monologue vs. dialogue). The studies in Section 2.2.1 were mostly early works that influenced the speech-gesture production models introduced in Section 2.3, yet the majority of later studies have also confirmed these early findings. The consensus in these studies has been that in the absence of visibility, interlocutors gesture at a lower rate and/or the form of gestures show less detail (Bavelas et al., 2008; Mol, Krahmer, Maes, & Swerts, 2009; Pine, Gurney, & Fletcher, 2010; Holler, Tutton, & Wilkin, 2011; Mol, Krahmer, Maes, & Swerts, 2011; Bavelas et al., 2014) (see Bavelas & Healing, 2013 for an overview). It has also been observed that forms, frequency and gestural space are adjusted in order to compete/cooperate with an addressee, taking into account the addressee's attentiveness (Jacobs & Garnham, 2007; Holler & Wilkin, 2011; Kuhlen, Galati, & Brennano, 2012; Peeters, Chu, Holler, Hagoort, & Özyürek, 2015). Furthermore, recent studies that have investigated the effect of mode of conversation have shown that interlocutors gesture at a higher rate in dialogue than in monologue (Bavelas et al., 2008; Holler, Turner, & Varcianna, 2013; Bavelas et al., 2014).

Taken together, gesture and information structure can be said to serve the same communicative intention. They both operate around a central principle of addressee design which suggests that interlocutors tailor their message in a way that best meets the particular communicative needs of their addressee. Gesture takes on the visual aspect of this task, and information structure takes on the verbal aspect. On the whole, it is plausible that gestural units and information structural categories assuming the same function in communication can be in some form of temporal coordination with each other.

Information structure categories in an utterance are determined by the discourse model that the utterance is a part of. Interlocutors build the discourse model together by continuously updating it with propositions. In this sense, the model of communication is a continuous change of what is established as common ground between interlocutors. The concept of common ground can be defined as the knowledge and experience that interlocutors share, combined with the awareness that they share such common ground (Clark, 1996). Several studies have investigated the effects of common ground on utterance design (Clark & Haviland, 1977; Fowler, 1988; Clark, 1996 amongst others). The common finding of these studies has been that discourse-new elements are made explicit and carefully articulated, whereas repeated use of these elements leads to shorter and less clear articulation.

More recently, researchers have taken interest in whether these effects of common ground on speech would be paralleled in gesture given that speech and gesture form a harmonious ensemble (see Section 2.2). However, many of the studies investigating this effect have interpreted common ground in a more general sense that is independent of information structural theories. For example, they have tested spoken and gestural behaviour in contexts where interlocutor pairs were either exposed or not exposed to the same information before they interacted (Jacobs & Garnham, 2007; Holler & Stevens, 2007; Holler & Wilkin, 2009; Campisi & Özyürek, 2013; Galati & Brennan, 2014;

Schubotz, Holler, & Özyürek, 2015; Hilliard & Cook, 2016). The idea was that if they received the same information prior to their interaction, this would count as having established a common ground, and the interaction would be less informative. The studies did not directly refer to new/given information distinction as defined in Section 2.6.1, but it can be understood that their interpretation of common ground is based on the view that in the common ground condition interlocutors would be using already known (given) elements only, whereas in the other condition, any information would be new to them.

Another group of studies has had a very different approach to common ground. In their designs, interlocutors were asked to describe the same information to the same addressee more than once (Gerwing, 2003; Gerwing & Bavelas, 2004; Jacobs & Garnham, 2007; Holler et al., 2011; Galati & Brennan, 2014; Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015). They presumed that each repetition would increase the common ground between interlocutors. It is important to note that all these studies had different designs and methods of analysis (see Holler & Bavelas, 2017 for an overview). However, they all investigated gesture rate and/or gesture form across these different common ground conditions - whether gesture rate or form changes when interlocutors share or do not share a common ground.

Perhaps unsurprisingly, these studies reported conflicting results, which can be partly attributed to discrepancies in their analyses. This was true for both techniques of creating a common ground. In general, most of these studies reported that interlocutors significantly reduced the number of words they used in the common ground condition. Gestural behaviour also mirrored this verbal behaviour in that when interlocutors shared a common ground they also produced proportionally fewer gestures.[8] For the effect of common

[8] Discrepancies in the analyses and how they should be interpreted are explained in detail in Holler & Bavelas, 2017, p. 221-222.

ground on gesture form, these studies reported that gestures were observed to be smaller, less accurate and less informative when interlocutors shared a common ground.

Overall, these studies observed some effects of common ground on gestures, which paralleled the effects of common ground on speech. However, although they seem to set out to test the effect of a concept that is highly related to information structure, what they consider to be common ground is very different from what it is from an information structural perspective, which affects the design and usefulness of these studies. Common ground is not just a pre-set condition that is either there or not prior to interaction. Rather, it is also constructed moment by moment through a series of exchanges which establish common ground within a conversation. It is a temporary state of mind that is changed constantly as a conversation unfolds. Therefore, regardless of the condition and design, a common ground was still created and managed by interlocutors in these studies. In essence, what these studies were testing was not the effect of shared knowledge itself but rather the effect of the interlocutors' awareness that they shared knowledge. In the way these experiments were set up, interlocutors were aware of whether or not they were sharing a common ground across conditions (e.g., repetition of the same conversation between the same interlocutors). Therefore, interlocutors knew that their interaction was not informative, which affects the well-formedness of discourse context (Büring, 2003). That is, it violates the informativity principle or the maxim of quantity that dictates not to tell others what they already know (Grice, 1975; Büring, 2003). It is plausible to assume that gestural behaviour represented such not so well-formed (or ill-formed) discourse contexts with smaller, shorter, and less accurate gestures which can also be seen as less well-formed gestures.

If the effect of common ground is to be checked, studies should focus on how information structural categories mark and manage the common ground within/across utterances. There have also been studies that have adopted such an idea. These studies have investigated the relationship between information status and gesture. They have shown that tracking of discourse referents is multimodally achieved, and that the retrievability of referents affects speech and gesture in similar ways (Levy & McNeill, 1992; Gullberg, 2006; So, Kita, & Goldin-Meadow, 2009). These studies used better defined discourse referents such as new (i.e., first introduction to discourse with rich forms of referring expression like full noun phrases), re-introduced (i.e., given elements that have been introduced earlier in discourse, again with rich forms of expression), and maintained (i.e., given elements existing in the immediate discourse which are maintained with pronouns or null pronouns).[9]

In general, these studies have shown that interlocutors gesture more often with (re-)introduced referents compared to maintained ones (Levy & McNeill, 1992; Debreslioska, Özyürek, Gullberg, & Perniss, 2013; Perniss & Özyürek, 2015). The studies have also shown a similar correlation in terms of the richness of expressions in that interlocutors were reported to gesture more for full expressions of referents than reduced forms of referents such as pronouns (Gullberg, 2006; Yoshioka, 2008; Azar & Özyürek, 2015; Perniss & Özyürek, 2015; Azar, Backus, & Özyürek, 2019; Debreslioska & Gullberg, 2019). Therefore, it can be said that the marking of referent retrievability in speech and in gesture go hand-in-hand, and that information status and gesture can be used in a cooperative way to manage discourse.

As can be seen, while studies have investigated some links between information structure and gesture, i.e. with common ground (as background knowledge) and information status, the link with information structure as a whole has largely been neglected. In particular, one thing that stands out in the literature is that an association of topic and focus (as they are defined in

[9] A similar re-introduced and maintained distinction also exists in Götze et al. (2007) as subtypes of given information (given-active and given-inactive) although the scheme does not make this categorisation based on form.

linguistic theories) with gesture has been largely overlooked although there have been hints implying that gesture and topic/focus dimension may be related.

As previously discussed for Growth Point theory in Section 2.3, McNeill (1992) and McNeill and Duncan (2000) have argued that speech and gesture stem from the same minimal idea unit (i.e. growth point) which aims to convey "the most noteworthy" information in context as a result of being born as a "novel departure of thought from the presupposed background" (McNeill, 1992, p. 220). This explanation for the origin of gesture overlaps a great deal with the definition of focus in Section 2.6.1. Focus updates already established information with a piece of newsworthy information, making discourse depart from the presupposed ground and moving it forward in order to reach a communicative goal. Although McNeill (1992) and McNeill and Duncan (2000) were not interested in information structure per se, their claim constitutes a possible justification for the existence of an association between information structure (through focus) and gesture. Another similar presumption was made in Cassell et al. (1994). The study was interested in 3D animation, and they selected focal areas of utterances as the domain where gestures would be anchored, based on the view that gestures are informative and can be used in cooperation with speech to carry discourse forward.

Regardless of what notion of information structure has been called upon in relation to gesture, hardly any studies have considered a possible synchronisation of information structural categories and gesture in the way they have looked at the synchronisation of prosody and gesture as presented in Sections 2.5.1 and 2.5.2. To the author's knowledge, there are only two studies that have investigated gesture synchronisation in relation to information structural concepts which are P. Wagner and Bryhadyr (2017) and Ebert et al. (2011).

P. Wagner and Bryhadyr (2017) is less relevant to the present study firstly because it dealt with co-speech actions rather than gestures (distinguished based on the target of the movement, i.e., actions are practical whereas gesture is communicative, see Novack & Goldin-Meadow, 2017). In the study, interlocutors engaged in a game of TicTacToe where they had to verbalize their actions. Secondly, it did not test synchronisation with information structural categories themselves, rather it tested the synchronisation of maximum pitch excursions and prosodic boundaries with the apexes of co-speech actions (i.e., movement endpoints) using German multimodal data. The relation to information structure only comes from the study's setting where the moves were categorised according to their informativeness. The first and the final moves of the game were given (they were pre-set), and the rest of them were informative. Furthermore, interlocutors played the game under two mutual visibility conditions (when they can see each other and not).

Much like some of the studies in Sections 2.5.1.1 and 2.5.2.1, no explanation was given of the function of these maximum pitch excursions in German, and no description of exactly which prosodic phrases were investigated. There was also no synchronisation criterion in the sense of the present study. Instead, the study statistically compared the distance between the gestural and speech units under different conditions. In general, the findings showed that informativeness of co-speech actions caused tighter synchronisation of co-speech actions and prosodic anchors. The findings for visibility, on the other hand, did not show a consistent pattern. Overall, as the authors acknowledged, the co-speech actions are different to gestures. Therefore, the findings are not generalizable to speech-gesture synchronisation. Although the authors commented that these actions resembled deictic gestures in terms of form and function, as previously mentioned in Section 2.5.1, the temporal coordination of co-speech actions has been shown to differ from gesture's (see Novack & Goldin-Meadow, 2017 and references therein). Regardless, the study is one of the two studies that considered an effect of information structural notions on synchronisation.

To the author's knowledge, Ebert et al. (2011) is the only study to assume information structure as the sole domain where gestural phrases are anchored. The study tested whether gesture phrases were synchronised with contrastive and non-contrastive foci. The synchronisation was tested separately for the onsets and offsets of the gesture phrases (as previously described in Section 2.5.2). The onset of a gesture phrase could either be the onset of the stroke or preparation phase if there was one. However, for the offsets, the study only used the stroke offsets. They disregarded retraction phrases claiming that they are semantically neutral and therefore not fit for synchronisation. They also disregarded post-holds following Loehr (2004) who found intermediate phrases were (inconsistently) synchronised with gesture phrases only when post-hold phases were disregarded. The study interpreted this as "... they [post-holds] seem to have a different status as the other phases of a gesture phrase" (Ebert et al., 2011, p. 7). The present study disagrees with this view and argues that post-holds, by definition, exist to maintain synchronisation with speech. Further, unlike retraction phases, they are not semantically empty as post-holds contain apical information because they are essentially apexes frozen in time. The relevance of this view will be presented in Chapter 6.

In brief, the study tested the synchronisation of gesture phrase onsets and offsets with the onsets and offsets of the focal area. The findings for neutral (non-contrastive) focus synchronisation with the onsets showed a systematic shift where gesture phrases started about 310ms earlier than foci (sd= 410ms). The study noted that most gesture-focus pairings occurred within one second of each other, and interpreted this as evidence for synchronisation. For the offsets, they observed a mean time distance of 150ms where gesture phrases ended after foci. However, they also observed a huge standard deviation of 1240ms, and reported that some gesture phrases ended several seconds after their corresponding foci. They interpreted the systematic precedence of gesture as evidence for synchronisation based on earlier reports of precedence as reported in Section 2.2. Unlike the onsets, the offsets of gesture phrases and foci were not considered to be synchronised although the mean time difference between the offsets was smaller.

The synchronisation patterns of gesture phrases and contrastive foci were not as clear because the study had fewer contrastive foci in their data (260 neutral, 56 contrastive). For the onsets, both the standard deviation and the mean time difference were quite high (m= -770ms, sd= 770ms) where gesture phrases seemed to precede their foci by a long margin. The study did not present any findings for the synchronisation at the offsets. The study finally concluded that gesture phrases were not synchronised with contrastive foci (at the onsets).

Overall, the findings of the study regarding contrastive foci remained inconclusive due to the small number of observations. However, the study observed a synchronisation of gesture phrases and neutral focus at the onsets but not at the offsets. Similar to almost all other studies presented so far, the study failed to clarify what counted as synchronisation. It seems to have based its synchronisation decisions on standard deviation. A high standard deviation, although not clearly defined how high it should be, meant inconsistency which was taken to be evidence for asynchrony. The inconsistency could be a result of gesture type, which was not tested in the study. It is possible that the semantic functions of gestures may cause tighter synchronisation or even asynchrony. Moreover, the asynchrony they observed at the offsets may possibly be a result of the exclusion of post-hold and/or retraction phases.

Present Study

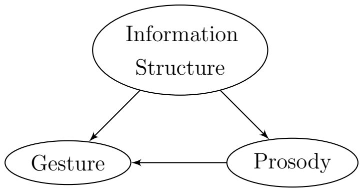


Figure 2.17: Three-way association of information structure, prosody, and gesture (modified from Figure 2.16)

As covered in Sections 2.5.1 and 2.5.2, the majority of studies in the literature have assumed that the synchronisation of gesture with speech is realised through prosodic structure. Considering the relationship between information structure and prosody and the studies that have shown an association between information structure and gesture (either implicitly or explicitly), the present study hypothesises that (1) gesture may be informed by information structural categories (2) gestural units may be synchronised with information structural categories. In order to test these hypotheses, the analyses in the present study focus on the synchronisation of gesture phrases with the information structural categories topic, focus, and background (i.e., IS units). The investigation of possible synchronisation with gestural units at other levels of the gestural hierarchy is left for future studies.

The first hypothesis means that gestural features associated with gesture phrases such as gesture type (e.g., deictics) may be sensitive to information structural categories (e.g., topic and contrast, excluding information status, see Section 3.4.3.1). It might be the case that particular gesture types are chosen to accompany parts of utterances with different discourse organisational functions, indicating a direct link between gesture and information structure.

It was already discussed in Sections 2.6.2.1 and 2.6.2.2 that prosody has a close relationship with both information structure and gesture. In light of these relationships, the second hypothesis claims a synchronisation of gesture phrases with IS units mediated by prosody in line with the implications of studies reviewed in Section 2.6.2. That is, gesture phrases may be synchronised with groupings of prosodic phrases as organised by information structure (the boundaries IS units are coextensive with the boundaries of prosodic phrases, see Section 3.4.3). Such a synchronisation would mean that gesture and information structure are synchronised, but indirectly through the medium of prosodic structure.

Through these two hypotheses, the present study hypothesises both direct and indirect links with gesture. It claims that gesture, prosody, and information structure form a three-way ensemble where both information structure and prosody govern gesture's synchronisation with speech. Figure 2.17 shows a basic schematic of this three-way synchronisation.

The analysis of synchronisation is conducted in the same way as the analysis of synchronisation for prosodic phrases outlined in Section 2.5.2.2. There are separate analyses for the onsets and offsets. The synchronisation of whole gesture phrases and IS units is assumed based on the synchronisation of their boundaries. First, gesture phrases are paired with these information structural categories depending on their semantic content (further explained in Chapter 6). Then, a series of analyses are done to reveal whether there are any pairing patterns between these categories and gesture phrases according to the topic/focus/background distinction, gesture type, and contrast (e.g., deictic gestures may often accompany topics). Finally, the actual time distances between onsets/offsets are calculated and statistically tested for synchronisation given the same synchronisation criterion (160ms) while also considering the effect of gestural and information structural context. On the whole, the final research questions of the investigation of macro level synchronisation between gesture and information structure are the following:

4. Are the onsets/offsets of G-phrases synchronised with the onsets/offsets of information structure units?

4.a Is the synchronisation of gesture phrases and information structure units affected by gestural and information structural contexts?

2.7 Summary

This concludes the review chapter. The chapter first gave an overview of our current understanding of speech-gesture interactions, and the psycholinguistic production models that were proposed to account for these interactions. Even though there have been many studies investigating synchronisation of gesture and speech (through prosody), it was highlighted that synchronisation has not been represented in these production models. Most of the studies investigating synchronisation focused on the synchronisation between gesture and prosody at the micro level. The consensus in these studies was that prominences in gesture and prosody are synchronised. However, these studies were not without shortcomings. A clear definition of synchronisation, a careful analysis of prosody as a system, and the potential effects of gesture type were overlooked in the majority of these studies. This was also true for the studies that investigated the synchronisation of prosodic phrases and gestural phrases (i.e., macro level). These studies showed a less strong synchronisation of phrases compared to the synchronisation of atomic anchors at the micro level. One common observation was that one phrase was often found to span over multiple phrases in the other modality. The weaker synchronisation of prosodic phrases as well as the implicit and explicit associations of gesture with information structure inspired a proposal that the synchronisation of speech and gesture may also be governed by information structure, which is also strongly linked to prosody. In line with this, the present study proposes a three-way synchronisation of gesture, prosody and information structure where anchoring of gesture is managed by both information structure and prosody.

The present study tests synchronisation between (1) apexes and tonal events (Chapter 4), (2) gesture phrases and prosodic phrases (Chapter 5), (3) gesture phrases and information structural categories (Chapter 6). It uses Turkish natural speech data elicited via a narrative task. It adopts a clear definition of synchronisation that is statistically tested, integrating the effects of gestural, prosodic, and information structural contexts.

3

Methods

The present study investigates the synchronisation of gesture with prosody and information structure using natural speech data. The research questions posed in Sections 2.5.1.2, 2.5.2.2, and 2.6.2 address the synchronisation of different units at different levels within the structures of gesture, prosody, and information structure (i.e., micro and macro levels). Moreover, the investigation also tests for the effect of other features of these units such as prominence, semantic function (gesture), and contrast (information structure). Therefore, the construction of a multimodal corpus with rich annotation was required to address the research questions posed by the present study. This chapter describes the methods employed in the construction of such a multimodal corpus for the present study. As we will see, various layers of annotation created in the corpus helped capture the complete gestural, prosodic, and information structural contexts in which speech and gesture synchronisation takes place. Furthermore, the richness of annotation in the corpus means that it can be used for other investigations related to the annotated modalities, which makes the creation of the corpus a contribution in its own right.

This chapter starts with a description of the participants, followed by descriptions of the design and the procedure of data collection. Next, the annotation schemes for gesture, prosody, and information structure are introduced. Finally, the statistical tests that are used throughout in the study are explained.

3.1 Participants

Ten (five female - five male) participants and one male confederate participant were recruited for the study. These participants were 18-26 year old native speakers of Turkish who were university students at a state university in Turkey. They were not recruited from a specific department - volunteers responded to ads posted on various locations within the campus. Upon responding, they were provided with an information sheet informing them about the study. Respondents who expressed further interest in participation also received a verbal description of the study from the researcher before signing a consent form allowing the collection and the use of multimodal data.

To avoid a possible effect of bilingualism on speech and gesture production, the participants recruited were as monolingual as possible within the Turkish context. English language teaching is practised as "foreign" language teaching (as opposed to second language teaching), and is a part of the primary school curriculum as a compulsory course starting from the age of ten (Kirkgoz, 2007). Therefore, none of the participants were truly monolingual. The participants were asked to fill in a background information sheet in order to provide more information about their linguistic background. None of the participants reported that they could speak a third language (the majority of them did not state that they could speak English either). Also, none of them reported having lived in another country for more than 2 years or having parents who could speak another language.

In addition to these participants, three actors (two female, one male) were recruited to assist with the preparation of the video stimuli used to elicit speech and gesture. These actors were the researcher's colleagues and friends living in New Zealand. They were 22-28 years old, and had Greek or Turkish backgrounds. They were informed about the study before recruitment, and also signed consent forms. Every participant and actor received a supermarket voucher for their contribution.

Only the confederate participant (the confederate hereon) and the actors knew about the actual topic of the study prior to filming. The participants did not know that the topic of the study was gestural behaviour. They were only told that the study was about how they conveyed information during narrations. This level of deception was necessary because being aware of the fact that their gestures would be examined could prevent natural gesture production. They only learned about the actual topic of the study after their filming session. At this stage, they were given a debriefing sheet, and it was also verbally explained to them why this information was withheld from them in the beginning of the session. They were also asked if they still wanted to participate after debriefing. They were informed that they could withdraw from the study at any time during or after the session but no participant withdrew at any point.

3.2 Design

The overall design of the study involved the participants watching 5 pairs of short videos and then one by one, recounting what they saw in the videos to the confederate who they believed had not seen the videos.

3.2.1 Video Stimuli

The videos the participants watched were designed by the researcher. They were shot in different environments with the help of the recruited actors. The actors acted out scenarios prepared by the researcher. The actors did not speak in the videos in order to ensure the re-usability of the same stimuli for speakers of different languages in future projects. The actors performed basic daily activities such as reading a newspaper, which added up to form a story. Each video had a scenario and told a full story with a sense of completion. The video lengths varied between 47-169 seconds.

The scenarios were written by the researcher in order to create a paired video design which enabled a methodical elicitation of a variety of information structural constructions. The videos in pairs told similar stories with the same actors. However, they had minor differences, therefore they were not identical to each other. For example, in the first video of one of the pairs, one of the actors was reading a book, whereas in the second video of the pair she was reading a newspaper. These minor differences were intended to elicit contrastive constructions from the narrations of the participants (see Sections 2.6.1 and 3.4.3.3). That is, the intention was that the participants would realise these differences (and similarities too) between two consecutive videos in a pair and then contrast these pieces of information in their utterances. Of course, there was no way of reliably predicting whether and how these would be expressed by the participants. However, the scenarios were designed in a way that offered at least ten differences between the videos in pairs. These differences strategically affected the actions and the doers of actions at roughly equal rates in order to maximize the chance that approximately equal numbers of contrastive topics and foci would be elicited (assuming that agents will be assigned as topics). These numbers were further boosted by the participation of the confederate, which is explained in

Section 3.3. Overall, this design made sure that in addition to neutral (i.e., non-contrastive) information structural (IS) units, contrastive IS units were also produced (see Section 3.4.3.3 for the relevance of contrast for the analyses in the present study).

The scenarios were also designed in a way that ensured the production of words with non-final stress. This stress type is the irregular form which is less common than the default word final stress in general. In order to elicit such words, certain objects such as *gaZEte*, 'newspaper' and *tornaVIda* 'screwdriver' which had non-final stress (stressed syllable in capitals) were inserted into the scenarios. In the scenarios, the actors interacted with these objects in order to increase the chance that the participants included these in their recounts. The inclusion of non-finally stressed words is relevant because the location of the pitch accent (associated with stress, see Sections 2.5.1.2 and 3.4.2.1) within words can potentially affect the synchronisation of apexes and tonal events. This possibility is investigated in Chapter 4.

Finally, in order to give a better idea about what these videos showed, the following describes the scenario of the first video that the participants watched.

> There are three chairs by the wall in a room. A man (on the left) and a woman (on the right) are sitting on chairs next to each other. They appear to be waiting for an appointment in a quiet room. The man is reading a newspaper while the woman is reading a blue book. The woman is also listening to music on an MP3 player. Then, a phone rings. They do not know whose phone it is (same ringtone). They both stop reading and start looking for their phones. The woman searches her bag and the man searches his pockets. After checking their phones, the man realises that it is his phone and declines the call. He then ges-

tures to the woman who is still looking for her phone in a way to communicate that it was his phone ringing, and he is sorry for disturbing her. The woman nods, smiles and continues reading. The man puts the phone back in his pocket and continues reading his newspaper. After a couple of seconds, a phone rings again. This time the woman does not look for her phone but the man does. He realises that it is not his phone ringing this time. He taps the woman on the shoulder and points at her bag to let her know that it is her phone this time around. The woman searches her bag for her phone as the man puts his phone in his pocket again. She thanks the man with a gesture and the man smiles at her. The woman leaves the room to take the call. The man stays and continues reading his newspaper.

As for the differences in the second video of the pair, they sat at different chairs. This time, the woman was reading the newspaper and the man was reading the book. The man's phone was in the bag and the woman had her phone in her pocket. At the first ringing, the woman's phone was ringing and she left the room to take the call. She returned to the room after a while. At the second ringing, only the woman checked her phone but it was not ringing. After she gestured to let the man know that it was his phone this time, he took his phone from his bag and left the room without thanking her.

3.2.2 Pilots

The videos and their scenarios were optimised through a piloting process. A few colleagues of the researcher were asked to watch the early versions of the videos and recount them. Their feedback on the content and naturalness of the videos as well as the researcher's own observations led to the final versions. The final versions of the videos were piloted with two native Turkish speakers living in New Zealand, and with two participants with the desired profile in Turkey. These pilots were also done to test the filming equipment (e.g., sound and video quality and so on) as well as to test the participant reaction to their setup. The pilots were filmed but not included in the analyses.

3.3 Procedure

The data collection took place in Turkey. The participants were invited for a 30-40 minute filming session in a quiet room in the university. The room setup included two chairs for the participant and the confederate, and two stands for the microphone and camera (see Figure 3.1). The chairs were facing each other with approximately 3 metres between them. The camera stand was behind the confederate's chair facing the participant. The confederate did not appear in the recordings visually. A Sony FDR-X3000R action camera was used to record the sessions. The videos were recorded at 1920x1080 resolution with 60 frames per second. The camera had an inbuilt microphone but it was not ideal for recording the participant's voice from such a distance. For this reason, an external microphone, a Samson Q7 super-cardioid dynamic microphone, was connected to the camera. The microphone was positioned behind the participant's chair pointing down at the participant's head. This way,

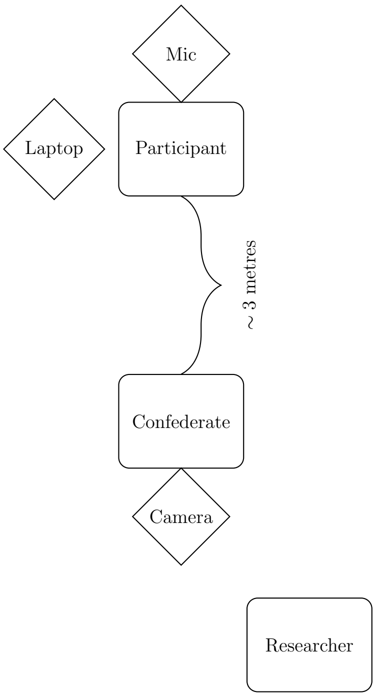


Figure 3.1: Room setup for filming

the microphone was not in the participant's vision during the recording, and its sound acceptance range was limited to the participant and the confederate.

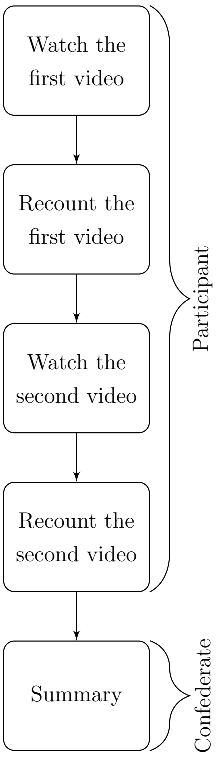


Figure 3.2: Task flow for one video pair during a data collection session

The researcher was sitting behind the confederate and was visible to the participant. He did not watch the participant's interaction with the confederate (was reading a book) and did not intervene during the recordings unless the participant had a question. The participants played the videos in the given order themselves (short to longer videos) on a laptop that was on a table next to their chair. The participants watched the videos one by one, and recounted what they saw to the confederate immediately after watching the video. They were allowed to play the videos and renew their recounts as many times as they wanted. In the verbal instructions they received, they were encouraged to pay attention to the differences between the videos in the pairs. This was to maximize the chance that they would produce contrastive constructions. After the recount of each pair, the confederate gave a summary of what he got from these recounts in order to check understanding for his own supposed task afterwards.

3.3.1 Role of the Confederate

The confederate's task was supposedly based on what he understood from the participant's recounts of the videos. This was done to make the participant's task more meaningful by giving the task a purpose and to encourage them to include as much detail as possible in their recount in order to help the confederate with his task. The confederate also functioned to offer the participant a natural communication target instead of just asking them to talk to a camera. In addition, the confederate helped elicit more contrastive constructions from the participant. The confederate was trained to make deliberate mistakes during his summary (e.g., confuse the order of events or the actor that performed certain actions). In the instructions, the participant was encouraged to correct the confederate if he misunderstood something about the video - it was important for the confederate to get everything right since he would have a task about his understanding of the recounts. The participants did not know about the confederate's real role, and that he actually had seen these videos before. They were told that the confederate was just another participant who would have another task after the participant left. Overall, these deliberate mistakes were intended to cause the participant to intervene and correct the confederate, which in turn would result in the production of replacement subtype of contrast (Götze et al., 2007), increasing the variety and the number of contrastive constructions in the data (see Section 3.4.3.3). The confederate decided where he would make the mistakes during the participant's recounts. It was not possible to make these mistakes about same points for every participant since they did not cover everything in the videos in the same way in their recounts. The confederate was also allowed to take the opportunity to elicit various forms of constructions by asking for clarification during the participant's recount in cases where they delivered information that was not very clear. The confederate could state that he did not understand and ask for repetition in these cases. The confederate was allowed to talk and nod during the participant's recounts, but his speech or gestures were not analysed. Figure 3.2 shows a summary of the procedure for one video pair. The procedure in the figure was repeated for each video pair.

3.4 Annotation

Six hours of multimodal speech data were collected across all data collection sessions. In order to address the research questions of the present study, the data had to be transcribed, translated, and manually annotated for gesture, prosody, and information structure. These included the annotation of all phrases (or units) within their respective structural hierarchy as well as the annotation of various features of these phrases (e.g., prominence, semantic meaning). As the reader will be able to infer from the annotation schemes described in this section, such multi-layered annotation of the data for all 10 participants would require immense effort and time. Therefore, the multimodal data from only four participants (two female - two male) could be annotated given the timeframe of the present study.

Not every utterance in the data was transcribed. First, the utterances that were accompanied by gestures were identified. Amongst these, the utterances that had verbal disfluencies and those that were accompanied by interrupted gestures were excluded. The utterances to be transcribed and annotated were sampled using a simple random sampling method. The present study did not use stratified random sampling in order to control and equalise the numbers of gesture types, prosodic phrase types or topics/foci in the samples. These have an imbalanced number of occurrences in natural speech since only some of them are obligatory and non-recursive within the structure they exist in (e.g., focus, nuclear intermediate phrase). Therefore, by not controlling the sampling in this way, the proportions of these phrases and features should be representative of natural speech, which is what the present study wanted to capture.

One constraint on sampling was the length of the sample. The minimal unit that could be sampled was the gesture unit (G-unit, see Section 2.1.1). This meant that the samples could not end before the gestural movements were completed and the gesturing hand was retracted to a rest position. One exception to this was that some G-units could reach unusually long durations where the hands were not retracted for 1-2 minutes. In these cases, the samples only contained approximately the first 30 seconds of the G-unit.

The transcribed and annotated data were comprised of utterances and gestures extracted from the monologues of the participants and their dialogues with the confederate. Only declaratives were annotated but the annotation was not limited syntactically to the annotation of the canonical word order (i.e., SOV). No connection between gesture, prosody and information structure was assumed while annotating them. The annotation of these was carried out separately and without access to the other annotations so that any findings of correlation that might be found was genuine.

The first step of the annotation was transcribing and translating the sampled speech. This process was done using Praat (Boersma & Weenink, 2018). Then the utterances were segmented and time-aligned at the word level using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017). The output of the segmentation and alignment process was hand-corrected. The annotation of prosody and information structure was done separately in Praat, making use of these transcriptions and segments. Gestures were annotated in ELAN (ELAN, 2019) without referring to any prosodic and information structural annotation. All the resulting annotations were transferred to ELAN which allowed them to be easily imported to R (R Core Team, 2018) for processing.

Sections 3.4.1 to 3.4.3 below describe the annotation schemes used to annotate the relevant phrases and features of gesture, prosody, and information structure as required for the analyses of the present study.

3.4.1 Gesture Annotation

This section provides a detailed account of the gesture annotation scheme used in the present study. It was built on the guidelines described in McNeill (1992), Kita et al. (1998), and Loehr (2004). The units of annotation described in this scheme are generally not very different from what was described in these earlier studies. However, the annotation scheme of the present study contains very detailed descriptions of the actual annotation practice including issues that may be encountered while annotating, and practical information about how these can be handled. This contributes to the replicability of the annotations and offers more comprehensive guidelines for future studies.

This scheme was used to annotate co-speech gesture which is any kind of body movement that is produced spontaneously during speech (Section 2.1.1). This definition only includes gesture that has affiliations with speech at semantic, pragmatic and discursive levels, and therefore classifies self-touch (e.g., itching, hair combing) and manipulation of objects (e.g., clothing) as non-gesture. Since the present study is interested in the relationship between speech and gesture in general, the kinds of gesture where speech was optional (i.e., emblems) or where its absence was obligatory (i.e., pantomime) were not annotated. The annotation of gestures was limited to hand/arm gestures only. This decision was made to reduce the immensity of effort of annotating different gesturing body parts at the same time. Limiting the annotation in this way enabled the researcher to annotate a greater number of hand gestures, which increased the strength of the analyses conducted. Secondly, eye, head or torso movements are kinematically more limited than the hand, which has implications for their segmentation and identification. For example, what defines phases such as preparation or retraction for the head or torso is less evident. As we will see in Chapters 5 and 6, these phases (and phrase structure in general) have an important role in the present study. For these reasons, only manual gestures were annotated. Lastly, differently from

McNeill (1992), hand shape and hand position were not annotated because these features do not help answer the research questions posed here because the present study is interested in the timing of gestures and not in the semantic content expressed through various hand shapes or positions.

Gestures were annotated without any reference to the audio files. The annotation of gesture relied on the speech transcriptions as well as the segmentation of the utterances at the word level. This was done in order to prevent bias that may be created by listening to the prosody of the accompanying speech (see Yasinnik et al., 2004). There were exceptions where listening to the accompanying speech was necessary. These exceptions happened when there were disruptions in speech caused by hesitation or self-correction. The interpretation of gestural movements during these disruptions without reference to the audio could sometimes be very challenging.

The practice of annotation consisted of two main steps which are the segmentation of gestural movements and the identification of these segments. These were briefly introduced in Section 2.1.1. In the next two sections, the annotation scheme used in the present study is described in detail.

3.4.1.1 Segmentation of gesture

There are different levels of organization in gesture. Figure 3.3 shows the hierarchy of organization of gesture (adapted from McNeill (1992)).

The original hierarchy in McNeill (1992) shows two more tiers above gesture unit (G-unit) being “consistent arm use and body posture”, and “consistent head movement”, which was not annotated in the present study as they were not sufficiently relevant for the study. In addition, McNeill’s (1992) hierarchy does not have the apex level (explained in Section 3.4.1.3).

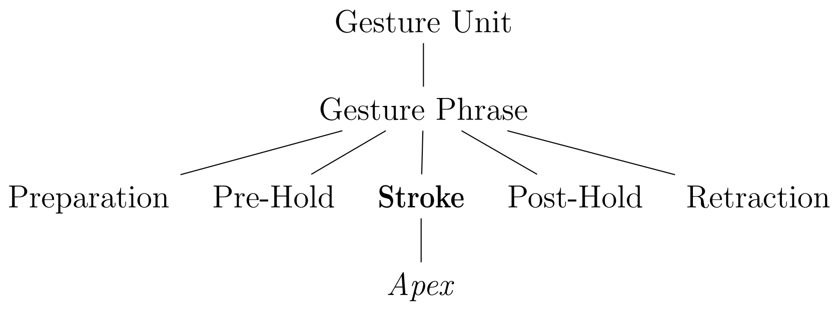
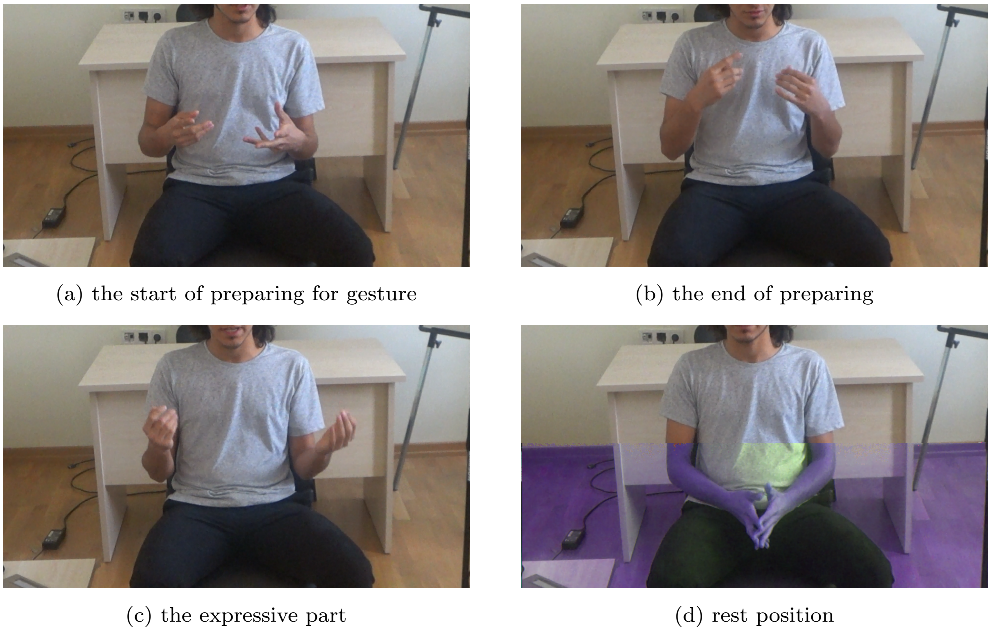


Figure 3.3: The hierarchy of phrasal organisation of gesture

Within the scheme described here, the largest annotated unit was the G-unit and the smallest was the apex. Apexes are different from the phrases in the hierarchy as they do not come together to form larger units unlike the others. Instead, they are dynamic events that take place during the stroke (this relationship is similar to the tonal event-prosodic phrase relationship which is explained in Section 3.4.2).

G-unit

A G-unit starts at the moment when the gesturing hand starts moving in order to depart from a rest position, and ends when it returns to a rest position. A rest position is a state where the hand is supported by an object (e.g., armrest or table) or a part of the body (e.g., lap, see Figure 3.4). Self-touch (e.g., adjusting hair) and touching clothes or other objects (e.g., ring or glasses) were also considered as rest positions as per McNeill (1992). The first moment when the hand touches the relevant body part or the object is the start of the rest. At the end of the G-unit, the hand did not always return to a fully supported rest position in the data. Instead, it might hang in the air in front of the chest, the neck or the stomach. In these cases, the rest was achieved by a default curved, relaxed state of the fingers and the palm which was sometimes accompanied by limp wrists close to the body, as seen in Figure 3.5.



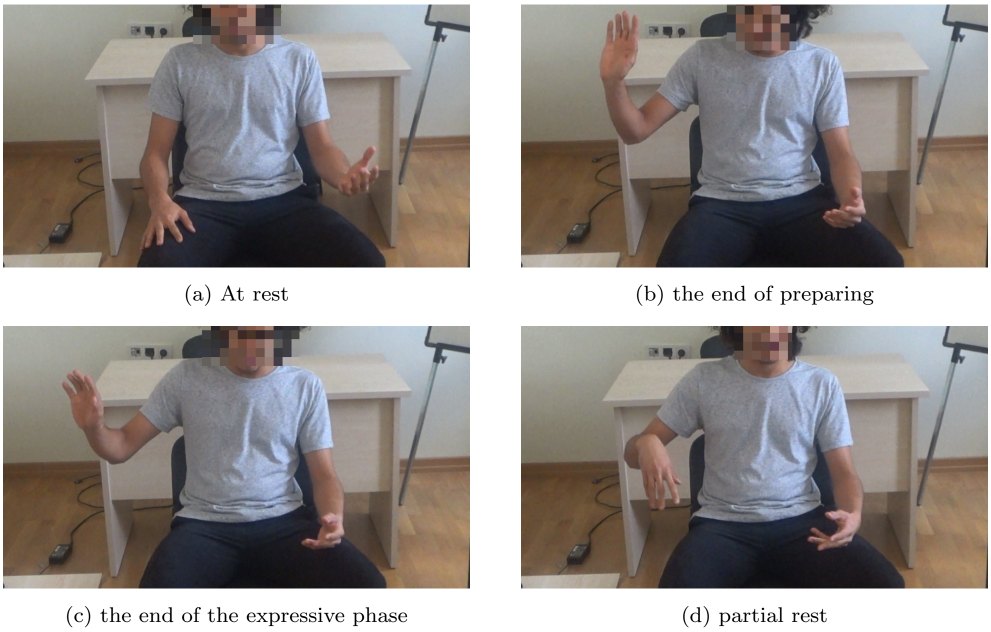
(a) the start of preparing for gesture

(b) the end of preparing

(c) the expressive part

(d) rest position

Figure 3.4: A gesture unit ending in full rest



(a) At rest

(b) the end of preparing

(c) the end of the expressive phase

(d) partial rest

Figure 3.5: A gesture unit ending in partial rest (the right hand)

In ELAN, there was a separate tier for the G-unit annotation. The tier was not associated with a gesture type (e.g., iconic or deictic). The annotation on this tier simply stood for a period of time where gestural movements took place. In the data, G-units usually had multiple gesture phrases (G-phrases) within them. The onset of a G-unit coincides with the onset of the first G-phrase that it contains, and its offset coincides with the offset of the last G-phrase in it. G-units were marked as "SGU" if they contained a single G-phrase and as "MGU" if they contained multiple G-phrases. Finally, ! was put at the end of SGUs and MGUs (e.g., SGU! or MGU!) if only a portion of the G-unit, not all of it, was annotated. This was a result of the sampling of gestures for annotation as explained previously. These annotations are referred to as "incomplete G-units".

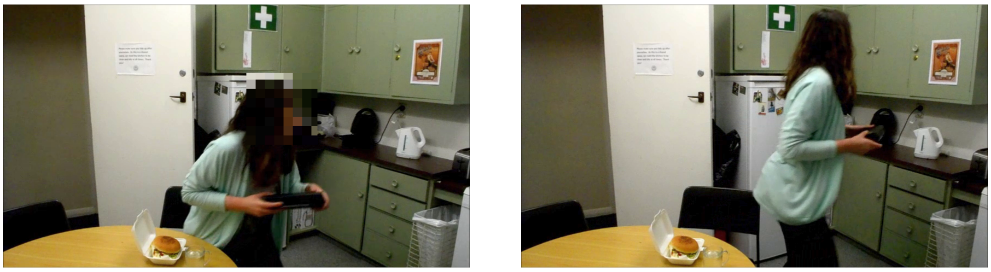
Available G-unit markings
on the tier "GUnit": SGU, MGU, SGU!, MGU!

G-phrase

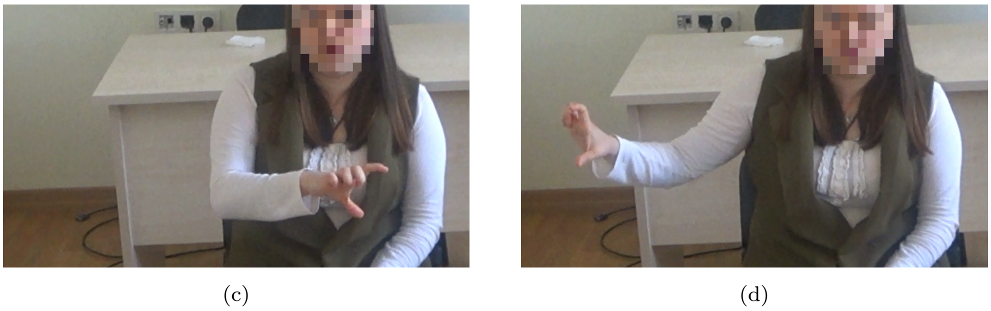
G-phrases occur within G-units. A G-phrase "corresponds to a single meaningful unit of bodily action such as pointing or a depiction" (Kendon, 2004, p. 108). The semantic functions of gesture (i.e., gesture type) were marked on G-phrases. It is actually the stroke phase that carries the semantic content of a gesture. However, since there can only be one stroke phase within a G-phrase, marking the gesture type at the G-phrase level made no difference. The classification of the gesture types based on whether or not the gestures demonstrated a discernible meaning and if so what kind, was done as per McNeill (1992). The gestures with discernible meanings (i.e., imagistic gestures) were "iconic", "metaphoric" or "deictic".

Iconic gestures portray imagery and have a close semantic relationship with the co-occurring speech. In the data, iconics often appeared as the re-enactments of actions or as the descriptions of object shapes or sizes that the participants observed in the stimuli. These re-enactments and descriptions did not need to capture all the details of the actions and objects to qualify as iconics. Gesture often selects and depicts one or two aspects of what is expressed in speech.

The marking of iconicity was not just based on what was encoded in speech but also on the researcher's (i.e., the annotator's) knowledge of the scene in the stimuli. Gesture may also complement speech by depicting an aspect of the idea that was not explicit in speech but present in the video stimuli - as both gesture and speech can express what was overlooked by the other (Section 2.2). Moreover, an iconic is not necessarily mimicking. In this scheme, the form and manner of an iconic did not need to match exactly what the participant said or watched in the video stimuli. For example, in the video stimuli, an action could be performed with both hands but the co-occurring iconic could relay the same idea with only one hand as in Figure 3.6. What was important here was that the gesture bore some similarity to what was in the video stimulus or speech.



(a) (b)



(c) (d)

Figure 3.6: (a) and (b) show the video stimuli that the participants watched. (c) and (d) show the gesture performed for that particular scene along with the utterance “she picked up the radio and put it on the counter”.

Metaphoric gestures also portray imagery. The difference between iconics and metaphorics is that an iconic portrays a concrete idea, whereas a metaphoric portrays an abstract concept. It creates a visible image for what is invisible, and we perceive that image as similar or logically connected to the invisible (McNeill, 1992). For instance, the right hand making circles in front of the chest while referring to a repetitive process of asking for a newspaper from somebody again and again is a metaphoric gesture. It visualises “the process” as a circle, and repetitiveness is communicated through multiple circles of the same size. In this scheme, the occurrence of a metaphoric gesture did not depend on the overt expression of abstract ideas in speech. Gestural metaphors could also exist as narrative tools. The presentations of new ideas, movements between discourse segments, and in particular, switching to a temporary extra/para-narrative style often attracted metaphorics without clear co-speech counterparts in the data. Although not very often, it was also possible to get metaphorics accompanying concrete ideas in speech. In these cases, the participants created a visual metaphor for a physical entity in speech in their mind, and only the gestural expression reflected that metaphor. For instance, the wiggling of all fingers along with the utterance “she was pressing the buttons” would be a metaphoric. The wiggling of fingers had no resemblance to the actual pressing action expressed in the speech or observed in the stimulus, so it was a metaphor for detail or fine-tuning, which was encoded in subtle finger movements. Metaphorics were more complex than iconics and harder to annotate because there was an extra step of finding a referent in speech and figuring out the metaphor for annotation. There were times when the researcher was not able to see the metaphor encoded in the gesture but the gesture was marked as a metaphoric regardless. In those cases, a process of elimination was followed. If the gesture was imagistic (i.e., not a beat) and not a deictic, and its iconicity matching with a concrete idea in speech or in the video stimulus was obscure, then the gesture was assumed to be a metaphoric. This method was plausible for the purposes of the present study because the research questions were not interested in actual gesture meanings but the general gesture type.

Deictic gestures are pointing gestures. They indicate the positions of objects or trajectories in space. The pointing may be at real targets that are present in the concrete world around the participant or at targets placed in the mental space of the participant during their narrations. In the data, there were many examples for both kinds. The most common example for the latter happened when the participants gestured for temporal expressions such as "before" and "after". In these situations, the participants pointed at a location usually behind their back for "before" and at an area in front of their chest for "after". Typically, a deictic gesture was performed with the index finger but using the whole hand with closed fingers was quite common as well. A deictic hand gesture usually had an arrow-like shape with the arm extending to a target from a location, giving the pointing a perspective - a reference point. However, this arrow-like hand/arm movement was not a criterion in the annotation of deictics. An outward sideways flap of the hand from the wrist could be a deictic when the gesture was referring to an "other". The reference point could be the participant's own body or the location of another object. In the before/after example, the participants used their own body as the reference when "before" was behind their back and "after" was in front of their body. The presence of a reference point or some form of covert referential information was crucial in deciding whether a gesture is deictic.

Beats are different from the gestures mentioned so far as they are non-imagistic (i.e., they do not have any discernible meaning). It was relatively easy to recognize a beat dynamically. A beat usually had one or two phases which are typically up and down movements or tiny circles (McNeill, 1992). There could be quite big and noticeable beats but they were usually small and quick flicks of hands. In this scheme, a beat did not have to be performed in its own gesture space. It could be performed at a rest position or during one of the phases of an iconic, metaphoric or deictic gesture. The latter was the most common occurrence in the present study. In particular, long preparation and post-hold phases tended to attract beats (sometimes multiple beats). While annotating beats, first it was decided whether the gesture had any meaning and what its movement characteristics (i.e., kinematic structure) were. The beat filter test (McNeill, 1992, p. 81) was applied when in doubt. The beat filter is a test with multiple questions that filters out imagistic gestures. If the answer to a question in the test is yes, a score of 1 is noted. The end score reveals the probability that the given gesture is a beat.

It is easy to provide general definitions of gesture types. However, fitting individual visual datum into one of these types can be challenging. A part of the problem was that a gesture rarely had the features of only one of these gesture types. They often appeared as conflations of iconic/metaphoric gestures with deictic gestures. The present study was not interested in this intersectionality of gesture types. Such an addition would make the annotation even more time-consuming and create extra variables that may or may not make a difference for its research questions. Therefore, only the most dominant gesture type was annotated in the present study.

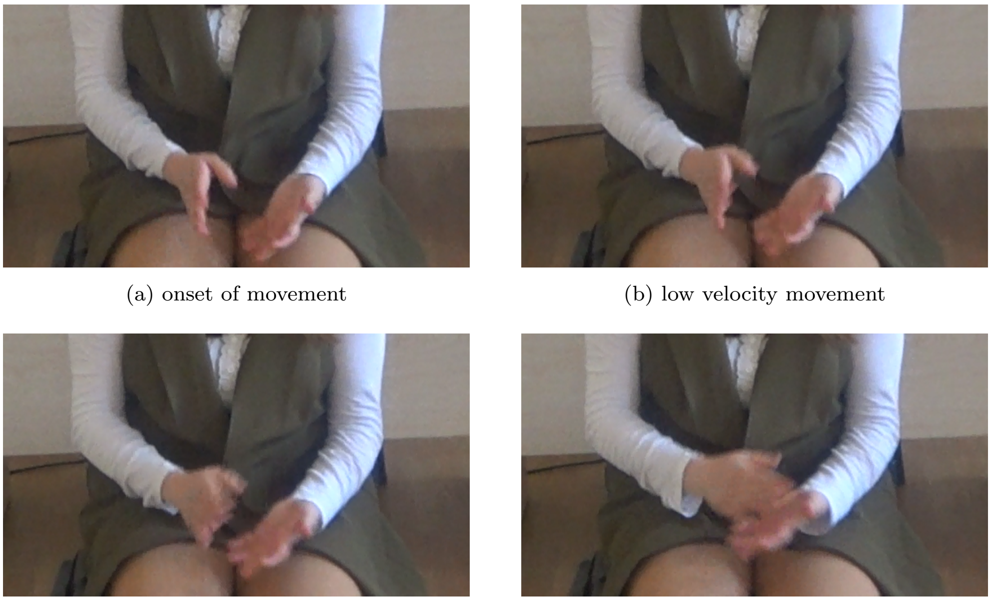
In ELAN, G-phrases were marked with the gesture types they contained. Imagistic gestures (iconics, metaphorics, and deictics) shared one tier and had the labels "Iconic", "Metaphoric", and "Deictic". Non-imagistic gestures (beats) had their own tier as they could be superimposed on other gesture phrases, which was usually the case in the data. The intervals where beat gestures took place were marked as "Beat".

Available G-phrase markings
on the tier "Gphrase": Iconic, Metaphoric, Deictic
on the tier "Beats": Beat

G-phase

The G-phrase is a sequence of discrete G-phases. The onset of a G-phrase coincides with the onset of the first G-phase that is within the G-phrase, and its offset coincides with the offset of the last G-phase within the G-phrase. There are five kinds of G-phases: "preparation", "pre-hold", "stroke", "post-hold", and "retraction" (McNeill, 1992). There can be only one of each within a G-phrase. The stroke is the only obligatory G-phase within a G-phrase. The other G-phases are organized around the stroke. The annotation of G-phases involves segmenting a G-phrase into qualitatively different phases with different kinetic characteristics. These characteristics can be organized under three main categories that are "movement to a target location", "hold in location", and "expressive movement (stroke)" (Kita et al., 1998).

The boundaries between G-phases were marked based on two criteria: (1) change in direction, (2) change of velocity profile (Kita et al., 1998). If the hand's velocity increases/decreases after a change in direction, this constitutes a change in the velocity profile, which marks a phase boundary. Within the scheme, the change in velocity did not need to be gradual; very brief stops/pauses between movements with distinct velocity also marked boundaries. These velocity changes and stops had minimal durations and they were impossible to catch when the videos were played at normal speed. Even with a frame-by-frame analysis, the velocity changes corresponded to the increase or decrease of blurriness of the hand in motion. This blurriness was the main perceptual cue for annotation - the blurriness nearly disappeared (never fully) at the phase breaks. Figure 3.7 shows how velocity and direction changes appear on video and the resulting phase boundary between (b) and (c).



(a) onset of movement

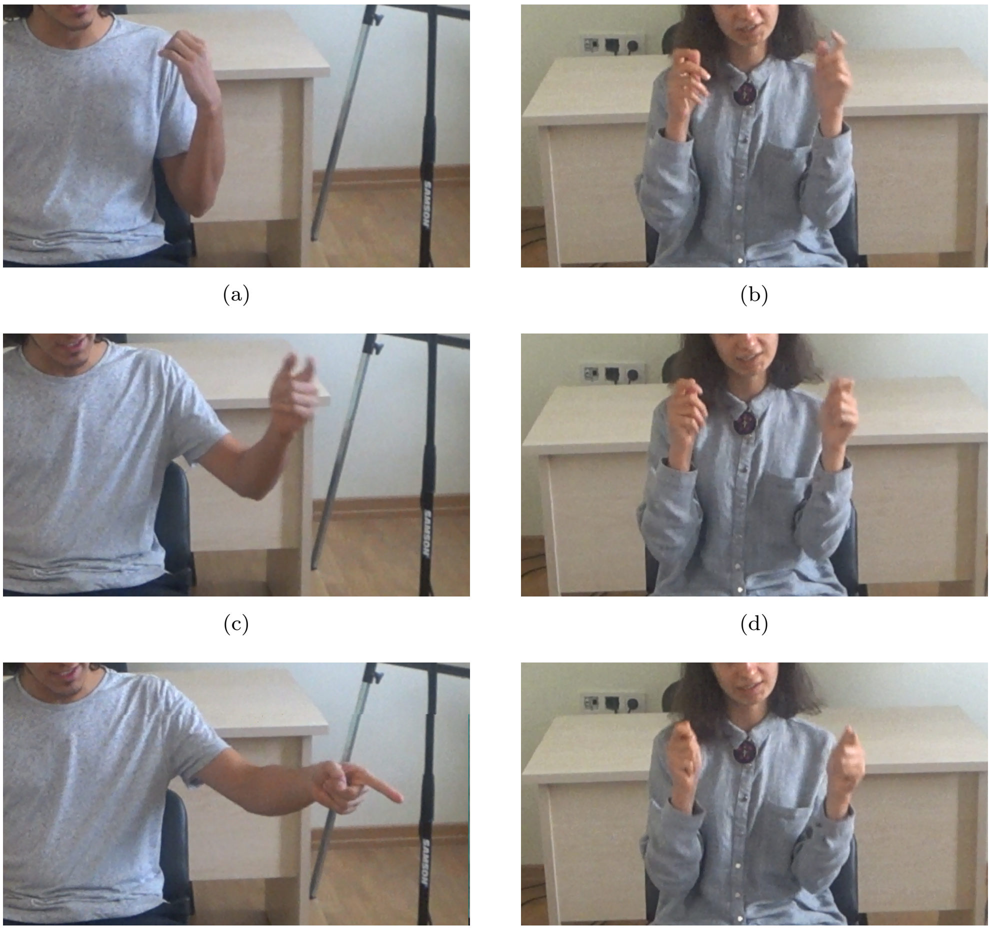
(b) low velocity movement

(c) direction change

(d) increased velocity

Figure 3.7: Four consecutive frames demonstrating a phase boundary between (b) and (c)

These two criteria applied both to the whole arm movements where the hand was relocated in the gestural space (on the left in Figure 3.8), and to within-hand movements where the hand was at a fixed location but there was a change in the orientation of the fingers or the palm (on the right in Figure 3.8). If the gesture included both at the same time; that is, if there were within-hand movements during a whole arm movement, these within-hand movements were disregarded for segmentation, and were considered as parts of the whole arm movement.



(a)

(b)

(c)

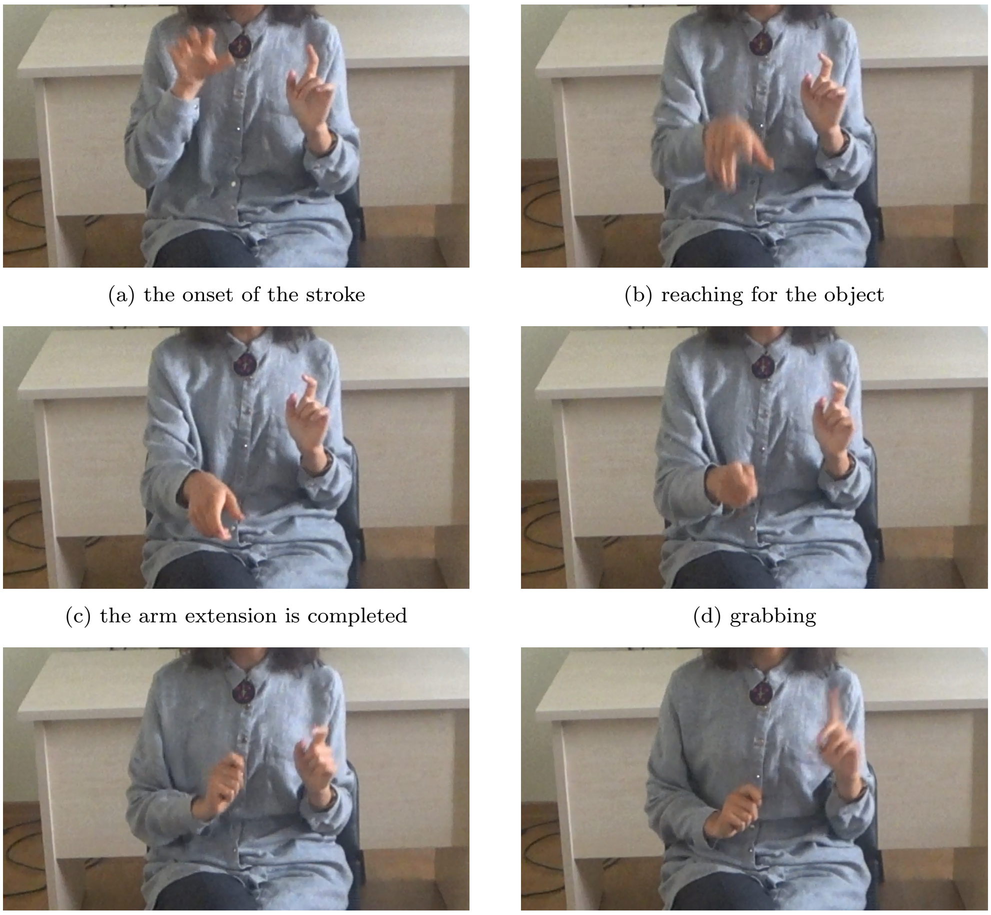
(d)

(e)

(f)

Figure 3.8: An example for whole-arm movement (left) and within-hand movement (right)

There were many cases in the data where the hand had multiple changes in direction but the velocity profile stayed the same in the stretches of movement before and after the direction change. These cases were marked as one G-phase. These are called "multi-segmented phases" (Kita et al., 1998). The most common examples in the data involved drawing the shapes of objects such as a table where the hand made a direction change at each corner of a rectangle. Another common example was grabbing gestures where the arm extended, got a hold of an entity, and pulled it back to the participant's torso or put it to another target location as seen in Figure 3.9.



(a) the onset of the stroke

(b) reaching for the object

(c) the arm extension is completed

(d) grabbing

(e) pulling the object back

(f) the offset of the stroke

Figure 3.9: The stroke of "She grabbed one from the middle"

The repetitions of the same movement (e.g., oscillations of body parts) were considered as one phase so long as the hand did not freeze/pause at a certain location (a pre- or post-hold) between the repetitive movements. For instance, the movements where the index finger traced a swinging trajectory back and forth multiple times were marked as one phase. In line with the annotation of within-hand movements during the whole arm movements, the repetitive movements occurring during the whole arm movements did not constitute separate phases but formed segments of one larger phase.

In addition to these kinematic qualities, the phases had functions such as expressing meaning and enabling the expression of meaning. These functions form separate phase types that are linearly organized. In the next section, these G-phase types are defined with their function within the G-phrase. These are based on the guidelines of Kita et al. (1998) and McNeill (1992).

3.4.1.2 Identification of G-phase types

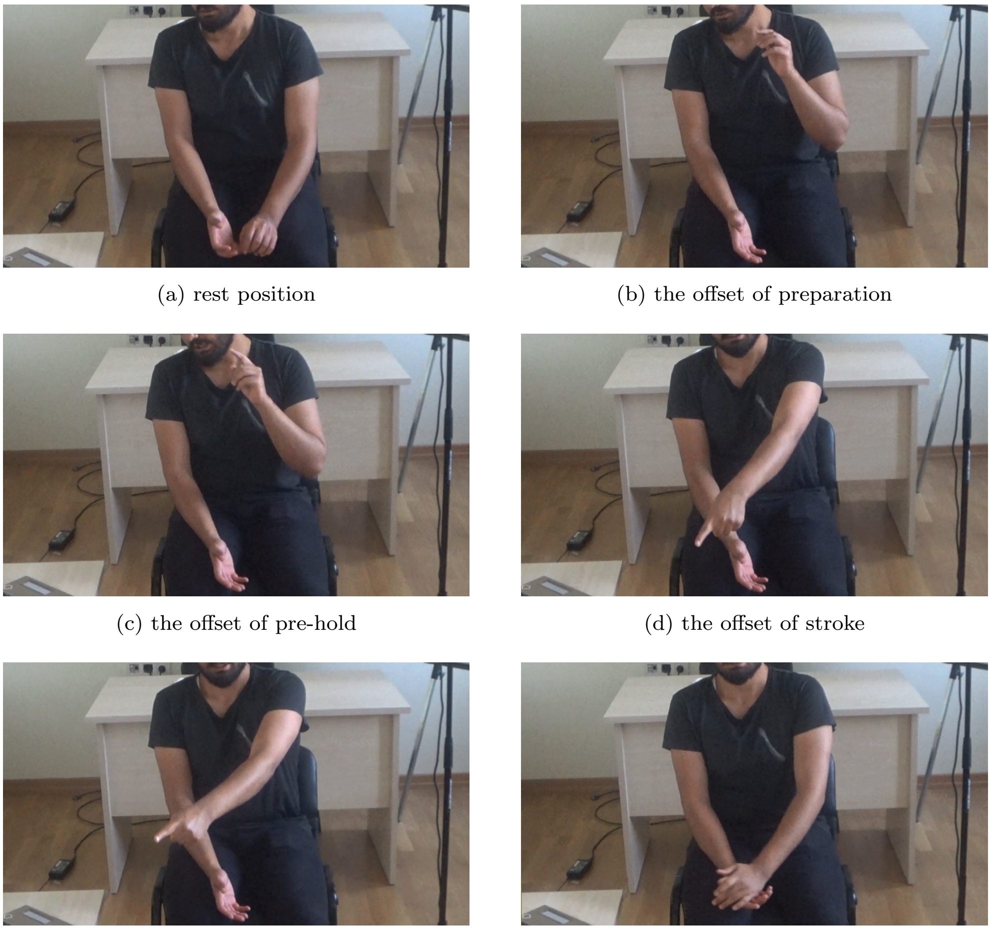
i. The *stroke* is the only obligatory phase in a G-phrase. It is where the meaning of the G-phrase is expressed and where the dynamic effort is at a maximum. The effort focuses on the form, orientation, and trajectory of the hand. In the data, the hand movement in the stroke was usually faster or more forceful than in the other types of G-phases. In rare cases, the movement could be slower than in the surrounding phases. The general idea was that the stroke was dynamically and semantically more prominent than the other phases around it. There might be minor changes in velocity within the stroke. The stroke often started a bit slower; the velocity peaked in the middle of the stroke and slowed down towards the end. This velocity change was different from what was observed at the phase boundaries as the blurriness of the hand motion did not decrease dramatically and was still very discernible. The stroke could consist of only one segment as well as multiple segments. Not every multi-segmented phase was a stroke (cf. Kita et al., 1998). For example, beats superimposed over another phase could make a phase look like it is multi-segmented - the boundaries of these beats constituted changes in direction and velocity of gestures. Also, the within-hand movements during the preparation phase described below could constitute a multi-segmented phase. Phases with repetition (oscillations) were always strokes as any repetition maximized the dynamic effort.

ii. In the *preparation* phase, the hand departs from a rest position and reaches a location where the stroke is going to be performed. Unlike the stroke, where the effort concentrates on the form/shape of the movement (i.e., focus on meaning), in the preparation, the effort of movement concentrates on the delivery of the hand to the onset of the stroke (i.e., focus on transition). A move from a rest position is not necessary - the preparation can start from the offsets of other phases. If there are multiple consecutive G-phrases, any hand movement that is not a beat between two strokes is also a preparation. In the data, a preparation phase often started before any actual hand movement. There might be a "liberating movement freeing the hand from a constrained position such as the undoing of the interlocking of fingers" (Kita et al., 1998, p. 29). The liberating movement could also be an elbow movement lifting up the hand and starting the actual hand movement. In this scheme, the onset of the preparation was marked at the onset of the liberating movement if there was one. While the hand was being carried to the stroke position, there might be within-hand preparation movements such as palm and finger orientations carrying the onset values for the stroke. These are considered as parts of the preparation. The offset of the preparation was the arrival of the hand at the stroke location. Any kind of within-hand adjustment after the hand reached the stroke location was a part of the stroke. The preparation phases were sometimes left out of G-phrase structure when there were multiple G-phrases lined up for production.

iii. A *retraction* is the movement where the hand reaches a rest position. Just like the preparation, the focus of the movement is to reach the offset of the phase. The retraction is semantically empty and there is not much effort involved. In the data, sometimes a retraction could look forceful if the hand was relaxed and fell on a rest position fast because of gravity, or if there was a beat during the retraction. As mentioned above, a rest position could be a position where the hand was supported by an object or by the body. Self-touching (e.g., hair-combing or adjusting clothing) were considered as retraction targets as well. If the interval between successive G-units was short, the hand returned to a partial retraction location in which the hand exhibited indications of relaxation by hanging freely in the air often with limp wrists or curled fingers and palm. Retraction phases were not marked unless the hand underwent changes compared to the preceding phase, regardless of its type. The first full touch of the hand (usually the side palm) at a rest location marked the offset of the retraction.

iv. A *pre-hold* exists between a preparation and a stroke where the hand stops moving in location. In the data, the hand was rarely in full standstill during these phases - it could drift slightly or twitch. There were not many pre-holds in the data. They usually happened when gestures were cancelled during preparation or there were increased pauses between utterances. These pauses might be caused by a speech error, hesitation or a delay in lexical retrieval. The pre-hold functioned to give speech production some time to catch up with gesture production (detailed in Section 7.2). There might also be beats during the pre-hold. In rare cases, the hand performed a mini-preparation right before the following stroke. This mini-preparation looked like another within-hand preparation movement where the starting values of the stroke (e.g., finger orientation) assigned by the preparation were re-adjusted. These mini-preparations were not annotated separately and were considered to be a part of the pre-hold phase.

v. A *post-hold* is dynamically the same as a pre-hold in that the hand performs a hold but this time right after the stroke. Similar to the pre-hold, the hand was not completely frozen and could drift about the gesture space slightly in the data. The most important thing about the post-hold is that almost all features of the hand form present in the last frame of the stroke are preserved throughout the phase. In this scheme, a beat during a post-hold did not end the post-hold if the hand shape was retained and the hand returned to the post-hold location at the end of the beat. If the hand moved to another gestural location for the beat and did not return to the post-hold location, or if the hand shape changed during the execution of the beat, the post-hold ended at the beginning of the beat.



(a) rest position

(b) the offset of preparation

(c) the offset of pre-hold

(d) the offset of stroke

(e) the offset of post-hold

(f) the offset of the retraction

Figure 3.10: All G-phase types within a G-phrase accompanying “the man points at it”

In ELAN, the G-phase annotations were on the tier “Gphase”. Figure 3.10 exhibits a G-phrase that has all G-phases possible in order from (b) to (f). For beats, G-phase types were not annotated since they had a standard biphasic organisation. Also, they were not used in the investigation of phrasal synchronisation (Section 2.5.2), therefore this segmentation would not be helpful in answering the research questions of the present study. Because they could be superimposed over other phases, they were annotated on a separate tier and treated as independent from the phrase structure. In brief, only the entire intervals of beats and their apexes were annotated.

Available G-phase markings
on the tier “Gphase”: Prep (=preparation), Pre-Hold, Stroke,
Post-Hold, Ret (=retraction)

3.4.1.3 Apex

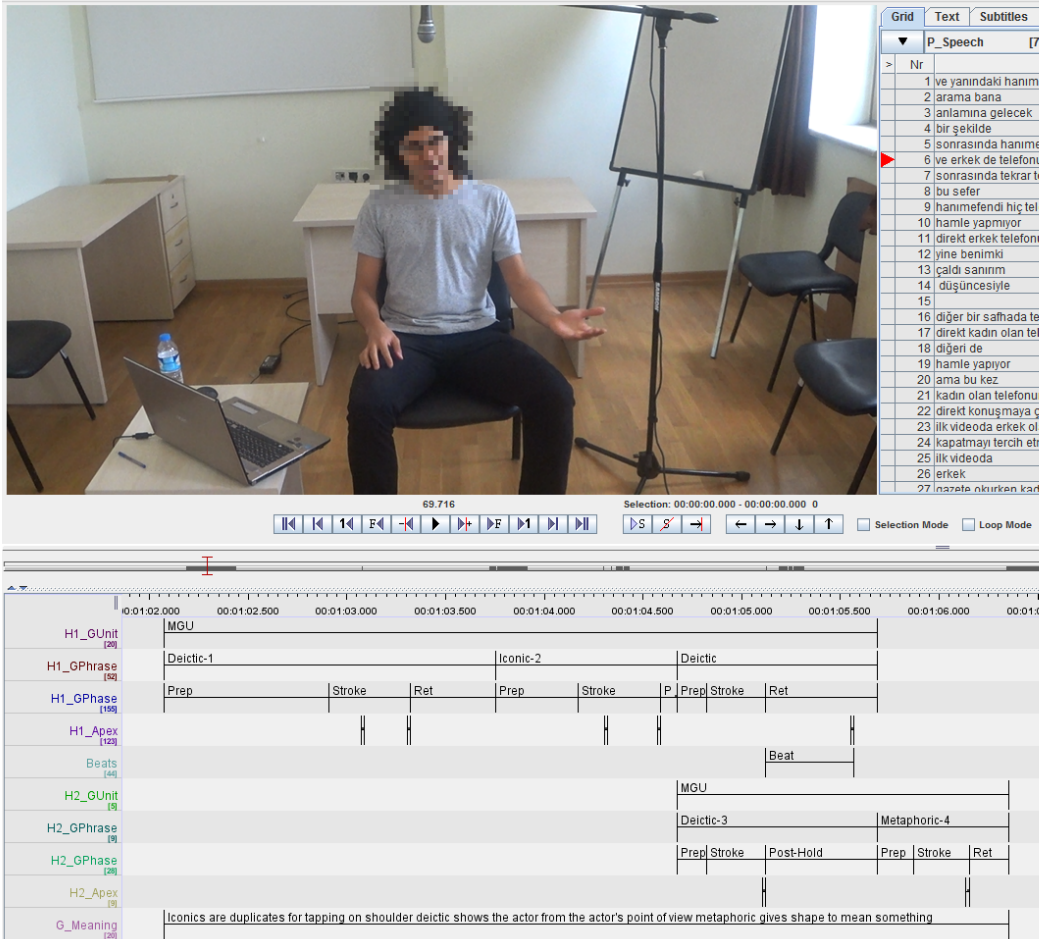
All of the gestural units described so far are relatively long intervals in time. In order to be used in the synchronisation of tonal events, a shorter, meaningful unit within gesture had to be identified. The present study adopted “the apex” for this purpose (Section 2.5.1.2). The apex is “a single instant which could be called *the apex* of the stroke, the *peak of the peak*, the kinetic *goal* of the stroke” (Loehr, 2004, p. 89). Syntagmatically, if the stroke has only one segment as in most deictics, then the apex is the offset of the stroke (the very endpoint of the last frame) where the arm completes its extension and the hand reaches its target. In multi-segmented strokes such as the grabbing gesture example (Figure 3.9) or the rectangle table gesture example, each change in direction is an apex (Shattuck-Hufnagel et al., 2007 define these as “hits”). Within-hand movements of the fingers and the palms could also be apexes. In the grabbing gesture example, the moment where the fingers closed on the palm to express the act of grabbing (Figure 3.9d) constituted an apex as the target of the grabbing action was a closed fist in addition to the frame where the act of pulling back was completed (Figure 3.9f). The position of the apexes could be hard to locate because of the nature of certain gestures. For example, circular gestures did not present such obvious direction changes or targets until their offset (which were marked as apexes). What was observed in the data was that at certain points during these gestures the hand slowed down, and the blurriness of the hand lessened and then accelerated again gaining blurriness in the video stream. These discontinuities had no apparent motivation and by no means signalled a phase boundary. Therefore, these were interpreted as indicators of an apex and annotated them as such in the present study.

In ELAN, the apex had its own tier. The apexes of beats and imagistic gestures shared the same tier. This did not constitute a problem because beats do not occur during the strokes of imagistic gestures. An apex annotation had the duration of one frame (17ms) as it was not possible to annotate single points in time in ELAN. The offset time of these annotations were used in the investigation of the present study.

Available apex markings
on the tier "Apex": A (=apex)

3.4.1.4 Handedness

The segmentation and identification of gestural units were annotated for both hands. Instead of distinguishing them as the right hand and the left hand, they were distinguished them as dominant (H1) and non-dominant hand (H2) depending on the involvement of hands. If one hand was more expressive than the other, it was considered as the dominant hand and the other hand as the non-dominant or the assisting hand. The expressive hand carried more semantic content, and the changes in direction and in velocity of the dominant hand were more pronounced. The assignment of dominance occurred at the G-unit level - the dominant hand could be reset across G-units. Each hand had its own G-unit, G-phrase, G-phase and apex tier.



>	Nr	
	1	ve yanındaki hanım
	2	arama bana
	3	anlamına gelecek
	4	bir şekilde
	5	sonrasında hanıme
▶	6	ve erkek de telefonu
	7	sonrasında tekrar t
	8	bu sefer
	9	hanımefendi hiç tel
	10	hamle yapmıyor
	11	direkt erkek telefon
	12	yine benimki
	13	çaldı sanırım
	14	düşüncesiyle
	15	
	16	diğer bir safhada te
	17	direkt kadın olan tel
	18	diğeri de
	19	hamle yapıyor
	20	ama bu kez
	21	kadın olan telefonu
	22	direkt konuşmaya ç
	23	ilk videoda erkek ol
	24	kapatmayı tercih etı
	25	ilk videoda
	26	erkek
	27	gazete okurken kad

Tier	Annotations
H1_GUnit [20]	MGU
H1_GPhrase [52]	Deictic-1 | Iconic-2 | Deictic
H1_GPhase [155]	Prep | Stroke | Ret | Prep | Stroke | P | Prep | Stroke | Ret
H1_Apex [123]	
Beats [44]	Beat
H2_GUnit [5]	MGU
H2_GPhrase [9]	Deictic-3 | Metaphoric-4
H2_GPhase [28]	Prep | Stroke | Post-Hold | Prep | Stroke | Ret
H2_Apex [9]	
G_Meaning [20]	Iconics are duplicates for tapping on shoulder deictic shows the actor from the actor's point of view metaphoric gives shape to mean something

Figure 3.11: A screenshot of ELAN that captures example gesture annotations on their relevant tier

The majority of gestures were produced by both hands in the data, so the tiers for the H2 were often empty. The hands were annotated separately when H1 and H2 performed different gestures for different contents. In these cases, any type of gestural units could overlap on the H1 and H2 tiers. For example, H1 could be half-way through the stroke when the preparation of H2 started or ended, or H2 G-phrase could be contained within the duration of the H1 G-phrase completely. During such overlaps, one hand usually went into a hold while the other hand performed its gesture. Figure 3.11 shows all gesture annotation tiers explained so far.

3.4.1.5 The Annotation of semantic relation

The final layer of annotation of gesture involved marking its semantic relationship with speech. In order to be able to check the synchronisation of gestural units with speech units, corresponding/co-occurring units in different modalities had to be paired (i.e., matched). One way of pairing is based on proximity. That is, the nearest speech unit (e.g., tones) to a gestural unit (e.g., apexes) can be paired, and the synchronisation can be checked only between these units. The alternative way is to mark the semantic relationship between gestural units and speech units, and check the synchronisation between units that are semantically related. Initial investigation of the data showed some overlap between these two approaches in that the semantically related units were often the nearest units to each other. However, this was not always the case, especially for units with short durations (e.g., apexes and tonal events) - depending on the relative length of paired units (and the other units that are around them), this overlap might not be observed. Within the present study, the semantic pairing method was used to decide between which units synchronisation should be checked. In general, theories of gesture production argue for a semantic relationship between speech and gesture starting from the early stages of production (see Section 2.3). For example, in the Growth Point theory (McNeill & Duncan, 2000), if gesture and speech co-occur they must cover the same idea unit. Accordingly, it is reasonable to assume that synchronisation of units in these modalities also reflects this relationship - interlocutors synchronise their gestures with targets within semantically related stretches of speech. A semantic method of pairing would account for such a relationship. Overall, the semantic co-occurrence of gesture and speech was taken into consideration in the present study when deciding between which units synchronisation should be tested.

Only in the investigation of the synchronisation of intermediate phrases with gesture phrases this method could not be used. This is elaborated on in Chapter 5.

The semantic relation between units was annotated at the word level for speech and at the G-phrase level for gesture in ELAN. First, each G-phrase was marked with a number. Then one by one, the semantically most related word on the word tier with the same number was identified and marked with the same number as the G-phrase. The matching numbers of units at these tiers helped in the pairing units at every level in gesture and speech. For example, in order to check the synchronisation of an IS unit with a G-phrase, the IS unit that contained the word with the same number as the G-phrase was selected for pairing/synchronisation. For apex-tone synchronisation, the semantic pairing process was more complex. Therefore, it is explained in detail in Chapter 4.

Establishing the semantic relation of an entire gesture with a word was not always easy. Some gestures encoded the semantic content that was encoded in multiple words by conflating their message within a single stroke (similar to gesture conflations in Section 3.4.1.1). In these cases, which information was more prominent in the stroke was identified in terms of the effort spent to encode it, and the word containing that information was marked as semantically the most related.

Available semantic relation markings
on the tier "Gphrase": "*xxxxx*-346" on the tier "Word": "*xxxxx*-346"

3.4.2 Prosody Annotation

The annotation of gesture was followed by the annotation of prosody. The prosodic structure of Turkish was introduced in Section 2.5.1.2. In this section, the prosodic structure of Turkish is further detailed, and the scheme used to annotate it is described. To the author's knowledge, there is no current complete intonational descriptions for Turkish within the Autosegmental-Metrical (AM) framework. However, there are partial descriptions (Özge & Bozsahin, 2010; Kamali, 2011; Ipek & Jun, 2013; Güneş, 2013, 2015). The annotation scheme used in the study was developed based on these earlier works. The scheme is mostly in line with them with variations that reflect the different nature of the data analysed in the present study (i.e., spontaneous natural speech as opposed to reading pre-set sentences). The marking of the phrases and events in the scheme uses Tones and Break Indices (ToBI) conventions (Beckman & Ayers, 1997).

In Praat, there were 8 tiers related to the annotation of prosody. Here, a brief summary of what these tiers contained is given. The details of the available markings on these tiers are given in Sections 3.4.2.1 and 3.4.2.2. The first 4 of the 8 tiers were transcription and translation tiers.

i. *P_Speech* tier contained the transcription of the participant's speech in Turkish. The transcription intervals on the tier were time-aligned with intonational phrases.

ii. *P_Trans* tier contained the translation of the transcribed speech into English.

iii. *Words* tier contained individual words and their durations. The Montreal Forced Aligner (McAuliffe et al., 2017) was used to time-align the orthographic transcriptions of words with the audio.

iv. *Words_Trans* tier contained the translations of words in the Words tier.

The confederate's speech was also transcribed and translated but it was not annotated for prosody and information structure.

v. The tier *C_Speech* contained the transcription of the confederate's speech.

vi. *C_Trans* contained the English translation of the *C_Speech* tier.

The next 2 tiers contained two layers of ToBI annotation.

vii. The *Tones* tier contained the time-stamps and the types of tonal events following ToBI conventions. There are two types of tones: H for high and L for low. The type of the tonal event was marked with an additional symbol after the tone markings: "*" for pitch accents, "-" for phrase accents, and "%" for boundary tones", e.g., L- for a low phrase accent. The scheme had two additional markings on this tier which were ! and -̂. ! was for the lowering of the peaks of sequential pitch accents within intonational phrases (i.e., downsteps). Unlike other symbols, it preceded the tone marking, e.g., !H*. -̂ was for the exceptional raising of pre-nuclear intermediate phrase accent (H- into H-̂) which cued the onset of the nuclear area (details in Section 3.4.2.2).

viii. The *Breaks* tier indicated the degrees of break or juncture between prosodic units. There were 4 degrees of juncture in the scheme. These junctures were marked with numbers which increased as the degree of juncture increased:

a. Break 0: Free clitic boundary.
b. Break 1: Normal inter-word boundary.
c. Break 3: Intermediate phrase boundary.
d. Break 4: Intonational phrase boundary.

There is also a Break 2 in the ToBI guidelines (Beckman & Ayers, 1997) which indicates intermediate phrase boundaries when there is discrepancy between intonation and pausing, i.e. when perceived pausing indicates phrasing but there is no intonational cue, or when there is an apparent intonational cue for phrasing but there is no pausing or juncture cue. Since the present study was only interested in delimiting intermediate phrases, not in the distinction between break types, Break 2 and Break 3 were grouped together under Break 3.

ix. The *ip* tier showed the extent of intermediate phrases as well as information about whether intermediate phrases are pre-nuclear, nuclear or post-nuclear within the intonational phrases.

x. The *IP* tier showed the extent of intonational phrases and information about whether an intonational phrase is separated from its syntactically parent intonational phrase.

The boundaries in all the tiers were marked at word boundaries as per the ToBI guidelines (see Figure 3.12).

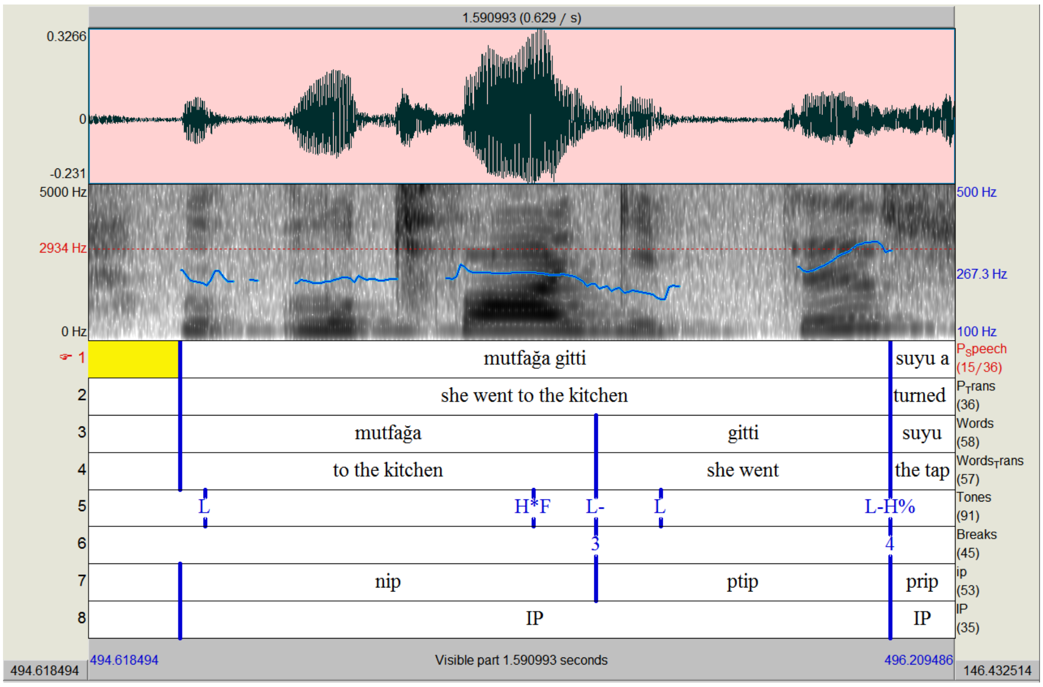


Figure 3.12: A screenshot of Praat that shows the relevant prosody annotation tiers

3.4.2.1 Prominence

Consistent with the earlier studies, prominence is signalled by pitch accents which are associated with the stressed syllables of prosodic words. Available pitch accent markings on the Tones tier were L*, H*, and !H*. L* was marked when the pitch movement on the stressed syllable created a dip or low flat trend, which was often the local minimum. H* was marked when there was a local pitch peak on the stressed syllable. Pitch accents in the nuclear region could be realised with pitch peaks that were lower than the peaks in the pre-nuclear area (i.e., downstepping). However, often no peaks were observed, and a flat pitch plateau was maintained over the nuclear ip (e.g., the nuclear ip marked as “nip” in Figure 3.13). This lowering was marked with !H* at the pitch peak if there was a discernible peak; if not, it was marked at the midpoint of the stressed syllable’s vowel. The syllables with !H* were perceptually salient and carried the local intensity peak. !H* could also be seen in the pre-nuclear area when there was a sequence of H*s and the H* after the first H* was realized with a lower pitch peak (see the

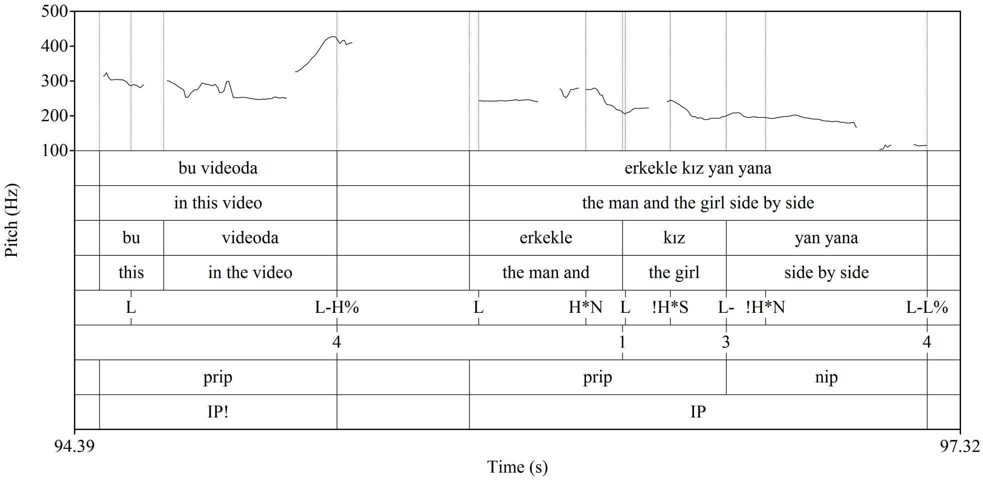


Figure 3.13: A pitch track showing a !H* in the nuclear ip with flat contour. !H* in “the girl” shows the lowering of the pitch peak from the preceding PW in the same pre-nuclear ip.

pre-nuclear area marked as “prip” in Figures 3.13 and 3.14). This lowering of peaks was considered as an event that can take place within and across intermediate phrases (see Ladd, 2008).

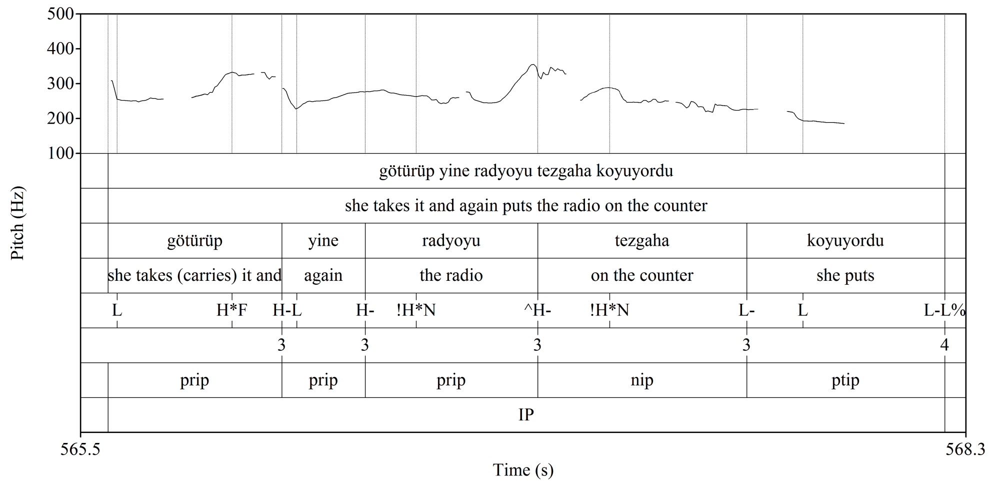


Figure 3.14: A pitch track showing three pre-nuclear ips where the final pre-nuclear ip shows a raised Hˆ-. The second pre-nuclear ip also shows an example for compressed ips. The first pre-nuclear ip has a pitch accent on the final syllable.

In the present study, the location of the pitch accent within the word was also annotated as this location can potentially affect synchronisation (investigated in Section 4.1). If the observed pitch accent was on the word-final syllable (i.e., regular stress), it was marked with an additional “F” after the pitch accent marking itself, e.g., H*F. If the pitch accent was not on the word-final syllable (i.e., irregular stress) it was marked with an “N” in the same manner, e.g., H*N. If the word had only one syllable, and because of this its finality could not be judged, “S” marking was used, e.g., H*S. Regular and irregular stress distinction in Turkish did not affect the type of pitch accent realised on the stressed syllable in the data. The annotation of pitch accents did not depend on the prescribed lexical stress locations of words (Sezer, 1981; Kabak & Vogel, 2001). That is, the pitch peaks were annotated exactly where they appeared on the pitch tracks, allowing shifts from their canonical locations. For instance, a word may have regular word-final stress, but if the intonational cue indicating a pitch accent is on another syllable for any reason, a pitch accent was marked where the intonational cue was. Not every word had a pitch accent - words in the post-nuclear area and words in accentless intermediate phrases (explained in the next section) did not have pitch accents.

Available pitch accent markings
on the Tones tier: “L*N”, “L*F”, “L*S”, “H*N”, “H*F”, “H*S”, “!H*N”, “!H*F”, “!H*S”.

3.4.2.2 Phrasing

As per Kamali (2011); Ipek and Jun (2013); Güneş (2013) and Güneş (2015), three prosodic phrases were adopted in the scheme: Prosodic Word (PW), Intermediate Phrase (ip), and Intonational Phrase (IP). “Utterances” were not annotated as they were not relevant to the research questions of the present study. Figure 3.15 illustrates the hierarchy of these three units. Note that the division of ips into pre-nuclear, nuclear, and post-nuclear ips is common to all previous analyses of Turkish although this division is not fully agreed in the AM framework generally (Ladd, 2008).

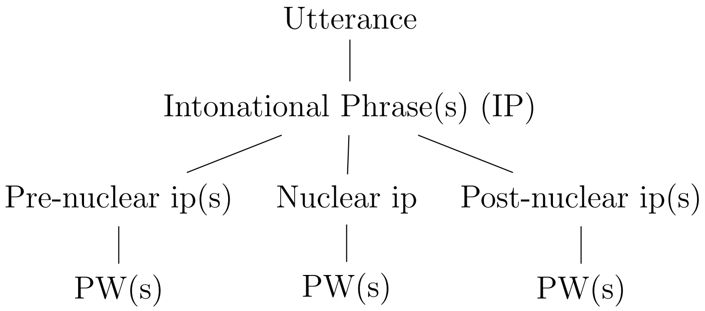


Figure 3.15: The prosodic hierarchy

Intonational Phrase (IP)

IPs are the largest phonological units (that utterances can be broken into) that have their own intonation patterns (i.e., contours). They are often followed by a clear pause or a break. I marked the onsets of these breaks or the completion of the IP contour with "4" on the *Breaks* tier in Praat.

In the data, a single IP could form an utterance on its own and could have one or more ips within. IPs did not always correspond to full sentences. They could contain more than one sentence. They could also be formed by single PWs/ips prosodically separated from their syntactically parent IPs, which happened mostly due to hesitation/planning in the data. Within this scheme, these separated IPs were marked with ! after the default annotation "IP" which marked the extent of the IP on a dedicated *IP* tier (see Figures 3.16 and 3.17). This marking was right-aligned, which means that in case of separation, all IPs got "!" except for the rightmost IP where the verb and the nucleus were.

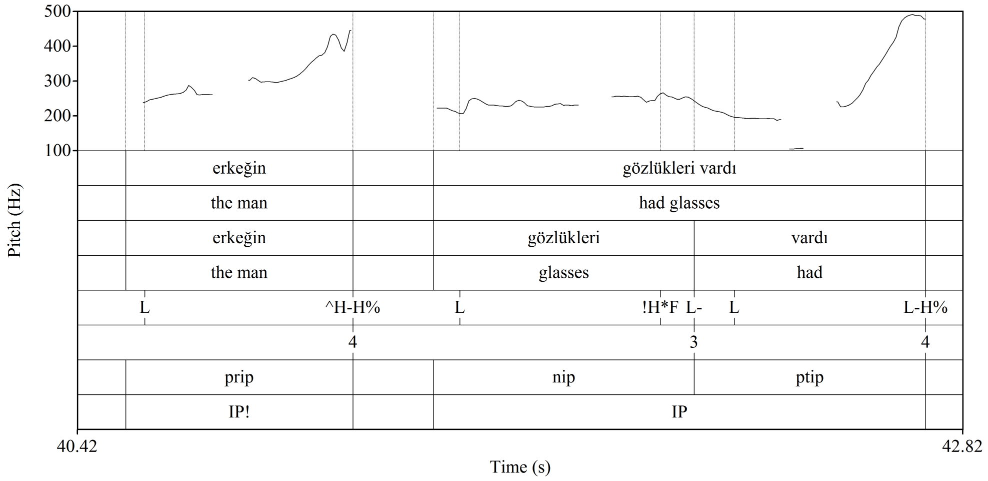


Figure 3.16: The pitch track of a separated IP "the man". The figure also shows the boundary tones and break indices marked.

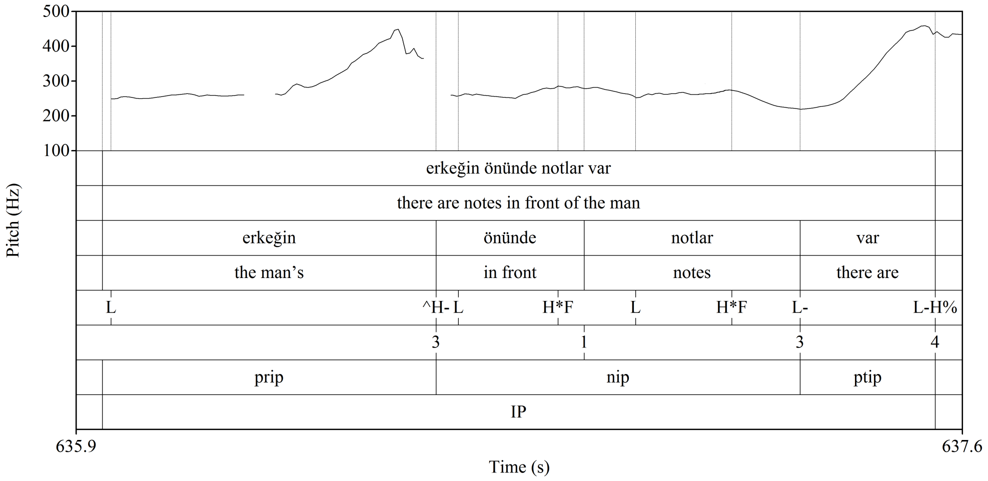


Figure 3.17: A pitch track of an IP with a single pre-nuclear ip with H-̂ which is raised as high as the continuation rise, L-H%.

IPs were marked at the right edge with a boundary tone, either L% or H% (see Figure 3.16) on the *Tones* tier. Since the right edge boundary of the IP coincided with the offset of the final ip within the IP, the marking of the last phrase accent and the boundary tone coincided, leading to combinations of L-/H- with L%/H% (e.g., L-L% or L-H%). All possible combinations of phrase accents and boundary tones were observed in the data (cf. Kamali, 2011 and Ipek & Jun, 2013).

The data consisted of narratives where the participants listed consecutive events. Therefore, there were plenty of IPs that gave a sense of incompleteness implying that "there is more to come". This sense of incompleteness was intonationally expressed with an IP final rise (as opposed to an expected low as in Kamali, 2011 and Ipek & Jun, 2013). These continuation rises were marked with H%. Since most IPs end with nuclear or post-nuclear ips, most continuation rises were realised as L-H% (Figure 3.17). H-H% occurred at the end of separated IPs that ended with pre-nuclear ips. If there was a sense of completeness and finality, a low boundary tone, L% was observed. The standard declarative boundary tone with the final phrase accent was L-L%

(Figures 3.13 and 3.14). In the data, IP-final plateaus were also observed. In these instances, the low boundary tone was raised by a high phrase accent into an intermediate level plateau, H-L% (Figure 3.20). However, there were not enough instances to determine the reason for this variation. There was another possible boundary tone marking, H-̂H%. This marking was essentially the same as H-H%. This is explained in the next section.

Nuclearity was marked on the ip tier. An IP can have multiple pre-nuclear and post-nuclear ips but only one nuclear ip. IPs could exist without a nuclear ip in cases of IP separation where the separated IPs only contained pre-nuclear or post-nuclear ips.

Available intonational phrase markings
on the tier "IP": "IP" and "IP!";
on the tier "Breaks": "4";
on the tier "Tones": "L-L%", "L-H%", "H-H%", "H-̂H%", "H-L%".

Intermediate Phrase (ip)

ips loosely correspond to syntactic constituents. They can have one or more PWs within them. Depending on factors such as the length of the phrase (e.g., genitive noun phrases), speech rate, and hesitation, there may be multiple ips within one syntactic constituent. ips are not followed by a pause or a break but there is a sense of juncture perceived at the boundaries of ips. These junctures or the completion of the ip were marked with "3" on the Breaks tier. ips were marked on their right edges with either H- or L- phrase accents.

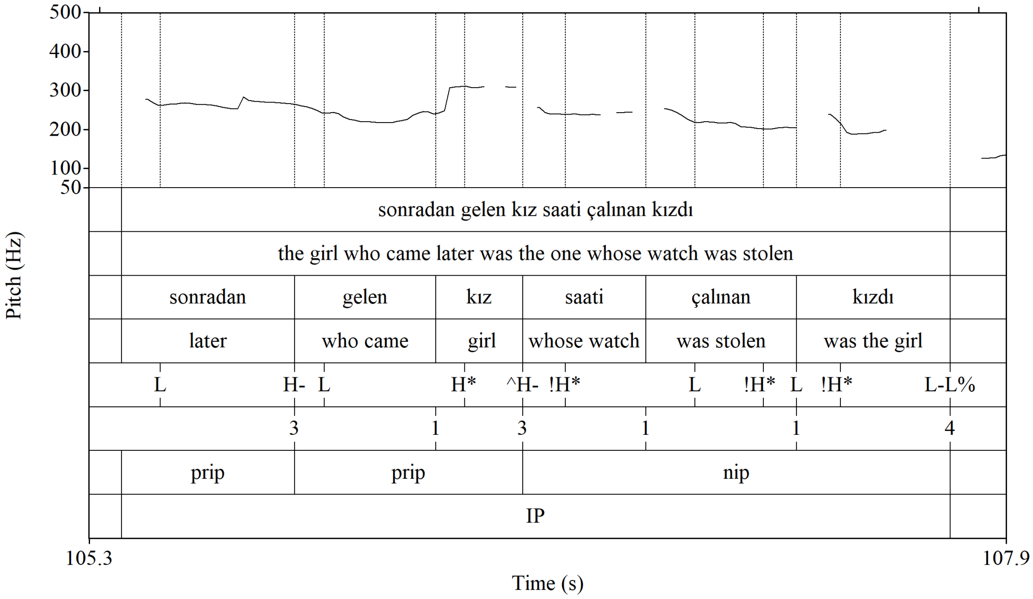


Figure 3.18: A pitch track that shows examples of one accentless *sonradan*, 'later' and one accented *gelen kız*, 'the girl who came' pre-nuclear ip.

Pre-nuclear ips were marked with "prip" on the *ip* tier. They mostly exhibited H* H- pitch contour, consistent with the earlier studies. With the inclusion of Ls (see next section), pre-nuclear ips showed a sequence of rising tones L H* H- (see *gelen kız*, 'the girl who came' in Figure 3.18 and the pre-nuclear area in Figure 3.14). As explained in Section 2.5.1.2, there were cases where a word-final H* and an ip-final H- coincided if the ip-final word had regular stress. In these cases, it was not clear if the H tone was a part of the pitch accent or the phrase accent. There are two views in the literature on this issue. One claims that the H tone is a part of the pitch accent that also functions as a phrase accent (Ipek & Jun, 2013). The other claims that the H is an independent event and is a property of the ip (Kamali, 2011; Güneş, 2015), implying that regularly-stressed words in Turkish are accentless. Within the present scheme, annotations supporting either claim were possible. If there was no perceptually salient pitch accent in the final PW of an ip but the final rise was clear on the pitch track (see *sonradan*, 'later' in Figure 3.18), the ip was marked with only H- on the right edge. If there was a pitch accent, the ip was still marked with H- but H* was also marked at the pitch peak location within the final syllable (see *kız*, 'girl' in Figure 3.18) rather than together with the phrase accent at the ip boundary (cf. Ipek & Jun, 2013). In these cases, the evaluation of perceptual saliency was also based on the intensity and duration of the syllable that a pitch accent would be associated with. Intensity and duration have been claimed to be correlates of stress in Turkish (although not as robust as pitch) (Levi, 2005). If these values on the syllable with the H tone were higher than those of the syllables in the immediate environment, a pitch accent was marked at the pitch peak.

Unlike what was found byKamali (2011); Ipek and Jun (2013); and Güneş (2015), sometimes a pre-nuclear ip could end with L-, especially if the ip was compressed between two other ips and had only one PW (often an adverb) in it. In these cases, there was often a non-final pitch accent and the fall started right after the pitch accent exhibiting a H* L- contour as in Figure 3.19. Alternatively, if there was an increase in the speech rate, pitch accents and PW initial L tones might be deleted leading to a high plateau, e.g., H- H- H- as in Figure 3.20.

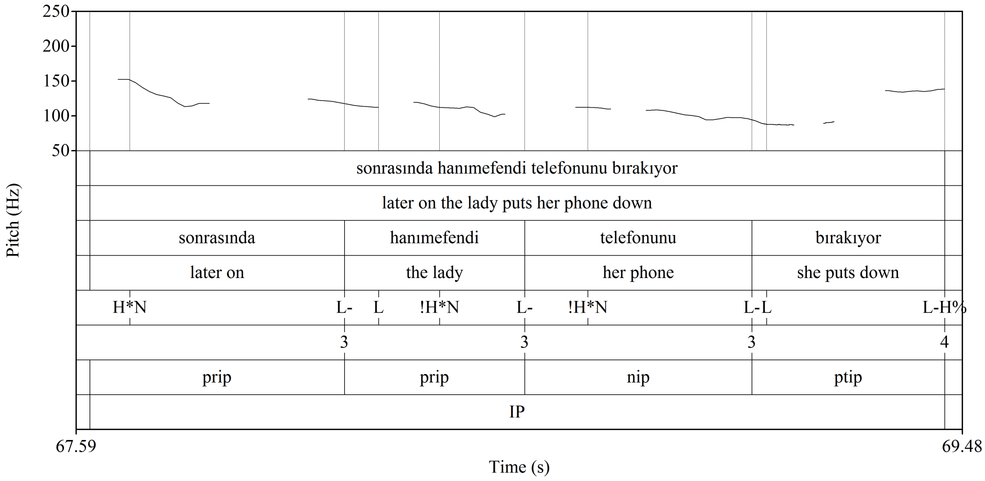


Figure 3.19: A pitch track showing two consecutive pre-nuclear phrases, *sonradan*, 'later on' and *the lady*, 'the lady', ending with L- phrase accents.

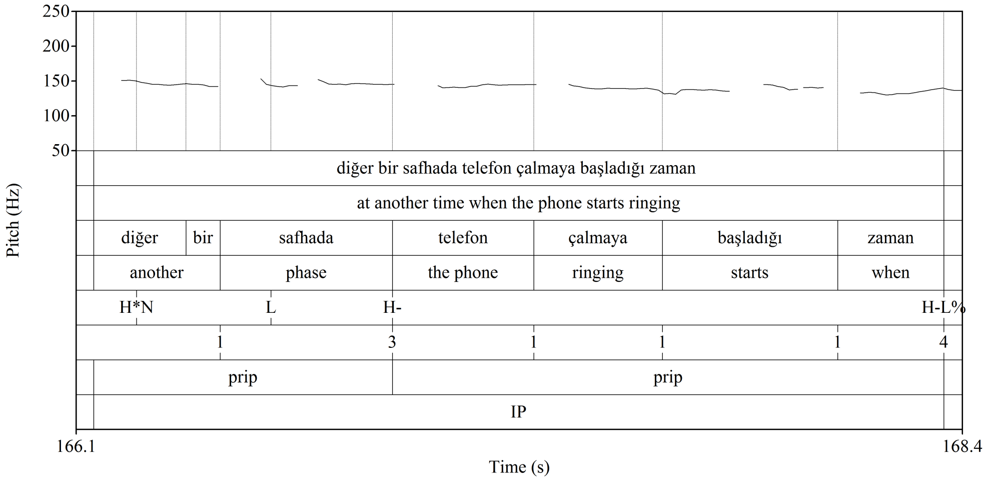


Figure 3.20: A pitch track showing a pre-nuclear high plateau

Another difference from some of the earlier works was that some pre-nuclear H- phrase accents could be raised dramatically compared to other pre-nuclear H- phrase accents (see *radyo*, 'the radio' in Figure 3.14, *kapağının*, 'door's' in Figure 3.21, and *ikisinin*, 'both' in Figure 3.22). These additional rises were not interpreted as H* coinciding with the phrase accent because pitch accenting never caused such a steep pitch rise across all ip types in the data - the pitch range used for accenting was relatively narrow. This raising, however, could reach the same levels in pitch as the continuation rises which got very close to the pitch ceiling values of the participants (see Figure 3.17). These raised H- phrase accents were usually observed immediately before the nuclear area. Therefore, they were interpreted as cues to the upcoming nuclear area (see "LHn" in Ipek & Jun, 2013). These were marked with H-̂. Unlike Ipek and Jun (2013), H-̂ was not consistently observed before every nuclear ip in the data.

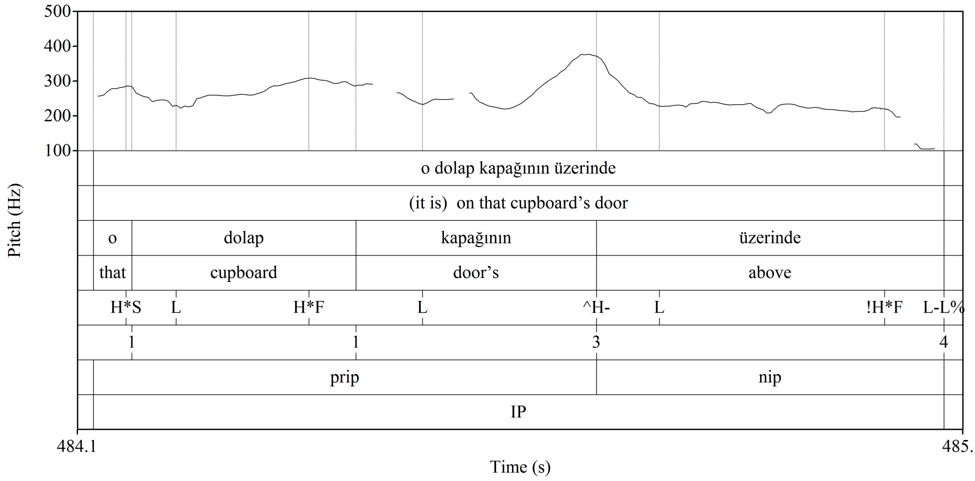


Figure 3.21: A pitch track showing pitch accents exhibit different pitch movement from raised H-̂. Pitch movement on pitch accents is not as sudden or steep.

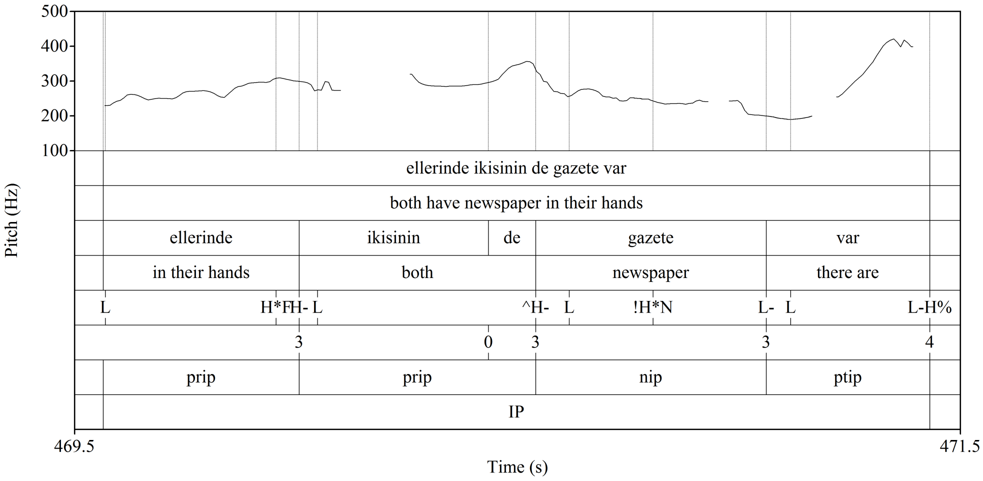


Figure 3.22: A pitch track showing the difference between the rise at the right edge of a pre-nuclear ip before the nuclear ip and another pre-nuclear ip that's away from the nuclear area.

Nuclear ips were marked with "nip" on the *ip* tier. On the *Tones* tier, they had L- phrase accent on their right edge. H*, !H*, and L* accents could all be seen in this area (unlike Kamali, 2011, Ipek & Jun, 2013, and Güneş, 2015, but see Özge & Bozsahin, 2010) but the most common accent was !H*. There was usually very little pitch movement over nuclear ips. With the inclusion of the L, the most common contour for the nuclear ip was L !H* L- as found by Kamali (2011); Ipek and Jun (2013), and Güneş (2015), which exhibited an intermediate level flat pitch plateau (see Figure 3.22). Even if there was an H* instead of a !H*, the pitch peak was at low-to-intermediate level pitch (as in Figure 3.17). The nucleus was canonically the pre-verbal element. The verb and the pre-verbal element could be in the same ip or the verb itself could form an ip on its own. If the nucleus was on the verb instead of the pre-verbal element, then it tended to get an L* (cf. Kamali, 2011; Ipek & Jun, 2013; Güneş, 2015). If not, the verb and any other following units formed post-nuclear ips.

Post-nuclear ips were marked with "ptip" on the *ip* tier, and had an L-phrase accent on their right edges (Kamali, 2011; Ipek & Jun, 2013; Güneş, 2015). In the data, the most common contour for these ips was [L L-] showing a low level plateau (see Figure 3.23). There could be multiple post-nuclear ips within an IP. By definition, there were no pitch accents in the post-nuclear ips. There were cases where the participants' voice quality dropped towards the end of the utterances (where the post-nuclear areas are) due to a creaky voice or mumbling. Visual pitch tracking could be very poor over these areas. In these cases, post-nuclear ips were marked relying more on the sense of juncture at their offsets. This was done in order to be able to mark the whole extent of the IP, which was important because the synchronisation of IPs and G-phrases was investigated in the study.

Available intermediate phrase markings
on the tier "Tones": "L-", "H-", "H-̂"
on the tier "Breaks": "3"
on the the tier "ip": "prip" (=pre-nuclear), "nip" (=nuclear),
"ptip" (=post-nuclear).

Prosodic Word (PW)

PWs had typical inter-word boundaries and were the domains of word stress. Both content words and function words could form PWs on their own. There was not a dedicated PW tier showing the extent of each PW in Praat. In Turkish, PWs are marked on their left edges with an L tone at the lowest point over the first syllable of the PW (Kamali, 2011; Ipek & Jun, 2013; Güneş, 2015), which was annotated on the *Tones* tier.

PW initial L tones were not discernible if there was a word-initial pitch accent within the PW (see Figure 3.19) or when the PW was in a pre-nuclear ip which was compressed due to fast speech rate (see Figure 3.20). In these cases, PW initial L tones were not marked.

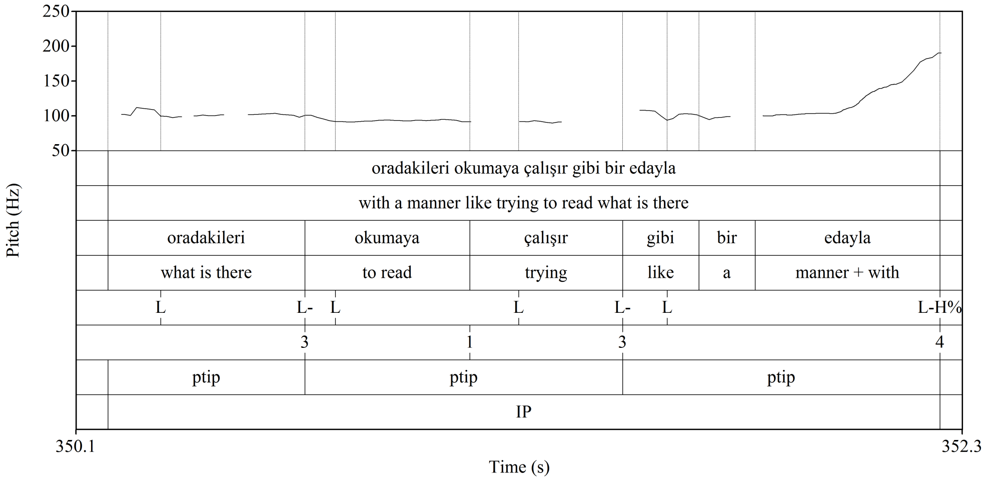


Figure 3.23: A pitch track showing a post-nuclear area where a low plateau is maintained throughout the IP. The PW initial L tones are marked at the offset of the first vowel of the PWs if there is no clear meaningful downtrend.

A theoretical issue about the annotation of these L tones had to do with L- phrase accent. In most cases, the preceding ip final H- made the L tone of the following ip's first PW very clear. However, if a PW followed an ip ending with L-, which usually happened in the post-nuclear area, there was a low flat contour which continued until the end of the IP. In these cases, PW initial L tones were not clearly visible because there was no contrasting pitch movement around them such as an H- or an H*. Moreover, the pitch never went lower within the post-nuclear area to create a terracing pattern in order to mark consecutive PWs (see Figure 3.23). Regardless, PW initial L tones were marked at midpoint of the first vowel within the word to stay consistent with the annotation of PWs in other types of ips.

Available prosodic word markings
on the tier "Tones": "L"
on the tier "Breaks": "1".

3.4.3 Information Structure Annotation

The final layer of annotation was information structure (IS). Information structural notions were introduced in Section 2.6.1. This section describes the annotation scheme that was used to annotate information structural features in Turkish. The scheme follows the definitions in Götze et al. (2007) with amendments to the layers of annotations and the levels within these layers. The annotation was done in Praat (without the prosody annotation being available) where there were three tiers relevant to information structure annotation. The following is a summary of what these tiers contained.

i. The *Infostat* tier contained the annotations of expressions according to their referential givenness in the discourse (i.e., information status).

ii. In the tier *Top/Foc*, the relation of each segment to others was annotated (i.e., topic, focus and background).

iii. In the tier *Contrast*, whether speech parts were contrasted with other parts in the discourse (i.e., contrastive or not) was annotated.

The confederate's speech was not annotated for prosody or information structure but it was used to make information structure related decisions since it also contained discourse related information. Information structure units usually contained multiple words in the data. The boundaries of information structure units were coextensive with ip boundaries. That is, the onset of the first ip within the information structural unit is the onset of the information structure unit, and the offset of the last ip is the offset of the information structural unit.

There were three layers of annotation in the scheme which were information status, topic/focus/background, and contrast.

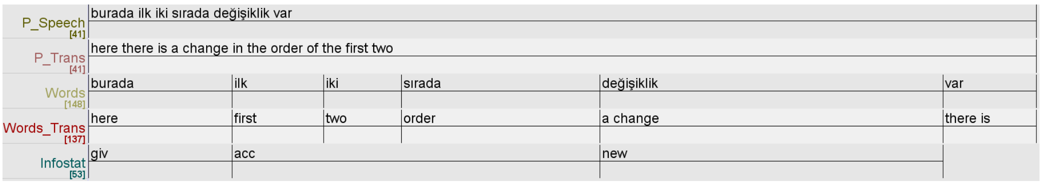
3.4.3.1 Information status

The information status layer marked the discourse referents according to their retrievability from the discourse. A discourse referent is a constituent that refers to an entity in a discourse. The annotation of information status was restricted to only referential noun phrases and post-positional phrases as well as their pronominal/adverbial counterparts as per Götze et al. (2007) (see adverb *burada* 'here' in Figure 3.24).

If a referent was available in the discourse, i.e, if it was mentioned explicitly, then the referent was *given.* Usually, the antecedents of the referents could be found in the last few sentences but sometimes the referential relation could extend across larger speech chunks. If the referent was not mentioned previously but was still accessible through/inferrable from the context, common sense or general world knowledge shared by the hearer, then the referent was *accessible.* If a discourse referent was not mentioned in the discourse previously nor could be accessed through inference, situational context or general knowledge, then it was *new*. The stepwise annotation procedure of Götze et al. (2007) was followed for the annotation of these units. Götze et al. (2007) also describe subcategories for these, but they were not annotated in the present scheme.

Figure 3.24 shows example annotations of information status. In the figure, *burada* 'here' refers to a second video in the pairs of videos that the participant watched and was recently mentioned. Therefore, it was annotated as given and marked with "giv" on the *Infostat* tier. *ilk iki sırada* 'in the order of the first two' was annotated as accessible and marked with "acc". No specific "order" had been mentioned explicitly but the participant had mentioned three events consecutively using adverbs of time such as "then" or "later on". Therefore, it was inferable from the situational context which order the speaker was referring to. In the same figure, *değişiklik* 'a change' was marked as "new" because it had no antecedents and was not a part of the previous discourse.

Although information status was annotated, it was not analysed as a factor that can affect synchronisation. The data collected was not suitable for such an investigation as the distribution of new/accessible/given elements was unbalanced (mostly given) as a result of the random sampling process.



P_Speech [41]	burada ilk iki sırada değişiklik var					
P_Trans [41]	here there is a change in the order of the first two					
Words [148]	burada	ilk	iki	sırada	değişiklik	var
Words_Trans [137]	here	first	two	order	a change	there is
Infostat [53]	giv	acc			new	

Figure 3.24: An ELAN screenshot that shows the Infostat tier and possible annotations on the tier

Available information status markings

on the tier "Infostat": "giv" (=given), "acc" (=accessible), "new"

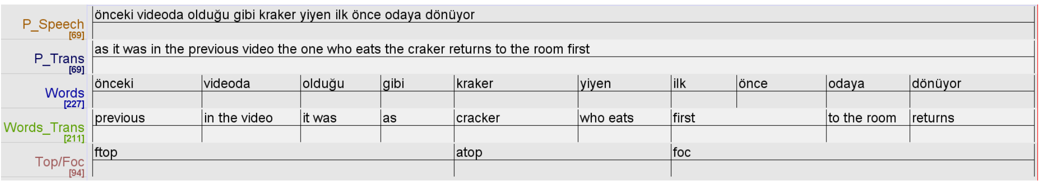
3.4.3.2 Topic/Focus/Background

This layer contains annotations that mark the main predication of a sentence (i.e., focus), the frame in which this predication should be considered or the entity that this predication is about (i.e., topic), and any backgrounded information that covers the rest of the clauses (i.e., background). These categories are mutually exclusive, e.g., an entity or a part of it cannot be a focus and a topic at the same time.

Topic

Following Götze et al. (2007), the topic is the part of a sentence that relates it to a previous discourse. Topics can achieve this in two ways forming two subcategories: aboutness topics and frame-setting topics. An *aboutness topic* (atop) indicates an entity about which a sentence makes a predication. Atops were fronted and typically noun phrases. The atop of a sentence could naturally be fitted in the position of X, if the sentence were transformed into a sentence beginning with expressions such as “let me tell you something about X, ...”, or “Concerning X, ...” (see Götze et al., 2007 for their aboutness topic test). A *frame-setting topic* (ftop) indicated the frame that the predication should be considered in. These frames usually consisted of temporal or spatial locations in which the event or state described in the predication takes place. In the data, ftops were typically post-positional phrases, adverbial phrases, or subordinate clauses that specified time or location. However, not every such phrase/clause was necessarily an ftop. For instance, the adverb *ilk önce*, ‘first’ in Figure 3.25 holds temporal information but was included under the main predication because the temporal information it presents was not a part of the discourse established between the participants.

Figure 3.25 shows example atop and ftop annotations. *önceki videoda olduğu gibi* ‘as it was in the previous video’ was marked as an ftop because it offered a frame existing in the shared background which the main clause



P_Speech [69]	önceki videoda olduğu gibi kraker yiyen ilk önce odaya dönüyor									
P_Trans [69]	as it was in the previous video the one who eats the craker returns to the room first									
Words [227]	önceki	videoda	olduğu	gibi	kraker	yiyen	ilk	önce	odaya	dönüyor
Words_Trans [211]	previous	in the video	it was	as	cracker	who eats	first		to the room	returns
Top/Foc [94]	ftop				atop		foc			

Figure 3.25: An ELAN screenshot that shows the Top/Foc tier and the annotations of topic

should be considered in. *Krakeri yiyen*, 'the one who eats the cracker' was an atop because the predication in the rest of the sentence (marked as "foc") was about the entity described in that relative clause.

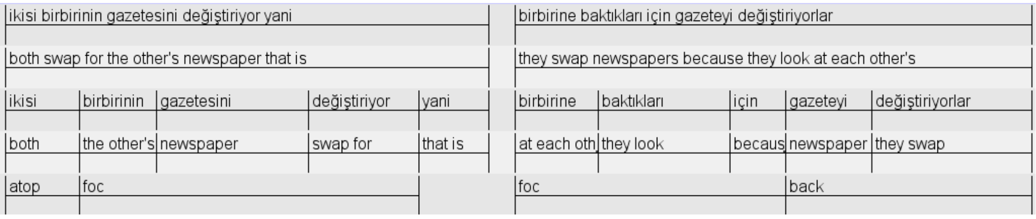
It was possible to have topicless sentences. A sentence describing an all-new event did not have a topic. Turkish has a highly inflected verbal morphology. Subject pronouns, which tend to be topics, can be dropped easily as the subject information can be inferred from grammatical inflection on the verb. Topics were not annotated in these cases. There could be more than one topic in a sentence. Overall, although atops and ftops were distinguished in the scheme, they were grouped together in the analysis. This was because initial investigation showed no difference in the pairing or synchronisation behaviours depending on the distinction between them.

Focus

Focus is the main predication of a sentence (Götze et al., 2007). It is the part that provides the most relevant information in a given context, and the information given in the focus carries the discourse forward. Götze et al. (2007) define multiple subcategories of focus in their extended scheme. These were not annotated in the present study as they do not help answer the research questions. Focus could expand over a constituent or a whole sentence. There was always at least one focus in a sentence. In this scheme, focus areas were marked with "foc" on the Top/Foc tier (see Figure 3.26).

Background

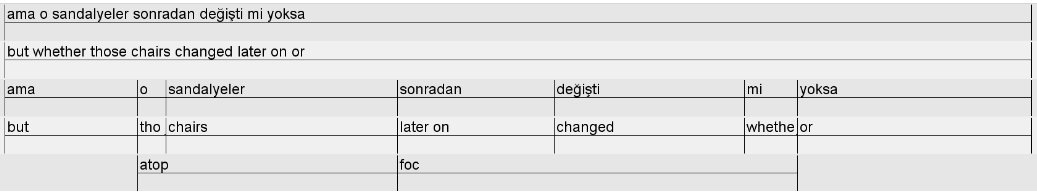
Unlike Götze et al. (2007), the scheme used in the study distinguishes between focus (as the most relevant information) and the rest of the information in a sentence that is not topic. These non-focal and non-topical units are called backgrounds. The information presented in a background can already be assumed from the discourse and does not move the discourse forward. However it "acts as an usher" for focus by updating already shared information as the speaker sees fit (see "ground" and its subcategories "link" and "tail" in Vallduví & Engdahl, 1996 for full description). By definition, both topic and background can be found in the previous discourse. The difference between them has to do with their positioning relative to the focus. Topics could not to appear after their foci in the current scheme. Any element that came after focus was a background (with the exception of contrasted cases explained in Section 3.4.3.3). Backgrounds were usually observed when there was repetition in consecutive sentences in speech (see the background marked with "back" in Figure 3.26).



ikisi birbirinin gazetesini değiştiriyor yani					birbirine baktıkları için gazeteyi değiştiriyorlar				
both swap for the other's newspaper that is					they swap newspapers because they look at each other's				
ikisi	birbirinin	gazetesini	değiştiriyor	yani	birbirine	baktıkları	için	gazeteyi	değiştiriyorlar
both	the other's	newspaper	swap for	that is	at each oth	they look	becaus	newspaper	they swap
atop	foc				foc			back	

Figure 3.26: An ELAN screenshot that shows examples of focus and background annotations

The present scheme allowed discontinuities within topic/focus/background markings. For instance, if a single topic in the discourse was separated into two IPs for any reason, each IP was marked as a topic. In this sense, the marking of information structure followed the prosodic structure. These were not seen as distinct informational structural units. However, they stood separately in time with a pause between them, which presented two separate targets for gestural synchronisation. This had to be accounted for in the annotations. Discontinuities could also be caused by fillers (e.g., *yani*, ‘that is’, *hani*, ‘well’), coordinating conjunctions (e.g., *ve*, ‘and’, *ama*, ‘but’), extra/para-narrative elements (e.g., *böyle kabul edersem* ..., ’if I presume that ...’), and contrast operators (e.g., *sadece*, ‘only’). These were not annotated for information structural features within this scheme (see Figure 3.27).



ama o sandalyeler sonradan değişti mi yoksa

but whether those chairs changed later on or

ama	o	sandalyeler	sonradan	değişti	mi	yoksa
but	tho	chairs	later on	changed	whethe	or
	atop		foc			

Figure 3.27: An ELAN screenshot that shows discontinuities in topic, focus, background annotation

Available topic/focus/background markings
on the tier “Top/Foc”: “atop” (=aboutness topic), “ftop” (=frame-setting topic), “foc” (=focus), “back” (=background)

3.4.3.3 Contrast

The final layer of information structure annotation was contrast. A contrasted element gives rise to a notion of contrast with a previous element in that it distinguishes the element from a set of alternatives made available by context. Contrast does not necessarily operate within focus (cf. Götze et al., 2007) - topics can be contrasted as well, but not backgrounds. Any contrasted element that is not focus is a topic which can be an atop or ftop. Whole focus/topic areas or only certain words within these areas can be contrasted.

Contrast causes information structural constructions of different lengths. For example, when an utterance is produced, the entire stretch of the utterance can be brought into focus (i.e., broad focus, see a. and b. in Example 6), or only one part of it can be selectively focused (i.e., narrow focus, see c. and d. in Example 6).

(6) a. What happened next?

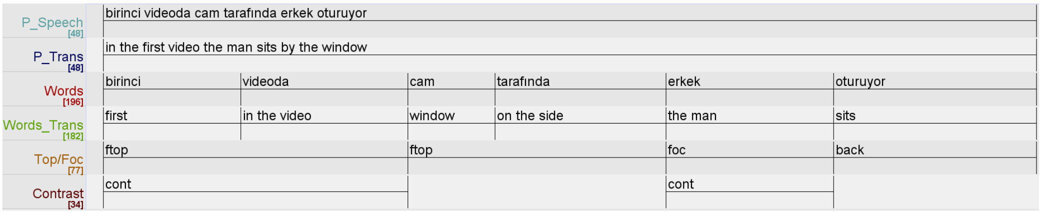
b. She slapped the bouncer.
Broad focus

c. Was it Jessica or Laura who slapped the bouncer?

d. Jessica slapped the bouncer.
Narrow focus

In a typical utterance in narrations, the focus is usually the whole predicate, resulting in broad focus. As seen in d. in the example, contrastive elements result in narrow foci because they bring only one item into focus to exhaust its alternatives. In turn, broad and narrow foci have different durations as a result of the number of elements they consist of. The design of the present study explained in Section 3.2 concentrated on the elicitation of contrast specifically so that the data contained IS units with varying durations because the testing of gesture synchronisation with IS units with different durations increases the reliability of investigation in addition to potentially revealing an effect of contrast on synchronisation.

In the scheme, contrasted elements were marked with “cont”. Figure 3.28 shows examples of contrast annotations. The participant said the sentence in the figure as a response to the confederate’s “the girl sits at the table by the window”. “In the first video” is contrastive because it excluded “the second video” as the only other option. “the girl” is rejected and substituted by “the man” which was available in the set of possible alternatives, so it also marked as contrastive. Subcategories of contrast described in Götze et al. (2007) were not annotated in the data.



P_Speech [48]	birinci videoda cam tarafında erkek oturuyor					
P_Trans [48]	in the first video the man sits by the window					
Words [196]	birinci	videoda	cam	tarafında	erkek	oturuyor
Words_Trans [182]	first	in the video	window	on the side	the man	sits
Top/Foc [77]	ftop		ftop		foc	back
Contrast [34]	cont				cont	

Figure 3.28: An ELAN screenshot that shows the Contrast tier and possible annotations on it

Available contrast markings

on the tier "Contrast": "cont" (=contrastive)

This concludes the annotation section. Table 3.1 shows all available markings for prosody, IS, and gesture.

Table 3.1: All available markings for gesture, prosody, and information structure

Unit	Marking
Tones	L, L*, H*, !H* **(PW)**; L-, H-, Hˆ **(ip)**; L%, H% **(IP)**
Breaks	1 **(PW)**, 3 **(ip)**, 4 **(IP)**
ip	prip **(pre-nuclear)**, nip **(nuclear)**, ptip **(post-nuclear)**
IP	IP, IP! **(separated)**
Information Status	giv **(given)**, new, acc **(accessible)**
Topic, Focus, Background	atop, ftop **(topic)**, foc **(focus)**, back **(background)**
Contrast	cont **(contrasive)**
G-unit	SGU **(single)**, MGU **(multiple)**; SGU!, MGU! **(separated)**
G-phrase	Iconic, Deictic, Metaphoric; sematic relation number
G-phase	Prep **(preparation)**, Pre-Hold, Stroke, Post-Hold, Ret **(Retraction)**
Apex	A **(Apex)**
Beat	Beat

3.4.4 Annotation Reliability

All the annotations were done by the researcher. As stated in the beginning of this chapter, the annotation of gesture, prosody, and information structure were done separately so that the annotations of these do not influence each other. The annotation process of each was iterative. First, an initial annotation process laid out recurring patterns in each layer of annotation. These patterns were often similar to what was described in the earlier studies, but there were also dissimilarities or observations that had not been addressed in these studies. After a discussion period with experienced annotators, how these should be addressed was decided, and the initial annotation schemes were amended accordingly. Then, the existing annotations were revised according to the amended schemes. Throughout the annotation processes, the researcher collaborated with the experienced annotators to constantly improve the schemes so that the resulting annotations represent verbal and non-verbal phenomena in the most comprehensive and accurate manner.

A formal test of annotator reliability was not conducted due to the limited timeframe of the present study and the lack of funding. Commonly used reliability tests for gesture, prosody, and information structure focus on testing inter-annotator reliability as to the categorical consistency of annotations rather than temporal consistency (see Calhoun, 2007 for an example check for prosody and Loehr, 2004 for gesture). First of all, because there was only one annotator (i.e., the researcher), an inter-annotator reliability test was not relevant. Secondly, in terms of categorical consistency, these tests would check, for example, whether or not there was a pitch accent in the phrase, and if so, what kind (e.g., !H* or H*). Similarly for gesture, such tests would check the identification of gesture types or phase types. The standard practices of annotator reliability with such aims would involve the comparison of two sets of blind annotations from different annotators (i.e., the annotators do not see the other's annotations). However, the present study aims to capture the synchronisation of units, and because of this, it is more interested in the actual placement of the annotations in time (e.g., exactly where a phrase accent or a G-phase boundary was marked in the prosodic signal). To the author's knowledge, there are no established methods of testing such timewise agreement of annotators in gesture, prosody, and information structure. Because of all these reasons, a formal test of annotation reliability was not employed in the present study. Instead, the study made use of informal reliability checks during the iterative development of its annotation schemes. In these checks, the experienced annotators did a blind annotation of a small sample of data. Their annotations were then compared to the researcher's and discussed between the annotators. These comparisons and discussions were used to further improve the descriptions of the schemes.

3.5 Statistical Analysis

After the annotation of gesture, prosody and information structure was completed, all of the resulting data was imported to R for analysis. The analysis had three main steps. The first two steps involved the pairing of units and the calculation of time distances between the paired units. These steps were specific to the type of units investigated (e.g., apex and tone, ip and G-phrase), therefore they are detailed in the beginning of the relevant chapters presenting the results of the analyses. The final step of analysis was the statistical analysis. The statistical methods and tests used in the present study were the same for every unit under consideration. Therefore, these methods and tests are described in this section.

As mentioned in Chapter 2, the present study utilises statistical tests to investigate whether various features of gestural, prosodic, and information structural units as well as different participants and scenarios involved had an effect on synchronisation. These features here refer to all kinds of annotations of gesture, prosody, and information structure explained in this chapter so far. These include:

1. Tone type (e.g., pitch accent)
2. ip type (e.g., pre-nuclear)
3. IP type (e.g., separated IP - IP!)
4. Gesture type (e.g., deictic)
5. IS unit type (e.g., topic)
6. Contrast (e.g., contrastive)

Within the present study, all these were considered as factors that could potentially influence the synchronisation of units. That is, the time distance between paired units might show variation depending on each feature listed or their interactions (e.g., pitch accents occurring in a pre-nuclear ip). In order to check whether these factors (relevant ones for the specific pairings) had a significant effect on the time the distances, the present study used linear mixed-effects regression modelling using *lmer()* function in the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R. Mixed-effect models incorporate both fixed effects and random effects, allowing for the measurement of the effects of fixed features as well as accounting for random effects such as individual differences between participants and scenarios.

A linear regression assesses whether a continuous dependent variable (i.e., time distance) can be predicted from a set of independent variables (i.e., features). At the start of the analysis, it was not clear what combination of these features would create the best prediction of the time distance for the synchronisation of units under consideration. To increase the efficiency of this prediction, it was necessary to determine which of these features would account for most of the variance on the time distances. For this purpose, a selection method was employed to find the optimal regression model that makes the best prediction. The goal of this selection was to reduce the set of factors to only those that significantly predicted the time distances. The standard procedure is to construct a full (i.e., maximized) model where all possible factors are in the model together. These factors are then dropped from the model depending on whether dropping them produces a significantly less explanatory model. Two separate selection tests for random effects (i.e., participant and scenario) and fixed effects (i.e., features) selection were used.

For random effects, an elimination method was employed using *ranova()* function in the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). The function uses a full model that includes all the relevant fixed effects and random intercepts/slopes fitted via restricted maximum likelihood (which fixed effects were included in the models is specified in each relevant results chapter). The elimination of terms is limited to the random effects, which takes place in cycles. At every test cycle, one random effect term is deleted if the term does not contribute to the estimation. This contribution can be defined differently - in the present study this contribution was measured through p-values (as opposed to AIC values). In other words, the function drops the least significant effect from the regression model (at $p< .05$) and reruns the new model going through iterations until a significant effect is computed. However, as we will see in the presentation of results in the following three chapters, there was no significant effect of any random effect terms on the time distances between any paired units paired in the study. This showed that the time distances between the paired units showed no significant variation across participants and scenarios. This finding was confirmed by comparing the full model against a linear model fitted with the same fixed effects but without any random effects using an analysis of variance comparison (i.e., ANOVA), and the results again showed no significant effect of the random effect terms.

A similar method was employed for the fixed effects. In order to select an optimal regression model, a backward elimination process was used. All relevant terms were included in the full model. In addition to the single terms, the two-way interaction terms of the features were also included since these interactions themselves could also have a significant effect on the time distances. Only two-way interactions could be fitted because there were not enough data points to test the effect of the interactions of three or more fixed effects with multiple levels. In the same way as the random effect elimination, the fixed effect terms (single and interaction) that do not significantly contribute to the estimation are removed from the model in several test cycles. *fastbw()* function from *rms* package (Harrell Jr, 2019) was used for the elimination of fixed effects. This meant that the regression model needed to be fitted by *ols()* function (Ordinary Least Squares regression) from the same package. ols() fits a standard linear regression model with the same fitting operations used by lm() (R Core Team, 2018), but also stores the variance-covariance matrix which is required for the elimination process (see package documentation in Harrell Jr, 2019 for details). In brief, ordinary least squares (OLS) is a method of performing a linear regression. OLS regression tries to find the linear regression coefficient in a way that minimizes the sum of the squares of differences between estimates and actual values (Moutinho & Hutcheson, 2011). Note that both lm() and ols() are basic models that do not incorporate random effects with fixed effects. ols() was used because all of the random effects being dropped in the elimination processes - having no random effects required the use of basic multiple linear regression models that contain fixed effects only.

After these two elimination processes for fixed and random effects, the remaining significant effects were fitted into a final linear regression that contained the significant terms. This final model is used in the actual synchronisation tests which are introduced in the next section.

3.5.1 Two One-sided T-test (TOST) Equivalence Test

The present study is not directly interested in the output of the regressions. The regression modelling was used to account for the effects of gestural, prosodic, and information structural features on the time distances between units. Rather, the present study introduces a synchronisation criterion, and statistically tests whether or not the time distances predicted by each regression model satisfies this criterion.

A perfect synchronisation of units means that the units occur exactly at the same time and the time distance between them is zero. However, that was never the case in the data. All of the units under consideration occurred near each other, and the time distances between them could be measured in milliseconds. Therefore, a synchronisation criterion which defined how near these units should be in order to be considered synchronised was needed. Naturally, this criterion should be a time interval centered on zero (as the perfect synchronisation point) and should set a tolerance zone around zero only large enough to predict meaningful timing relationships. The statistical testing of the fulfilment of this criterion would be by rejecting observations that are not in this area or in other words, not close enough to zero, hence not synchronised.

A simple and widely used approach in other disciplines (e.g., medicine) for similar purposes is to test for equivalence using a two one-sided t-tests (TOST) procedure (Lakens, 2017). In an equivalence test, the observed effect is statistically compared against two pre-specified equivalence bounds - a lower bound and an upper bound. The area between these bounds defines a tolerance zone (often containing zero) and the observations that fall within this zone are considered statistically equivalent to the centre point between the lower and upper bounds (e.g. zero). The test includes the calculation of confidence intervals (CIs) around an observed effect. If the CI does not cross the equivalence bounds then there is equivalence of the observed effect to zero. As a procedure, a lower (ε_{lower}) and an upper equivalence bound (ε_{upper}) are set (see dashed lines in Figure 3.29). Two null hypotheses are tested using two one-sided t-test where t is the CI of the observed effect (see the horizontal line in Figure 3.29): (1) $t \leq \varepsilon_{lower}$, and (2) $t \geq \varepsilon_{upper}$. If both of these null hypotheses are rejected, meaning $\varepsilon_{lower} < t < \varepsilon_{upper}$ is true (as it is in Figure 3.29), the test concludes that the observed effect falls within the equivalence bound and is close enough to the centrepoint to be statistically equivalent to it.

The TOST procedure was suitable for the analyses of the present study. The equivalence bounds could decide how near the units should be in order to be considered synchronised. In this case, the equivalence bounds could be set at either side of zero at equal distances. The observed effect within this equivalence zone would be statistically equivalent to zero, hence synchronised. The observed effects in this case would be the predictions of time distances extracted from the model.

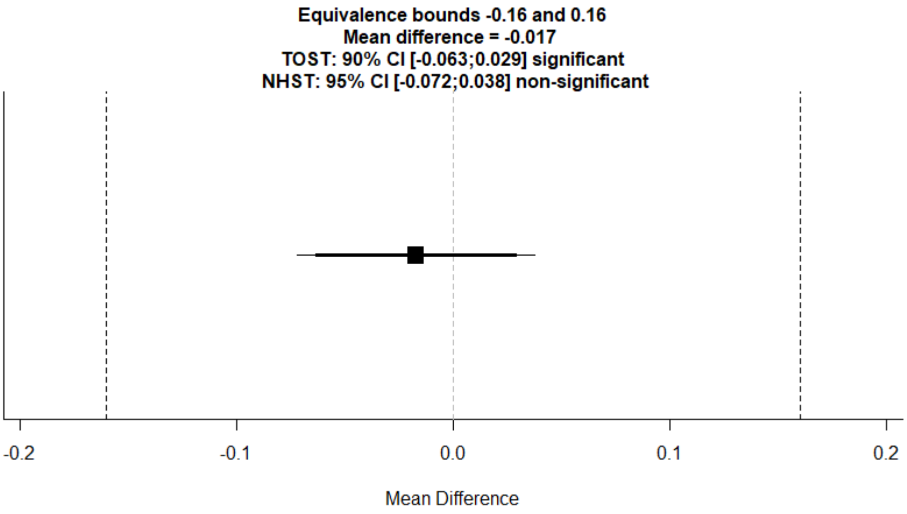


Figure 3.29: An example equivalence test output

The final issue was about setting the equivalence bounds. There is no set number in the literature that shows how near gestural and speech units should be in order to be considered synchronised. Loehr (2004) used a method in which he calculated the average distance between any kind of gestural marking and any kind of prosodic marking in his data, and found 275ms. He considered this average to be the synchronisation criterion. That is, if the distance between two units was less than this average, then they were synchronised. The present study was stricter with this duration, and used a phonologically relevant duration that is meaningful outside of its own annotation schemes. For this purpose, the average syllable duration, 160ms, was used as the synchronisation criterion. Centred on zero as the perfect synchronisation condition, 160ms on either side of zero set the equivalence bounds (i.e., lower bound -160ms and upper bound +160ms). Namely, this meant that if the CI of the time distance estimate extracted from the regression models did not overlap with and occurred within ±160ms equivalence bounds, then the estimate was statistically equivalent to zero and the synchronisation was achieved.

In R, the *TOSTER* package (Lakens, 2017) was used to run TOST procedures. The estimates of time distances were extracted from the regressions using the *effects* package (Fox & Weisberg, 2018). Figure 3.29 shows an example output taken from the TOSTER package documentation. The dashed lines indicate the equivalence bounds at -0.16s and +0.16s. The square is the mean difference (i.e., the observed effect) and the horizontal line is the CI. In the example, the CI of the estimate fits within the bounds, therefore it is statistically equivalent to zero (NHST is the null hypothesis test - it is not detailed here as it is not relevant for the purposes of the present study).

4

Synchronisation with Tonal Events

Chapters 1 and 2 have shown that there is a natural pairing of structural hierarchies of gesture and prosody when they are compared side-by-side as in Figure 2.13. This chapter focuses on the synchronisation of the smallest units, i.e., apexes and tonal events (referred to as tones hereon). Note that as stated in Sections 2.5.2.2 and 2.6.2, the present study also investigates the synchronisation of gesture with prosody at the phrasal level and with information structure. These are presented in Chapters 5 and 6 respectively.

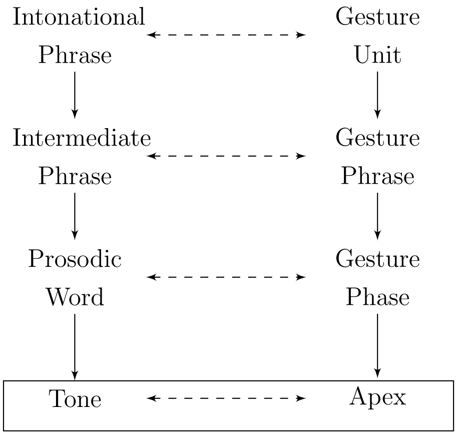


Figure 2.13: Pairing of structural hierarchies (repeated from page 75)

Section 2.5.1 has shown that the synchronisation of apexes has been tested using various prosodic/acoustic anchors. Despite their differences and shortcomings, these studies have observed a consistent synchronisation of apexes with the anchors they considered to be "prominent", e.g., pitch accents. This has led to the claim that prominences in gesture and prosody are synchronised. In this chapter, the present study interrogates these claims by in investigating the synchronisation of apexes with tonal events in Turkish.

Turkish presents a challenge for this claim because not every phrase contains a pitch accent, and for such phrases this synchronisation claim is inapplicable. Furthermore, prosodic phrases are tonally crowded as there are multiple tones associated with prosodic words (PWs) (i.e., PW initial low tones, Ls from hereon, and pitch accents) in addition to other non-PW associated tones, i.e., phrase and boundary tones, all occurring within short durations of each other. Considering all these, it is not clear which tone(s) apexes may be synchronised with under these conditions - it may be the case that apexes may also be synchronised with other events in the prosodic signal.

Previous studies have also overlooked the fact that the prosodic context in which synchronisation takes place may influence synchronisation. Prosodic context here is used to refer to both the tonal environment (i.e., other available tones in the phrase for synchronisation) and the phrasal environment (i.e., which prosodic phrases the tones are in). To account for the effect of the prosodic context, the present study tests the synchronisation of apexes with tonal events - as tones are sensitive to the prosodic structure of utterances (see Section 2.5.1.2). Tones can be prominence marking or boundary marking. Within the present study, gesture apexes can be synchronised with any tone, enabling an investigation of a general claim that apexes are synchronised with all tonal events, rather than more narrowly with pitch accents.

Previous studies have investigated apex synchronisation for only one gesture type, disregarding the possibility that apexes of different types of gesture can be synchronised differently. The present study tests apex synchronisation for all types of gesture to show whether or not synchronisation depends on the semantic function of the gesture containing the apex. Taken together, the research questions addressed in this chapter are the following:

1. Which tonal events are apexes synchronised with in Turkish?
   1a. Is the synchronisation of apexes and tonal events affected by prosodic, gestural, and information structural contexts?

In answering these questions, this chapter follows a step-by-step presentation of results. Within the chapter, a series of questions are asked and then answered in order to get to the answer of the general research question of the chapter. The questions addressed with regards to apex-tone synchronisation in this section are the following:

1. Are there tones near apexes?

2. Which tones do apexes tend to be paired with?

3. Does the pairing pattern of apexes and tone depend on:

   a. whether or not there is a pitch accent in the PW?
   b. the types of available tones in the PW?
   c. the type of intermediate phrase involved?

4. Does forced pairing affect the pairing patterns?

5. Which tones are apexes paired with in accented versus accentless prenuclear intermediate phrases?

6. Are apexes synchronised with their nearest tone?

The presentation of results starts by describing how apexes and tones were paired prior to any analysis (1). A pairing of units means that they are temporally nearest to each other, e.g., an apex and the nearest tone to it in terms of time form a pairing. Also in (1), in order to give an overall view of how close the paired units are to each other the general time distances between these are presented. This step is necessary to show that there are tones within short distance of apexes, and therefore, there is a plausible environment for the testing of synchronisation. In (2), through a series of analyses, the study checks for potential pairing patterns (i.e., are apexes paired up with a specific type of tones), followed by checking whether these pairing patterns (if any) are consistent in different prosodic contexts (3). This is important because it directly relates to the general research question which aims to find whether apexes are synchronised with specific tones over others. By checking if any patterns found persists in various prosodic conditions, the analyses test for the effect of prosodic contexts as indicated in the general research question. The analysis in (4) tests whether the pairing method employed in the study influenced the pairing patterns discovered. These four questions fully describe the pairing patterns in the data. In (5), the observed pairing patterns are used to comment on the debate about the accentlessness of words with final stress in Turkish (Section 2.5.1.2). Finally, (6) focuses on synchronisation of the pairings. It is important to note that a pairing of two units does not necessarily guarantee synchronisation - a pairing indicates proximity only. The members of a pairing are considered synchronised only if they systematically co-occur within the average syllable duration of each other (i.e., according to the synchronisation criterion defined in Section 3.5.1. Statistical tests are carried out to determine whether or not a pairing fulfils this criterion given the possible effects of varying gestural, prosodic, and information structural contexts (as explained in Section 3.5).

4.1 Are There Tones Near Apexes?

Before addressing the questions listed in the previous section, the method of pairing apexes and tones and the calculation of time distances between them must be explained. It was mentioned in Section 2.4 that a pairing consists of two units that are nearest to each other, e.g., an apex and the nearest tone to it form a pairing. Section 3.4.1.5 discussed two approaches to the pairing process which are the proximity approach and the semantic approach. In brief, the distinction between them is about whether the pairings should be between the apexes and tones of semantically related units (i.e., the semantic approach) or between the temporally nearest apexes and tones regardless of any semantic relation (i.e., proximity based approach). The present study adopted the semantic approach since it is stricter as a result of the addition of another constraint on pairing. Most theories of gesture production also argue for a semantic relationship between gesture and speech (Section 2.3) at the early stages of production. Therefore, it is reasonable to have the semantic relation of speech and gesture as a constraint on the synchronisation.

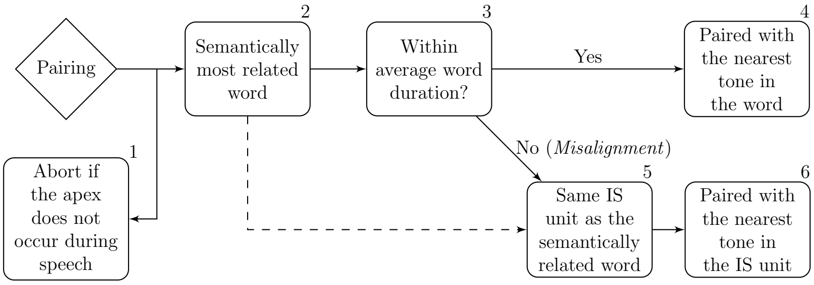

Pairing
2
Semantically most related word
3
Within average word duration?
Yes
4
Paired with the nearest tone in the word
1
Abort if the apex does not occur during speech
No (*Misalignment*)
5
Same IS unit as the semantically related word
6
Paired with the nearest tone in the IS unit

Figure 4.2: Semantic pairing task flow

The process of pairing of apexes and tones using the semantic approach involved several steps which are described in Figure 4.2. The first step was the exclusion of apexes that did not occur during speech. The next step (2) stated that an apex was only allowed to pair with a tone in the semantically most related word to the accompanying gesture (see Section 3.4.1.5).

The following step (3) calculated the time distances between apexes and tones. If the nearest tone in the semantically related word was within the average word duration (340ms), then that tone got paired with the apexes (4). The average word duration was selected because earlier studies have argued that gesture and its lexical affiliate are often within one word distance of each other (Section 2.2). If it was not within 340ms, this was considered as "misalignment". Cases of misalignment usually occurred when the full gestural content was encoded in more than one word and where deciding which of these words is semantically most related was challenging. These cases of misalignments manifested themselves as increases in the time distances between apexes and their nearest tones. To address this issue, the domain of the semantic relation was shifted to a higher level than the word level. That is, in the cases of misalignment, the pairing process first located the IS unit (topic, focus or background) containing the semantically most related word (5), and then paired the apex with the nearest tone within that IS unit (6). The pairing domain was chosen to be the IS unit and not the ip (i.e., the ip containing the semantically related word) because the full gesture meaning was often encoded over multiple ips (a result of the short ip durations in Turkish, see Section 2.5.1.2). Such a shift in the pairing domain assured that the semantic relation of the gesture and speech is preserved (represented by the dashed line in Figure 4.2).

After the pairing process, the time distances between apexes and tones were calculated in order to have an overview of the general time distances between apexes and their nearest tone. The calculation of the time distances was a straightforward process because both apexes and tones were marked as points in time within the annotation schemes of the present study (see Sections 3.4.1.3 and 3.4.2). Consequently, the time distance between these units were calculated using a simple formula where the time-stamp of

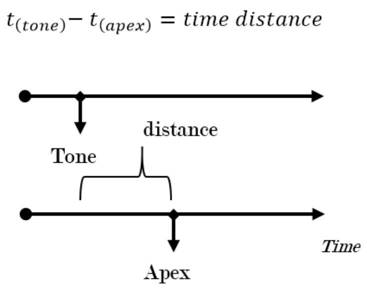


Figure 4.3: Calculation formula and a co-occurrence example with negative time distance between units

the apex was subtracted from that of the tone. Figure 4.3 shows an illustration of an apex and a tone occurring near each other and the formula used to calculate the time distance between them. As per the formula, if the resulting distances in milliseconds (ms) were negative that meant that apexes occurred after their paired tones (Figure 4.3 is an example of this). If the distances were positive that meant that apexes occurred before their tones (this layout is consistent in every relevant figure presented from here on). Figure 4.4 shows scaled distributions of these time distances.

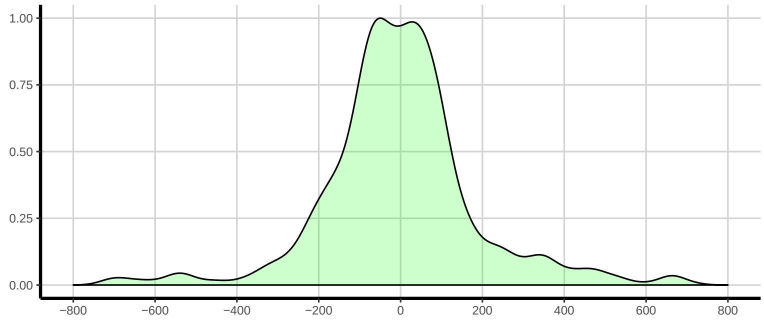


Figure 4.4: Time-normalized histogram of the time distances of each apex from its nearest tone

It can be seen in the figure that apexes and tones paired using a pairing approach that takes the semantic relationship into consideration presented a compact distribution of time distances (N=820, M=-30ms, SD=300ms). Most of the pairing instances were observed within -200ms and 200ms range - there were only a few pairing instances outside of this range. The distribution was also tested for bimodality using Hartigan's dip test of unimodality (Hartigan, Hartigan, et al., 1985), and the results showed no evidence against unimodality (D = 0.009, p = 0.933).

Overall, the aim of this process was to check whether or not there actually are tones around apexes and how far tones are from apexes. The distributions presented here indicated that there were tones available nearby (mainly ±200ms) in the speech stream for apexes to be paired/synchronised with, enabling further analyses. The next step in the investigation looks for pairing patterns by checking whether or not apexes tended to be paired with a specific type of tone.

4.1.1 Which Tones do Apexes Tend to be Paired With?

The findings in the previous section showed that there were tones near apexes and that temporally proximate pairing relations can be established between them. Following on from these findings, this section investigates whether apexes were paired with certain types of tones more than others, indicating a pattern. Sections 2.5.1.2 and Section 3.4.2 defined four different tone types in Turkish: prosodic word-initial low tones (L), pitch accents, phrase tones, and boundary tones, which were all available in the speech stream in the data. In order to show if apexes tended to be paired with one of these, all pairing instances were grouped by tone type. Table 4.1a shows the total number of tones annotated for each type and what percentage of these annotated tones were paired with apexes. Table 4.1 shows the actual numbers of paired tones and what percentage they constituted out of 820 pairing instances.

Table 4.1: The number of tones paired with apexes and what % they constitute out of all pairing instances

	Annotated	% paired
Pitch.acc	1030	36.6%
Phrase.tn	687	15.1%
Boundary	679	8.2%
L	1301	21.7%
Total	3697	22.1%

(a) The total number of tonal annotations in the data and what % of those that paired with apexes

	Paired	%
Pitch.acc	374	45.4%
Phrase.tn	107	13.1%
Boundary	61	7.4%
L	278	33.9%
Total	820	

(b) The number of tones and the frequency at which they paired with apexes

In the data, 22% of tones were paired with apexes within intonational phrases that were gestured. 3% of pitch accents were paired with apexes, and these pairings constituted almost half (45%) of all pairing instances. Most of apexes were paired with pitch accents, closely followed by Ls. 22% of Ls were paired, which constituted 34% of all pairing instances. Based on these numbers, it is possible to argue for a preference in the pairing behaviour which is that apexes tended to be paired with pitch accents and Ls. The claim that apexes are synchronised only with pitch accents (as the only prominence markers) was not fully supported here since it seems that apexes can also be paired with Ls roughly as frequently as with pitch accents. However, it is difficult to reach a clear pattern of pairing just by looking at these distributions. It might be the case that apexes are paired with pitch accents as well as with Ls, or there might be different modes of pairings which alternate depending on the tonal or phrasal environment that tones and apexes existed in.

The next section investigates the possibility that there might be alternate modes of pairing. This is done by looking at the pairing patterns within the tonal context of the PW. As detailed in Section 2.5.1.2, pitch accents, the most preferred tonal events paired with apexes, were not always observed in

PWs in the data. It is possible that the absence of a prominent anchor might be a factor effecting the pairing patterns.

4.1.1.1 Does the pairing pattern of apexes and tones depend on whether or not there is a pitch accent in the PW?

The rule that apexes are paired (or synchronised) with pitch accents depends on the existence of pitch accents. Therefore, whether the pairing pattern in Table 4.1b would be the same in the absence of a pitch accent in the PW must be checked. Table 4.2 breaks down the distribution in Table 4.1 by whether or not there was a pitch accent in the PW that the paired tone was in.

Table 4.2: Is there a pitch accent in the prosodic word that the paired tone is in?

	No pitch accent	Pitch accent
Pitch.acc	NA	71.7%
Phrase.tn	15%	11.4%
Boundary	8.2%	6.1%
L	76.9%	10.8%
n	294	526

The table demonstrates that when there was a pitch accent in the PWs, apexes were overwhelmingly paired with pitch accents. If there was no pitch accent in the PWs, apexes were overwhelmingly paired with Ls. In fact, 80% of the pairing instances with Ls in the data occurred when there was no pitch accent in the PWs. As a result, this revealed two modes in the pairing behaviour of apexes. Apexes were paired with pitch accents when a pitch accent was available in the PWs. However, when there was no pitch accent available, Ls were greatly favoured for pairings. Note that both pitch accents and Ls are associated with the PW, showing that the pairing of the structural hierarchies in Figure 2.13 was preserved. That is, both pitch accents and Ls are associated with PWs, and the pairing preference shifts from one to the other depending on the accentlessness of the PW - apexes were not paired with phrase tones and boundary tones which are associated with phrases at higher levels in the prosodic hierarchy.

This change in the pairing behaviour depending on the tonal content of the PWs establishes the fact that the tonal context can have an effect on apex-tone pairings (and therefore on synchronisation). As a result, a further investigation is needed of the tonal contexts in which these apex-tone pairings take place. The investigation in the next section explains the pairing process in detail by showing whether the pairing of apexes depends on what other tones occur within the duration of the PWs.

4.1.1.2 Does the pairing pattern of apexes and tones depend on the type of available tones in the PW?

An issue that has been overlooked by previous studies on synchronisation is the availability of other tones for pairing within prosodic phrases (see Section 2.5.1). The general approach has been that apexes were pre-emptively associated with prominence markers (e.g., pitch accents), and if they consistently occurred "near" each other, then this counted as synchronisation. To the author's knowledge, no previous study has fully described the actual pairing preference when all potential candidates are considered (e.g., if there is an L and a boundary tone available, which one is preferred?). Such an analysis is useful because it can create and solve either-or scenarios of pairing for every combination of tones that can exist over the duration of a PW. Another use is that it reveals the pairing instances where there is only one tone available, making any pairing preference irrelevant. Looking at the pairing distributions without these forced pairing instances can potentially change the observed pairing patterns. In brief, an investigation that describes which tone is chosen for pairing and in preference to which other available options can reveal multiple alternative pairing patterns. Table 4.3 shows the results of such an investigation within the present study. In the table, for every paired tone type (represented in columns), a list is given of the other tones that were available over the duration of the same PW (represented in rows).

Table 4.3: The types of tones preferred for pairing (columns) over other available tones within a PW (rows)

Other available tones	Chosen tone type			
	Pitch.acc	Phrase.tn	Boundary	L
Only choice	74	3	10	44
2 tones				
Pitch.acc	-	31	18	18
Phrase.tn	71	-	-	88
Boundary	68	-	-	111
L	70	40	18	-
3 tones				
L+Pitch.acc	-	**33**	**14**	-
L+Phrase.tn	**72**	-	-	-
L+Boundary	**45**	-	-	-
Pitch.acc+Phrase.tn	-	-	-	**31**
Pitch.acc+Boundary	-	-	-	**20**

15% of apexes were paired with tones that were the only option in the PW. The numbers in this category were not useful for checking for preference, but they reflect the overall pattern presented in Table 4.1 (i.e., Pitch accent > L > others). The colouring in the “2 tones” grouping in the table indicates the pairing preferences when there were only two tones in a PW. For instance, the red cells highlight the instances when apexes could be paired with either a pitch accent or a L. For these two cells, the column names indicate the selected tone type within the pair, and the row names indicate the alternative that was not selected for pairing. The cells in red shows that when there was only a pitch accent and a L in a PW, apexes tended to be paired with pitch accents with 79% (70 over 18). This confirms the favouring of pitch accents for pairing in an either-or scenario between the two most likely candidates, as demonstrated previously. The preference of pitch accents was also clear in all tonal contexts in which pitch accents were involved: 70% versus phrase tones (71 over 31), and 79% versus boundary tones (68 over 18). In instances where there was no pitch accent involved, apexes were paired with Ls. When apexes could be paired with either a phrase tone or an L, they tended to be paired with Ls with 69% (88 over 40). The pattern was the same for Ls versus boundary tones with 86% (111 over 18). Overall, these findings support the rule that apexes are paired with pitch accents when they are available, and when they are not, the pairing preference shifts to Ls, the other PW associated tonal event.

The grouping “3 tones” in Table 4.3 shows the instances where there were 3 tones (i.e. the maximum number of tonal events) available over the course of a PW for pairing. Yellow highlighted cells show the apex preference when there was a L, a pitch accent and a phrase tone in the PW. The pairing preference did not show any difference in these multitonal environments. Apexes were again paired with pitch accents with 53% (**72**) followed by phrase tones with 23% (**33**), and Ls with 24% (**31**). The instances in which a boundary tone was involved were highlighted in orange. Although the number of instances was lower for these, the pairing preference remained the same: 60% with pitch accents (**45**), 25% with Ls (**20**), and 15% with boundary tones (**14**).

The analyses so far have aimed to determine whether the observed pairing patterns were consistent for different tonal contexts within the PW. They have shown that there were two modes of pairings alternating between pitch accents and Ls depending on whether there was a pitch accent in the PW (pitch accents are preferred in general). This finding highlights the importance of prosodically informed investigation of synchronisation. It shows that comprehensive investigation of prosodic structure in relation to synchronisation can reveal more about the nature of synchronisation at the micro level.

The investigation in this section has shown the pairing preference of apexes in all possible tonal combinations. The investigation has not found any additional pairing patterns under different tonal contexts, but affirmed the initial bi-modal pairing preference between pitch accents and Ls. Starting from the next section, the investigation looks further into the effect of prosodic structure on synchronisation by checking whether the pairing decision is influenced by the intermediate phrases in which the paired tones occurred. The intermediate phrases in Turkish show varying contours depending on their relative position in relation to the nucleus (Section 2.5.1.2). The analyses in the next section test for a possible effect of this higher phrasal environment on the pairing patterns.

4.1.1.3 Does the pairing pattern of apexes and tones depend on the type of intermediate phrase involved?

As explained in Sections 2.5.1.2 and 3.4.2.2, intermediate phrases (ips) can be categorised depending on their position relative to the nucleus, resulting in three types of ips which are pre-nuclear ips (prips), nuclear ips (nips), and post-nuclear ips (ptips). How tonal events are realised within ips depends on the type of ips. For instance, nips and ptips are usually marked with low phrase tones at their right edge, whereas prips are marked with high phrase tones. In terms of pitch accenting, ptips do not have pitch accents by default, and the realisation of pitch accents varies between prips and nips (i.e., downstepping, see Section 3.4.2.1). Overall, it is clear that tonal events are sensitive to their position in the prosodic structure. The different realisations of tonal events may influence apex-tone pairing patterns, potentially resulting in different pairing patterns depending on the type of the ip that the pairings take place in. This section considers whether the pairing patterns observed in Table 4.2 change depending on the ips.

To begin with, Table 4.4 shows the total number of annotated ip types in the data for comparison purposes. 58% of prips (N=374) contained tones that were paired with apexes. This percentage was 73% (N=363) for nips and 37% (N=83) for ptips. It can be said that most of the pairing instances happened within prips and nips, and not within ptips, which can be attributed to the lack of prosodic prominence in them. Another point to note is that apexes tended to be paired with the tones in nips more than prips despite the difference in their numbers. The more common pairings of apexes with nuclear tonal events also meant that gesture strokes tended to occur near nips. This was the first indication in the analyses of the present study that gesture might be sensitive to prosodic prominence not only at the micro level (i.e., apex-tone) but also at the phrasal level - maximally prominent prosodic phrases attract G-phrase strokes, triggering a coupling of smaller prominent units at the same time.

Table 4.4: The number of pre-nuclear, nuclear, and post-nuclear intermediate phrases. There are 1363 intermediate phrases in the corpus overall.

Pre-nuclear	Nuclear	Post-Nuclear
648	491	224

Next, in order to be able to reveal whether ip types have an effect on the apex-tone pairings, the pairings in Table 4.2 were regrouped according to the ip types. Table 4.5 shows whether there was a pitch accent in the PW of the paired tone in columns and the type of the ip that the PW was in.

Table 4.5: The categorisation of the paired tones according to whether or not there is a pitch accent in the prosodic word, and the type of the ip that these prosodic words are in

	PW accented?	
	No	Yes
Pitch.acc	-	68.6%
Phrase.tn	22.5%	15.2%
Boundary	4.6%	6.3%
L	72.8%	9.9%
n	151	223

(a) Pre-nuclear

PW accented?	
No	Yes
-	73.9%
11.7%	8.6%
8.3%	5.9%
80%	11.6%
60	303

(b) Nuclear

PW accented?
No
-
3.6%
14.5%
81.9%
83

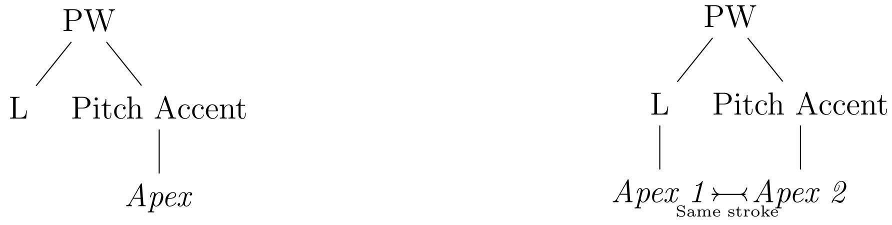
(c) Post-nuclear

For prips in Table 4.5a, in the condition where there was no pitch accent in the PW, the pairing preference was Ls with 72%; and in the pitch accent condition, the pairing preference was pitch accents with 68%. For nips in Table 4.5b, the preferences and percentages were similar to prips (nips can contain more than one PW and some of these may not contain pitch accents). There cannot be a pitch accent condition for ptips (see Table 4.5c), but the pairing preference in absence of pitch accents was with Ls. Overall, ip type resulted in no change to the rule that apexes tend to be paired with pitch accents if pitch accents are present, and with Ls if they are not. However, it was observed that more apex pairings took place within nips, which hinted at the possibility that G-phrases may be attracted to ips that carry the maximal prosodic prominence. This possibility is investigated in detail in Chapter 5.

This section investigated the pairing patterns of apexes and tones in the context of ips and found no variation depending on the position of the ip relative to the nucleus. One potential issue in looking at the pairing patterns of atomic units within larger phrases is what the present study calls “forced pairing”. Within the study, apexes and tones were paired on a one-to-one basis - a tone could be paired with only one apex. This meant that if there was more than one apex within an ip, the pairing of one of the apexes would have an effect on the pairing of the others since it made one of the tones unavailable for pairing. This could potentially skew the results of the analyses of the present study, and is therefore addressed in the next section.

4.1.1.4 Does forced pairing affect the pairing patterns?

One issue with demonstrations like Table 4.5 might be that they ignore the possibility that apexes might share tones for pairing when there is more than one within a phrase. In the present study, strokes can have multiple apexes (multi-segmented apexes in Section 3.4.1.3), and these apexes might occur during the same PW or ip. Since an apex could only be paired with a single tone, a second apex happening over the same ip/PW would have to be paired with another tone that had not already been paired with an apex (compare (a) and (b) in Figure 4.5 for an example).



(a) A single apex

(b) Multiple apexes

Figure 4.5: A single apex (a) and multiple apexes (b) being paired with a tone within a single PW

In the distributions in Table 4.5, instances such as Figure 4.5b register as independent pairing instances disregarding the pairing pressure caused by neighbouring apexes. *Apex 1* is forced to be paired with the L because there is already another apex paired with the pitch accent.

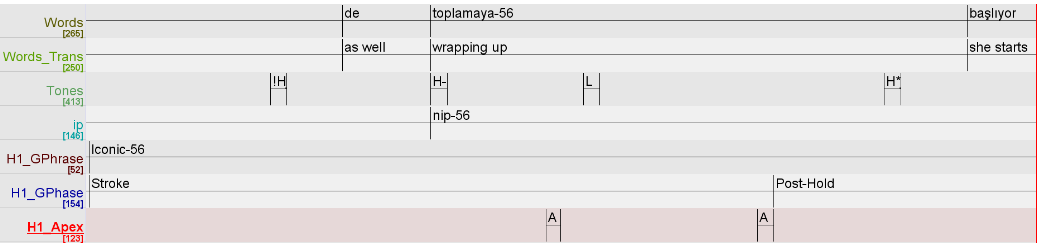


Figure 4.6: An ELAN screenshot showing a forced pairing example. The H1_Apex tier shows the apex positions and the Tones tier shows the tone types.

Figure 4.6 shows an example of forced pairing extracted from the data. In the example, an iconic gesture accompanies the utterance "... she starts wrapping up as well". The gesture encodes the same meaning as *toplamaya* 'wrapping up' as the numbering (56) indicates. Both apexes take place during the same PW and are paired with the L and the pitch accent respectively. The forced pairing here matches the pitch accent and the stroke final apex, but it also matches the L with the non-final apex as it is the only other tone available in the PW. Note that the non-final apex is already very close to the L, and there is no problem with the pairing mechanism itself. However, the preferentiality of pairing is lost because of the elimination of options by the other apex in the same manner as the "only choice" pairings in Table 4.3. The unavailability of choice is what makes the pairing "forced" in this sense.

These forced pairing instances could potentially skew the pairing patterns observed by registering pairings that did not actually represent any preference. In order to test if there was such an effect, the cases where apexes shared tones within prosodic units as in Figures 4.5b and 4.6 were excluded, and then the resulting distribution was compared to the original distribution in Table 4.5. The domain of forced pairing was selected as the ip (as opposed to the PW) as in order to be stricter in the analysis (since it is the larger unit). That is, cases where there were multiple apexes paired with the tones within the same ip were excluded. As a result, 134 instances were removed from the distribution in Table 4.5. The distribution of pairing instances without the forced pairing instances is shown in Table 4.6.

Table 4.6: The same distribution as Table 4.5 but excluding the forced pairing cases

	PW accented?	
	No	Yes
Pitch.acc	-	72.3%
Phrase.tn	18.5%	12.2%
Boundary	4.8%	5.3%
L	76.6%	10.1%
n	124	188

(a) Pre-nuclear

PW accented?	
No	Yes
-	72.3%
8.6%	9.4%
11.4%	6%
80%	12.4%
35	267

(b) Nuclear

PW accented?
No
-
4.2%
11.1%
84.7%
72

(c) Post-nuclear

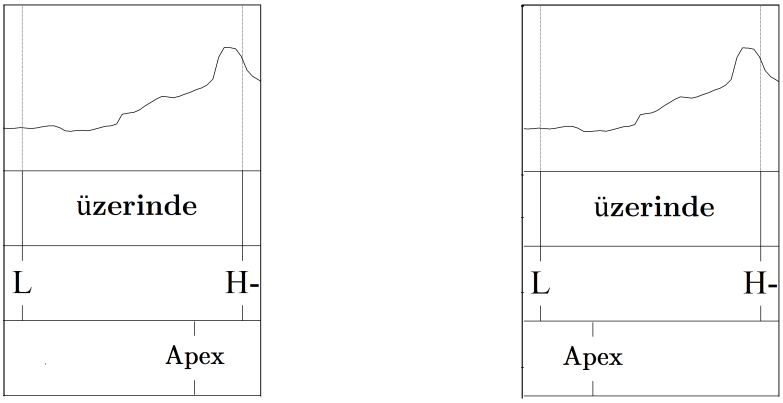
When compared to Table 4.5, the distribution in Table 4.6 did not show any major difference - pitch accents were preferred when they were available; otherwise Ls were preferred for pairing. As a result, it was concluded that the sharing of tones by multiple apexes occurring within the same ip did not cause a major effect on the pairing patterns. The present study disregarded forced pairing as a factor that may affect the pairing of apexes and tones in the subsequent investigation.

The investigation so far has found no effect of the ip types on pairings. In relation to the ip types, Sections 2.5.1.2 and 3.4.2.1 also made an important distinction between accented and accentless prips in Turkish. These accented and accentless prips have a similar rising contour, but the accentless ones lack prominent pitch accents. In line with the findings so far, this situation predicts different pairing behaviours for apexes within these prips. The next section investigates the pairing of apexes in accented and accentless prips.

4.1.1.5 Which tones do apexes pair with in accented versus accentless pre-nuclear intermediate phrases?

The absence of pitch accents as a factor affecting pairing has so far been investigated within PWs. However, larger prosodic phrases such as prips or even separated IPs containing only prips (see Section 3.4.2.2) can also have no pitch accents. As set out in Sections 2.5.1.2 and 3.4.2.1, previous studies on Turkish phonology argued for and against the claim that accentlessness is a feature of PWs with word-final stress. In the case of prips which are marked with a high tone at the right edge, some argued that this final rise is not a pitch accent and only marks the end of the prip (Kamali, 2011); whereas others claimed that this rise has a double function marking both a pitch accent and the end of the prip (Ipek & Jun, 2013).

It is argued in the present study that the pairing patterns of apexes can bring some multimodal insight into this discussion. In Section 2.5.1.2, the present study has hypothesised two possible pairing/synchronisation scenarios regarding accentless prips as illustrated in Figure 2.12.



(a) H- phrase accent

(b) PW initial L

Figure 2.12: Two hypothetical gesture apex synchronisation cases (repeated from page 74)

The apex pairing rule has shown consistency throughout the analyses so far - apexes are paired with pitch accents, and if there is no pitch accent available, then they are paired with Ls. Following this rule, regarding the pairing preference within the prips, one hypothesis was that if the final rise has a double function then the apexes should be paired with the phrase tones at the end of the pitch rise because of double functioning as a pitch accent (Figure 4.7a). The alternative hypothesis was that there is no double function of the final rise, and therefore the phrase lacks prominence, in which case the apexes should be paired with Ls (Figure 4.7b), as dictated by the pairing pattern.

Before the pairing pattern in accentless prips, the pairing pattern in accented prips is presented for comparison. There were 405 accented prips in the data (out of 648) and 69% (N=278) of these contained tones paired with apexes. Table 4.8 shows the pairing pattern of apexes in the accented prips. Boundary tones were grouped together with phrase tones since boundary tone markings also contain a phrase tone marking according to the prosody annotation scheme in Section 3.4.2. Note that prips ending with a boundary tone are the result of separated IPs ("IP!").

Table 4.8: The types of paired tones in pre-nuclear ips with at least one pitch accent

	N	%
Pitch.acc	153	55%
L	68	24.5%
L-X%	30	10.8%
H-X%	27	9.7%
Total	278	

As seen in the table, the pairing preference was consistent in the accented prips - apexes were paired with pitch accents when present and with Ls if pitch accents were not present. The rate of pairing with pitch accents was lower than the tables presented previously in this chapter. This was because the accentlessness condition was checked within a larger domain that is the ip (as opposed the PW) where there can be multiple PWs with or without a pitch accent.

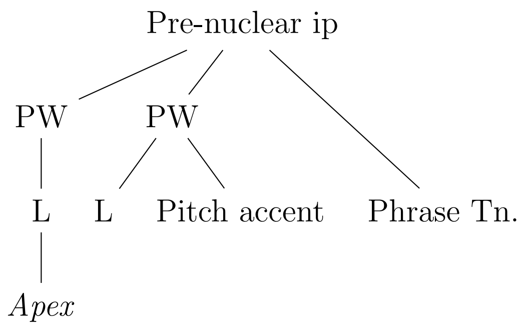


Figure 4.8: An illustration of apex pairing with L in an accented pre-nuclear ip

There can be multiple PWs in an ip and not all PWs have to contain a pitch accent. Using the ip as the pairing domain meant that if an apex was paired with a tone in a PW without a pitch accent, and if another PW in the same ip had a pitch accent (see Figure 4.8), the pairing would be registered to have happened within an accented ip. This was likely to affect the distribution in the table. In fact, the increased number of pairings with Ls was a direct effect of such an approach. 68% of the pairings with Ls (N=46) in Table 4.8 occurred when there was no pitch accent in the PW but there was one somewhere in the same prip (e.g. Figure 4.8). Nevertheless, despite this handicap, apexes were paired with pitch accents more than Ls in accented prips.

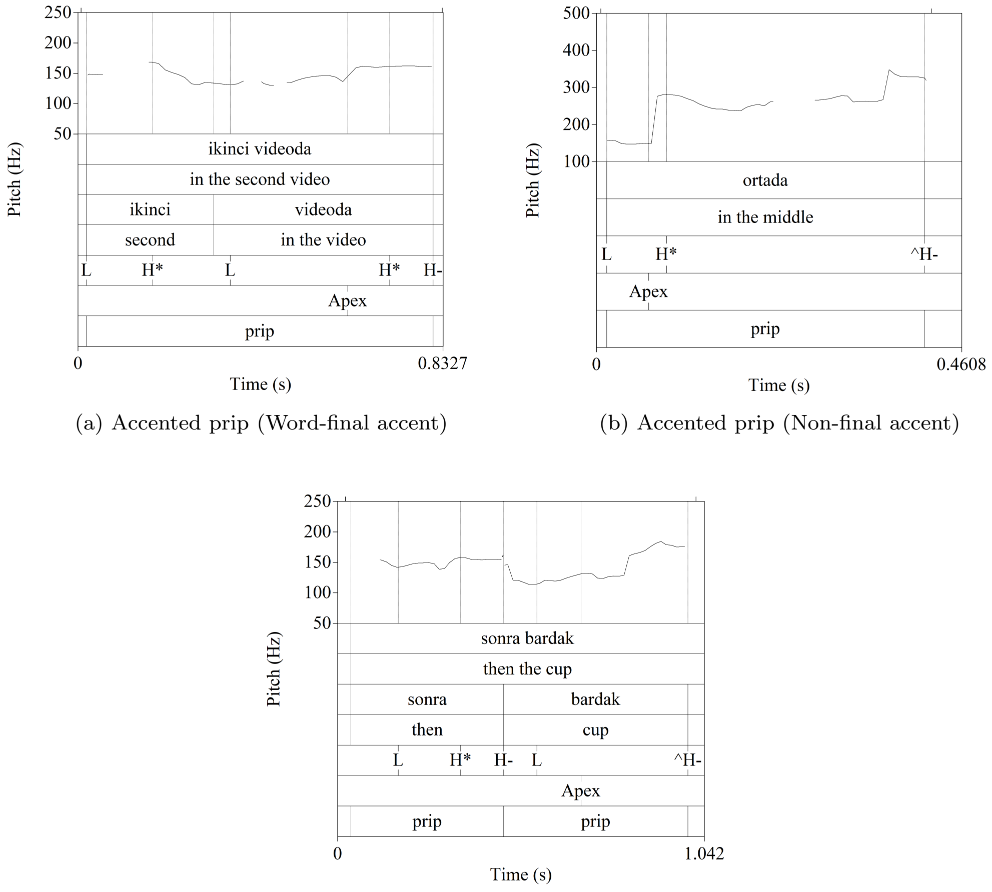
At this stage, whether the pairing pattern in accented prips was affected by the location of the pitch accents was also tested. It was explained in Section 2.5.1.2 that words can be stressed on their final syllable or on a non-final syllable and that pitch accents are associated with these stressed syllables. The position of pitch accents within PWs could potentially have an effect on the distributions because in the non-final condition, pitch accents would be further away in time from the phrase tones at the edges of prips

Table 4.9: The effect of the location of pitch accent on the pairings in accented prips. Only instances where there were one of each tone type are listed (see Figure 4.8).

	Pitch.acc	H-X%	L-X%	L	Total
Final	57.4%	11.9%	11.9%	18.8%	101
Non-final	64.9%	7.8%	23.4%	3.9%	77

compared to the prips with word-final accents (compare the distance of the apex to the final rise in Figures 4.9a and 4.9b). This increased distance in time amplifies the preferentiality of tones for pairing. That is, if apexes are paired with pitch accents even when they were further separated from the ip-final phrase tones, this would support that they were not attracted to the ip final rises per se but to the pitch accents.

Table 4.9 breaks down Table 4.8 by whether the pitch accents in prips were word-final or non-final (pairing instances as illustrated in Figure 4.8 were excluded). No major effect of pitch accent location on the pairing preference could be observed in the table - the preference was pitch accents in both conditions. As shown in the pairing examples in Figures 4.9a and 4.9b, when pitch accents moved away from the phrase final rises (non-final condition), the apex locations tended to move away from the prip final rises along with pitch accents. This finding reinforces the claim that pitch accents and apexes are tightly coupled in that apexes coordination is responsive to the shifts of stress locations within words. Similarly, the findings also show that apexes are not necessarily attracted to pitch peaks but to prominence unlike some of the studies reviewed in Section 2.5.1. Phrase and boundary tones in prips often presented higher pitch values than pitch accents, yet apexes were paired with pitch accents instead of phrase/boundary tones.



(a) Accented prip (Word-final accent)

(b) Accented prip (Non-final accent)

(c) Accentless prip

Figure 4.9: The apexes pairing with the tones in accented and accentless prips

The dispreference for such phrase final rises was even more apparent in accentless prips. In these prips, the phrase final rises still persist in the absence of pitch accents, marking the phrase boundary. Therefore, apexes occurring during these prips could be paired with either Ls or H- phrase tones (see Figure 4.9c). If the final rises have a double function in which they mark a pitch accent as well as a phrase boundary, apexes should be paired with phrase tones. On the other hand, if there is no double function, then apexes should be paired with Ls, dispreferring pairing with phrase tones as they belong to a higher-level prosodic unit (see Figure 2.13).

Table 4.10: The types of paired tones in pre-nuclear ips with no pitch accent

	N	%
L	64	66.7%
H-X%	25	26%
L-X%	7	7.3%
Total	96	

There were 244 accentless prips in the data (out of 648) and only 39% (N=96) of these contained tones that were paired with apexes. This revealed the first difference between accented and accentless prips, which was that accentless prips were gestured less frequently than the accented ones (nearly half the time) - phrases bearing prosodic prominence attracted gestures more.

Table 4.10 shows the pairing preference of apexes in accentless prips. Apexes tended to be paired with Ls rather than with pitch rises (H-X%). Figure 4.9c shows an example of this preference. In such cases, the apex locations consistently shifted away from the phrase final rises (where pitch accents would have been) towards Ls. As can be understood, the pairing preference observed in these cases was deliberate in that apexes did not stay at the default pitch accent location when no pitch accent was there - otherwise this would cause more pairings with pitch rises (H-X%).

As was found for accented prips, using the ip as the pairing domain can produce biased pairing patterns in the distributions. That is, there may be multiple PWs within the accentless prips as in Figure 4.10, creating additional Ls as potential targets for pairing. In this illustration, the apex can be paired with the L of the first PW without having to show any preference between the L and the H- of the prip in the second PW. However, there were only 12 instances in the data, and the exclusion of these instances did not affect the distribution (63% for PW initial Ls over 27% H-X%).

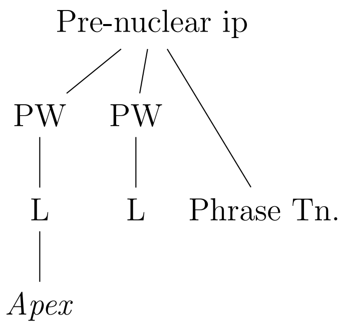
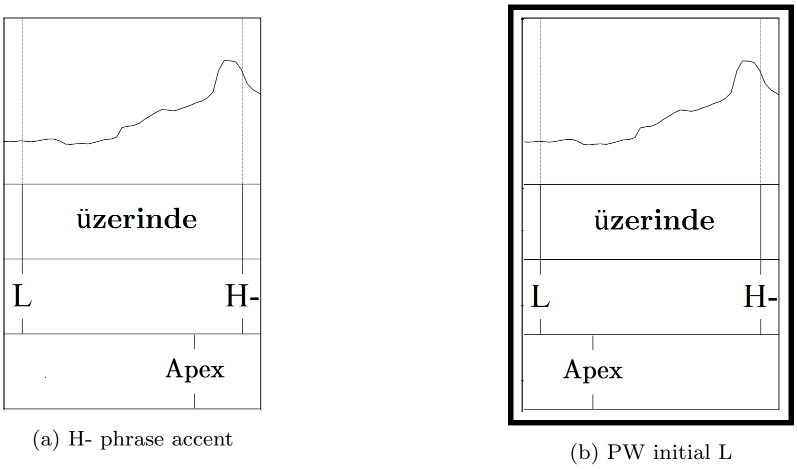


Figure 4.10: Pairing with L in a pre-nuclear ip with a pitch accent



(a) H- phrase accent

(b) PW initial L

Figure 4.11: Two hypothetical gesture apex synchronisation cases (modified from Figure 2.12).

Overall, the results have shown that given the options illustrated in Figure 4.11, apexes tended to be synchronised with Ls (Figure 4.11b). This finding indicates that there were no pitch accents at the location of the phrase final rises, which would result in more pairings with H-. Instead, the pairing preference shifted to Ls, as has been shown to happen in the absence of prominence. The present study interprets this finding as a support for the accentlessness claim in Kamali (2011) since the claim that proposes a double-function for the phrase final rises (Ipek & Jun, 2013) cannot predict the apex synchronisation in these cases.

The results of the investigation of pairing phenomena in different prosodic contexts have been consistent. The possible effects of gestural context (i.e., gesture type) and information structural context (e.g., whether the pairings took place in foci or topics) were also investigated in the present study. However, no different pattern was observed depending on these variables. Therefore, those are not reported here. Overall, there is a bimodal pairing of apexes with tones in which apexes are paired with pitch accents if there are pitch accents in the phrase, if not, they are paired with Ls. Current claims about synchronisation (Section 2.5.1) cannot account for the pairings with Ls because they are not prominent events nor they are acoustic peaks. In line with this, the present study considers pitch accents and Ls as the preferred targets of apex pairings, and phrase/boundary tones as dispreferred targets. This preference that apexes tend to be paired with tones within PWs and not within other prosodic phrases implies that prosodic structure is also a factor in the anchoring of apexes in addition to prominence. This is further detailed in the next section.

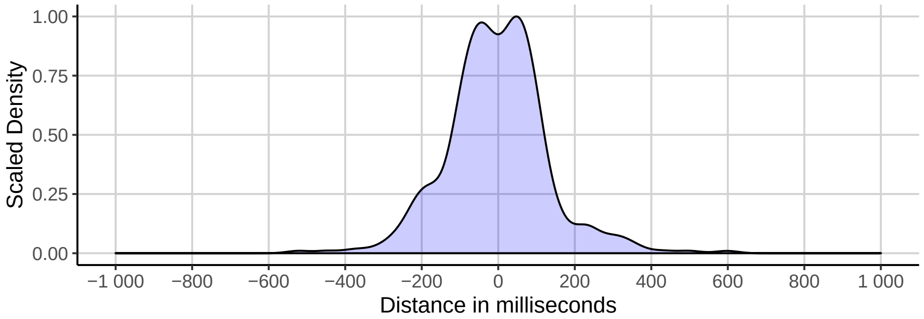
4.2 Are Apexes Synchronised with Their Nearest Tone?

The present study has already made a distinction between pairing and synchronisation. The pairings investigated so far indicate a proximity and semantic based relation of an apex with a tone - an apex and the nearest tone to the apex form a pair (also accounting for the semantic relation). Consequently, there is no limit on how near an apex and a tone should be in order to be paired. Synchronisation, on the other hand, deals with the actual measurements of time distances between paired units. It introduces a synchronisation criterion (the average syllable duration, see Section 3.5) in order to define what it considers "synchronised". Namely, if the time distance between the paired units is less than the average syllable duration, then the pairing achieves synchronisation; if not, the members of the pairings are not synchronised. So far, this chapter only has dealt with pairings. However, this section investigates whether the pairings of apexes and tones are synchronised and whether synchronisation is affected by prosodic, gestural, and information structural factors.

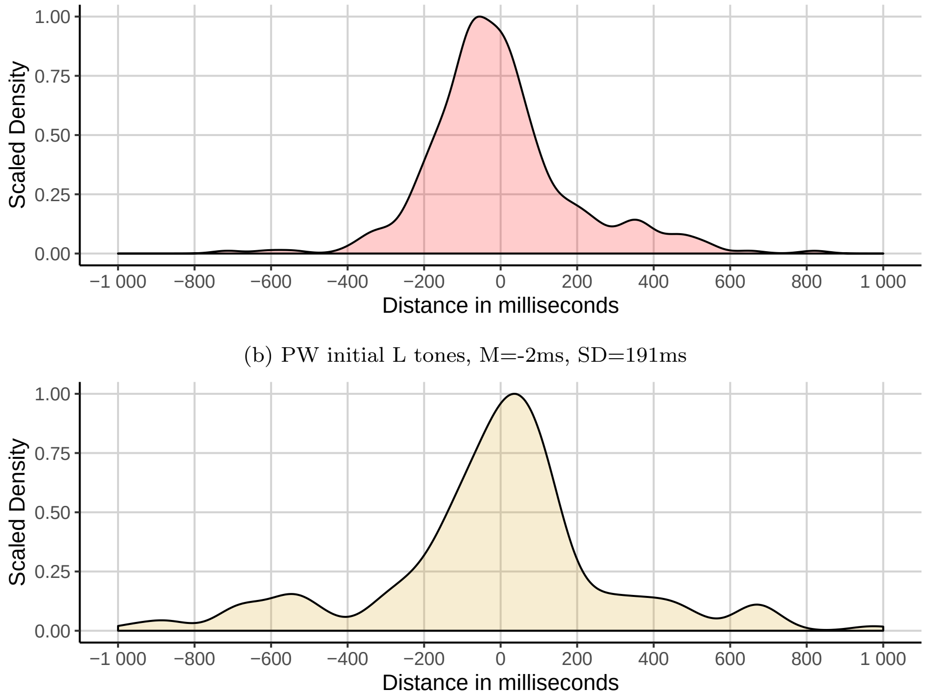
Figure 4.4 in Section 4.1 showed how near apexes were to their nearest tone, regardless of tone type. The pairing patterns presented up to this point indicated preferences depending on the tone type. Consequently, it may be the case that the time distances between apexes and tones may reflect this preference. That is, the time distance between the paired units may be greater or smaller depending on the tone type. In order to get a general view if that was the case, the time distance between the paired units for each tone type were calculated following the method explained in Figure 4.3. Phrase and boundary tone pairings were grouped together because the pairing instances were not evenly distributed across tone types, and therefore, there were not enough observations for these tones to enable meaningful comparisons. The grouping of phrase and boundary tones is also meaningful from a phonological perspective in that boundary tone annotations also included a phrase tone within the annotation scheme employed in the present study (Section 3.4.2).

Figure 4.12 shows the resulting normalised (scaled) distributions of time distance. In the figures, the mean distance for pitch accents and Ls were similar to each other and very close to zero. The distribution for phrase/boundary tones showed a higher mean and standard deviation. For each type condition, most of the observations were between ±200ms range. The distribution for pitch accents in Figure 4.12a showed the most compact peak followed by Ls in Figure 4.12b which had a slight spread into the positive direction on the x-axis. The distribution for phrase/boundary tones had smaller peaks further away from the main peak, reaching time distances as far as 800ms. Following on this, it was checked whether these small peaks and spreads were indicators of bimodality using Hartigan's dip test (Hartigan et al., 1985). However no significant evidence of bimodality was found (a: D = 0.018, p = 0.481; b: D = 0.015, p = 0.935; c: D = 0.016, p = 0.992).

Looking at the descriptives, it might be possible to talk about an effect of the tone type on the calculated time distances. Pitch accents and Ls had their means nearly on zero with standard deviations about the average syllable duration. Phrase and boundary tones exhibited a lead of tones over apexes about a syllable duration (i.e., tones occur before their paired apexes) but with a high standard deviation. These findings overlapped with the pairing preferences in that the time distances between apexes and the preferred tonal targets were closer to zero than the dispreferred targets. Therefore, the type of tones involved in the pairings might be a factor that affects the time distances, and hence synchronisation.



(a) Pitch accent, M=-7ms, SD=161ms



(b) PW initial L tones, M=-2ms, SD=191ms

(c) Phrase and boundary tones, M=-124ms, SD=556ms

Figure 4.12: Time-normalized histograms of the time distances of the nearest tone from an apex for each tone type

There might also be other factors related to gesture, prosody and IS that could affect these distances. As explained in Section 3.5, the present study used linear mixed-effect regression modelling to test whether (1) tone type, (2) ip type, (3) gesture type, (4) IS unit type, and (5) contrastiveness (as well as all two-way interactions between these factors) had an effect on the measured time distances. Within the tone type factor, phrase and boundary tone pairings were grouped together as in the earlier presentations (Figure 4.12). This resulted in three levels of tone types which were pitch accent, L, and phrase and boundary tones (referred to as "phrasal tone" from hereon). In addition, pairings within ptip level from the ip type factor, and background level from the IS type factor were excluded because of lack of data - the pairings tended not to take place within these areas. This led to an overall exclusion of 102 pairing instances in total. In brief, for each apex-tone pairing, these five features related to the phrases the tones and apexes were in were obtained from the data. These constituted the fixed effects in the model. The random effects included the participant and scenario information (both intercepts and slopes for single-term fixed effects). Overall, the full model was the following:

$$\begin{aligned} time\ distance \sim (Tone\ type + ip\ type + Gesture\ type + \\ IS\ type + Contrast) \wedge\ 2 + (Random\ effects) \end{aligned}$$

The backwards elimination process of insignificant effects (see Section 3.5) was applied on the successful model. The elimination of the random intercepts and slopes showed no significant effect of different participants and scenarios on the observed time distances. The elimination process for the fixed effects revealed tone type, the interaction of tone type and ip type, and the interaction of tone type and gesture type as factors that significantly affected the time distances between apexes and tones. A final model was fitted including only these terms. Table 4.11 shows the matrix of significant effects in the final model.

Table 4.11: The matrix of significant fixed effects on time distances between apexes and tones

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tone type	2	1.48	0.74	13.96	0.0000
Tone type:ip type	3	0.80	0.27	5.06	0.0018
Tone type:Gesture type	9	1.17	0.13	2.45	0.0093

The synchronisation analysis in the study is not directly interested in the results of this model. This model is fitted to get the estimates of time distances accounting for the effect of the factors included in the model (i.e., the effect of prosodic, gestural, and information structural contexts on synchronisation). In order to be able to decide whether these significant effects caused synchrony or asynchrony, the TOST procedure (or equivalence tests, see Section 3.5.1) was employed using the estimates acquired from the regression. To briefly recap this procedure, the process calculates the CIs of the estimated means taken from the model and tests whether or not these CIs fit within the set equivalence bounds via two one-sided t-tests.

Effect of gesture type

In Figure 4.13, the equivalence test output is plotted for the significant interaction of tone type and gesture type where the dashed lines indicate the equivalence bounds at ±160ms (the average syllable duration). From the equivalence test, it can be concluded that for iconic and deictic gesture apexes, the time distances of pairings with pitch accents and Ls were statistically equivalent to zero (see Table 4.12), meaning that apexes were synchronised with these tones given the ±160ms equivalence bounds. However, the same was not true for phrasal tones. For these, the observed effect was statistically different from zero, therefore apexes were not synchronised with these tones when they were paired with them.

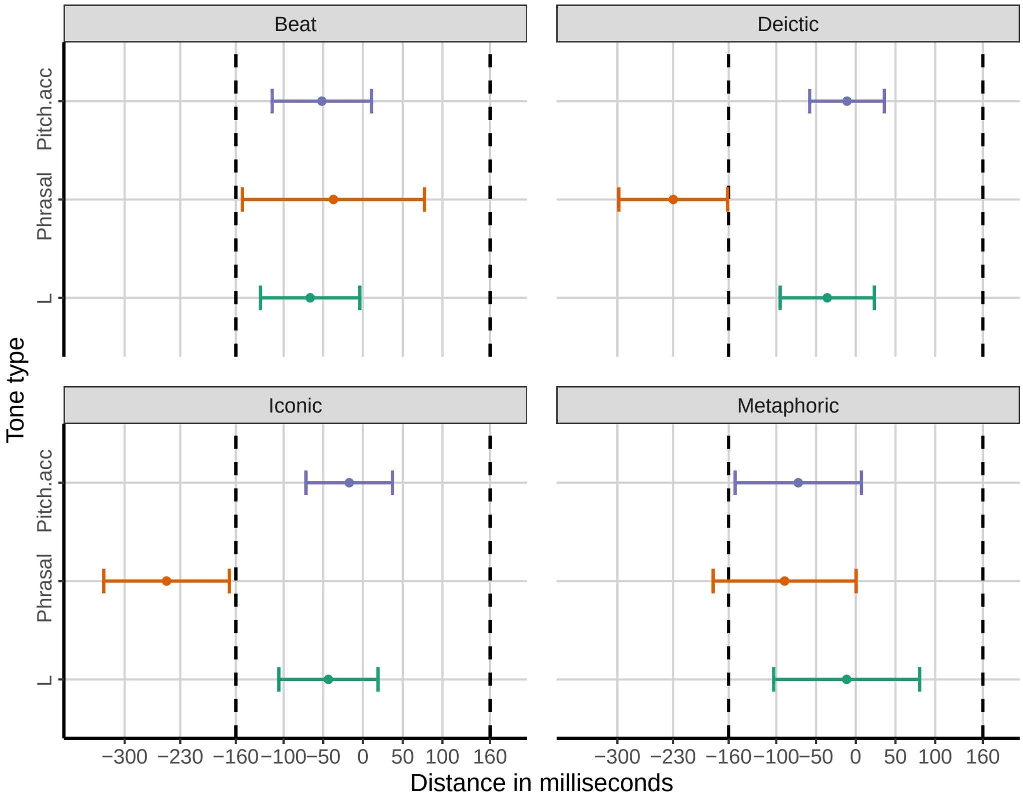


Figure 4.13: The estimated means of Tone type:Gesture type interaction with confidence interval at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

The estimates of phrasal tones (i.e., phrase and boundary tone grouping) in iconics and deictics stands out in the figure. When apexes were paired with phrasal tones, the estimates showed an average lead for tones of about -252ms in iconics and -226ms in deictics. In line with this, the equivalence tests were non-significant and confirmed that the CIs of the phrasal tone estimates do not fit between the equivalence bounds in iconics and in deictics (see Table 4.12). For pitch accent and L pairings, the tone lead was a lot less with -11ms/-43ms in iconics and -9ms/35ms in deictics, and the equivalence test results were significant for these tone types.

Table 4.12: The equivalence test results for apex-tone pairings in metaphorics and beats

	t	df	p
Pitch.acc	5.139	115	$< 0.001*$
L	3.665	74	$< 0.001*$
Phrasal	-2.167	33	0.982

(a) Iconics

	t	df	p
Pitch.acc	6.226	159	$< 0.001*$
L	4.109	66	$< 0.001*$
Phrasal	-2.001	52	0.975

(b) Deictics

	t	df	p
Pitch.acc	2.164	31	$0.018*$
L	3.175	24	$< 0.001*$
Phrasal	1.537	24	0.067

(c) Metaphorics

	t	df	p
Pitch.acc	3.395	65	$< 0.001*$
L	2.937	49	$< 0.01*$
Phrasal	2.103	15	$< 0.05*$

(d) Beats

For metaphoric and beat apexes, all estimates were within the equivalence bounds ranging between -89ms and 58ms, and the equivalence test were significant for all conditions except for the pairings of the apexes of metaphoric gestures with phrasal tones (see Table 4.12). These findings meant that beat apexes were synchronised with their nearest tone regardless of type. For metaphorics, the pairings with pitch accents and Ls satisfied the synchronisation criterion but phrasal tones did not because the lower bound (-160ms) was crossed for these as seen in Figure 4.13. However, despite the non-significant equivalence result, the present study interprets metaphoric apex pairings with phrasal tones as a successful synchronisation because the bound was crossed by only 23ms, and most of the CI fell within the bounds. In this type of situations, a deviation as small as one video-frame length (i.e., 17ms) can be considered acceptable. Therefore, it was concluded that the apex-tone pairings for metaphoric and beat gestures achieved synchronisation in the present study.

Taken together, the equivalence test results revealed two patterns of synchronisation with tone types depending on gesture type. The synchronisation in iconics and deictics mirrored the pairing preference - pitch accents and Ls were the preferred targets of apexes for pairing, and these pairings achieved synchronisation. The dispreferred targets, i.e., phrasal tones, were not synchronised with apexes when they were paired. On the other hand, this situation was not the same for the apexes of metaphorics and beats. The pairing preference was not mirrored in synchronisation since it was achieved with all tonal targets regardless of type. The present study interprets the synchronisation pattern in iconics and deictics as the standard apex synchronisation behaviour (further detailed later in this section) and the pattern in metaphorics and beats as a deviation from the standard as a result of the rhythmic behaviour of apexes observed in metaphorics and beats in the data.

Beats carry no narrative content and are in tune with the rhythm of speech (see Section 3.4.1.1). The relationship between the rhythmic feature of beats and the synchronisation patterns has to do with that fact that beats have been shown to occur consecutively at regular intervals, forming clusters (Tuite, 1993) (see Sections 2.1.1 and 3.4.1.1) . This observation was also true within the present study. Figure 4.14 shows an example from the data where there are three consecutive beats with relatively similar durations superimposed on the post-hold phase of a deictic gesture. Notice the pairings of apexes in the example. The apex of the deictic gesture pairs with the pitch accent as predicted. The first and the final beat apexes pair with the Ls in the absence of pitch accents and the apex in the middle pairs with the phrase tone.

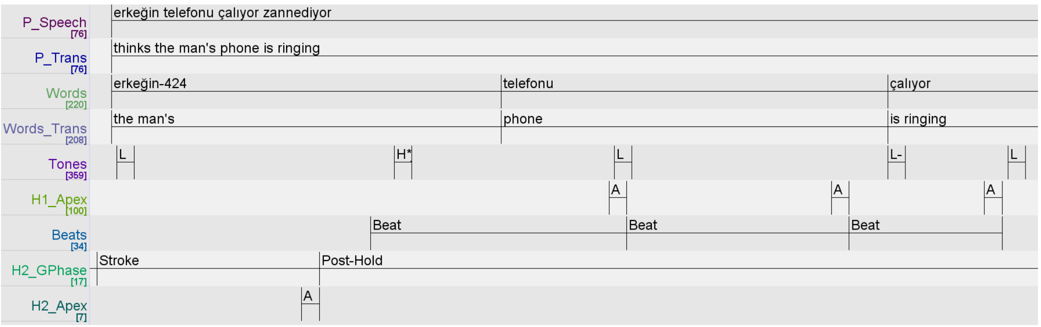
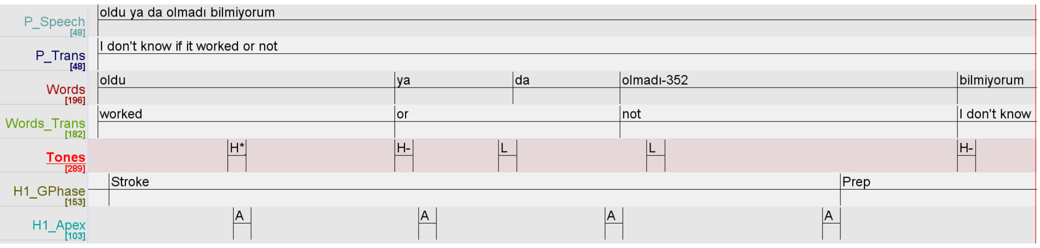


Figure 4.14: An ELAN screenshot showing that rhythmic pattern of apexes in consecutive beat gestures

This type of production of apexes in rhythmic sequences can force pairings with the dispreferred tones since apexes have to occur at a location that is imposed by the rhythm, and the nearest tone to that location will form a pairing with that apex regardless of preference. In terms of synchronisation, the time distances between apexes and tones are more likely to be shorter in rhythmic productions because Turkish can offer several tonal events over short durations as potential targets. Consecutive apex productions at fixed distances are able to find a tone occurring nearby in every case, ensuring synchronisation. Note that the standard pairing pattern could still be observed for beats as only 10% of beat apexes were paired with phrasal tones. It was only the case that the synchronisation of dispreferred pairings was easily made possible by the rhythmic production of beats. All things considered, the present study attributes the synchronisation of phrasal tones with beat apexes to the rhythmic and consecutive production of beats over short durations.

There was also a similar synchronisation behaviour of metaphoric apexes. Metaphoric gestures are produced to represent abstract concepts. Because the gesture elicitation was done through a narrative task of real-life action events in the present study, the number of metaphoric gestures was relatively low (14% of all annotated gestures). In the data, most of the metaphoric gestures were for paranarrative elements in line with McNeill (1992). Paranarrative elements in speech are the elements where narrators speak as themselves outside of the plot of the stimuli, often expressing thoughts on the narrative content. These paranarrative elements usually occurred when participants wanted to express uncertainty, continuity (of events) or repetition in the data. Figure 4.15 shows an example of an metaphoric gesture expressing paranarrative uncertainty.



P_Speech [48]: oldu ya da olmadı bilmiyorum

P_Trans [48]: I don't know if it worked or not

Words [196]: oldu | ya | da | olmadı-352 | bilmiyorum

Words_Trans [182]: worked | or | not | I don't know

Tones [289]: H* | H- | L | L | H-

H1_GPhase [153]: Stroke | Prep

H1_Apex [103]: A | A | A | A

Figure 4.15: An ELAN screenshot showing the rhythmic pattern of apexes within a metaphoric gesture

In the example, the metaphoric gesture accompanies a statement where the participant expressed that she did not remember whether the screwdriver the actor picked up did the job for her or not. The gesture was formed by two hands in front of the body and the palms facing each other. The stroke was executed by the flapping of hands sideways asynchronously, which encoded the expression of uncertainty. Gestures of this type created multiple apexes because they contained multiple abrupt stops and direction changes within the stroke (multi-segmented strokes in Section 3.4.1.3). This was also true for other paranarrative metaphorics - the continuity of events/processes was usually expressed through multiple small circles (at the same location or moving away from the body), and repetition was usually expressed by the hand going back forth between two apex locations. In brief, in the data, paranarrative elements in speech attracted metaphorics, and participants usually encoded these by producing strokes with multiple apexes. These strokes caused the production of rhythmic apexes, much like the beats in Figure 4.14 where multiple apexes occurred over a short period time at regular intervals.

In Figure 4.15, the first apex paired with the pitch accent. The following apexes took place at roughly 180ms intervals after the first stroke, each pairing with their nearest tone (H-, L, H- respectively). Note that both in metaphoric and beat gesture examples, the first apex in the series was paired with the preferred target, and the following ones occurring at fixed distances were paired at random. This shows the first apex follows the regular pattern, but the following rhythmically produced apexes do not. As with the beat synchronisation, the present study attributes the synchronisation of metaphoric apexes with phrasal tones to the rhythmic occurrence of apexes over a short duration, which bypasses any pairing/synchronisation preference.

In summary, tone type-gesture type interaction term revealed different synchronisation patterns. Iconic and deictic apex synchronisation reflected the overall pairing preference - synchronisation was not achieved with the dispreferred tonal targets. This behaviour was interpreted to be the standard synchronisation pattern. The metaphoric/beat apex synchronisations deviated from this as a result of the common rhythmic formations of apexes over short periods of time.

Effect of ip type

The next significant effect was tone type:ip type. Figure 4.16 plots the equivalence test output for this effect. For nuclear ips (nips), the estimates of pitch accents and Ls were only about 1 frame duration away from zero (20ms, -18ms respectively). The phrasal tone estimate was slightly further away from zero with -76ms. Consequently, the equivalence test confirmed that CIs of each fell within the set bounds, and therefore the observed time distances under this condition were statistically equal to zero (see Table 4.14).

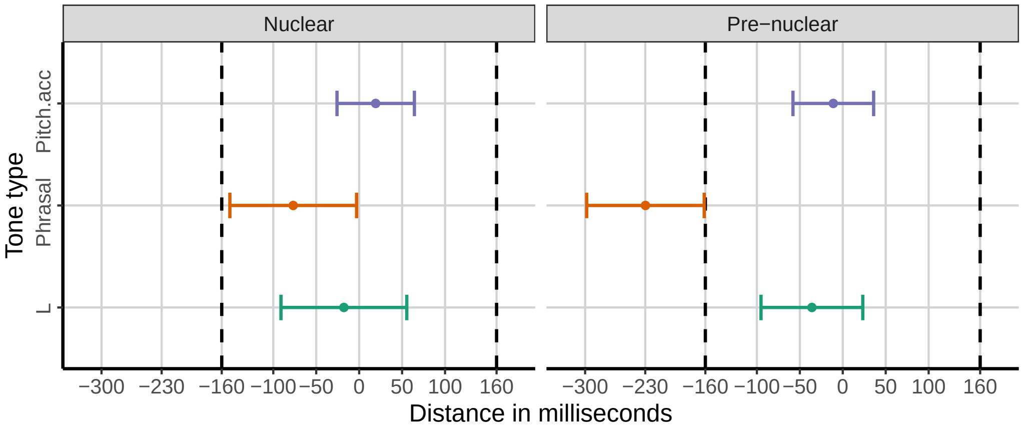


Figure 4.16: The estimated means of tone type:ip type interaction with CIs at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

Table 4.14: The equivalence test results for apex-tone pairings in pre-nuclear and nuclear ips

	t	df	p
Pitch.acc	6.226	151	$< 0.001*$
L	4.109	141	$< 0.001*$
Phrasal	-2.001	80	0.976

(a) Pre-nuclear

	t	df	p
Pitch.acc	6.490	221	$< 0.001*$
L	3.811	74	$< 0.001*$
Phrasal	2.214	46	< 0.05

(b) Nuclear

For pre-nuclear ips (prips), the estimated means of pitch accents and Ls were again closer to zero - they were only 1-2 frames away from it (-9ms and -34ms respectively). However, the phrasal tone estimate was much further away with -229ms, and fell outside of the lower bound. Consequently, the equivalence test was not significant for phrasal tones. However, the results were significant for pitch accents and Ls, showing that the time distances between apexes and pitch accents/Ls were not statistically different from zero.

The expected pattern was observed for prips where the synchronisation results reflected the preference indicated in the pairing pattern - pitch accents and Ls were synchronised with the apexes they were paired with. The less common pairings with phrasal tones exhibited a lead for tones with a distance that was more than the average syllable duration, and therefore these failed to synchronise. Unlike in prips, the apex-tone synchronisation was successful for all pairings in nips. One possible explanation might be that the prominence that nips carry had an effect on the distance between the members of the pairings. Regardless of the tone type, apexes were more tightly coupled with tones if they were in the nuclear area. This effect was not clearly observed for pitch accents and Ls as these were already tightly synchronised with apexes. However, the phrasal tone estimate moved almost an average syllable duration (153ms) closer to zero in nips compared to prips, achieving synchronisation. The effect of nuclearity presented here can be seen as evidence that the apex-tone synchronisation is sensitive to phrasal prominence as well. Note that previously in Section 4.1, it was shown that apexes tended to be paired more with the tones in nips, which was interpreted as an indication of apexes' sensitivity to phrasal prosodic prominence. The finding that this pairing preference was also mirrored in synchronisation (i.e., phrasal prominence both attracts apex-tone pairings and ensures their synchronisation) reinforces the claim of sensitivity. Further evidence supporting this claim is presented in Chapter 5.

The final significant effect in the model was the simple effect of tone type on the time distances. Figure 4.17 plots the equivalence test results for this effect. Note that the estimates here were averaged over the levels of gesture type and ip type effects that were presented in previous sections since the tone type was the common variable in all significant terms. Therefore, the results of this effect present a general view of synchronisation for apexes and tones, showing the overall (i.e., standard) synchronisation pattern of apexes and tones. The estimates for pitch accents and Ls were again 1-2 frames away from zero (-9ms and -34ms). However, the estimate for phrasal tones was at -229ms, falling beyond the lower bound. Consequently, the equivalence test results were significant for pitch accents and Ls, but not significant for phrasal tones.

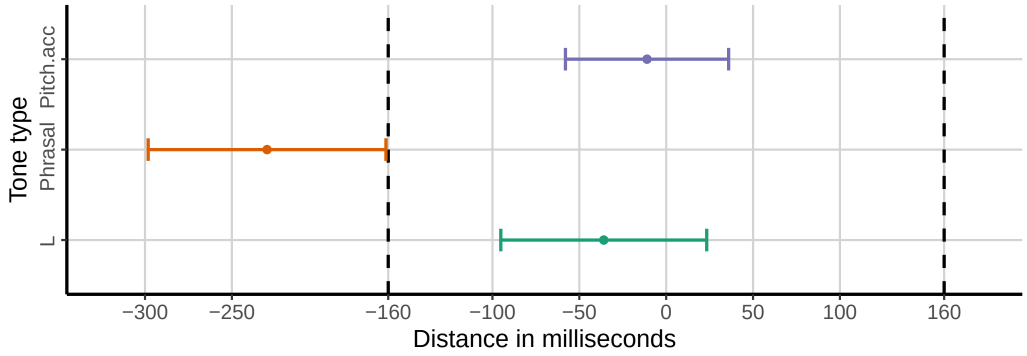


Figure 4.17: The estimated means of tone type term with the CIs at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

Table 4.16: The equivalence test results for apex-tone pairings for each tone type

	t	df	p
Pitch.acc	6.226	373	$< 0.001*$
L	4.109	216	$< 0.001*$
Phrasal	-2.001	126	0.976

Once again, apexes showed synchronisation with pitch accents and Ls, but not with phrasal tones, paralleling the pairing preference. In all significant effects presented, the linear predictions of tone type demonstrated the same tendency. The estimated means of pitch accents and Ls were 1-2 frame away from zero, while the estimated mean of phrasal tones was far into the negative direction on the x-axis showing that these tones occurred over an average syllable duration before the apexes they paired with. This tendency of increased distance between apexes and phrasal tones could also be observed in the conditions where these pairings achieved synchronisation. For example, in the nuclear ip condition in Figure 4.16 and in the metaphoric gesture condition in Figure 4.13, the phrasal tone estimates were 30-50ms away from those of pitch accents and Ls leaning towards the lower equivalence bound.

4.3 Summary

This chapter investigated the micro level synchronisation of gesture and prosody. Unlike the previous studies reviewed in Section 2.5.1, it accounted for the possible effects of prosodic and gestural contexts as well as the impact of the methodology it employed (e.g., pairing method, and forced pairing) on synchronisation through a series of detailed analyses. The consideration of these as factors on synchronisation has offered a novel and valuable insight into synchronisation phenomenon that has not been reported before.

The results have shown that apexes were chiefly attracted to prosodically prominent pitch accents, and in their absence, they were attracted to Ls. This meant that the pairing preference of apexes stayed within the PW domain, and did not shift to the markers of prosodic phrases higher in the hierarchy. As for synchronisation, the findings were in parallel with the pairing patterns. That is, pitch accents and Ls were more likely to be near apexes in different tonal environments, and they were near enough to achieve synchronisation within the average syllable duration. In light of this, the present study has shown for the first time that in addition to prominence, prosodic structure is also a constraint on synchronisation of units at the micro level. Prosodic structure first sets the domain of synchronisation, constraining which tones apexes can be synchronised with. In the case of Turkish, this domain was the PW, which was in line with the pairing of hierarchies presented in Figure 2.13. This domain enables pitch accents and Ls as potential anchors for apexes. The prominence constraint then dictates pitch accents as the anchor. In cases where there are no pitch accents, the only other available tone within the set domain (Ls) takes over the anchor role. This is further discussed in Section 7.1.2.2.

Although less likely, phrase and boundary tones could be found near apexes, but these co-occurrences did not achieve synchronisation in general.

It was found that rhythmic apex productions in metaphorics and beats could force pairing and synchronisation. Another important finding that will be more relevant in the next section was that the apex-tone pairing and synchronisation were sensitive to phrasal prominence. The nuclear areas attracted apexes more, and the time distances between apexes and tones were reduced under the effect of nuclear prosodic prominence.

The results have also shown that the consistency of gesture-prosody synchronisation can be used to complement phonological investigations. In the case of Turkish, the apex synchronisation pattern could be used to test the accentlessness of words with final stress (Section 2.5.1.2). The results presented here do not support the claim that words with final stress have pitch accents, since apexes were not attracted to the syllables where pitch accents would be. Instead, apex synchronisation operated as it does in the absence of prominence - apexes were synchronised with Ls.

Overall, the results presented here have shown evidence that gesture and prosody are synchronised at the micro level. The coupling of these two modalities is very strong as gesture responds to various aspects of prosodic organisation. This means that gesture production is highly informed by the generation of prosody. Consequently, the implications of these findings must be represented in unified speech-gesture production models. This is covered in detail in Section 7.2.

The present study also hypothesised that the synchronisation of gesture and prosody persists at the macro level. Therefore, it has extended its analyses of synchronisation to this level and tested whether gestural phrases show synchronisation with prosodic phrases. Chapter 5 presents the results of this investigation.

5

Synchronisation with Prosodic Phrases

The analysis in Chapter 4 has shown that gesture and prosody are temporally coordinated at the micro level. The results were partly in line with the findings of previous studies which reported a synchronisation between apexes and prosodically prominent events (Section 2.5.1). However, in the present study, the detailed analysis of synchronisation which accounted for the effects of the prosodic, gestural, and information structural contexts revealed that gesture and prosody are more connected than previously assumed. It has been shown in Chapter 4 that gesture is informed by the prosodic phrasing, lexical prominence, and phrasal prominence in terms of how it is anchored to speech. All these indicate a deeper systematic connection between gesture and prosody. However, the claims in the literature of interaction/synchronisation between gesture and prosody have mostly stemmed from the observation of synchronisation at the micro level only. There have not been many studies that looked at synchronisation at the phrasal level, and the ones that did were qualitative or showed inconclusive quantitative results (Section 2.5.2.1). Therefore, it is unclear whether gestural phrases are synchronised with prosodic phrases. This possibility should be investigated in order to decide the scope of gesture-prosody synchronisation. That is, it has not been established yet whether gesture-prosody synchronisation is limited to the synchronisation of apexes and tones or whether these modalities show some form of synchronisation between corresponding units at the phrasal levels within their structural organisations.

The analyses in this chapter address this need in gesture studies and look at a possible synchronisation of gesture phrases (G-phrases) with prosodic phrases. The synchronisation of gesture phases (G-phases) and gesture units (G-units) with prosodic phrases is left for another study. In brief, the investigation's aim is to find the best prosodic phrase anchor for G-phrases for synchronisation. In Section 2.5.2.2, the structural hierarchies of gesture and prosody have been paired. In line with this pairing, the present study considers intermediate phrases (ips) and intonational phrases (IPs) as the most likely candidates that could be synchronised with G-phrases (see Figure 2.13). Therefore, the synchronisation of G-phrases with both of these types of prosodic phrases are investigated in this chapter.

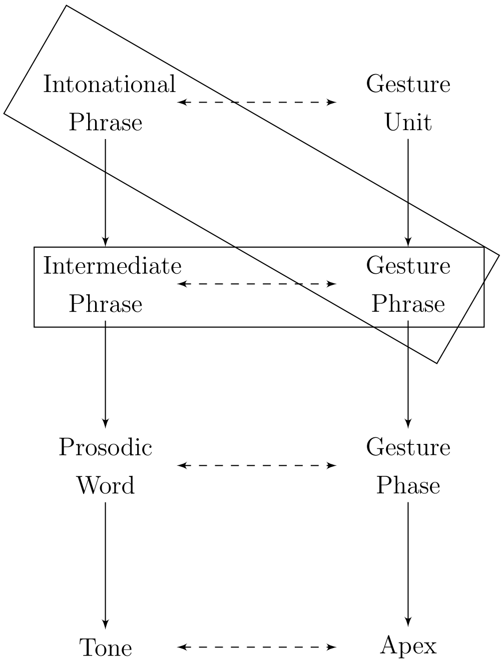


Figure 2.13: Pairing of structural hierarchies (repeated from page 90)

The presentation of results in this chapter follows a similar organisation to that in Chapter 4. It starts with explaining how the phrases under consideration were paired and how the time distances between them were calculated. Next, a description is given of patterns in the pairing of phrases that depended on their features (e.g., do G-phrases tend to be paired with nuclear ips more?). Finally, analyses are presented of the synchronisation of the paired phrases given the effects of prosodic, gestural, and information structural contexts. Overall, the analyses in this chapter answer the following questions (Section 2.5.2.2):

2. Are the onsets/offset of gesture phrases synchronised with the onset/offsets of intermediate phrases?

   2.a Is the synchronisation of gesture phrases and intermediate phrases affected by prosodic, gestural, and information structural contexts?

3. Are the onsets/offsets of gesture phrases synchronised with the onsets/offsets of intonational phrases?

   3.a Is the synchronisation of gesture phrases and intonational phrases affected by prosodic, gestural, and information structural contexts?

5.1 Pairing and Calculation of Time Distances

In Chapter 4, the pairing and calculation of time distances between tones and apexes consisted of measuring the distance between two points only. However, phrases (both prosodic phrases and G-phrases) are intervals in time that have an onset and an offset. This meant that there were four points in time (i.e. two pairs of interval boundaries) that needed to be synchronised in order to able to assume a synchronisation of two phrases. The analysis of synchronisation for phrases was based on checking the synchronisation of the onsets and offsets of these phrases as two different sets of measures. Therefore, the results of pairing and synchronisation in this chapter are presented separately for onsets and offsets.

As in Chapter 4, before any actual synchronisation tests, the onsets and offsets of corresponding phrases had to be paired. For the pairing of apexes and tones a semantic pairing method was used - synchronisation was tested between apexes and tones that existed in semantically related stretches of gesture and speech. The same method could also be employed in the pairing process G-phrases with IPs, but not with ips.

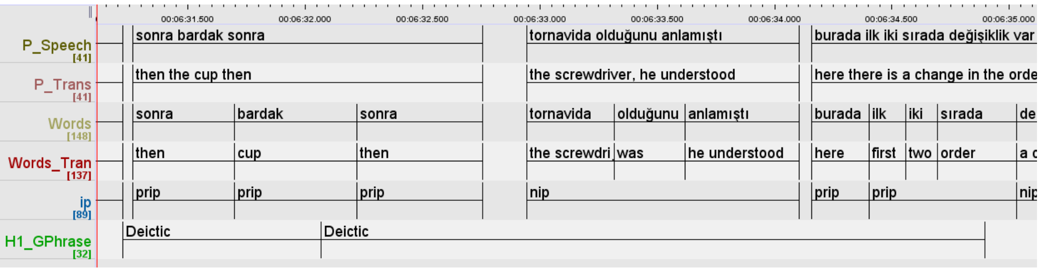
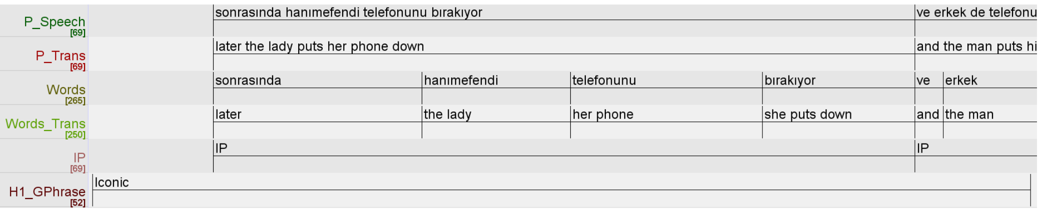


Figure 5.2: An ELAN screenshot showing the onsets and offsets of G-phrases (on the H1_GPhrase tier) and ips (on the ip tier). G-phrases usually spanned over multiple ips which can be constituents of different IPs.

The reason why the semantic method could not be employed for ips and G-phrases had to with the difference in duration and in number between ips and G-phrases in the data. The number of ips annotated (N=1363) was more than double the number of G-phrases (N=589) in the data. It was usually the case that there were multiple ips occurring over the duration of a G-phrase (see Figure 5.2). As reviewed in Section 2.5.2, such overlapping or containment phenomena was also observed in other languages for which the synchronisation of gestural phrases with prosodic phrases has been investigated. In the case of Turkish, this was seen as a natural result of the prosodic structure, in which ips have a short duration because many PWs form their own ips. Therefore, a semantic pairing process would pair a G-phrase with only one of the ips contained within the duration of the G-phrase, and any synchronisation test would predict asynchrony because of the durational difference between these (see Figure 5.2). These factors made such a one-to-one semantic pairing redundant as a qualitative look at the data had already revealed that a one-to-one synchronisation of single G-phrases and single ips was not observed in the data.

Although single G-phrases and ips were not completely synchronised, they could still show some form of temporal sensitivity to the other, which could occur as a synchronisation of independent boundaries. A G-phrase which spans over consecutive multiple ips forms a grouping of ips. It might be the case that G-phrases were synchronised with these groupings. In these cases, the onset of a G-phrase would be synchronised with the onset of the first ip, and its offset would be synchronised with the offset of the final ip. In other words, this would mean that whenever a G-phrase started or ended, there would be an ip starting or ending around the same time. Systematic observations of the sensitivity of boundaries in this manner would imply synchronisation.

In order to be able to test this, a proximity based method of pairing was employed. In this method, a G-phrase onset/offset was paired with the nearest ip onset/offset, regardless of their semantic relation. This meant that for one G-phrase, the paired ip onset and offset might belong to different ips. For example, in Figure 5.2, the pairing of the first deictic onset would be with the onset of the first prip (*sonra*), but the pairing of the deictic offset would be with the offset of the second prip (*bardak*). In that example, the G-phrase spans two ips and the G-phrase pairing is with the independent outer boundaries of this ip grouping. On the whole, the results of synchronisation tested between G-phrases and ips can predict a sensitivity of boundaries of these phrases by testing whether there is systematically an ip starting or ending around the same time a G-phrase starts/ends.



P_Speech [69]	sonrasında hanımefendi telefonunu bırakıyor				ve erkek de telefonu	
P_Trans [69]	later the lady puts her phone down				and the man puts hi	
Words [265]	sonrasında	hanımefendi	telefonunu	bırakıyor	ve	erkek
Words_Trans [250]	later	the lady	her phone	she puts down	and	the man
IP [69]	IP				IP	
H1_GPhrase [52]	Iconic					

Figure 5.3: An ELAN screenshot showing the onsets and offsets of an overlapping G-phrase (on the H1_GPhrase tier) and an IP (on the IP tier)

It was possible to use the semantic pairing method for the pairing of IPs and G-phrases. IPs usually consisted of multiple ips and had a longer average duration as a result. In addition, the number of IPs in the data (N=675) was close to the number of G-phrases annotated (N=589). In line with this, there was a better overlap in time between IPs and G-phrases in terms of duration (see Figure 5.3) and one-to-one semantic pairing was possible. This also meant that a one-to-one pairing of G-phrases with IPs would not predict synchrony or asynchrony by default as in the case of ip and G-phrase pairings. Consequently, IPs and G-phrases were paired using the semantic method - the synchronisation was tested between the IPs and G-phrase which carried the same semantic content.

Regardless of the pairing method used, the measuring of time distances was done between paired onsets and offsets as two different sets of measures (1. the time distances between onsets; 2. the time distances between offsets). These led to two realisations of synchronisation where the onset synchronisation and offset synchronisation were tested separately.

Figure 5.4 shows how the time distances between the paired onsets and offsets was calculated. Note that this calculation method was also used in the analysis of synchronisation of G-phrases and IS units (Chapter 6) since IS units can also be considered as phrases (i.e., longer intervals in time). Consistent with the approach used for apex/tone pairings (see Figure 4.3), the time-stamps of G-phrase onsets/offsets were subtracted from the IS/prosodic phrase onsets/offsets. As a result of the direction of the subtraction in the formula, any negative time distances meant that the onsets/offsets of G-phrases occurred later in time compared to the onsets/offsets of IS/prosodic phrases (e.g., offsets in Figure 5.4), and the positive values of time distances meant the opposite (e.g., onsets in Figure 5.4).

$$\frac{Onset}{Offset}\,t_{\left(\substack{IS/Prosodic\\ phrase}\right)} - \frac{Onset}{Offset}\,t_{(G-phrase)} = time\ distance$$

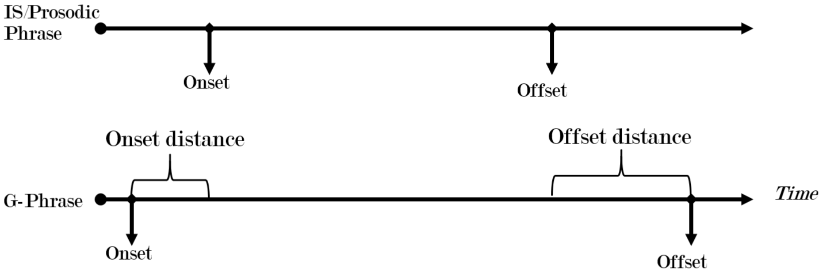


Figure 5.4: Distance calculation formula and a co-occurrence example

As stated previously, the analyses in this chapter aim to find the best phrasal prosodic anchor for G-phrases, and the best candidates for this were determined to be ips and IPs. This aim requires a comparison of the synchronisation of G-phrases with ips to their synchronisation with IPs. One potential problem in such a comparison would be that in the data, IPs could sometimes contain only one ip. In these cases, the onsets and offsets of these two phrases naturally coincide, therefore, any positive and negative result of synchronisation can be attributed to both types of prosodic phrases. The present study addressed the issue by excluding these cases (N=109) from the analyses.

5.2 General Time Distances and Pairing Patterns

This section investigates whether G-phrases (imagistic gestures excluding beats, see Section 2.5.2.2) were synchronised with intermediate phrases (ips). The previous section explained the pairing process of G-phrase onsets and offsets with ip onsets and offset, where the nearest boundaries (of the same kind) were paired, regardless of their semantic relation.

After the pairing process, the first step of the analysis was to calculate the time distance between the members of each pair and plot the distribution in order to get an understanding of how the compared markings were distributed around each other. The following list summarises the steps of the analysis. This process was largely the same for all analyses of phrasal synchronisation (including IS):

a. Find the nearest ip onset/offset for every G-phrase onset/offset.

b. Extract the time-stamps of the onsets/offsets and subtract the G-phrase time-stamp from the ip time-stamp (see Figure 5.4). In other words, measure the distance between onsets/offsets in negative milliseconds if the ip marking precedes the G-phrase marking, and in positive if otherwise.

c. Extract other features related to these phrases such as gesture type, ip type, IS unit, and contrast as well as participant and scenario information.

d. Plot a histogram.

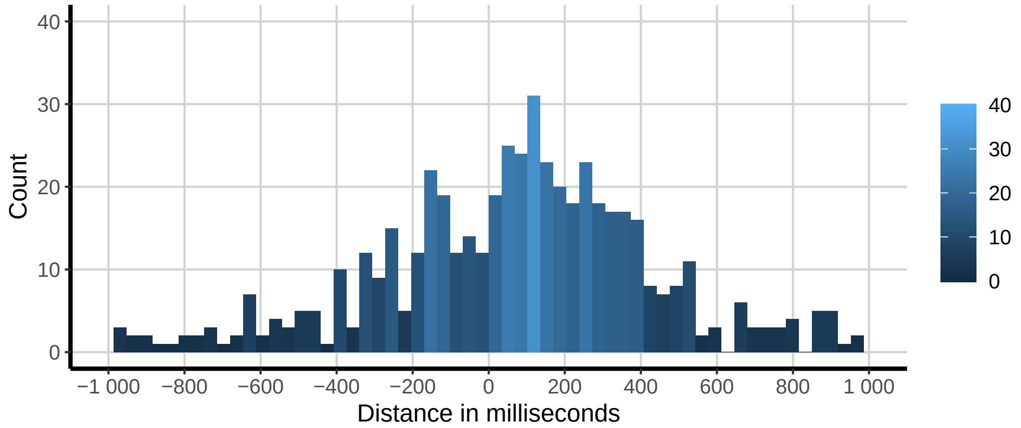


Figure 5.5: A histogram showing the time distances between ip onsets and G-phrase onsets. Negative values show that G-phrase onsets occurred after ip onsets. Binwidth equals to the duration of two frames (34ms) in every histogram presented from here on.

Figure 5.5 shows the resulting distribution of time distances for ip and G-phrase onsets. In the distribution, the standard deviation was 365ms, meaning that the majority of pairings (i.e., the nearest onsets) tended to take place within 365ms. On average, the distance between onsets was 67ms - quite close to zero. A mean very close to zero by itself did not necessarily reveal much about the nature of the distribution because a random distribution of negative and positive distances may also average to zero. However, in that case, the histogram would look relatively flat and not form a peak close to zero as in Figure 5.5. This clustering around the perfect synchronisation condition and the formation of an apparent compact peak meant that these onsets tended to occur very close to each other in time.

Figure 5.6 shows the distribution of offset distances. The distribution of time distances was very similar to the one of onsets in Figure 5.5, with 365ms as the standard deviation and 15ms as the mean. There was again a clear peak very near zero, indicating that the offsets of these units tended to occur very close to each other. Both onset and offset distributions did not show any evidence supporting bi-modality as per Hartigan's test (Hartigan et al.,

1985) (onsets: $D < 0.012, p = 0.937$; offsets: $D < 0.015, p = 0.656$).

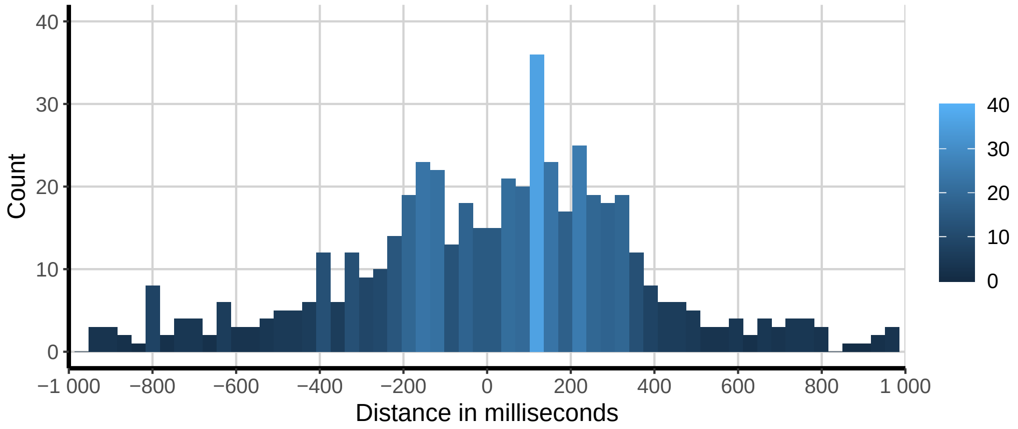


Figure 5.6: A histogram showing the time distances between ip offsets and G-phrase offsets. Negative values show that G-phrase offsets occurred after ip offsets.

Overall, these two distributions show that there was a tight temporal relationship between the boundaries of ips and G-phrases. The means of onset and offset distances were not located away from zero. The mean was within 4 frame distance for onsets and only 1 frame (17ms, the smallest resolution possible for the analysis of movement) for offsets. These durations were phonologically very short - shorter than the average syllable duration (160ms). In addition, the bulk of the calculated distances occurred within 21 frames which was roughly the average word duration (340ms) in the data. Even though a single ip did not seem to fully synchronise with a single G-phrase as shown in Figure 5.2, the clear and compact clustering of distances near zero for both onsets and offsets indicated that the boundaries of ips and G-phrases were sensitive to each other (although this is to be confirmed by the TOST procedure below). In other words, whenever a G-phrase started or ended, there was an ip starting or ending not too far away.

As it was with the apex-tone synchronisation in Chapter 4, one criticism of the previous studies dealing with phrasal synchronisation was that the prosodic, gestural and information structural context were neglected in their analyses (see Section 2.5.2). In Chapter 4, it has already been shown that these factors could reveal alternative modes of pairing (and therefore synchronisation). Similar observations might also be observed for the pairings of ips and G-phrases - there might be preferences in the pairing of ip and G-phrase boundaries depending on the gestural, prosodic, and information structural contexts the pairings took place in, e.g., the onsets of G-phrases might tend to be paired with the onsets of pre-nuclear ips. In the next step, the pairings of G-phrases and ips were investigated in different contexts in order to see whether there were patterns of pairing.

In Table 5.1, the ip/G-phrase pairings were categorised depending on the type of the paired ip (Table 5.1a), the type of the IS unit the paired ip was in (Table 5.1b), and whether that IS unit was contrastive or not (Table 5.1c).[1] In the table, the first two rows show what percentage of the total number of pairings (N=515) which constituted that pairing condition for onsets and offsets respectively. The third row in the tables show the overall number of annotations in the data for the conditions given in the columns.

Table 5.1a shows that for onsets, most of the pairings occurred with prips followed by nuclear ips (nips). For offsets, prips and nips were preferred equally although there was a decrease in the percentage of pairings with the prips and an increase for the nips when compared to the percentages of onsets. The general preference can be interpreted to be that G-phrases tended to start with a prip and end with either a prip or nip. Ptips seemed to be paired less frequently with G-phrases. These findings were especially highlighted when the total number of annotations of each were considered.

[1] The categorisation according to gesture type was excluded because this categorisation can only reflect the overall of number of annotated gestures - every G-phrase has to have an onset and offset co-occurring with ips.

Table 5.1: The pairings of ips and G-phrase categorised depending on the prosodic and information structural contexts that the pairings existed in

N=515	Condition		
	Pre-Nuclear	Nuclear	Post-Nuclear
Onset	56.8%	30.8%	12.4%
Offset	38.8%	38.3%	22.9%
N (ip)	648	491	224

(a) ip type

N=515	Condition		
	Topic	Focus	Background
Onset	35.7%	58.8%	5.5%
Offset	18.4%	64.5%	17.1%
N (IS unit)	387	540	133

(b) IS type

N=515	Condition	
	Neutral	Contrasted
Onset	46.9%	53.1%
Offset	55.9%	44.1%
N (Contrast)	602	458

(c) Contrast

It was already mentioned in Sections 2.5.1.2 and 3.4.2 that the duration of ips in Turkish is very short (PW $\approx$ ip), and that G-phrases span over multiple ips. Taking the findings in Table 5.1a into consideration, it seemed to be the case that a full G-phrase synchronisation was most likely to be with an element that would contain a combination of prips and/or the nip. Ptips are utterance final and usually contain the verbs in Turkish. In line with this, the prip+nip combinations in the immediately preverbal position were the main locus where the G-phrase pairings took place. This pre-verbal area has long been considered to be the default syntactic focus position in Turkish (Section 2.6.2). This finding was the first hint in the data implying that single G-phrases might be synchronised with single IS units, focus in particular. This observation was further supported by Table 5.1b which also demonstrated a pairing preference of G-phrases with foci in that for both onsets and offsets, the ips in the pairings were mostly in the focal area (however also note that there were more foci than topics in the data). The combined findings of Tables 5.1a and 5.1b indicate focus as a probable synchronisation domain in speech for single G-phrases.

Another finding in Table 5.1b was the increased percentage of onset pairings with topics (36%), which was not paralleled for offsets (18%). In a similar fashion, there was an increase in the percentage of offset pairings in backgrounds (17%) compared to the onsets (5%). These were mostly the spillover effects resulting from the pairings with focus. In the annotation scheme of the present study (Section 3.4.3), IS units had a linear ordering which was topic > focus > background. In some cases, when a complete G-phrase was paired with the focus of an utterance, the G-phrase was positioned in a way to contain the focus, leaving the first and the final phases of the G-phrase out of the pairing process. This caused the G-phrase onsets/offsets to pair with the ip onsets/offsets that were peripheral to the focus (topic and background). This constituted the first indication that IS units might actually be synchronising with G-phrase medial-phase(s) instead of full G-phrases. Overall, both Tables 5.1a and 5.1b provided implications for the pairing/synchronisation of G-phrases with foci. This possibility is investigated further in Chapter 6. Finally, in Table 5.1c, no clear patterns of pairings were observed when contrast was considered.

5.3 Testing of Synchronisation

As in Chapter 4, after the explanation of the pairing process and the investigation for pairing patterns, the next step in the analysis consisted of the actual testing of synchronisation.

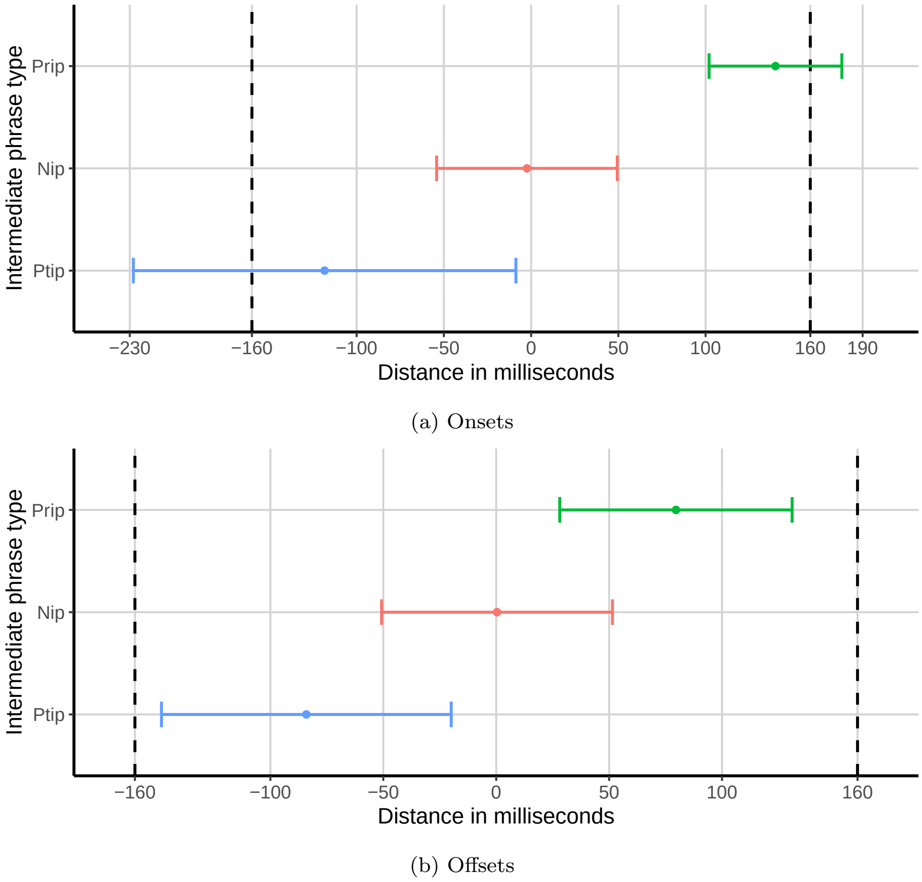
In Chapter 4, it has already been shown that prosodic, gestural and information structural factors could reveal alternative modes of pairing and synchronisation. Although not as striking as the apex-tone pairings, there also seemed to be patterns in the pairings of ips and G-phrases when these factors were considered, which implies that the synchronisation of G-phrases with ips can also be influenced by these factors. In line with this, the analysis tested whether the features in Table 5.1, as well as gesture type had an effect on the calculated time distance between paired ips and G-phrases. This was done using the same method as in Section 4.2 where a mixed-effects model was fitted. As onset and offset synchronisations were treated as two different sets of measures, two mixed-effects models were fitted (one for onsets and one for offsets) for the synchronisation tests of ips and G-phrases.

One consideration in the fitting of these models involved backgrounds. In the data, there was a lack of observations for the pairings within backgrounds. In general, only 5% of gestures (N=27) were for backgrounded elements, and it has already been shown in Table 5.1b that ip/G-phrase pairings were not common within backgrounds. However, notice that in the table, there was a difference in the number of pairings between the onsets and offsets in the background condition. The number of onset pairings within backgrounds was only 26, whereas this number was 106 for the offset pairings. This clear difference between the onsets and offsets suggested that backgrounded ips might be a valid target for the offset synchronisation but not for the onset synchronisation. Following this interpretation, the background level of IS units was excluded from the model for onsets, but it was kept in for offsets. The full models tested the effects of ip type, gesture type, IS unit type, and contrast (and their two-way interactions) as well as participant and scenario (random intercepts and slopes for single-term fixed effects). The following shows the full models that were fitted:

$$\begin{aligned} &time\ distance \sim (ip\ type + Gesture\ type + \\ &\qquad IS\ type + Contrast) \wedge\ 2 + (Random\ effects) \end{aligned}$$

First, the elimination of random effects showed no significant effect of different participants and scenarios on the time distances (as in apex-tone synchronisation). After the exclusion of random effects, the backward elimination of the fixed effects revealed a significant effect of only ip type on the distances for both onsets ($F = (2)15.8, p < .001$) and for offsets ($F = (2)7.5, p < .001$). This meant that the calculated time distances between paired ip and G-phrase onsets/offsets showed significant variation depending on which type of ip was involved in the pairing. Whether this significant effect caused synchronisation was next tested using equivalence tests.

As in Chapter 4, the output of the final model was only used to extract estimates of time distances taking into account the significant effects in the model. These estimates were then used in the equivalence tests in order to determine synchronisation given the significant effect on the time distances. In Figure 5.7, the outputs of the equivalence tests for onsets and offsets were plotted. The figures show the estimates and CIs of each level of the significant factor ip type (i.e., prips, nips, and ptips), showing their relation to zero and to the equivalence bounds at ±160ms.



(a) Onsets

(b) Offsets

Figure 5.7: The estimated means of ip types with confidence intervals at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

Starting with the onsets in Figure 5.7a, all the estimated means were within the equivalence bounds (but not the CIs). The estimate of nips was virtually on zero with 2ms. The estimate of prips was in the positive axis with 136ms, which meant that G-phrase onsets started less than an average syllable duration earlier than their nearest ip onsets. The estimate of ptips was -118ms, indicating a lead of ip onsets over G-phrase onsets. The equivalence test was only significant for nips, but not for prips and ptips whose CIs fell outside of the equivalence bounds. However, for prips, the upper bound was only crossed by 18ms (the duration of one video frame) and most of the CI was still within the bounds (see Figure 5.7a). In line with the decision about metaphorics in apex-tone synchronisation in Section 4.2, the pairings of G-phrase onsets with prip onsets were also considered as synchronised due to the marginal violation of the criterion. However, the same was not assumed for ptips where the lower bound was crossed by 67ms while also showing inconsistency in the measured distances as indicated by a CI span of 235ms. In brief, the onsets of G-phrases were synchronised with the onsets of prips and nips but not ptips.

The estimates for offsets in Figure 5.7b showed a similar pattern except all values were closer to zero. The estimate of nips was again very close to zero with 2ms. This value was 78ms for prips and -90ms for ptips. The CIs of the estimates all fell within the equivalence bounds for offsets, and the equivalence test was significant for all ip types (see Table 5.2), meaning that the offset of G-phrases are synchronised with the offsets of ip types, regardless of the significant effect of the ip type on the time distances between the paired offsets.

Table 5.2: The equivalence test results for onset/offset pairings for each ip type

	t	df	p
Prip	-1.031	273	0.152
Nip	5.984	147	$< 0.001*$
Ptip	0.747	32	0.230

(a) Onsets

	t	df	p
Prip	-3.070	192	$< 0.01*$
Nip	-6.135	195	$< 0.001*$
Ptip	2.325	117	$< 0.05*$

(b) Offsets

The asynchrony with ptips for onsets was attributed to the pairing preference. It was shown previously in Table 5.1a that ptips were dispreferred in the pairings of onsets. This was because ptips occurred in the utterance-final positions bearing no prosodic or discursive prominence (Sections 2.2 and 2.3). Their utterance-final position also meant that they were not ideal candidates for onset synchronisation because G-phrases rarely started that late relative to the utterances. A synchronisation of the onsets of G-phrases and ptips would mean that the last part of the G-phrase would either occur in silence, i.e., outside of the boundaries of the IP, or spill over to the following IP (if there was one), both which rarely happened in the data. These rare instances were considered as arbitrary pairings - the onsets of ptips were not legal targets for the onsets of G-phrases, therefore synchronisation between these was not achieved. Also, note that this asynchrony was only observed for onsets. The offsets of ptips were valid candidates for the offsets of G-phrases. A G-phrase could be synchronised with a prip at the onset and be synchronised with a ptip at the offset spanning over a whole IP, without having to occur in silence or spill over to the following semantically unrelated IP. In brief, these findings indicated that the pairing preference was mirrored in the synchronisation results - the less likely pairing combinations also failed to synchronise, given the synchronisation criterion of the analysis.

In the results of the apex-tone synchronisation in Section 4.2, it was reported that the nuclear prominence caused a reduction in the time distances between paired apexes and tones, regardless of type (see Figure 4.16). A similar effect of the nuclear prominence was also observed on the paired onsets and offsets of G-phrases and ips. When compared against the other ip types, nips carrying the prosodically prominent nucleus were synchronised with G-phrases almost perfectly (see Figure 5.7). Taken together, prominence can be said to be a factor in the synchronisation of gesture and speech at the phrasal level (as well as in the micro level) in that the maximally prominent ip usually tended to take place within the duration of the G-phrase (i.e., prip+nip combinations), and if G-phrases were paired with these nips, then the boundaries of these phrases synchronised near perfectly. Moreover, the ips that lack prominence markers (ptips) were not preferred for the pairings with G-phrases (especially for onsets), which was also reflected on the synchronisation results.

5.4 Summary

There was no full synchronisation between single ips and single G-phrases, because one G-phrase usually spanned over a grouping of consecutive multiple ips. However, there was a strong sensitivity between independent boundaries of ips and G-phrases - the boundaries of G-phrases were synchronised with the boundaries of ip groupings. This confirms the prediction in the pairing of gestural and prosodic structural hierarchies in Figure 2.13. This finding shows that gesture production is evidently informed by the prosodic organisation (explained further later in Section 5.7). Overall, Figure 5.8 shows a simple representation that summarises the preferred pairing and synchronisation patterns observed in the investigations so far by placing a G-phrase at its most likely temporal location in relation to ips. The vertical lines indicate the boundaries of the relevant phrases. The distances between the nearest onsets/offsets is based on the estimates calculated.

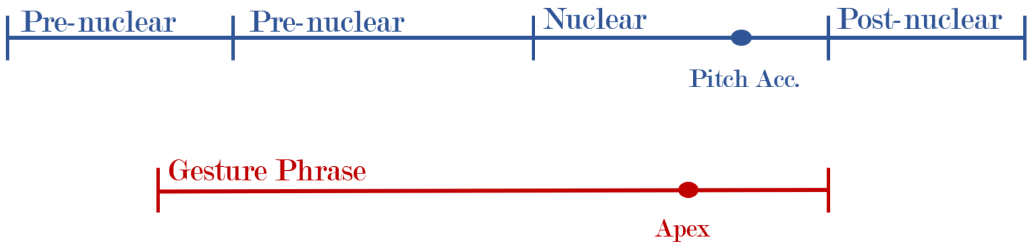


Figure 5.8: A simple representation of the synchronisation of G-phrases and ips

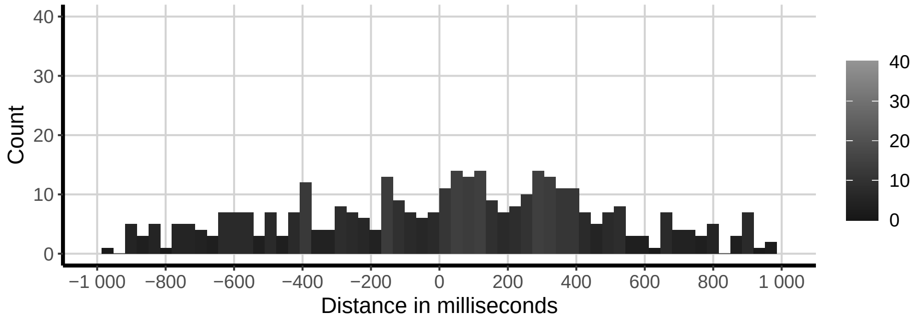
The goal of the analysis in this section has been to find the best prosodic anchor for G-phrases. There was evidence of synchronisation between ips and G-phrase, but G-phrases spanning over multiple ips imply a potentially larger domain governing the synchronisation of gesture and speech. The analysis has shown that most of the pairing instances were with the combinations of prips and nips. In line with this, it has been noted that G-phrases may be synchronising with information structural units, i.e., focus, which supports the original claim of the present study that gestures are also informed/synchronised with information structure. This possibility is further investigated in Chapter 6.

5.5 Gesture Phrase - Intonational Phrase Synchronisation

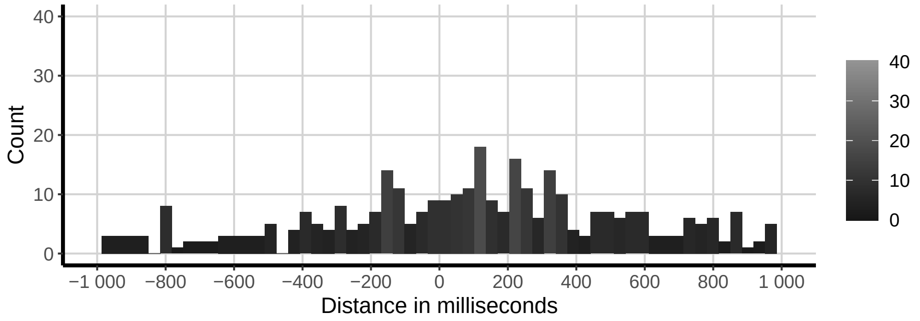
The finding that G-phrases are synchronised with ip groupings may also be an indication of the fact that G-phrases are synchronised with a larger prosodic constituent. As previously hypothesised in Section 2.5.2.2, the next best suitable candidate within the prosodic structure of Turkish is the IP. IPs usually contained multiple ips, and in terms of duration, the average IP duration (1310ms) was close to the average G-phrase duration (1453ms). These made IPs valid candidates for G-phrases to be synchronised with. Such a synchronisation would mean that the relatively short duration of ips in Turkish forces a shift in the hypothesised synchronisation of hierarchies (Figure 2.13), assigning the IP as the prosodic anchor for the synchronisation of G-phrases. This possibility is investigated in this section.

The analysis here is similar to the analysis of the synchronisation of G-phrases and ips. The pairing and synchronisation tests of IPs and G-phrases consisted of checking the time distances between their onsets and offsets using the same formula as in Figure 5.4. The same analysis steps were followed as in the ip/G-phrase synchronisation tests. One exception was that G-phrases and IPs were paired semantically before the calculation of the time distance for the reasons explained in Section 5.1. In other words, the synchronisation was only tested between the onsets and offsets of phrases that carried the same semantic content. As previously noted in Section 5.1, some of the IP/G-phrase pairings were excluded if the IPs contained only one ip (n=109) because the coinciding boundaries of ips and IPs in these cases meant that any kind of synchronisation results could be attributed to both of these prosodic phrases.

As the first step of the analysis, the time distances between the onsets and offsets of the semantically paired G-phrases and IPs were calculated. Figure 5.9 shows histograms of the calculated time distances for these.



(a) Onsets



(b) Offsets

Figure 5.9: Two histograms showing the time distances between IP onsets/offsets and G-phrase onsets/offsets. Negative values show that G-phrase onsets/offsets occurred after IP onsets/offsets.

In the distribution of onset distances in Figure 5.9a, the mean was reasonably close to zero with -106ms but the standard deviation was very high with 878ms. In the distribution of offset distances in Figure 5.9b, the mean was more than the average syllable duration away from zero in the positive axis with 250ms. The standard deviation was again very high with 942ms. In these figures, the clear and compact peaks observed between ips and G-phrases were not present (compare with Figures 5.5 and 5.6). There was no clustering that would indicate perfect synchronisation, i.e., the means were further away from zero. The high standard deviations also flattened the distributions, forming a wider plateau with no peaks that could be interpreted as meaningful. Moreover, in both distributions almost 25% of the instances occurred outside of the ± 1000ms range displayed in the figures. Overall, the distributions of time distances in these figures were vastly different from the distributions related to ips and G-phrases, and the signs of a possible synchronisation were absent. Therefore, the synchronisation of G-phrases with IPs did not seem plausible based on these distributions.

In the next step, the possibility of possible pairing patterns was investigated to see whether there was a preference in how G-phrases were paired with IPs. The features extracted that were related to these phrases were only gesture type and IP type (i.e., whether the IP was separated from its syntactic parent or not, see Section 3.4.2.2). IS unit type, contrast, and ip type were not relevant because a single IP naturally contains different levels of these features.

Table 5.4: Two-way frequency tables of paired IPs and G-phrases

	IP	IP!	N (G-phrase)
Deictic	48.7%	46.9%	244
Iconic	35.2%	39.6%	186
Metaphoric	16.1%	13.5%	85
N (IP)	401	114	515

Table 5.4 shows the numbers of paired IPs and G-phrases. The percentages in the columns show what percentage of the IPs (numbers shown in row "n (IP)") were paired with specific gesture types. The column "n (G-phrase)" shows the numbers of gesture types involved in the pairings. Table 5.4 did not reveal any specific pattern in the pairing of G-phrases with IPs. The distinction between IP and IP! (separated IPs) did not relate to preferences for a specific gesture type, and the percentages mirrored the number of gesture types involved.

Next, the analysis tested whether these phrasal features as well as different participants and scenarios had a significant effect on the calculated time distances. The full models that were fitted looked like the following:

$$time\ distance \sim (Gesture\ type * IP\ type) + (Random\ effects)$$

For random effects (intercepts and slopes for single-term fixed effects), the results of the elimination process showed no significant effects for both onsets and offsets, which was consistent with the results of analyses related to G-phrases and ips. Then, the elimination process for the fixed effects (i.e., features) also showed no significant effect of any terms for both onsets and offsets - neither gesture type nor IPs being separated or not caused any variation in the calculated time distances between paired G-phrases and IP onsets/offsets. In the absence of significant effects, the input needed for the equivalence tests were directly extracted from the data. The outputs of the equivalence tests are shown in Figure 5.10, which shows the estimates of the time distances for the onsets and offsets as well as their CIs.

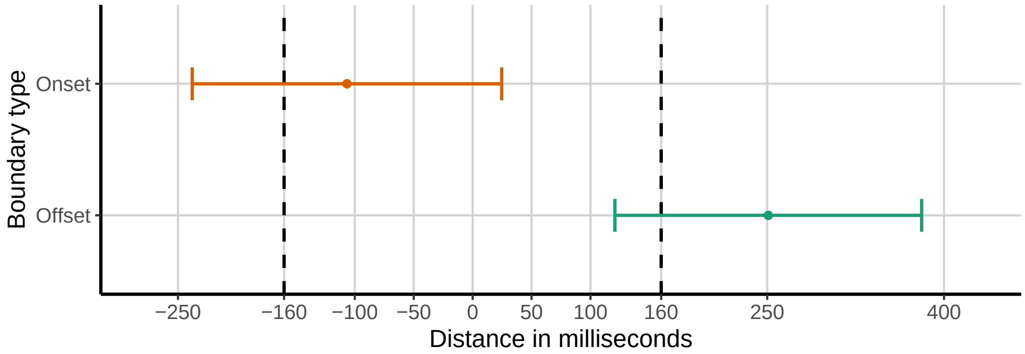


Figure 5.10: The estimated means of time distances between IP and G-phrase onsets/offsets with confidence intervals at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds.

No synchronisation of onsets or offsets for was observed these phrases. Unlike Figure 5.7, neither of the CIs fitted within ±160ms equivalence bounds. In addition, the CIs spanned over 300ms, which signalled a great deal of variation in the time distances between the boundaries of these phrases. However, one relevant finding of the TOST procedure might be that the synchronisation trend for the onsets was in the negative area, but it was in the positive area for the offsets. This meant that G-phrases tended to be contained within the durations of IPs - IPs started earlier and ended later than G-phrases. Overall, the results of the equivalence tests for IPs were predicted by the distributions in Figure 5.9 in which no evidence that could suggest a synchronisation of G-phrases with IPs was found.

5.6 Summary

The strong sensitivity between the boundaries of ips and G-phrases was not present between the boundaries of IPs and G-phrases. The analyses concluded that G-phrases were not synchronised with IPs. Figure 5.11 shows a representation that places a G-phrase to its mostly likely temporal location in relation to both ips and IPs, summarising the synchronisation patterns observed so far.

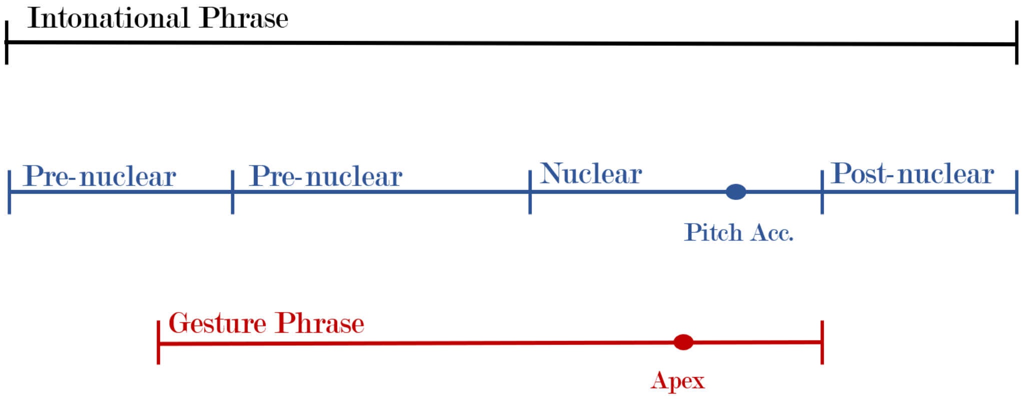


Figure 5.11: A representation of the pairing and synchronisation models after IP synchronisation results. This is expanded from Figure 5.8.

5.7 Summary of Phrasal Synchronisation

This chapter has investigated the macro level synchronisation of gesture and prosody. Unlike the relevant earlier studies (Section 2.5.2), the analyses have accounted for the possible of effects of gestural, prosodic, information structural factors in phrasal synchronisation as statistically defined in the present study. Considering the limited number of studies on the phrasal synchronisation of gesture and prosody as well as the inclusion of these factors in the analyses, the present study contributes to our current understanding of the synchronisation of gesture and prosody by showing that gesture and prosody are also synchronised at the phrasal level and that prosodic structure has an effect on this synchronisation.

The results show that the boundaries of G-phrases were synchronised with the boundaries of ip groupings. Following that, the analyses looked for what these ip groupings might be corresponding to within the linguistic structure of Turkish. It has been found that the synchronisation with ip groupings is not an indication of synchronisation with IPs - no evidence has been found to support this.

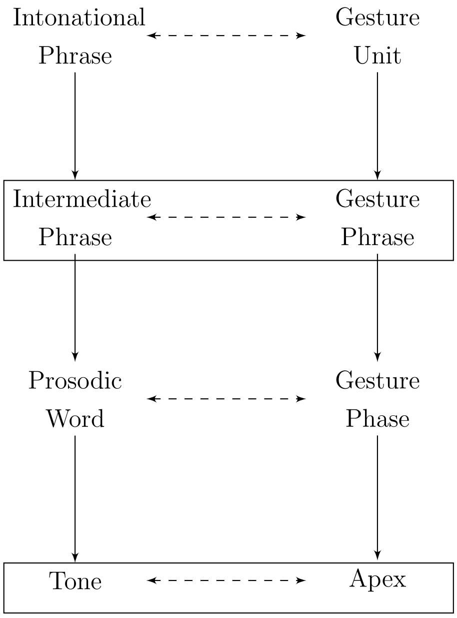


Figure 5.12: Pairing of structural hierarchies (modified from Figure 2.13)

The results of synchronisation analyses have confirmed the predictions about the pairing of structural hierarchies in Figure 5.12. That is, G-phrases are organised in a way that to be synchronised with ips. The durational discrepancies between ips and G-phrases did not cause a shift in the synchronisation target of G-phrases from the ip level to the IP level although IPs might be seen as better targets for G-phrases when their durations are considered. This suggests that gesture production is informed by the prosodic phrasing of its co-occurring speech. The boundaries of ips within the speech signal are made available to gesture production so that the phrasing of gestural movements can be organised and timed accordingly. Moreover, the boundaries of ips bearing the maximum prosodic prominence have been shown to be synchronised more tightly. This implies that in addition to the prosodic phrasing, prosodic prominence is also a factor influencing the synchronisation of gesture and prosody. These implications for gesture production overlap perfectly with the implications of the apex-tone synchronisation for gesture production in which prosodic phrasing and prominence have also been shown to be constraints on synchronisation. Taken together, these findings have implications for unified models of speech and gesture production (Section 2.3). These implications are further discussed in Chapter 7 where an extended model of speech-gesture production is proposed.

Initial observations in the pairing of G-phrases with ips have also revealed that G-phrases may be synchronised with information structural units, i.e., focus, since there was a tendency that the paired G-phrases and ips took place in the focal areas. This was in line with the hypothesis of the present study which claimed that gesture is also informed by/synchronised with information structure. Chapter 6 investigates this possibility in a systematic way.

6

Synchronisation with Information Structure

In the majority of earlier studies, prosody has been seen as the main driver of gesture and speech synchronisation (Sections 2.5.1 and 2.5.2). The findings of the present study in Chapters 4 and 5 have revealed that gesture and prosody are indeed connected; more deeply than previously assumed in these earlier studies. However, the findings in these chapters have also implied that gesture, through prosody, may be synchronised with information structure (IS). In Section 2.6.2, it was shown that prosody has a very close relationship with IS in addition to its relationship with gesture. Therefore, it is possible that prosody may be mediating the synchronisation of gesture and IS.

There have also been a number of studies showing more direct forms of association between gesture and IS. An earlier model of speech and gesture production implied that gesture and IS may be associated in the high-level planning of speech and gesture (see Growth Point Theory in Section 2.3).

Moreover, Section 2.6.2 also reviewed more recent studies that associated gesture and IS. In general, these studies claimed that information structural properties such as common ground and retrievability have an effect on gesture frequency and gesture form.

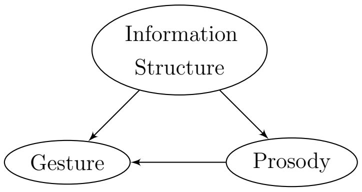


Figure 2.17: Three-way association of information structure, prosody, and gesture (repeated from page 111)

Despite all these suggestions in the literature, investigations of synchronisation have been mostly limited to the synchronisation of gesture and prosody. Synchronisation with other speech related components such as information structure has been largely overlooked. Taking all the proposals and previous associations into consideration, the present study has hypothesised that gesture may also be synchronised with IS. In line with this claim, it has been postulated that gesture, prosody, and information structure form a three-way synchronisation where the anchoring of gesture is governed by both prosody and information structure (see Figure 2.16).

In order to test this hypothesis, this chapter investigates whether gesture phrases (G-phrases) were synchronised with topics, foci, and backgrounds (i.e., IS units). In Section 2.6.2, it was explained why the topic, focus, and background dimension is relevant - these categories play the main role in building of a discourse which gesture contributes to. Therefore, G-phrases, i.e., single meaningful units of bodily action, might be synchronised with these. Note that, in the present study, IS units are defined on ips. That is, in the annotation scheme of the study each IS units is made up of one or more ips (see Section 3.4.3). The investigation of synchronisation between G-phrases and IS units therefore provides a potential answer to the issue found in Chapter 5 that ip and G-phrase boundaries are synchronised but not coextensive, and that G-phrases seem instead to be synchronised with a unit intermediate between ip and IP. Overall, this chapter aims to answer the following questions:

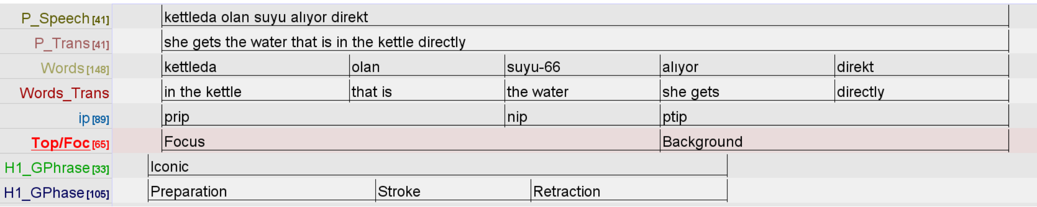
4. Are the onsets/offsets of G-phrases synchronised with the onsets/offsets of information structure units?

   4.a Is the synchronisation of gesture phrases and information structure units affected by gestural and information structural contexts?

The analytical steps and methods used in this chapter are similar to those used in the analysis of synchronisation between G-phrases and prosodic phrases since both G-phrases and IS units are intervals in time (see Section 5.1). This means that the synchronisation of onsets and offset are presented as two different sets of measures, as in Chapter 5. The presentation of results is also done in a similar fashion. First, the pairing process and the calculation of time distances between the paired G-phrases and IS units are described, followed by an analysis of whether or not there were patterns in the pairings (e.g., did iconic gestures tend to be paired with foci?). At this point, any observed patterns could be seen as evidence that gesture production is sensitive to the type of IS unit it accompanies, showing a direct association between gesture and IS. This is followed by an analysis to see whether synchronisation was achieved between paired units given the potential effects of gestural and information structural contexts.

6.1 Pairing and Calculation of Time Distances

The numbers of annotated IS units and G-phrases presented a similar ratio to the number of annotated ips and G-phrases (see Section 5.1), in that the number of IS units (N=1060) outnumbered the number of G-phrases (N=589) almost two to one. However, this did not constitute a problem for the semantic pairing process. Utterances typically contained two IS units (e.g., topic + focus), whereas they were usually accompanied by only one G-phrase. Initial observations showed that G-phrases overlapped with only one of these IS units and did not span over both topic and focus. Such overlaps in time made it clear that G-phrases and IS units are suitable candidates for pairing and synchronisation since they tended to start and end somewhat close to each other as seen in Figure 6.2. The figure shows an overlap of a G-phrase with a focus that contains a prip and a nip in the preverbal area.



P_Speech [41]	kettleda olan suyu alıyor direkt				
P_Trans [41]	she gets the water that is in the kettle directly				
Words [148]	kettleda	olan	suyu-66	alıyor	direkt
Words_Trans	in the kettle	that is	the water	she gets	directly
ip [89]	prip		nip	ptip	
Top/Foc [65]	Focus			Background	
H1_GPhrase [33]	Iconic				
H1_GPhase [105]	Preparation	Stroke	Retraction		

Figure 6.2: An ELAN screenshot showing the onset and offset of a G-phrase (on the H1_GPhrase tier) relative to the onset and offset of a focus (on the Top/Foc tier)

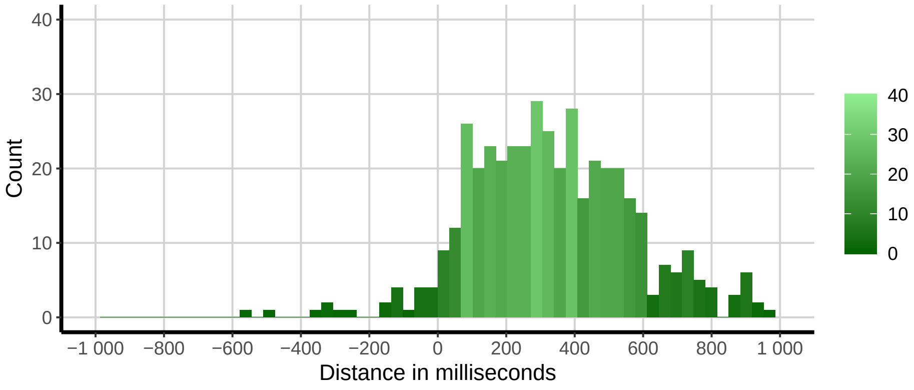
These clear overlaps and durational similarity between G-phrases and IS units meant that they could be paired using the semantic pairing method. That is, G-phrases were paired with IS units which contained the word that was semantically most closely related to the content of the G-phrases, and synchronisation was tested between these semantically related units (unlike ip/G-phrase synchronisation).

The calculation of time distances between the onsets and offsets of paired units was done using the same formula used in the analyses of synchronisation between G-phrases and prosodic phrases (i.e., IS units time - G-phrase time, see Figure 5.4).

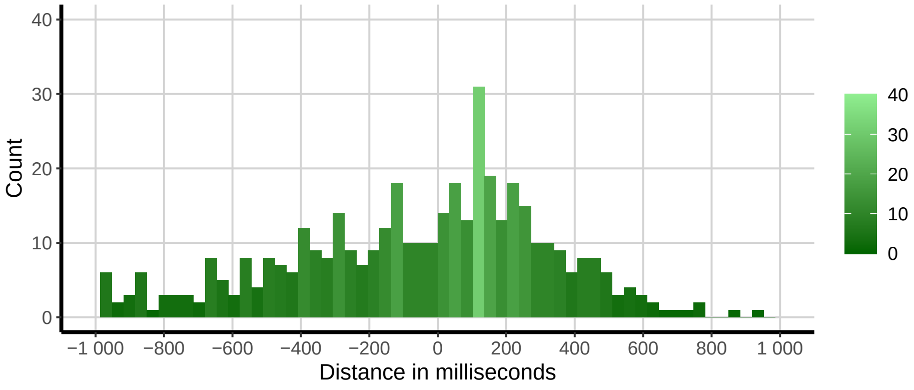
6.2 General Time Distances and Pairing Patterns

As the first step of the analysis, the time distances between (a) onsets and (b) offsets of G-phrases and IS units were plotted in order to get an understanding of how closely aligned these were.

For onsets, Figure 6.3a showed a clear clustering of observations that was centred, on average, 413ms away from zero with a standard deviation of 373ms. The steep peak reflects a distribution that is very different from the distributions of IPs/G-phrases in Figure 5.9, but similar to those of ips/G-phrases in Figures 5.5 and 5.6. The difference was that in the distributions of ips/G-phrases, the peak was quite close to zero, whereas here, it was about a median G-phase duration (400ms) away from zero in a positive direction on the x-axis. This meant that G-phrases tended to start about a G-phase duration earlier than IS units.



(a) Onsets



(b) Offsets

Figure 6.3: Two histograms showing the distances between the onsets/offsets of G-phrases and IS units. Negative values show that G-phrase onsets/offset occurred after IS unit onsets/offsets.

In the offset distribution in Figure 6.3b, there was also an apparent peak around 150-200ms, but the mean of the distribution was -195ms. This indicated a skew in a negative direction on the x-axis. In addition, the spread of instances in this negative direction was wider, which was reflected on the high standard deviation, 675ms. A negative mean indicated a trend where G-phrases tended to end more than an average syllable duration after their paired IS units. The overall temporal positioning of G-phrases relative to IS units seemed to be that G-phrases tended to start early and end later than their paired IS units (this is further discussed in the next section).

In the next step of the analysis, the gestural and information structural contexts of the pairings of IS units and G-phrases were described. That is, the analysis laid out how many and what kind of G-phrases and IS units were paired in the data, in order to reveal if there were any patterns in the pairings of units under consideration. As mentioned in Section 3.4.3.1, information status was not included in the analysis. Consequently, the extracted features that are related to the pairings involved gesture type, IS unit type, and contrast. In Table 6.1, the pairings of G-phrases and IS unit are grouped according to these features, and their percentage share of all pairings is listed. The second row in each sub-table shows the percentage share of each level of these features of all annotations in the data.

Table 6.1: The pairings of G-phrase and IS units categorised depending on the gestural and information structural contexts that they existed in

	Deictic	Iconic	Metaphoric	N
Gesture type	50.8%	33.8%	15.4%	515
Annotated	49.7%	35.4%	14.9%	589

(a) Gesture type

	Neutral	Contrasted	N
Contrast type	55.5%	45.5%	515
Annotated	56.7%	43.3%	1060

(b) Contrastiveness

	Focus	Topic	Background	N
IS type	68%	26.8%	5.2%	515
Annotated	51.0%	36.5%	12.5%	1060

(c) IS type

At a first look, Table 6.1a shows that deictics (51%) were paired with IS units more than iconic (34%) and metaphorics (15%). However, when the total number of annotations in the data is considered, it becomes clear that these gesture types were paired with IS units around the same rate as each other, and the difference in the percentages is a result of there being more deictics than iconics in the data. In other words, the number of gesture types in the pairings mirrors their total number in the data. Table 6.1b shows a similar scenario. The almost half-and-half sharing of pairings between contrastive and neutral (i.e., non-contrastive) units paralleled their total numbers in the data - contrastive and neutral units were paired with G-phrases at similar rates. However, the distribution in Table 6.1c indicates a tendency. Foci were more likely to be paired with G-phrases. That is, the semantic content expressed in G-phrases were more likely to be contained within foci with 68%. Taken together, these show that G-phrases tended to accompany foci, which supported the assumptions of Growth Point theory (see Section 2.3) where gesture was claimed to accompany structurally highlighted and newsworthy information. In contrast, background areas attracted as little as 5% of G-phrases. This implies that the speech chunks that were deemed already established and shared by interlocutors (i.e., not newsworthy) were not gestured as “the highlighter” function of gesture was reserved mostly for foci. Overall, the findings showed that gesture was informed by the organisation of information within utterances, supporting the claim of the present study that the anchoring of gesture relative to speech is governed by IS in addition to prosody.

Next, the analysis checked whether there were also patterns of pairings between the different levels of each feature, for instance whether iconic gestures tend to be paired with foci more than with topics and backgrounds. Table 6.2 shows a two-way frequency tables between the levels of gesture type and IS type.

Table 6.2: Two-way frequency table between the levels of Gesture type and IS type

	Focus	Topic	Background	N (G-phrase)
Deictic	58.7%	38.6%	2.8%	259
Iconic	82.2%	12.4%	5.3%	174
Metaphoric	67.5%	19.5%	13.0%	82

Table 6.2 shows that the majority of pairings was with foci for all gesture types in line with the pattern in Table 6.1. An interesting observation in the distributions was that the pairing rate with foci increased as the iconicity of gesture increased. Iconics have the highest level of iconicity, followed by the metaphorics which can be argued to bear less iconicity as they refer to abstract concepts that may or may not be easily accessible in the speech stream. Deictics, on the other hand, have the least amount of iconicity since they are location markers which do not necessarily carry any formal resemblance to their speech referent. This ranking of iconicity was paralleled in the percentages of pairings with foci: 82% of iconics were paired with foci, followed by 68% for metaphorics and 59% for deictics. In terms of the pairings with topics, deictics seemed to be attracted to topics more than iconics and metaphorics with 39%. It can be said that in order to establish links to entities in the previous discourse, interlocutors preferred pointing at the locations of those entities in their mental space.

Table 6.3: Two-way frequency table between the levels of IS type and contrast

	Neutral	Contrasted	N (IS type)
Focus	56.5%	43.5%	345
Topic	42.5%	57.5%	139
Background	100%	-	31

Table 6.4: Two-way frequency table between the levels of gesture type and contrast

	Neutral	Contrasted	N (G-phrase)
Deictic	41.3%	58.7%	259
Iconic	68.0%	32.0%	174
Metaphoric	71.4%	28.6%	82

Table 6.3 shows that IS units paired with G-phrases carried contrast at equal rates. Consequently, contrast did not seem to be a factor affecting the pairing rates of G-phrases with IS units. However, it was observed to have an effect on the type of G-phrases in the pairings. Table 6.4 shows the two-way frequency table between gesture type and contrast. Iconics and metaphorics showed similar percentages for their pairings with neutral (i.e., non-contrastive) and contrastive IS units, mostly favouring pairing with non-contrastive units. On the other hand, deictics were paired with contrastive units more than iconics and metaphorics were. Table 6.5 further breaks down the pairings for contrastive IS units depending on whether the IS units were topics or foci (i.e., contrastive topics or contrastive foci).

Table 6.5: A table showing the frequency of different gesture types paired with contrastive IS units

	Contrastive	
	Focus	Topic
Deictic	56.8%	84.4%
Iconic	31.8%	9.1%
Metaphoric	11.5%	6.5%
n	148	77

57% of contrastive foci were paired with deictics, followed by iconics with 32%. For contrastive topics, this percentage was 84% for deictics and only 9% for iconics. These showed that deictics were overwhelmingly used to exhaust/differentiate the alternatives for elements that existed in the previous discourse. Similarly, the pairing pattern in Table 6.2 also showed that deictics tended to be paired with topics which relate an utterance to the previous discourse. These two findings indicated that deictic gestures were preferred when establishing relations with the previous discourse. On the other hand, the function of carrying a discourse forward was reserved for gestures that bear higher levels of iconicity, i.e., iconics and metaphorics. The high level of representativeness of their speech referents allowed them to capture the newsworthy content in foci more comprehensively while also highlighting this content by copying it (or supplementing/complementing, see Section 2.2) into a visual modality. In brief, deictics were preferred for establishing backwards relations in a discourse, whereas iconics and metaphorics were preferred for establishing forward (or progressive) relations. These findings offered support for the claim that gesture is informed by IS since the selection of gesture type for production showed sensitivity to information structure. This implies an effect of IS on gesture production, and unified models of speech and gesture production should account for this effect. This is further discussed in Sections 7.1.3 and 7.2.1.

6.3 Testing of Synchronisation

The next step in the analysis consisted of testing whether the calculated time distances between the onsets and offsets of G-phrases and IS units were affected by the features of phrases involved, i.e., gesture type, IS units type, and contrast. As in Chapters 4 and 5, such contextual effects on synchronisation were tested via mixed-effects regression models. One consideration before fitting the relevant models involved backgrounds. As was the case in Section 5.2, there insufficient pairings with backgrounds for background to be included in the model (N=31). Backgrounds can also be excluded on the basis that G-phrases tended not to accompany backgrounded units which have no active informational structural functions (i.e., units that are not topical or focal, see Tables 6.1 and 6.2). Subsequently, two separate models for onsets and offsets were fitted. These included gesture type, IS unit type, and contrast as fixed effects (and their two-way interactions), and participant and scenario as random effects (intercepts and slopes for single-term fixed effects). The initial full models had the following structure:

$$time\,distance \sim (Gesture\,type{+}IS\;type{+}Contrast)\wedge 2{+}(Random\;effects)$$

Starting with the elimination of random effects, no significant effect of participant and scenario was observed, which was consistent with the results in Chapters 4 and 5. Consequently, the random effects were dropped from the models. Then, the elimination of fixed effects also did not reveal a significant influence of these effects on the time distances - the time distances between the onsets and offsets of G-phrases and IS units were not influenced by gesture type, IS unit type, and contrast.

The final step of the analysis was to test whether the paired G-phrase and IS units were synchronised given the synchronisation criterion of the study.

Since there were no significant effects on the time distances, the equivalence tests were used directly on the raw data. Figure 6.4 plots the output of the equivalence tests where the mean time distances and their CIs for the onsets and offsets have been positioned relative to zero and the equivalence bounds.

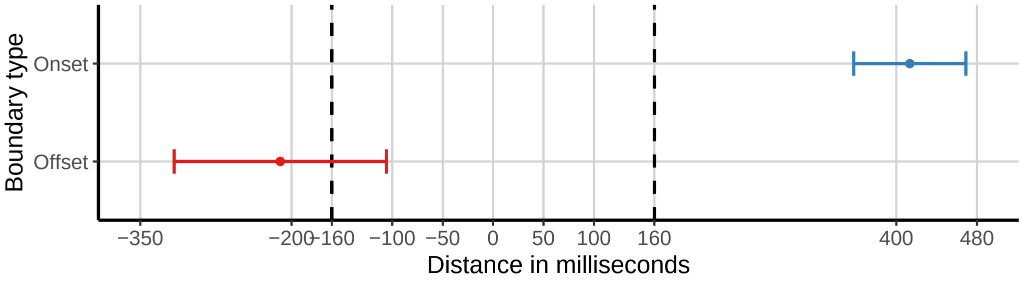
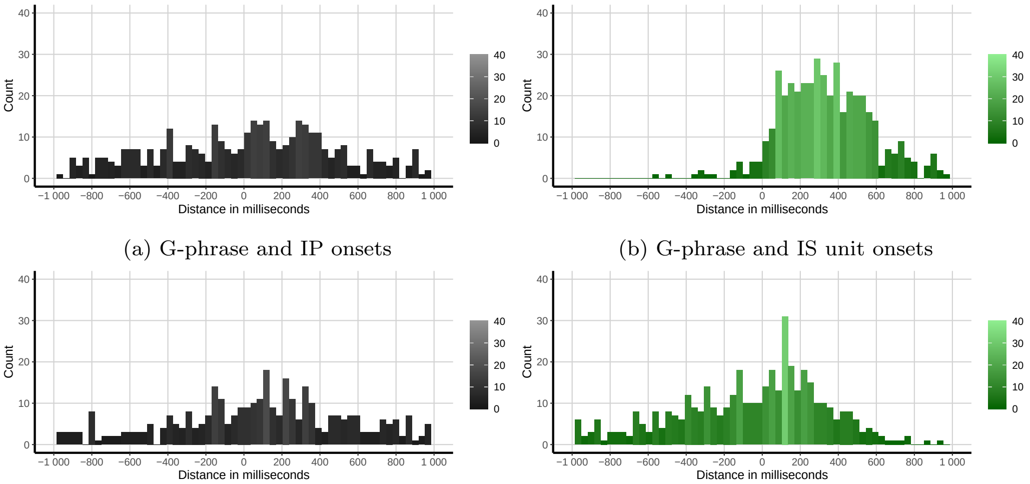


Figure 6.4: The estimated means of time distances between IS unit and G-phrase onsets and offsets with confidence intervals at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds (the average syllable duration is 160ms).

Based on the test, the CIs of both onsets and offsets did not fit within the set bounds. Therefore, the time distances between the onsets/offsets were statistically not equivalent to zero (onsets:$t(493) = 14.748, p = 1.000$, offsets:$t(493) = -1.600, p = 0.945$). G-phrases were not synchronised with IS units within the bounds of an average syllable duration. However, the mean distances of onsets and offsets suggested a different temporal relation for G-phrases and IS units compared to G-phrases and IPs. It is informative to compare these two results because they both resulted in asynchrony, but the distributions of time distances indicated different behaviours.

First of all, Figure 5.10 showed that G-phrases were contained within the span of IPs. However, in Figure 6.4, IS units were contained within the span of G-phrases - for onsets, the CI was in the positive area on the x-axis (i.e., gesture onsets occurred earlier), for offsets, it was in the negative area on the x-axis (i.e., gesture offsets occurred later).



(a) G-phrase and IP onsets

(b) G-phrase and IS unit onsets

(c) G-phrase and IP offsets

(d) G-phrase and IS unit offsets

Figure 6.5: A comparison of the distributions of time distances between G-phrases/IPs pairings (repeated from Figure 5.9) and G-phrase/IS unit pairings (repeated from Figure 6.3).

More importantly, the distributions presented in the histograms in Figure 6.3 were greatly different from those of IPs in Section 5.5. No clear peaks were observed in the distributions of time distances between G-phrases and IPs (see Figures 6.5a and 6.5c). However, there were clear peaks in the distributions of G-phrases and IS units (see Figures 6.5b and 6.5d). As stated previously, the present study considers the systematic co-occurrence of two units as synchronisation (Section 2.4). Therefore, the compact clustering of observations, although not centred on zero, indicated systematicity in the time distances between the onsets and offsets of G-phrases and IS units, hence synchronisation in the sense given in Section 1.1.

Taken together, there are two ways of interpreting these observations related to G-phrases and IS units. One would be that these delayed peaks were systematic, and G-phrases were synchronised with IS units with a delay that is about a median G-phase duration at the onsets and more than an average syllable duration at the offsets. The basis of this interpretation would be that two units need not start and end exactly at the same time to be considered synchronised. A systematic delay, as indicated by clustering at a certain distance, would also imply synchronisation. The other interpretation would be that the presence of peaks and the finding that IS units were contained within the duration of G-phrases might indicate that IS units were actually synchronised with a gesture phase (G-phase) or a combination of G-phases within the G-phrase structure. In this interpretation, the observed peaks would not be considered as systematic delays justifying the synchronisation of G-phrases with IS units. Instead, they would indicate that IS units were synchronised with G-phrase medial phases. The present study tests the latter interpretation in its analysis in the next section.

6.4 Apical Area - Information Structure Synchronisation

As discussed in Section 2.6.2, Ebert et al. (2011) tested the synchronisation of focus with strokes which are typically G-phrase medial (i.e., strokes usually follow preparations and precede post-holds or retractions). They argued that strokes would be ideal candidates for synchronisation with foci because they considered the stroke as the only phase with any semantic value. Therefore, preparation, hold, and retraction phases were excluded in their synchronisation tests. They reported a synchronisation of strokes and foci at the onsets, but not at the offsets. Based on the interpretation of the findings in the previous section, it might be the case that a similar synchronisation behaviour exists in Turkish as well. IS units were shown to be contained within G-phrases, and the clustering of observations meant that this containment was systematic. Therefore, it might indeed be the case that the synchronisation of IS units could be refined to a synchronisation with the semantically valuable gesture components which were G-phrase medial as opposed to a synchronisation with complete G-phrases.

The analysis in this section tests this possibility. However, unlike Ebert et al. (2011), the present study argues that post-holds also bear semantic value because they contain apex related information as a result of being the extensions of the gestural form at the end of the stroke. Consequently, they should be involved in the tests of synchronisation, for G-phrases in which there are post-holds. So, instead of testing synchronisation only between IS units and strokes, the present study tests the synchronisation of stroke+post-hold combinations with IS units. These combinations are referred to as *apical areas* in the rest of the present study.

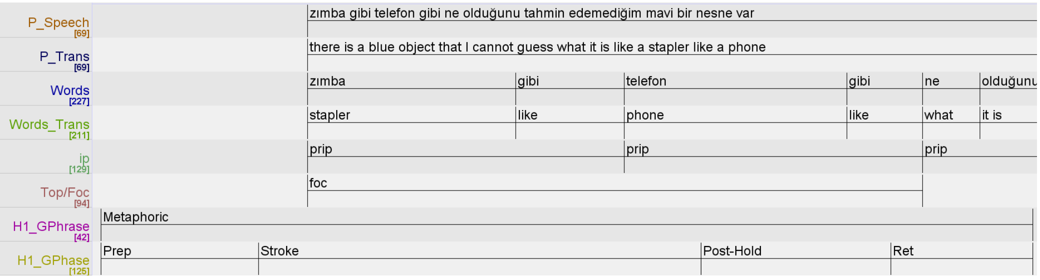
The inclusion of post-holds to synchronisation was semantically motivated. Not all phases in a G-phrase bear a semantic value. The organization of phases is organically linear where these phases are ordered to bring about the stroke and then end it. The preparation phase solely exists to enable the execution of the stroke, and the retraction exists for physical recovery after the execution of the stroke. Therefore, the gestural information that is sensitive to the organisation of information within utterances is not encoded in these parts. Consequently, it was possible that the synchronisation might not be sensitive to the full structure of a G-phrase. Rather, IS units might be synchronised with the elements that convey meaning within the G-phrase rather than with other parts of the G-phrase that enable these elements. Post-holds were considered to be one of the phases that convey meaning because they are not as semantically empty as preparation and retraction because of their parasitic relation to the stroke (cf. Ebert et al., 2011; Loehr, 2004). That is, a post-hold is fundamentally an apex suspended in time. The apex was defined as the kinetic goal (i.e., target) of the stroke that is usually stroke-final (Section 3.4.1.3), and in post-holds, the stroke-final gestural forms are retained (Section 3.4.1.2). Apexes are further relevant because they are also prominent units which have been shown to be synchronised with prosodically prominent units (Chapter 4). For all these reasons, the present study sees post-holds as meaning conveyers, distinguishing them from preparations and retractions for synchronisation purposes. Therefore, apical areas tested for synchronisation often consisted of stroke + post-hold combinations, but in the absence of a post-hold (they are optional, see Section 3.4.1.2), strokes themselves made up the apical areas.

Note that pre-holds were not included in apical areas. Pre-holds can also carry stroke related information as the hands are frozen in the stroke initial position. The semantic information in a pre-hold is only relevant for providing a starting point relative to which the action or the state in the stroke is described. The pre-hold does not contain apical information as it is not related to the target of the stroke. Moreover, the pre-hold functions to let the speech catch up with the gesture (see Section 3.4.1.2). Therefore, the synchronisation starting point would be expected to be at the end of the pre-hold since the pre-hold is there to enable the following speech and gesture to be synchronised. A final practical reason for the exclusion of pre-holds was their low number of occurrences in the data. Table 6.6 shows the percentages of each G-phase type over all G-phase annotations in the data.

Table 6.6: The percentages of G-phase types in the data

Preparation	Pre-Hold	Stroke	Post-Hold	Retraction	n
28.6%	2.2%	33.4%	23.6%	12.2%	1690

Interestingly, Karpiński et al. (2009) also noted similar percentages for the annotations of G-phases in their data where pre-holds were the rarest G-phase type. The low number of occurrences for pre-holds can also be interpreted to mean that they are not essential tools for synchronisation. Figure 6.6 shows an example of the organization of these G-phases in a G-phrase accompanying a focus in the data.



P_Speech [69] zımba gibi telefon gibi ne olduğunu tahmin edemediğim mavi bir nesne var

P_Trans [69] there is a blue object that I cannot guess what it is like a stapler like a phone

Words [227] zımba | gibi | telefon | gibi | ne | olduğunu

Words_Trans [211] stapler | like | phone | like | what | it is

ip [129] prip | prip | prip

Top/Foc [94] foc

H1_GPhrase [42] Metaphoric

H1_GPhase [125] Prep | Stroke | Post-Hold | Ret

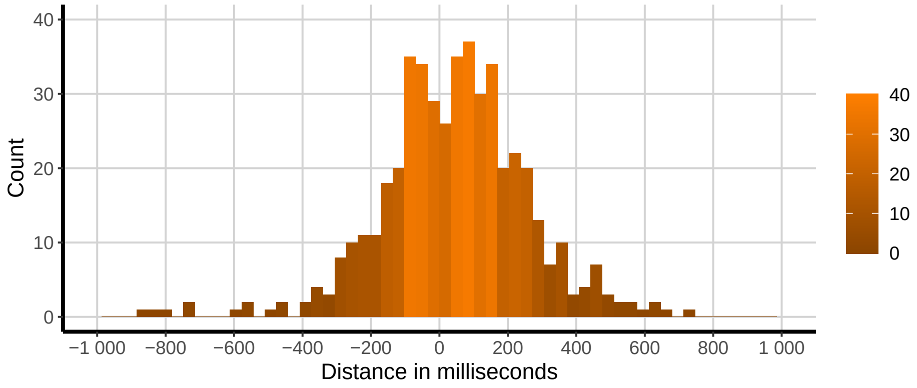
Figure 6.6: An ELAN screenshot showing the organisation of G-phases within a G-phrase paired with a focus

Figure 6.6 shows a typical example of how a G-phrase and its internal phase structure were temporally positioned in relation to an IS unit. As in the figure, a typical G-phrase consisted of a linear combination of preparation + stroke + post-hold + retraction. In line with the distributions of time distances between the onsets and offsets of paired G-phrases and IS units presented in the previous section, the onset of the G-phrase in the figure preceded the onset of the focus by about the duration of the preparation phase, and the offset of the G-phrase occurred after the offset of the focus by about the duration of the retraction phase. Based on such examples, the apical areas introduced in this section can be seen as suitable candidates for synchronisation with IS units, which also explains the consistent delays observed in the distributions for G-phrases and IS units in Figures 6.3a and 6.3b.

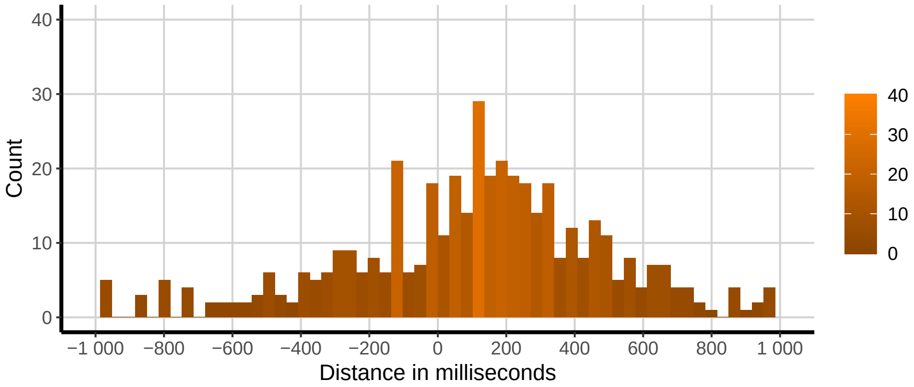
6.4.1 Pairing and Calculation of Time Distances

The semantic pairing process and the calculation of time distances between paired onsets/offsets were exactly the same as the analysis of G-phrase/IS units synchronisation. The onset of the stroke was the onset of the apical area, and its offset was the post-hold's offset if there was one within the G-phrase. If there was no post-hold, the stroke's offset was the offset of the apical area. For further comparisons, apical areas were also labelled depending on whether or not they contained post-holds. If an apical area did not have a post-hold, it was labelled as *stroke-only apical area*, and if it had a post-hold, it was labelled as a *full apical area*. This labelling will be relevant later in the chapter.

As the first step of the analysis, Figure 6.7 shows the calculated time distances between a) the onsets and b) the offsets of paired apical areas and IS units. The distribution for the onsets in Figure 6.7a shows the tightest distribution observed in the present study. There is a clear clustering of observations with a mean quite close to zero (M=43ms, SD=219ms). If we compare this to the onset distribution of full G-phrases in Figure 6.3a, we see that the peak is now centred on zero instead of being a G-phase duration away, and the distribution is more compact with the majority of the observations occurred within 12 frames (219ms) rather than roughly 21 frames (365ms) for the ip/G-phrase onsets. This distribution made a convincing case for the synchronisation of the onsets of apical areas (i.e., the onsets of strokes) with the onsets of IS units.



(a) Onsets



(b) Offsets

Figure 6.7: Two histograms showing the time distances between the onsets/offsets of IS units and apical areas. Negative values show that apical area onsets/offset occurred after IS unit onsets/offsets.

Figure 6.7b shows a tighter alignment of offsets as well. There was a clear peak nearly centred on zero with 47ms as the mean (sd=476ms). The figure does not show whether IS unit offsets were paired with the stroke offset or the post-hold offset. However, it does show that apical areas, as by defined the present study, offered a better anchor for the offsets of IS units than was provided by the offsets of G-phrases (see Figure 6.3b). Overall, the indicators of synchronisation were present for both onsets and offsets where clearer clusterings of observations centred on zero were observed. This implies that meaningful apical areas were found to be better gestural anchors for the synchronisation with IS units than complete G-phrases (including preparation and retraction).

The next step in the analysis would be to test whether there were patterns in the pairings of apical areas with IS units. However, this step was not necessary for apical areas. Apical areas are formed by meaning bearing phases (which determine gesture type) contained within G-phrases. Therefore, in terms of the features involved (e.g., gesture type and IS unit type), the pairings of apical areas with IS units were not different to the pairings of G-phrases with IS units. As a result, the pairing patterns observed between G-phrases and IS units also apply to the pairings of apical areas and IS units (see Tables 6.1, 6.2, 6.3, and6.4). One exception to this was about the G-phrase structure - whether apical areas included a post-hold or not (i.e., stroke-only and full apical areas). This labelling was relevant for the offset synchronisation because it was used to test whether it is offset of the stroke or of the post-hold that is better synchronised with the offset of IS units, in order to assess the inclusion of post-holds into the gestural anchor for synchronisation. Consequently, it was also investigated whether there were pairing patterns depending on whether apical areas contained a post-hold or not. However, no different patterns were observed.

6.4.2 Testing of Synchronisation

Because of the additional feature of offset type, it was necessary to test whether gestural and information structural contexts influenced the time distances between the onsets/offsets of apical areas and IS units.

One thing to note in relation to this test was that there was an additional feature related to the offset synchronisation which was the offset type - whether there was a post-hold or not in apical areas. In the data, 70% of apical areas (N=361) contained a post-hold and 30% (N=154) did not. This distinction could have an effect on the time distances. There were therefore two types of offset: (1) the offset of the stroke , (2)and the offset of the hold The synchronisation of these two types with the offsets of IS units were compared in the relevant regression models.

Two models were fitted as the onset and offset synchronisation were two different sets of measures. The full model for the onsets was the same as the model for G-phrases and IS units as the tested features were not different. For the offsets, the only addition was offset type to the fixed effects (gesture type, IS unit type, contrast, and their two-way interactions). Note that the background level was also excluded from the IS unit type factor due to the lack of observations (as in the previous section). The random effects in the models consisted of participant and scenario (intercepts and slopes for single-term fixed effects). The full model for the offsets had the following structure (the model for onsets did not have offset type):

$$time\ distance \sim (Gesture\ type + IS\ type + Contrast\ + Offset\ type) \wedge\ 2 \\ + (Random\ effects)$$

The process of eliminating random effects returned no significant effect of participant and scenario for both onsets and offsets once again. After dropping the random effects from the model, the elimination of fixed effects revealed no significant effect for the onsets. However, there was a significant effect of offset type on the time distances between offsets ($F(1) = 102.39, p < .001$), showing an average of 422ms difference between the time distances of the stroke offsets and the post-hold offsets to the offsets of IS units. Next, the estimates of time distances were extracted from the models, and the equivalence tests were applied in order to determine whether synchronisation was achieved for onsets and offset given this significant effect. The output of the equivalence tests are plotted in Figure 6.8, which shows the placement of the CIs for the onsets as well as for the two significantly different offset types. The figure also shows the equivalence bounds (± average syllable duration)

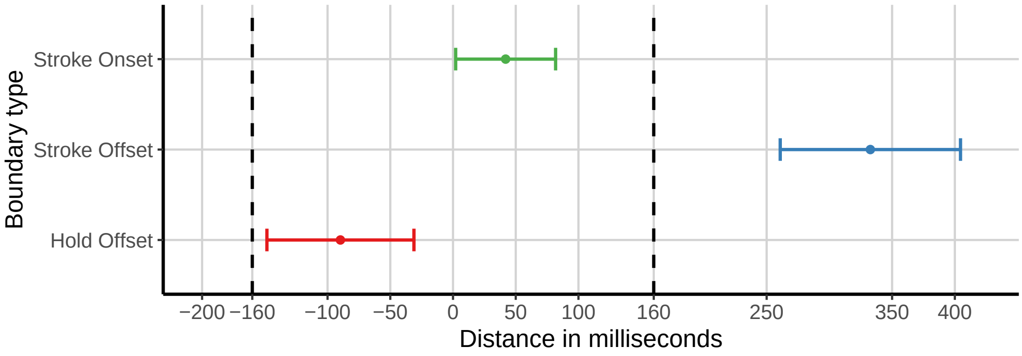


Figure 6.8: The estimated means of time distances between apical area and IS unit onsets/offsets with confidence intervals at 95%. The dashed lines are the upper (160ms) and lower (-160ms) equivalence bounds

The results of the equivalence tests were in line with the distributions. The CI of the onsets fitted within the equivalence bounds at ±160ms, meaning that the observed time distances between apical area onsets and IS unit onsets were statistically equivalent to zero ($t(493) = -7.532, p < .001$). For the offsets, the CI of the stroke offsets did not fit within the equivalence area ($t(493) = 4.956, p = 1.000$), and was about 332ms away from zero, showing that the IS unit offsets ended about a two syllable duration after the stroke offsets in general. On the other hand, the CI of the post-hold offsets fell within the equivalence bounds. The test showed that the time distances between IS unit offsets and apical area offsets were statistically equivalent to zero when the offsets of apical areas belonged to post-holds ($t(488) = 2.910, p < 0.01$). Overall, the onsets of apical areas were synchronised with the onsets of IS units. For the offsets, synchronisation was only observed for the offsets of full apical areas (stroke+post-hold) and not for stroke-only apical areas. These findings conclude that full apical areas were synchronised with IS units.

One open question concerns the nature of offset synchronisation for stroke-only apical areas (i.e., no post-hold in the G-phrase). Strokes could form apical areas by themselves since strokes are the source of apical information.

In line with this, the original prediction of the present study was that when there were no post-holds, strokes themselves would be synchronised with IS units. However, based on the results of the equivalence tests in Figure 6.8, strokes were not synchronised with IS units at the offsets, which was in line with the findings of Ebert et al. (2011).

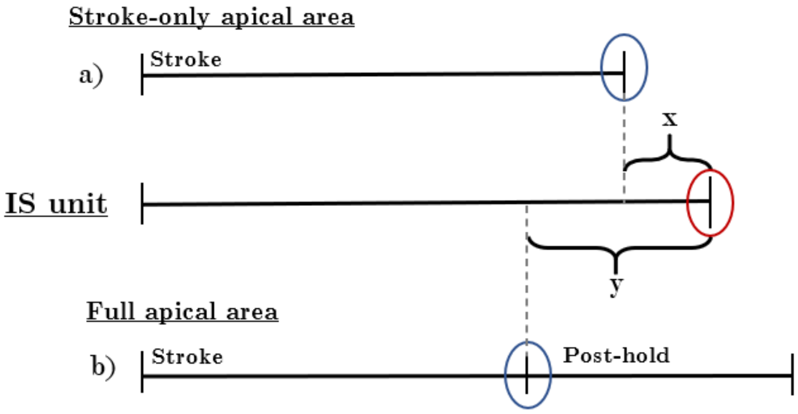


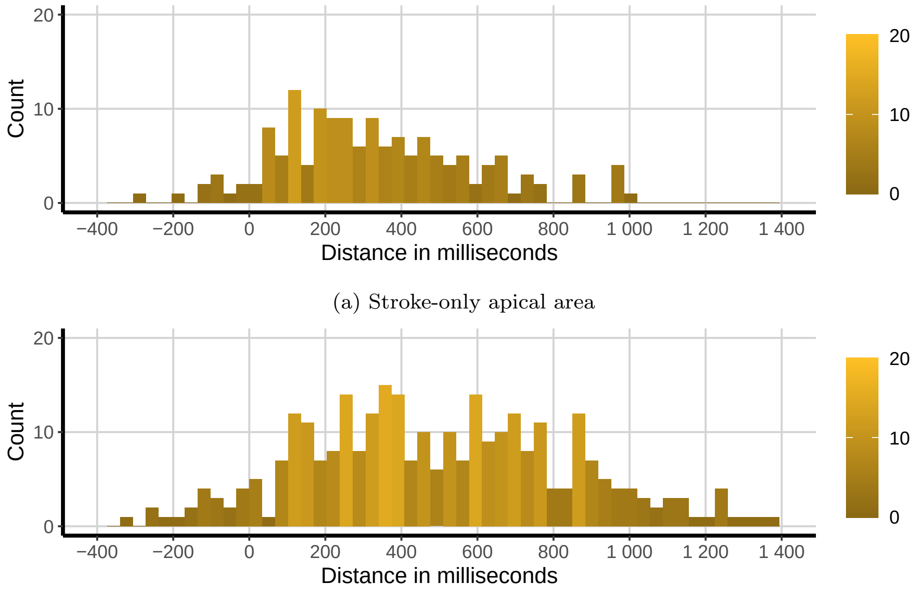
Figure 6.9: A stroke-only apical area and a full apical area where the offsets of the strokes were circled. In (a) the offset of the stroke is the offset of apical area. Since there is a post-hold in (b), the offset of the stroke is apical area medial, therefore not the offset of the apical area.

In the light of this original prediction, the present study investigated whether at least there was a synchronisation attempt at the offsets in the absence of post-holds. This could be done by comparing the time distances of the offsets of strokes to the offsets of IS units under two conditions as explained in the items (a) and (b) below:

a. In cases of stroke-only apical areas as in (a) in Figure 6.9 (i.e., no post-hold), the offset of the stroke was the offset of the apical area.

b. In cases of full apical areas as in (b) in Figure 6.9, the offset of stroke was apical area medial, and consequently, not the offset of the apical area.

Following the claim that apical areas are synchronised with IS units, a shorter time distance between the offset of the stroke in (a) and the offset of the paired IS unit was expected, compared to the offset of the stroke in (b) (i.e., $x < y$ as represented in the figure). This was because in (a), the offset of stroke is the offset of the apical area which, according to the claim, should be the target of synchronisation. On the other hand, for the offsets in (b), the time distance between the stroke and the paired IS units would be longer because there would be no synchronisation pressure on the offset of the stroke since the offset of the post-hold was already available as the offset of the apical area which would be the synchronisation target for the offset of the paired IS unit. Figure 6.10 plots the distribution of the time distances between the offsets of IS units and the offsets of strokes in stroke-only apical areas and full apical areas.



(a) Stroke-only apical area

(b) Full apical area

Figure 6.10: Two histograms showing the time distances between the offsets of IS units and the offsets of strokes when there was no post-hold in the apical area (a) and when there is one (b). Negative values show that the offsets of strokes occurred after the offsets of IS units.

It was found that the mean and standard deviation in the stroke-only condition (M=332ms, SD=258ms) were smaller than in the full apical area condition (M=527ms, SD=391ms). Based on these, there seemed to be a better clustering of observations in Figure 6.10a, whereas Figure 6.10b showed a very wide plateau of observations ranging from 100ms to 800ms. In order to see the difference between these clearly, the CIs of the mean time distances were calculated and plotted in a similar manner to the equivalence tests, showing their positioning relative to zero and to each other. Figure 6.11 shows the resulting grid.

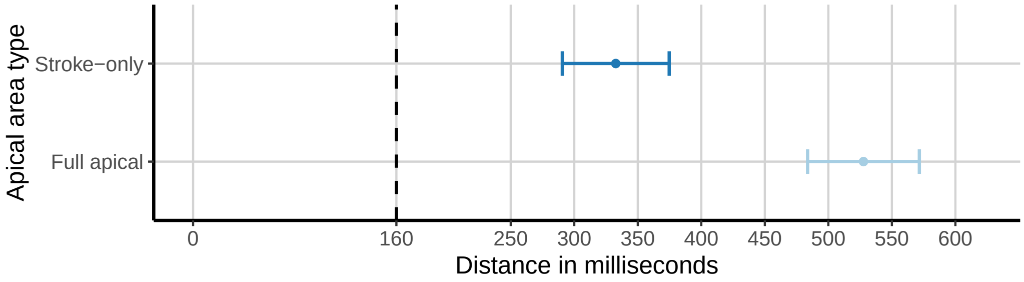


Figure 6.11: The confidence intervals (95%) of the time distances between the offsets of IS units and the offsets of strokes in stroke-only and full apical areas

There was no overlap between the CIs as there was about 100ms between their edges, and 200ms between their means. The figure shows different trends for the stroke-only and full apical area conditions. For stroke-only apical areas, the time distances between the offsets of strokes and the offsets of IS units was smaller than those of full apical areas. In other words, the offsets of strokes were closer to the offsets of IS units when there was no following post-hold. When there was a post-hold, the distances between the offsets of strokes and the offsets of IS units increased (i.e., $x < y$ in Figure 6.10 was found to be true). In the present study, this was interpreted as an attempt at synchronisation, supporting the claim that apical areas were the targets of synchronisation for IS units. For stroke-only apical areas, the offsets of IS units were closer to the offsets of strokes because strokes constituted apical areas by themselves. For full apical areas, the offsets of strokes were not the target of synchronisation because the offsets of apical areas were marked by the offsets of post-holds which was already found to in synchrony with the offset of IS units.

The explanation of why post-holds were not generated in order to ensure synchronisation within the average syllable duration in cases of stroke-only apical areas was beyond the scope of the present study. It is possible that such productions were results of conversation management strategies between interlocutors (e.g., turn-taking) or other kinds of production pressure, which were not investigated in the present study.

6.5 Summary

Earlier studies had suggested an association of gesture with IS, without providing a systematic test of their synchronisation. The findings presented in Chapter 5 suggested a synchronisation of G-phrases with foci. The analysis in this chapter showed that G-phrases indeed tended to accompany foci overwhelmingly. Further analyses of pairing patterns also revealed that gesture types were sensitive to the type of the IS unit they accompanied. Iconics and metaphorics tended to be paired with foci, and deictics were paired with topics and contrastive elements. This showed that the establishment of relations to the previous discourse was gesturally achieved through deictics, whereas the representational strength of iconics and metaphorics caused them to be preferred for the function of carrying the discourse forward. Overall, these findings suggested that there is a direct association of gesture and IS where the anchoring location of a gesture as well as its type can be predicted by the organisation of information within an utterance.

Tests of synchronisation showed that there was a delayed synchronisation of G-phrases with IS units. That is, IS units were contained within the duration of G-phrases but started and ended at fixed distances away from the G-phrase boundaries. This was interpreted to be an indicator of synchronisation with G-phrase medial phases. The present study introduced apical areas which were stroke+post-hold combinations that carry apical meaning within G-phrases, and the synchronisation of IS units was tested with these apical areas. The findings indicated that apical areas were synchronised with IS units. Figure 6.12 shows a summary of the general synchronisation behaviour observed in the present study by placing a G-phrase in its most likely temporal position in relation to an IS unit.

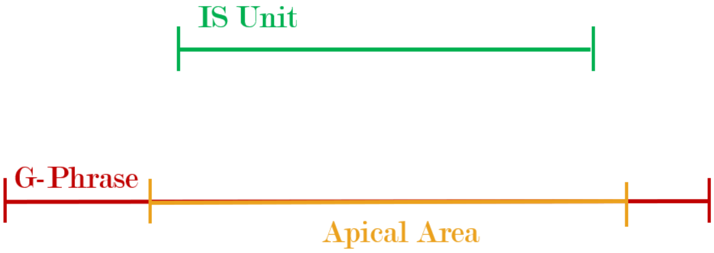


Figure 6.12: Stroke-only apical area and full apical area

In the present study, the boundaries of IS units were coextensive with the boundaries of prosodic phrases (Section 3.4.3) - IS units corresponded to a single or a grouping of prosodic phrases. Therefore, the results of synchronisation presented here mean that prosodic phrase groupings, as organised by IS, were synchronised with gesture. Consequently, the present study has shown evidence for the fact that the synchronisation of gesture with IS is mediated

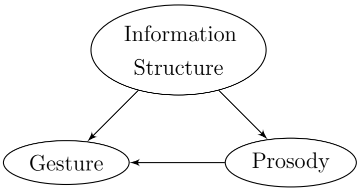


Figure 2.17: Three-way association of information structure, prosody, and gesture (repeated from page 111)

by prosody, which constitutes an indirect link between gesture and IS (see Figure 2.17). A direct link between gesture and IS was already established which was manifested through categorical sensitivity between gestural and information structural features (i.e., pairing patterns). Taken together, these findings support the present study's claim that gesture, prosody, and IS form a three-way ensemble in the synchronisation of speech and gesture. In the light of this, psycholinguistic models of speech and gesture production should account for the fact that gesture production is informed by the generation of prosody and information structure over the course of speech production. This is further discussed in Chapter 7.

7

Discussion

The present study has investigated the synchronisation of gesture with prosody and information structure in Turkish natural speech data making use of a statistically tested synchronisation criterion. Synchronisation was treated as a phenomenon that can be affected by gestural and prosodic and information structural contexts. Analyses were conducted at both the smallest gestural unit level and larger phrasal levels. At the smallest level, a series of analyses looked at which tones apexes tended to pair with (the nearest tone to an apex is referred to as a pairing), followed by analyses into whether these pairings were synchronised or not. Through similar analytical steps, consideration of phrasal levels looked at pairing patterns between G-phrases and ips/IPs, and whether these pairings were synchronised given the same synchronisation criterion. Finally, the present study investigated whether a synchronisation relationship existed between information structure and gesture at the phrasal level. It tested whether G-phrases and apical areas were synchronised with topics, foci or backgrounds. Section 7.1 briefly recaps the results of these analyses and discusses their significance and implications in relation to previous studies. In Section 7.2, these results are interpreted within the frame of psycholinguistic speech-gesture production models, and some modifications of existing models are proposed. Finally, Section 7.3 concludes the present study and discusses possible avenues for future research.

7.1 Summary of Results and Implications

This section is organised around three questions which the analyses in Chapters 4, 5, and 6 aimed to answer:

1. Are apexes synchronised with their nearest tone?
2. Are gesture phrases synchronised with prosodic phrases?
3. Are gesture phrases synchronised with information structure units?

In each of the following subsections, first the relevant results are summarised, followed by a discussion of their significance and implications.

7.1.1 Are apexes synchronised with their nearest tone?

The analysis of synchronisation between gesture and prosody first focused on the micro level. Chapter 4 tested the synchronisation between apexes and tones as atomic units. There were two steps in the analysis. In the first step, it was investigated whether there were patterns in the pairings of apexes with tones prior to testing for synchronisation. The second step consisted of the statistical testing of actual synchronisation. Equivalent steps were used in the analyses in the other two results chapters as well.

The pairing analysis in Chapter 4 showed that apexes tended to be nearest to prominent pitch accents amongst other possible tonal events (i.e., phrase tones and boundary tones). If pitch accents were not available in the prosodic phrase (i.e., for accentless phrases), then apexes were paired with Ls which are the only other tonal event at the level of the PW. Pairings with phrase accents and boundary tones which are associated with prosodic phrases above the level of the PW were systematically avoided. This pattern was found to be consistent across different prosodic contexts. Based on these observations, pitch accents and Ls were identified as the preferred anchors of apexes for pairing, and phrase accents and boundary tones as dispreferred anchors.

In the second step of the analysis (Section 4.2), it was shown that synchronisation patterns were in line with these pairing patterns. That is, if an apex was paired with a preferred tonal event, then these were synchronised given the synchronisation criterion. In other words, apexes tended to occur within an average syllable duration of pitch accents and Ls. In cases where apexes were nearest to dispreferred anchors, apex-tone synchronisation was not achieved, meaning that apexes tended to occur farther than an average syllable duration away from these tone types.

Including the possible effects of gestural and prosodic context in the analysis revealed exceptions to the general synchronisation pattern. The first exception was observed in the apexes of beats and metaphoric gestures, where rhythmic apical productions imposed synchronisation of these apexes with tones regardless of tone type. Beats often occurred in clusters where consecutive apexes were a fixed distance from each other as dictated by rhythm. In addition, metaphorics produced for paranarrative statements expressing uncertainty or repetition often presented similar rhythmic kinematic features. These rhythmic apical productions caused apexes to be synchronised with the nearest tone to their location regardless of the type of the tone.

The other exception was related to the nuclearity of ips in which the tones were found. If apexes were paired with a tone in nuclear ips, these apex-tone pairings tended to be synchronised regardless of the tone type. Note that the general pairing pattern was still observed here - a great majority of apexes were paired/synchronised with pitch accents. However, under the effect of maximum prosodic prominence, apexes and tones were more tightly coupled in time.

7.1.1.1 Implications

As reviewed in Section 2.5.1, the synchronisation of gestural and prosodic anchors at the smallest unit level has been investigated in a small number of languages using natural speech and experimental data. The studies using natural speech data have usually made qualitative and/or inconclusive observations about synchronisation without defining what exactly constituted synchronisation in their analyses. On the other hand, the studies using experimental data have also failed to establish a stable synchronisation criterion for deciding how near two anchors should be in order to be considered synchronised. These studies have also lacked ecological validity due to altering the spontaneous and unrestricted nature of co-speech gesture through heavy experimental manipulations. Despite their methodological differences and shortcomings, however, both groups of studies have reported a prominence-based synchronisation between gesture and prosody where dynamically prominent apexes were claimed to be synchronised with certain acoustic events that were considered to bear prominence.

The methodology employed within the present study has aimed to address such shortcomings, especially the ones about synchronisation decisions. Using a clear synchronisation criterion that is statistically testable gives a frame of reference for the judgements of synchronisation. Equivalence tests create a zone of tolerance for synchronisation so that minor changes in the time differences between units under the effect of various factors (e.g., gestural, prosodic, information structural) do not register as asynchronisation automatically. The present study has used the average syllable duration as its synchronisation criterion. Syllable duration is phonologically meaningful as the syllable is the smallest phonological unit that carries prominence, and prominence has been shown to be a key factor that synchronisation is built on in the present study. The findings also have shown that this duration is tolerant to spurious effects which might have led to a finding of asynchronisation otherwise while also being strict enough to reveal fine synchronisation patterns at every level of the synchronisation analyses.

The present study interrogated the prominence-based synchronisation theory where apexes were claimed to be synchronised with certain acoustic measurements that were considered to bear prominence. The selection of prosodic anchors in earlier studies have been independent from the phonological theory of prominence, which casts doubts on the accuracy of this claim. In order to address this, instead of selecting a seemingly random acoustic measure for synchronisation tests, the present study used tonal events (i.e., F0 turning points) which can be seen as the most likely correlates of prominence from the point of view of well-established, empirically-tested prosodic theory. Tonal events function to mark prominence as well as to mark boundaries of prosodic phrases. Within the present study, the analysis of synchronisation considered any tonal event, regardless of function, as a potential anchor for an apex. This meant that any consistent observation of synchronisation would be the result of a genuine systematic behaviour.

The results of the present study were in line with the prominence-based synchronisation claims. It was shown that gesturally prominent apexes were chiefly coordinated with prosodically prominent pitch accents, indicating that prominence was a constraint for synchronisation, as in earlier studies. However, further analyses revealed another synchronisation pattern where apexes were synchronised with boundary marking Ls in the absence of pitch accents. To the author's knowledge, the possibility that apexes may also be synchronised with boundary marking tonal events has only been mentioned recently in Rohrer et al. (2019) (see Section 2.5.1.1). The results of the present study are the first to offer evidence that apexes can also be synchronised with boundary marking tonal events.

This observation in the present study was also important because it showed for the first time that synchronisation at the micro level was managed by the prosodic hierarchy in addition to prominence. Pitch accents and Ls are both associated with PWs in Turkish. When the highest priority target for synchronisation (i.e., a pitch accent) was not present, apexes stayed anchored at the PW level by synchronising with Ls, and this result was found to be consistent across various prosodic contexts. This systematic pattern indicated that the synchronisation of apexes was not arbitrary even in the absence of prominence which had been seen as the sole guiding principle behind synchronisation in previous studies. In fact, apex synchronisation displayed sensitivity to the prosodic hierarchy by not synchronising with the markers of other prosodic constituents (i.e., phrase accents and boundary tones) that are higher in the prosodic hierarchy (also higher in pitch) (see Figure 7.1). This result is interpreted to mean that both prosodic prominence and hierarchy are active agents that play a role in the anchoring of gesture to speech.

The observed synchronisation pattern also has implications for phonological theories about word stress in Turkish. Section 2.5.1.2 reported a disagreement in the literature about whether words with final stress bear a pitch accent (Kamali, 2011; Ipek & Jun, 2013; Güneş, 2015). It was noted that particularly in pre-nuclear ips, the rise in pitch that marks the end of ips (H-) coincides with the pitch accent (H*). Following this observation, some authors have argued that the final pitch rise in this context is just the ip-final H-phrase accent and there is no pitch accent in the phrase (Kamali, 2011; Güneş, 2015), while others have argued that it functions as an independent H* coinciding with the H- (Ipek & Jun, 2013).

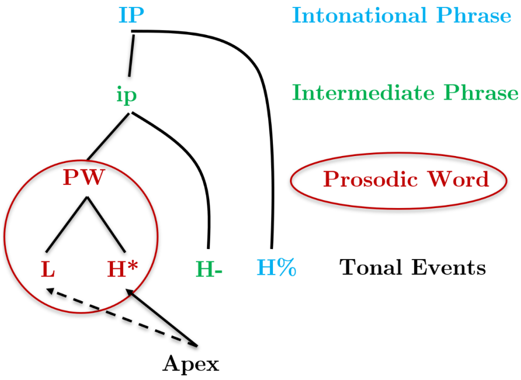
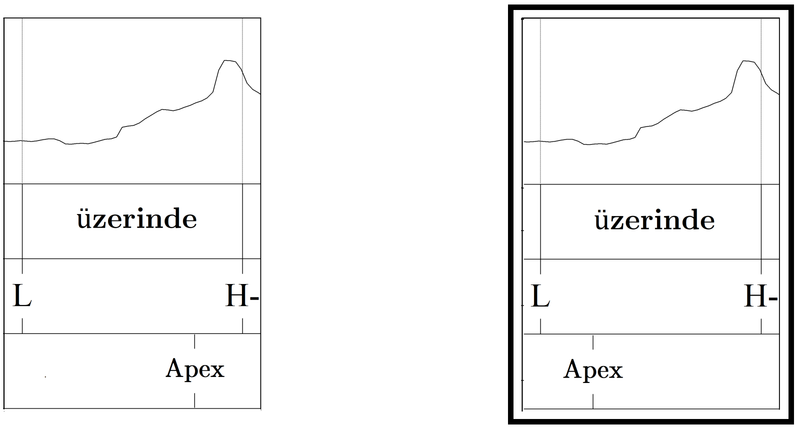


Figure 7.1: A schematic that shows the domain of apex synchronisation as the prosodic word. Apexes tend to be synchronised with pitch accents. In cases where there are no pitch accents, the synchronisation tends to be with prosodic word initial low tones (indicated by the dashed arrow).



(a) H- phrase accent

(b) L

Figure 4.11: Apexes were typically synchronised with Ls in accentless pre-nuclear ips as in (b), not with H- phrase accents as in (a). (repeated from page 206)

In line with the finding that apexes are synchronised with pitch accents, the present study argued that looking at apex synchronisation can shed light on the accenting of these words. If apexes were synchronised with phrase-final rises (H- in Figure 7.2a), this would suggest that these rises function as pitch accents as well as phrase accents since pitch accents attract apexes as per the prominence constraint. However, this was not found to be the case. Instead, apexes were consistently synchronised with Ls (Figure 7.2b) obeying the prosodic constituency constraint as explained above. This synchronisation behaviour indicated an absence of prominence in these phrases, which implied that the H tones under consideration do not function as a part of a pitch accents but rather only mark the end of the ip as a part of the phrase accent supporting Kamali (2011) and Güneş (2015).

Note that the present study did not aim to explain why some pre-nuclear ip final words had pitch accents on their final syllable and some did not. Accentless realisations of these words can be a result of a variety of factors, which is out of the scope of this study to analyse. The present study only showed that apexes were not synchronised with the pitch accent argued by some to be a constituent part - together with the phrase accent - of the tonal events at the end of the ip. Rather, apex synchronisation behaved as if there was no pitch accent in these words, offering multimodal evidence in support of the claim that words with final stress are not accented.

The consideration of all tonal events in Turkish as potential synchronisation anchors for apexes was the main factor enabling these findings and interpretations. This highlights the importance of a linguistically informed selection of anchors for synchronisation tests. Using measurements such as jaw displacement, vowel onsets, and acoustic peaks may provide methodological convenience (especially for experimental studies); however, the integration of prosodic structure as an inter-connected system into the analyses of synchronisation can add valuable insight into our understanding of how speech and gesture are temporally coordinated. Furthermore, the prosodic structures of languages exhibit different features cross-linguistically (e.g., lack of pitch accents in phrases in Turkish). As shown by the present study, such different prosodic features can lead to different synchronisation patterns, which can even be used to supplement phonological analyses taking advantage of the consistency of synchronisation behaviour. Overall, the present study emphasises that if our aim is to study the synchronisation between gesture and prosody, the methodologies we adopt should be grounded in prosodic research.

One main methodological approach that separated the present study from earlier studies was that it tested the effects of gestural, prosodic and information structural contexts on synchronisation. This approach was based on the understanding that apexes and tonal events exist as members of complex systems where events, constituents, and their positioning influence each other. In general, even when such possible effects were accounted for, the synchronisation patterns observed in the present study were found to be largely consistent. However, the semantic function of some gestures and the nuclearity of ips caused exceptional synchronisation behaviours between apexes and tonal events. Rhythmicity in beat and metaphorical gestures resulted in multiple consecutive apexes occurring fixed distances away from each other. It was prosodically unlikely that there would be a pitch accent at every consecutive apex location, especially when the distances between apexes were so short. As a result, such rhythmic apical productions resulted in the synchronisation of apexes with the nearest tones regardless of type within the present study. On the whole, it can be said that rhythmicity in gesture caused by the pragmatic needs of interlocutors can override the constraints of synchronisation. After all, gesture exists primarily to address the communicative needs of interlocutors, and synchronisation is only a means to reach this end.

Another exception to the general synchronisation behaviour was observed within nuclear ips. As predicted, apexes occurring during these phrases tended to be paired/synchronised with pitch accents overwhelmingly more than other tones. However, even when they were paired with dispreferred tonal targets, the synchronisation was achieved. The present study interpreted this finding as evidence that the synchronisation at the micro level is also sensitive to phrasal prominence. Under the effect of maximum prosodic prominence encoded in nuclear ips, apexes were more tightly coupled with tones. Note that phrasal prominence also played a similar role in the synchronisation of the boundaries of ips and G-phrases where nuclear ip boundaries were shown to be more tightly synchronised with G-phrase boundaries compared to pre-nuclear and post-nuclear ips (see Section 7.1.2).

To summarise, the results of analyses of synchronisation confirmed that prominences in gesture and prosody are synchronised. However, linguistically informed analyses of a statistically defined synchronisation revealed the prosodic hierarchy as another constraint on synchronisation in addition to prominence. This showed that synchronisation of atomic anchors in gesture and prosody was informed by the prosodic structure of the language investigated.

7.1.2 Are gesture phrases synchronised with prosodic phrases?

The next step in the investigation of gesture-prosody synchronisation was concerned with phrases (see Chapter 5). The present study aimed to find the best prosodic phrase anchor for G-phrase synchronisation. For this purpose, the synchronisation of G-phrases with ips and IPs was tested using the same method of analysis of synchronisation as for apex synchronisation.

Starting with pairings of G-phrases with ips, in Section 5.2, it was observed that one-to-one pairing of single G-phrases with single ips was not possible because G-phrases often contained multiple ips. This durational difference between the phrases led to the finding that a single G-phrase was not synchronised with a single ip in Turkish. Consequently, the analysis instead paired the boundaries (i.e., onsets and offsets) of these phrases based on proximity rather than semantic relation (see Section 5.1). That is, a G-phrase boundary was paired with the nearest ip boundary, meaning that the left and right edge boundaries of one G-phrase could be paired with the boundaries of different ips. As a result of this pairing process, two clear pairing patterns were observed. In the first, it was seen that G-phrases tended to span over ip combinations containing pre-nuclear and nuclear ips. These pre-nuclear and nuclear ip combinations correspond to the pre-verbal area in Turkish, which make up the default focus position. The second observed pattern was highly related to the first one. This pattern revealed that most pairings of ip and G-phrase boundaries took place within focus areas as defined in Section 3.4.3.

The results of synchronisation tests showed that the boundaries of ips and G-phrase were synchronised. G-phrases were temporally sensitive to ip boundaries - whenever there was a G-phrase starting/ending, there was an ip starting/ending at the same time.

The nuclearity of ips significantly affected the time difference between ip and G-phrase boundaries. When the boundaries of nuclear ips were synchronised with G-phrase boundaries, the synchronisation observed was near perfect with almost no delay between the corresponding boundaries. This was seen as another example for the effect of phrasal prominence on synchronisation where under the effect of maximum prosodic prominence, the pairings were synchronised more tightly. However, the boundaries of pre-nuclear and post-nuclear ips still showed synchronisation given the synchronisation criterion employed in the present study.

For the synchronisation of IPs and G-phrases, one-to-one pairing of these phrases was possible due to these having closer average durations. However, no specific pairing pattern was observed. Further, no evidence was found for the synchronisation of G-phrases with IPs. The most relevant result of this synchronisation analysis was that G-phrases tended to be contained within the duration of their corresponding IP.

7.1.2.1 Implications

Compared to the micro level, there have been fewer studies investigating the synchronisation of gesture and prosody at the phrasal level (see Section 2.5.2). These previous studies have tested the synchronisation between different combinations of gestural phrases (i.e., G-phrase and G-unit) and prosodic phrases (ips and IPs). They have shown similar methodological shortcomings as the studies at the micro level where a definition of synchronisation, and the effects of gestural and linguistic contexts were often neglected. Moreover, their results remained mostly observational, which was especially true for the studies that investigated the synchronisation of G-units with either IPs or some form of groupings of IPs.

The present study has highlighted the fact that gesture and prosody share structural similarity where both are hierarchically structured with nested phrasal constituents that are linearly organized around a prominent constituent. The study has argued that if we want to make a general claim that gesture and prosody are synchronised, then the analysis of synchronisation should also be extended to phrases in these modalities. Accordingly, the present study selected G-phrases and systematically tested whether they were synchronised with prosodic phrases in Turkish.

The results indicated that single G-phrases were not synchronised with single ips or IPs. A single ip was not found to be a valid target for synchronisation for a single G-phrase because of the durational mismatch between ips and G-phrases. In Turkish, single PWs often form ips by themselves, which means that these ips are too short (they consist of a single orthographic word) to match G-phrases. Instead, the most common pattern in this study was that one G-phrase spanned over multiple ips. Note that such a pattern where a phrase in one modality (gesture or prosody) contains multiple phrases in the other modality was also a common observation in most of the studies covered in Section 2.5.2. Figure 7.3 shows a comparison of spanning/overlapping phenomena in Turkish with what was observed in earlier studies.

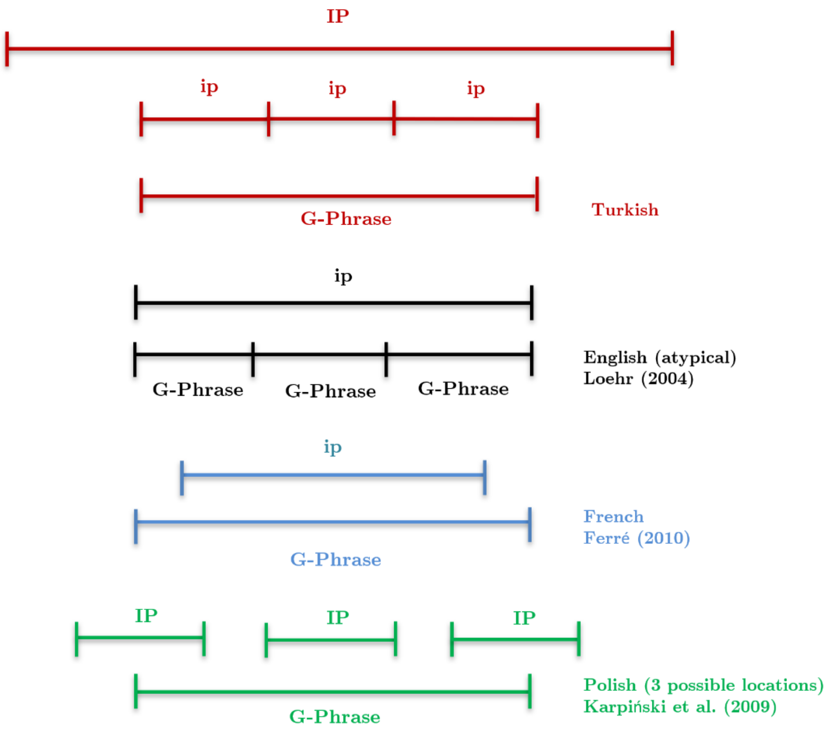


Figure 7.3: The comparison of G-phrase synchronisation with ips and IPs in Turkish to those in English, French, and Polish (extended from Figure 2.14)

None of the earlier studies in Figure 7.3 reported a clear synchronisation, except for Loehr (2004) who added that the phrasal synchronisation was not as strong as the synchronisation of apexes with pitch accents. However, the present study revealed synchronisation of phrase boundaries. G-phrases started at the same time as an ip and also ended at the same as an ip. It was only the case that these boundaries might belong to different ips for the same G-phrase due to the durational discrepancy between phrases.

At a first glance, the findings that G-phrases are sensitive to ip boundaries, and that they span over multiple ips might be taken to be a consequence of a synchronisation with IPs instead. IP boundaries are also ip boundaries (of the first and last ip in the IP), and IPs usually contain multiple ips because IPs are higher in the prosodic hierarchy. However, a synchronisation of G-phrases with IPs would go against the predictions of the present study which predicted a synchronisation of G-phrases with ips given their positions within their respective hierarchies (see Figure 7.4). In agreement with this, no evidence was found for an IP/G-phrase synchronisation, meaning that durational mismatch did not cause a shift from the ip to the IP for G-phrase synchronisation. This finding showed that the synchronisation of boundaries of ips and G-phrases was genuine and not a by-product of an IP/G-phrase synchronisation. Despite the durational mismatch, a temporal sensitivity was expressed through the

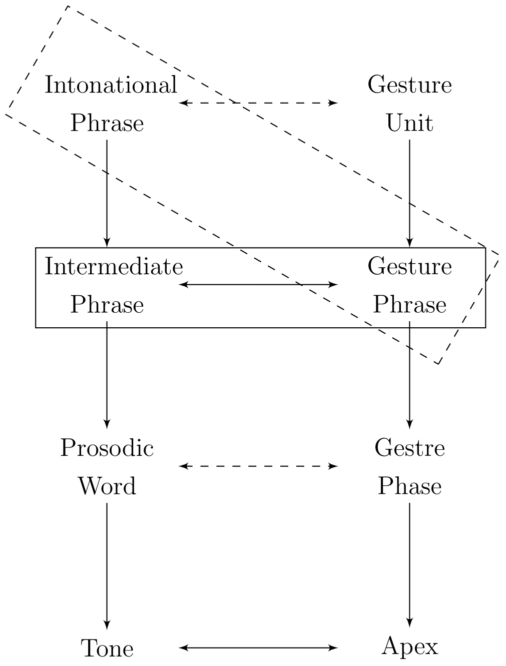


Figure 7.4: Pairing of structural hierarchies (modified from Figure 2.13)

synchronisation of onsets and offsets for ip/G-phrases. The present study interprets these findings as further evidence that gesture synchronisation is informed by the prosodic hierarchy. Notwithstanding the durationally better suitability of IPs, G-phrase synchronisation did not shift up to the IP level but stayed anchored at the level of ip.

The analysis of linguistic context once again uncovered key information about synchronisation. In Section 7.1.1, it was discussed that apexes were synchronised with tones more tightly under the effect of nuclear prominence. The same effect was also observed for the synchronisation of phrase onsets/offsets. Both the onsets and offsets of nuclear ips were almost perfectly synchronised with those of G-phrases. This meant that prosodic prominence is a factor that influences synchronisation behaviour at the macro level as well.

Taken together, these findings establish that gesture and prosody showed synchronisation at both micro and macro levels. At both levels, synchronisation was seen to obey the same constraints, i.e., prosodic prominence and prosodic hierarchy. These findings have implications on the speech-gesture production models, which are discussed in Section 7.2.

The analysis of linguistic context also revealed findings that strengthened the present study's proposal that G-phrases might also be synchronised with IS units, especially focus. G-phrase boundaries were sensitive to ip boundaries, but the whole G-phrase as an interval was not found to be synchronised with one whole prosodic phrase. In terms of basic durational comparison, ips were too short and IPs were too long to be able to be consistently synchronised with G-phrases. This implied that an ideal candidate for G-phrase synchronisation would exist between the ip and the IP levels. In line with this, the findings showed that G-phrases were observed to span over multiple consecutive ips. Thanks to the integration of the analysis of linguistic context, the present study was able to identify what kind of an organisation these ips was a part of. It was found that these multiple ips tended to be made up of pre-nuclear and nuclear ip combinations, which correspond to default focus position in Turkish. The association of G-phrases with foci was also supported by the finding that most ip/G-phrase pairings took place within focal areas. All of these findings are interpreted to be evidence for gesture's synchronisation with prosody which also mediates the synchronisation of gesture and information structure. The proposal about the three-way interplay of gesture, prosody, and information structure in the synchronisation of speech and gesture is further discussed in Section 7.1.3.

7.1.2.2 Implications for synchronisation

The present study has found two constraints on synchronisation: prosodic hierarchy and prominence. The present study assumes a rank-ordering of these constraints in terms of their application sequence in which the prosodic hierarchy constraint is applied before the prominence constraint. At the first glance, the findings of synchronisation at the micro level might be taken to mean that the prominence constraint applies first (i.e., synchronisation with pitch accents), and only when this constraint is not applicable prosodic constituency functions to govern the anchoring of gesture (i.e., synchronisation with L). However, the findings at the macro level suggest otherwise. Through their boundaries, G-phrases can be synchronised with any ip regardless of whether or not the ip is prominent in its utterance. The prominence constraint only sets a preference in which post-nuclear ips lacking in prominence are dispreferred for synchronisation. This implies that for the anchoring of gesture, prominence operates within the domain set by the prosody hierarchy.

The micro level synchronisation is interpreted to work in the same order in the present study. First, the prosodic hierarchy sets the PW as the synchronisation domain, and only then the prominence constraint sets pitch accents as the apex anchor. In cases when the prominence constraint cannot function, the prosodic hierarchy constraint itself appoints Ls as the anchor to preserve the domain of synchronisation. Moreover, the effect of nuclear prosodic prominence is constant at both levels where under this effect anchors were closer to each other in time in nuclear ips.

These two constraints and their order of application have a lot in common with the speech production flow taking place in the formulator in Levelt's (1989) speech production model which was integrated into gesture production models reviewed in Section 2.3. This is further discussed in Section 7.2.1.

7.1.3 Are gesture phrases synchronised with information structure units?

The present study hypothesised that gesture may also be directly informed by IS, and gestural units may be synchronised with IS categories. As shown and summarised in Chapter 5 and Section 7.1.2, a possible synchronisation of G-phrases and IS units was implied by the finding that G phrases span over multiple ips which tended to be in the focal area. Accordingly, in order to test this, the study investigated whether G-phrases were synchronised with topics, foci and or backgrounds. The method of analysis was consistent with the analyses of gesture-prosody synchronisation.

First of all, G-phrases were paired with IS units carrying the same semantic content. There were specific patterns observed in these pairings (Chapter 6). As predicted, G-phrases tended to pair with foci more than with topics. They only paired with backgrounds in a few cases. This pairing pattern could be further broken down according to gesture type. Topics and contrasted elements were more likely to be paired with deictics. On the other hand, foci were more likely to be paired with iconics and metaphorics.

However, these pairings of G-phrases and IS units were not found to influence actual synchronisation. G-phrases and IS units displayed a general synchronisation behaviour that was different to previous observations of synchronisation between gesture and prosody. The boundaries of these units did not occur at the same time as one another, instead there was a systematic delay between their boundaries. This systematic delay showed that IS units were contained within G-phrases. The systematic delays at the onsets and offsets as well as the containment of IS units within G-phrases implied that IS units might be synchronising medial G-phase(s) within the G-phrase.

In light of earlier findings showing that apexes are relevant units for synchronisation, the present study postulated that IS units might be synchronising with G-phrase medial stroke + post-hold combinations that contain apex-related semantic content, i.e., apical areas. It was indeed found that IS units were in strong synchrony with apical areas.

In order to confirm that G-phrases were synchronised with apical areas and not with strokes, the study further compared whether G-phrase offsets were synchronised better with stroke offsets or apical area offsets. It was indeed the case that post-hold offsets were preferred over stroke offsets for synchronisation. Moreover, when strokes were not followed by a post-hold, their offsets were closer to IS unit offsets compared to when they were followed by a post-hold. This supported the claim that apical areas are the actual targets of the synchronisation and not strokes.

7.1.3.1 Implications

Section 2.6.2 reviewed studies that have shown a link between IS and gesture. A group of these studies has reported a frequency-based co-occurrence of focus/contrast and gesture (usually non-manual gestures). Another group has established that the apexes of non-manual gestures were synchronised with prosodic prominences within foci. Others have associated notions related to IS, such as common ground and information status, with gesture, and have claimed that gesture frequency and form are affected by the retrievability from the discourse. Although different forms of association between gesture and IS have been proposed, earlier studies have not considered a temporal coordination between them in the way that they have done for gesture and prosody. Especially at the phrasal level, the possibility that entire topical or focal areas might be synchronised with gestural phrases has not been investigated systematically, except for Ebert et al. (2011) which was limited in its analysis because it only investigated the synchronisation of focus and preparation+stroke combinations. The present study has investigated a potential synchronisation of IS and gesture by testing whether G-phrases are synchronised with topics, foci, and/or backgrounds while systematically testing for whether these IS categories, gesture type, and contrast have an effect on synchronisation.

The results have indicated certain pairing patterns between G-phrases and IS units. As predicted, more than two thirds of G-phrases tended to pair with foci. This result can be seen as empirical evidence for McNeill's (1992) claim that gestures are novel departures of thought from the presupposed background. Gestures indeed accompany foci that update the presupposed background between interlocutors. The present study attributes this categorical sensitivity between gesture and foci to discursive prominence. Foci are core elements in the building of a discourse because they contain the semantic content that carries a discourse forward. Therefore, they can be considered discursively more prominent than topics and backgrounds in terms of reaching a communicative goal. This interpretation was also supported by the observation that virtually no G-phrases were paired with backgrounds. Backgrounds are elements that can be assumed from the discourse, which do not serve an active IS (topical or focal) function. As a result, they do not bear any discursive prominence, and gestures are not attracted to these elements. This type of behaviour is similar to what has been observed in the analyses of gesture-prosody synchronisation in which prominent units were found to be paired/synchronised. This shows that prominence as a constraint is also at play in the anchoring of gesture relative to the IS of utterances.

The pairing patterns of G-phrases and IS categories were also affected by gesture type. Topic-deictic pairings were common. This showed that the function of relating to elements in the previous discourse was gesturally realised by pointing at their locations in the mental space of interlocutors. Contrasted elements also tended to pair with deictics. Pointing is a non-verbal demonstrative which is naturally used to highlight one entity over other available ones, thereby exhausting the alternatives to that entity. For instance, pointing at a green bottle amongst a number of bottles with different colours along with the utterance “I want the green bottle” can perhaps be seen as the default gestural behaviour for such a context. In this sense, it can be said that the demonstrative nature of deictics causes a natural functional overlap between deictics and contrast. Interestingly, what is common in the pairings with deictics is that deictics are used to establish links within a discourse. A topic links an utterance to a previous discourse, and a contrasted element gives rise to a notion of contrast with a previous element, which in a sense links those contrasted elements (Vallduví, 2016). These suggest that establishing backward links within a discourse attracts deictics more than other gesture types.

The other pairing pattern observed was that iconics and metaphorics were more likely to be paired with foci. This can be attributed to iconics and metaphorics having high levels of representativeness of concepts (concrete or abstract) used in speech. Through this representativeness, iconics and metaphorics are able to convey more descriptive content which can complement the focal speech content carrying the discourse forward. This interpretation suggests that forward relations in the discourse attract more iconics and metaphorics than deictics.

Deictics have been reported to be used for topical functions in conversation before. For example, interlocutors can point to another interlocutor who brought up a certain topic in an earlier part of the conversation to indicate which part of the discourse they are referring to (Bavelas et al., 1992). Deictics are also used to assign referents locations in space so that each time they are referred in a conversation, interlocutors' expressions make use of that particular space (Gullberg, 2006). The results of the present study are in line with these findings - deictics can have topical roles in a conversation. The present study further contributed to our knowledge about the discursive functions of gesture by revealing that the roles of establishing backward and forward relations in discourse are shared by different gesture types depending on their power of representativeness. In general, this shows that gesture is informed by these information structural categories during its production.

It must be noted that these pairings are in no way deterministic. Gestures also encode the semantic content of topical and focal elements outside of these patterns. For example, as presented in Section 6.2, approximately 58% of deictics were paired with foci, which is in line with the general pattern where a great majority of G-phrases were paired with foci. However, when all G-phrase-topic pairings were investigated, it was seen that most of these G-phrases were deictics, hence the pairing pattern. All things considered, these patterns have indicated *tendencies* showing that in general gestures tend to accompany discursively prominent elements, and that certain gesture types are favoured over the others depending on the IS category that gestures accompany. This suggests that gesture is sensitive to the organisation of information in utterances. It was mentioned in Section 2.6.2 that gesture has already been shown to be sensitive to information status dimension of IS (i.e., new/given distinction). The present study contributes to the limited body of research on gesture and IS by showing evidence for the association of gesture and the other two IS dimensions: topic/focus/background and contrast.

In terms of synchronisation, these pairing patterns did not have a significant effect on the synchronisation. That is, the synchronisation of IS units and G-phrases was the same for all gesture type, topic/focus/background, and contrast combinations. The general synchronisation pattern was similar for both onsets and offsets where there were apparent clusters of observations in the distributions of time differences, but they were not centered on zero. There are two ways of interpreting this finding. The first one is that although not on zero, the clustering of observations indicate a systematic co-occurrence of G-phrases and IS units. As indicated in Chapter 1, the present study considers systematic co-occurrences of units as synchronisation. Accordingly, this interpretation suggests that G-phrases are IS categories are synchronised but with a systematic delay.

Ebert et al. (2011) reported a similar finding of a systematic delay which they interpreted as evidence for synchronisation. They only reported a synchronisation of onsets of G-phrases with focus (m=310ms, sd=410) but not for offsets (m=-150ms, sd=1240ms) due to the high standard deviation observed in the offset synchronisation. Note that the mean and standard deviation of time differences are somewhat close to what was reported in the present study (onsets: m=413ms, sd=373ms; offsets: m=-195ms, sd=657ms), but the standard deviation was about half of that in their study. The difference in offset synchronisation is likely to be a result of what they considered to be the offset of G-phrases. They disregarded the post-holds and retraction phases of G-phrases, and only checked the synchronisation of stroke offsets and focus offsets. The present study, on the other hand, did not disregard any G-phases and tested synchronisation with the actual G-phrase boundaries. This approach seems to have caused a more compact distribution of time differences of offsets compared to Ebert et al. (2011).[1] Overall, the present study has reported a systematic delay for both onsets and offsets.

The second interpretation, synchronisation with apical areas, has to do with where these systematic delays between onsets and offsets place IS units relative to G-phrases. The findings have shown that G-phrases started about a median G-phase duration earlier than IS categories and ended more than a syllable duration after them (high standard deviation indicated even longer delays were possible). This meant that IS units were contained within G-phrases. However, as described in Sections 2.1.1 and 3.4.1, G-phrases are a combination of G-phases. Therefore, it might be the case that the clustering of observations might not be showing a systematic delay but instead might be indicating a synchronisation of G-phrase-medial phases.

The present study has probed into such a possibility in Section 6.4. It has defined apical areas, as the potential anchor of IS categories, and shown that IS categories were very tightly synchronised with them. Further analysis aimed to demonstrate how well stroke offsets would have synchronised with IS category offsets if they were selected as the anchors of offset synchronisation as in Ebert et al. (2011). The findings of this analysis demonstrated that post-hold offsets were synchronised better with IS category offsets, confirming that apical areas are the valid anchors for IS synchronisation.

Overall, both interpretations show that gestures are synchronised with IS units. However, in the first interpretation, the motivations behind those systematic delays must be explained for a more meaningful synchronisation

[1] Notice that the offset synchronisation still shows a higher standard deviation than the onset synchronisation. My observation while annotating the data was that some post-holds had unusually long durations compared to other post-holds in their immediate context. I suspect that these post-holds with longer durations served a different function in the speech. They might have been used for the regulation of conversation, e.g., turn-taking which can even occur in monologues (Sacks, Schegloff, & Jefferson, 1979) Regardless, this is beyond the scope of the present study and is not discussed here.

scenario. In the cases of onsets, the most common explanation that has been offered in studies that investigated gesture-prosody synchronisation (see Section 2.5.2) is that gesture generally precedes speech counterparts, and the findings of systematic delays reflect this. While this interpretation may be accurate, it still leaves us with the ambiguous concept of precedence. What counts as an acceptable amount of precedence and what does not for different types of events and phrases has not yet been defined clearly. The precedence criterion also falls short in explaining offset synchronisation since it seems that gesture offsets often occur at varying distances after their corresponding speech part (for examples in synchronisation with prosody see Loehr, 2004; Ferré, 2010, and see Ebert et al., 2011 for information structure). Nevertheless, the systematicity of delays indicate synchronisation. It is only the case that what motivates these delays has not been sufficiently explained yet.

The second interpretation can be seen as a stronger indicator of gesture and IS synchronisation because it shows that both onsets and offsets of the tested units tended to co-occur within as little as an average syllable duration. This finding implies that it is not necessary to consider precedence as an indicator of synchronisation. Rather, gesture-speech synchronisation might be with particular phases or phase combinations within gesture. Such an approach also opens up new avenues for the analyses of gesture synchronisation.

7.1.4 Summary

Earlier studies have argued that the synchronisation of gesture and speech at the micro level is managed by prosody. The same argument has also been made about synchronisation at the macro level although there have only been few studies investigating this possibility, which have offered inconclusive results. The present study has systematically investigated the synchronisation at both levels. It has revealed consistent synchronisation of gestural and prosodic units, confirming that gesture production is informed by prosody.

Further, the present study has explored the possibility that gesture may also be synchronised with information structure. The findings have shown a categorical sensitivity where certain gesture types tend to accompany specific IS categories more than the others. More importantly, it has found that gesture phrases are synchronised with both topics and foci, but gestures are much more likely to accompany foci. The consistent synchronisation of full topical and focal areas with gestural constituents (either G-phrases or apical areas) can be interpreted to mean that amongst its other functions in communication, gesture is a cue to information structure in addition to prosody and syntax. This is especially true for focus as it is the preferred anchor. Gesture phrases clearly mark where the entire focal domains starts and ends within an utterance - they distinguish foci from topics and backgrounds. All things considered, to the author's knowledge, the present study is the first study to fully establish a synchronisation relationship between gesture and information structure through a systematic investigation.

On the whole, all of these findings have indicated that gesture is informed by IS in addition to prosody, confirming the three-way association of these as represented in Figure 2.16. Within this view, gesture is linked to prosody. Its production is informed by the encoding of prosodic prominence and hierarchy at both micro and macro levels. In addition, gesture is directly linked to information structure. Gesture is sensitive to IS categories, and this sensitivity is expressed through the selection of certain gesture types for certain IS categories as well as by tight synchronisation between these units.

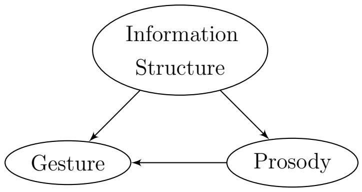


Figure 2.16: Three-way association of information structure, prosody, and gesture (repeated from page 97)

The present study also shows that gesture and IS are synchronised but this synchronisation is mediated by prosody thanks to the strong connection of prosody and information structure. Because prosody is a cue to information structure (see Section 2.6.2), the synchronisation of gesture with prosody naturally contains an element of information structural influence. However, the present study does not agree with the view that gesture-prosody synchronisation is only an epiphenomenon of gesture and IS synchronisation as argued in Ebert et al. (2011). This is based on the fact that the present study has found no significant effects of IS categories (including contrast) on the synchronisation behaviour of apexes and tones or gesture phrases and intermediate phrases (Sections 4.2, 5.2, and 5.5). The only exception here is that nuclear prosodic prominence, which is linked to focus, has been shown to make synchronisation between units tighter. The present study considers this as evidence for the indirect link between IS and gesture because it is not a deciding factor for synchronisation by itself - gestural units are synchronised with prosodic units outside of the bounds of nuclear prosodic prominence. For example, in the apex-tone synchronisation, apexes can be synchronised not only with nuclear pitch accents which are on focused elements (see Section 2.6.2), but also with pre-nuclear pitch accents. In fact, as was presented in Chapter 4, almost 60% of pre-nuclear ips contained such synchronisation cases. In addition, perhaps more strikingly, the synchronisation of apexes with the onsets of prosodic words in the absence of pitch accents cannot be attributed to IS but to prosodic structure. These results rule out that prosody is epiphenomenal to IS-gesture synchronisation.

Instead, the results of the present study have indicated that gesture and IS are synchronised through prosody. Within the present study, IS units are also groups of prosodic phrases (including recursive prosodic phrases), and gesture synchronisation has been shown to be with these groups framed by IS rather than single prosodic phrases or prosodic phrase groupings spanning over both topics and foci. In brief, the direct link between gesture and IS stems from gesture's sensitivity of different IS categories. These findings, including gesture-prosody synchronisation, indicate a three-way synchronisation between gesture, prosody and information structure (Figure 2.16).

As previously pointed out in Chapter 1 and Section 2.4, synchronisation has not been represented in psycholinguistic models of speech and gesture production. The findings discussed in this section have implications for current gesture-speech production models in terms of how and at which stages of production speech and gesture are linked to each other. Section 7.2 discusses these implications and suggests modifications for their representations.

7.2 Synchronisation in Speech-Gesture Production Models

Previous research has revealed various conceptual links between speech and gesture (Section 2.2). Numerous speech and gesture production models have been proposed to represent these links. Section 2.3 introduced the four most influential models, i.e., the Growth Point theory (GP theory), the Lexical Retrieval Theory (LR theory), the Sketch Model (SM), and the Interface Hypothesis (IH) (see Section 2.3 for an overview).

Synchronisation in general has not been represented in models of speech-gesture production. They have only integrated synchronisation in very broad sense or have not considered it at all. In brief, in terms of synchronisation, there is not enough information about how synchronisation would be managed during production in the GP theory. It mostly makes the case that gesture and speech originate from the same idea unit, therefore, any kind of synchronisation between the two can only be attributed to this. The LR theory does not predict that speech and gesture production processes are connected through synchronisation. These production processes are only linked to enable gesture to facilitate lexical retrieval, which is achieved by linking gesture to speech modules that have access to the lexicon. The SM is similar to the GP theory in that gesture production diverges from speech production at the level of the conceptualiser, which presumably manages synchronisation as well.

The IH does not set out to explain the synchronisation of speech and gesture. However, its highly interactive design allows one to make claims related to synchronisation. Because of this interactivity, it is also able to account for linguistic effects on gesture, which was one of the main motivations of behind its development. Section 2.3 presented a simplified schematic of the model. Figure 7.6 shows the original in Kita and Özyürek (2003, p. 28).

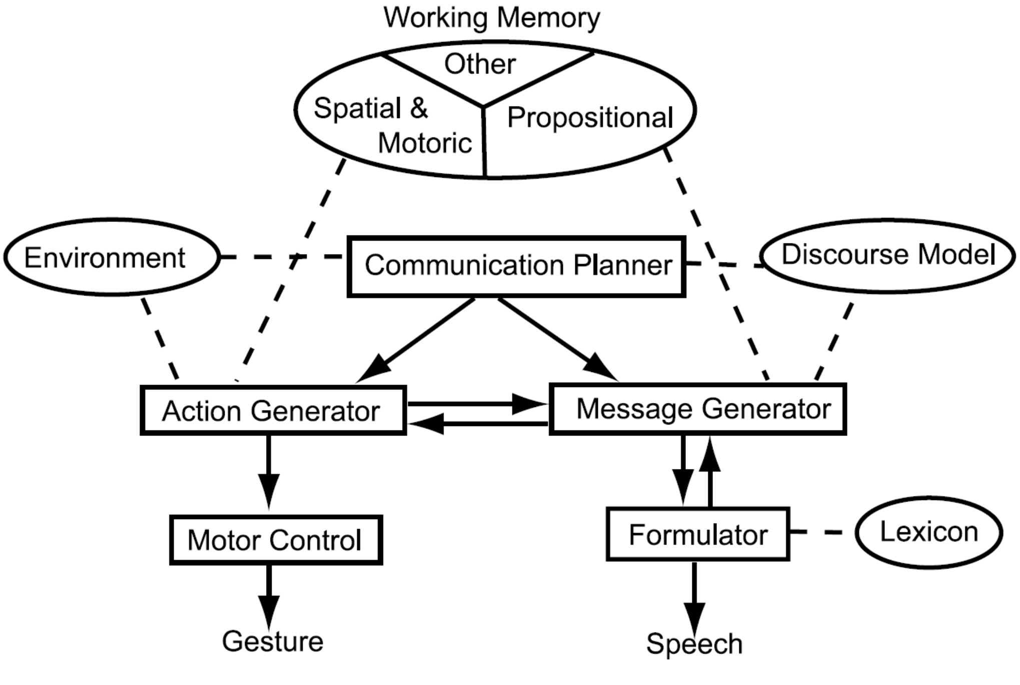


Figure 7.6: The model of speech and gesture production in the Interface Hypothesis (Kita & Özyürek, 2003, p. 28)

It can be understood from Figure 7.6 that the IH mostly focuses on the interaction of speech and gesture production in the early stages (i.e., the conceptualiser level). There is not much detail about the processes taking place after the action generator and the message generator. However, since the model has its roots in Levelt (1989), the present study assumes that the IH adopts Levelt's (1989) descriptions of the inner workings of the formulator and articulator (although the model does not show the articulator module).

The present study considers the IH to be the most comprehensive model in terms of representing the associations between gesture and speech. Based on the results of the present study, synchronisation can be integrated into this model with certain modifications. Gesture and IS synchronisation can be explained within the existing interactions of modules at the conceptual level. However, the results of the present study are also concerned with prosodic hierarchy and prominence which are encoded at the formulator level. This extension requires certain modifications to the model in Figure 7.6. Accordingly, the present study aims to extend the IH beyond the conceptualiser stage by proposing new modules and connections between these in order to integrate synchronisation. The model including the proposed modifications is referred to as *Extended Interface Hypothesis* (EIH), which is detailed in Section 7.2.1.

7.2.1 Extended Interface Hypothesis (EIH)

Section 2.3 already summarised the IH. The description of the EIH in this section further details the workings of production processes while highlighting the proposed modifications. Note that a modified version of the IH has also been proposed by Kopp, Bergmann, and Wachsmuth (2008). Their model seems to consider speech-gesture synchronisation as the timing relationship between each gesture and its lexical affiliate, and does not give details of any other temporal constraints between speech and gesture. Its architecture is built around a central module called the "blackboard" which all modules are connected to. There is some lack of clarity in the interpretation of the functions of modules and in the input-output flow between them resulting from this type of design where each module is connected to every other. The EIH proposes more specific links between its modules based on the empirical findings of the present study and others.

The EIH inherits the view that speech and gesture production are separate but highly interactive processes that parallel each other. The production processes show more parallelism in the EIH compared to the IH because the IH lacks a formulation stage for gesture production after the action generator. The EIH also adopts the view that gesture is generated from a general mechanism based on the claim that gesture and action have a common source (Jürgen, 1996). However, the EIH mostly comments on the production of co-speech gestures. The generation and synchronisation of actions with speech is left for future study (see Hostetter & Alibali, 2019).

The EIH covers the production of all gesture types in its explanations. The IH only dealt with representational gestures (iconics, metaphorics, deictics) and excluded beats from the model. Beats were included in the analysis of micro level synchronisation in the present study, and it was concluded that there was no major difference from the other gesture types in their synchronisation behaviour. In contrast, at the macro level, beats were excluded from the analysis on the basis that in general, they do not have durations comparable to the durations of prosodic phrases investigated (i.e., intermediate and intonational phrases). Yet, it is possible that they may be synchronised with prosodic words, which was not investigated in the present study. The EIH presumes that beats too would show phrasal synchronisation with a prosodic phrase, and generalises its explanations to all gesture types at both micro and macro levels.

7.2.1.1 Architecture

The architecture of the EIH is illustrated in Figure 7.7. Similar to the IH schematics, the rectangles represent modules in charge of processing information, and the arrows indicate the flow of input and output between these modules. Circles are information storages that the modules have access to. The connection between the modules and information storages is marked with dashed lines.

Conceptualisation

In the model, production starts with two distinct planning stages at the conceptual level. The conceptualiser in Kita and Özyürek (2003) is divided into two separate modules. These are the *Communication Planner* and the *Message Generator*. The Communication Planner takes on Levelt's (1989) macro-planning which decides which information will be communicated and generates communicative intentions. The present study has discussed how gesture functions as a communicative tool in Section 2.2.1. There have also been several studies that have shown that gestures can convey redundant information, which has been interpreted as a challenge to the communicativeness of gesture (Section 2.2.2). Within the EIH, the redundancy of gestural content does not pose a threat to the communicative value of gesture. By being synchronised with the prominent events and phrases in speech, gesture assumes a highlighter function - it marks parts of the speech that are the most communicatively valuable for an interlocutor. Therefore, this type of synchronisation behaviour alone makes co-speech gesture evidently communicative regardless of the content it bears (cf. Krauss et al., 2000).

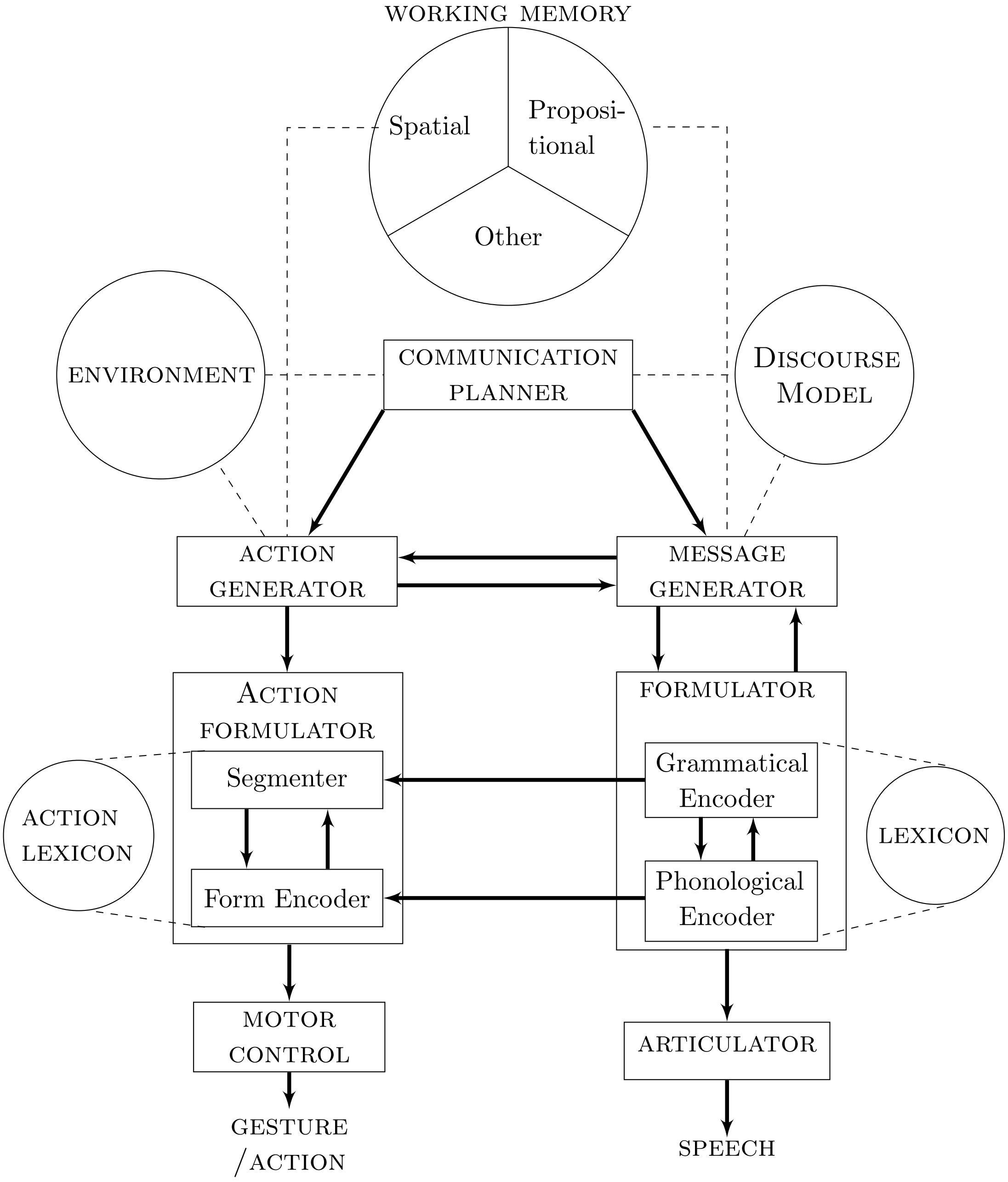


Figure 7.7: A schematic of the Extended Interface Hypothesis

Another function of the Communication Planner is to decide on which modalities of expression, i.e., speech and gesture, will be used in communication. This implies that speech and gesture may be used to communicate with or without each other. Therefore, the Communication Planner is not seen to be reserved for speech production only. Rather, it generates a general multimodal plan of communication and determines which information will be expressed in which modality. As in the IH, this distribution of information between speech and gesture is informed by the *Environment* which holds information about the environment the communication takes place in. It has been shown that depending on environmental factors (e.g., the distance of a referent) interlocutors distribute content between speech and gesture differently (Bangerter, 2004; Van der Sluis & Krahmer, 2007). To account for such effects, the Communication Planner has access to the Environment which stores this information.

One difference between The EIH and the IH has to do with the Discourse Model in the IH. The Discourse Model in the IH has two submodules called the Interaction Record and the Addressee Model (Kita & Özyürek, 2003; Kita, 2010, 2014). The Interaction Record tracks what has been communicated and not communicated in a conversation. This includes which information in the discourse has been gestured and how specifically this information has been gesturally encoded. The availability of this information has been shown to affect speech-gesture production. For example, if a gesture has already provided a piece of information, the subsequent verbal descriptions are less likely to contain that information (Melinger & Levelt, 2004). Moreover, gestures for semantically related referents at different parts of a discourse bear similar form features (McNeill et al., 2001; McNeill, 2005). Also, when different interlocutors gesture for the same referent in a discourse, the form features of their gestures converge (Kimbara, 2008). These findings show that along with speech, gesture too has access to what has been expressed verbally and gesturally in a discourse. The other submodule in the Discourse model, the Addressee Model, stores what interlocutors know about their addressee. The information stored here includes the addressee's visibility, interactivity, attentiveness, and awareness that interlocutors share some knowledge. All of these have been shown to have an effect on gesture production as previously mentioned in Section 2.6.2.

The IH accounts for all of these effects on gesture production by giving the Communication Planner access to the Discourse Model (see Figure 7.6). In this way, it can pass the information in the Discourse Model to other modules dealing with gesture production. The EIH fundamentally agrees with this representation in the model. However, unlike earlier gesture production models, it emphasises that keeping information about the addressee and what has been communicated or not communicated in a conversation are functions related to the concept of common ground within IS (see Section 2.6.2), it also gives further details on the influence these concepts on gesture production. Common ground has been a subject to many studies concerning information structure (Chafe, 1976; Clark & Haviland, 1977; Fowler, 1988; Clark, 1996; Krifka, 2008 amongst others). It is generally seen as a means "... to model the information that is mutually known to be shared and that is continuously modified in communication" (Krifka, 2008, p. 245). The notion of common ground, in essence, serves to create a distinction between what is known/presupposed and what is newly asserted/proffered. Such a distinction is the basis for organising information within utterances. As set out in Sections 2.6.1 and 3.4.3 topics and backgrounds contain the shared information in each proposition that already exists in the common ground, whereas foci update the common ground. Since the content of the common ground is continuously updated in a conversation, information in speech, and therefore gesture, has to be organised in accordance with the common ground. Moreover, the common ground does not only consist of what has been established and accepted by interlocutors but also keeps a record of entities in the discourse based on whether or not they have been previously introduced to the common ground. This type of classification of entities in the common ground enables new/given distinctions in the information status dimension of information structure (see Sections 2.6.1 and 3.4.3).

In Section 2.6.2, a number of studies were reviewed showing that the retrievability of referents affects gesture frequency. In addition, it was shown that several studies were reviewed that investigated whether the awareness that interlocutors share some common ground knowledge affected gesture. These studies have also reported effects on gesture frequency as well as gesture form. The present study has contributed to this body of research by showing that gesture production is informed by other information structure (IS) categories namely topic, focus, background, and contrast. It has been shown that there is a categorical sensitivity in that the type of gesture that is selected tends to depend on the IS category it accompanies. More importantly, gesture phrases have been found to be synchronised with these categories. In the light of all these findings, the present study argues that a valid model of speech and gesture production must also integrate information structure. In the EIH, this integration is achieved through links to the *Discourse Model.* It inherits the function of keeping a record of communication and the addressee information. Being linked to long term memory (not represented in Figure 7.7), the Discourse Model contains all presupposed and previously accepted information as well as a set of entities that the interlocutor believes are explicitly or implicitly available to interlocutors, which are continually updated by the exchanges between interlocutors during conversations. The Communicative Planner has access to the Discourse Model so that a communicative intention can be generated in order to satisfy the informational needs of the addressee. Through its link to the Discourse Model, the Communication Planner also prepares a frame which grossly specifies in what order information should be organised in an utterance as per Levelt (1989), thus creating the primitives of IS categories. This ordering of IS categories is also passed to the Action Generator, establishing the initial step of IS-gesture synchronisation.

The *Message Generator* equates to micro-planning in Levelt's (1989) model. It receives the communicative intention and the decision about which modalities are to be involved from the Communication Planner. It retrieves from working memory the propositional representations of the concepts to be verbally expressed. In correspondence with the Discourse Model, these representations are assigned with IS categories based on their availability in the Discourse Model, and they are ordered according to the frame provided by the Communication Planner. After all these concurrent processes, the final product of the Message Generator is the pre-verbal message which contains the high-level planning of syntactic structure, as well as prosodic structure, which are generated as requisites of IS category assignment.

The EIH fully adopts the module of *Action Generator* in the IH. Similar to the Message Generator, it assumes the role of conceptual planning of gestures and actions. The Action Generator works with the Communication Planner, the Message Generator and the Environment to prepare the action/gesture content (unless a distinction is explicitly made, the term gesture is used to refer to both gesture and action for the sake of simplicity from hereon). It has access to the Environment in order to take into account the physical space available for gesture. There may be too little gestural space which would naturally impact the size of gestures, or there may be other physical obstacles in the way of potential gestural movements (De Ruiter, 2000). Moreover, actions may involve touching or object manipulation where the gesturing body parts come into contact with objects. These environmental factors are available to the Action Generator through the Environment and are included in the gestural plan.

In line with the communicative intention it receives from the Communication Planner, the Action Generator determines the content of a gesture by accessing the relevant parts of working memory. The feedback from the Message Generator also has an effect on this content in the IH. It allows for an interaction between the Action Generator and the Message Generator as well as between the Message Generator and the *Formulator*. These links are mainly established to enable the gestural content to be shaped by linguistic formulation that takes place in the Formulator (detailed later in this section). The IH bases this link on the findings of studies that have shown that the mapping of motion event components (i.e., manner and path) onto syntactic clauses is also paralleled in gestural encoding (Özyürek, 2002; Kita & Özyürek, 2003; Kita & Lausberg, 2008; Özyürek et al., 2008). That is, if these components are expressed in one clause, gesture also conflates these within a single gesture, and if they are expressed in separate clauses then gesture only encodes one of these components. These findings show that gesture production is also informed by the syntactic encoding taking place in the Formulator. The linkage of the Formulator to the Action Generator through the Message Generator also accounts for this effect on gesture production.

The interaction between the Message Generator and the Action Generator is also claimed to be in line with the idea that gesture facilitates conceptualisation for speaking (Bock & Cutting, 1992; Kita, 2000; Alibali, Kita, Bigelow, Wolfman, & Klein, 2001; Melinger & Kita, 2007; Kita & Davies, 2009; Alibali & Kita, 2010; Alibali, Spencer, Knox, & Kita, 2011). This claim is not covered here but there is a growing body of research in agreement with this claim. Due to the link between the Action Generator and the Message Generator, the IH is also compatible with this claim.

The EIH acknowledges the implications of these earlier studies and adopts the same interactions between these modules as the IH. One addition to the interactions between these modules in the EIH involves the association of gesture and IS. The interaction between the Action Generator and the Message Generator allows gesture production to be informed by the IS categories encoded in the pre-verbal message. Gesture production in the Action Generator is also informed by which parameters of IS dimensions (i.e., topic/focus/background, information status, and contrast) are assigned to which parts of pre-verbal message. This information helps the Action Generator plan the gestural content (i.e., the semantic function of gesture) along with the high-level planning of the gestural syntax (i.e., gestural phrasing). This also means that the Action Generator plans where the gestural content is roughly anchored relevant to the IS categories in the pre-verbal message. The precise synchronisation does not take place at the conceptual level since neither gestural nor prosodic structure have been formulated yet, and the rigorous synchronisation behaviour observed in the present study requires such formulation.

In brief, the final output of the Action Generator is the gestural equivalent of the pre-verbal message. The EIH uses The Sketch Model's terminology (De Ruiter, 2000) and uses the term "sketch" to refer to the output of the Action Generator. The sketch contains the abstract blueprint of the gestural content and information about how the gestural content relates to the preverbal message. Concrete parameters related to motor control such as size, shape, speed, and location are underspecified in the sketch.

The modules and interactions introduced so far detail the conceptual planning in the production of multimodal expressions. The next stage in production is the formulation of these plans. In speech, this entails the linguistic encoding of the pre-verbal message, and in gesture, it entails the segmentation of gestural movement and the assignment of form features to the sketch. Studies on speech planning seem to agree that speech planning is an incremental process where interlocutors plan their verbal expressions in smaller chunks instead of planning whole utterances (for an overview see Wheeldon, 2013). This implies that there is an overlap between planning and articulation during speaking - the articulation of previous parts of utterances takes place in parallel with the planning of later parts (Levelt, 1989). How far ahead interlocutors can plan, i.e., the scope of the planning, has been controversial although recent studies have argued that the scope of planning varies in different situations (Konopka & Meyer, 2014). These situations include the goal of interlocutors (Ferreira & Swets, 2002), word order of phrases (Brown-Schmidt & Konopka, 2008), the availability of cognitive resources (V. Wagner, Jescheniak, & Schriefers, 2010; Konopka, 2012), and the information status of events (Ganushchak & Chen, 2016).

The gesture production models reviewed here do not give details about an incremental production of gesture, however it has been previously assumed in studies on gesture synthesis (Salem, Kopp, Wachsmuth, & Joublin, 2009; Van Welbergen, Reidsma, & Kopp, 2012). Given the overall parallelism of speech and gesture production, the EIH also assumes that both gesture and speech planning take place incrementally. Although the scope of planning units seems to be subject to variation, the EIH adopts prosodic words as the basic planning units as per Levelt (1989). For gesture, it suggests gesture phases (i.e., G-phases) as the basic units of planning because they are the minimal units that are combined into larger gestural phrases.

Formulation

The processes taking place after the conceptualiser stage are not very detailed in the IH as its main focus is on explaining the high-level planning of speech and gesture. The EIH extends the IH and explains the inner workings of the formulation stage for both speech and gesture production (compare Figures 7.6 and 7.7).

Starting with speech production, the module that processes the output of the Message Generator is the *Formulator*. It has two interacting submodules called the *Grammatical Encoder* and the *Phonological Encoder*. The processes involved in these submodules are based on Levelt's (1989) descriptions. Formulation starts with the Grammatical Encoder which takes the pre-verbal message as input and generates a surface structure as output. It projects the propositions and their relations in the pre-verbal message onto a grammatical phrase structure. This process is lexically driven; therefore, the grammatical encoder has access to the *Lexicon* (see Figure 7.7). The Lexicon stores lexical items which carry specifications for semantic and syntactic information, i.e., lemmas. The Grammatical Encoder retrieves a lemma from the Lexicon if it matches that part of the pre-verbal message (amongst a number of activated lemmas, see Levelt, 1989, chapter 6 for lexical activation). The syntactic properties available in the lemma trigger syntactic building procedures. These properties include syntactic categories (e.g., verb), the grammatical functions required (e.g., objects), the relations between functions (e.g., complements), and thematic roles. The lemmas also contain "diacritical variables" such as person, number, tense, aspect, mood and pitch accent as well as lexical pointers, which are important for phonological encoding (detailed later in this section).

The Grammatical Encoder builds a surface structure for the expression which consists of lemmas organised into phrases according to their semantic and syntactic properties. These phrases form the constituents of the surface structure (e.g., verb phrase and noun phrase). The Grammatical Encoder is also tasked with assigning IS categories to these syntactic constituents. Focus status is assigned to constituents that contain a prominent representation in the pre-verbal message, and topic is assigned to constituents that link the interpretation of the utterance to a previous one. Background is assigned to post-focal constituents containing highly accessible representations in the pre-verbal message. The assignment of focus has prosodic and syntactic requirements. In the model, words have accents depending on their position in the prosodic structure. Foci align with nuclear accents, usually occurring towards the end of sentences (Calhoun, 2010b). Subsequently, a syntactic structure must be chosen in way that ensures the focused constituent occurs at the desired utterance-final location (e.g., just before the sentence-final verb in Turkish). The overall implication of such a constraint is that the planning of prosodic structure is concurrent with syntactic planning. The EIH adopts the claim that the planning of prosodic structure takes place in the Grammatical Encoder following Calhoun (2010b). In terms of production, this means that surface structure also encodes information related to (de-)accenting and prosodic phrase breaks which are further processed by the Phonological Encoder.

The lemmas contain lexical pointers which "point to addresses where the corresponding word-form information is stored" in the Lexicon (Levelt, 1989, p. 180). These lexical pointers trigger the phonological encoding process in the Phonological Encoder which starts with the selection of specific morphological and phonological forms. With input from the lexical pointers, the prominence structure (i.e., metrical form) for each word is retrieved from the Lexicon. This metrical information is then integrated into the main prosodic structure of the utterance in line with the surface structure input, e.g., focused constituents are pronounced in a prosodically prominent way. Feedback from the Phonological Encoder to the Grammatical Encoder is possible in Levelt's (1989) model which minimally consists of the revision of syntactic frames in case the Phonological Encoder runs into trouble. The final product of the Formulator is the phonetic plan which is ready for articulation.

The *Articulator* is the final speech production module in the EIH (the IH ends with the Formulator). The motor execution of the phonetic plan, involving anatomical systems such as the respiratory, laryngeal, and supralaryngeal systems takes place in the Articulator. The final product of the Articulator, and the overall speech production, is overt speech.

The EIH diverges from the IH in terms of how it designs the gesture production post-conceptualiser stage. In the IH, the Action Generator seems to be tasked with both generating a sketch and formulating a motor action plan according to it at the same time. This plan is then executed by the following module, the Motor Control (see Figure 7.6). The EIH separates these two functions of the Action Generator in the IH and introduces an intermediate formulation module between the Action Generator and the Motor Control. This module is referred to as the *Action Formulator*. The main motivation behind introducing another module is that, as we will see, the functions realised by this module are thoroughly different from those realised in the Action Generator.

The Sketch Model (SM) (De Ruiter, 2000) also made use of such a formulation module (i.e., Gesture Planner). Different from the SM's Gesture Planner, the Action Formulator in the EIH is also involved in the planning of actions, and it does not have direct access to the Environment. There are certain overlaps between them in terms of the processes involved but the Action Formulator provides additional details for gesture phrasing and synchronisation.

In the EIH, the only information storage that the Action Formulator has access to is the *Action Lexicon*. The Action Lexicon contains action templates, which assist in the production of emblems, pantomimes, and actions as well as co-speech gestures (i.e., gesticulations, see Section 2.1.1). These action templates are abstract motor programs on which the content of sketch is fitted. They do not hold strict and complete motor programs for any given content. They come with a degree of freedom which enables the application of modifications to fit the content of the sketch as intended. The templates also contain form pointers which indicate where specific motor instructions for a given template can be found in the Action Lexicon. These more concrete instructions are referred to as action schemata.

The templates can be highly conventional (i.e., culturally specified such as the OK sign), practical (i.e., matching the goal of an action such as driving a screw), habitual (i.e., stylistic), or general representation techniques (e.g., pointing, shaping, drawing). The choice of which template to use depends on the goal of the gestural message containing environmental and discursive information. For example, if the aim is to describe an object, shaping as a representation technique is a likely choice. In the EIH, gesture types are not associated with particular templates stored separately in the Action Lexicon (unlike the Gestuary in the SM, cf. De Ruiter, 2000). All templates are available to all gesture types, and multiple templates can be conflated within a gesture type so long as they can represent the content of the sketch. This is exactly why gestures can show a continuum of gesture types as explained in Section 3.4.1.1. The EIH can also account for the production of emblems and pantomimes thanks to these templates, but the interactions between speech and gesture production only apply to emblems, actions, and co-speech gestures, and not to pantomimes. This is because pantomimes, by definition, require the absence of speech. The accompaniment of speech is optional for emblems and actions, which can have implications for their synchronisation. Whether emblems and actions are synchronised with speech in the same way as co-speech gestures is beyond the scope of the present study. Therefore, the EIH's explanations of synchronisation mainly concern co-speech gestures.

Within the EIH, the synchronisation of gesture with speech takes place through simultaneous interactions between the Action Formulator and the Formulator. The present study has shown evidence for the synchronisation of gesture and IS mediated by prosody. Groupings of prosodic phrases with the same IS category are synchronised with G-phrases. This implies that gesture production is informed by the generation of prosodic structure and IS category assignment. In addition, in terms of the synchronisation of prosody and gesture, two constraints have been observed: (1) the prosodic hierarchy constraint, (2) the prominence constraint. There is a rank-ordering of these constraints in terms of their application sequence - the prosodic hierarchy constraint is applied before the prominence constraint. That is, the prosodic hierarchy constraint first decides the domain of synchronisation, which is the prosodic word for apex synchronisation and the intermediate phrase for G-phrase synchronisation in the case of Turkish. It is only after this that the prominence constraint comes into effect to fine-tune the synchronisation of gestural and prosodic anchors (see Section 7.1.2.2). Note that this is in line with the generation of prosodic phrasing and prominence in the Formulator - the prosodic phrasing is generated in the Grammatical Encoder first, and then the prominence structure is woven onto it. Unlike any other speech-gesture production models, the EIH hypothesises that each of these processes simultaneously informs the corresponding submodules in the Action Formulator, which manages the synchronisation of speech and gesture.

The first of these submodules is the *Segmenter*. The Segmenter receives the sketch as input which contains the gestural message that is already labelled for the IS unit it is intended to accompany. The sketch is also indexed for the rough specifications of where gesture should take place (i.e., gesture space) in accordance with the environmental factors (through the Environment). Upon receiving the sketch, the Segmenter retrieves templates from the Action Lexicon that match the gestural message in the sketch. This constitutes the high level meaning-form mapping for gestures, i.e., gesture type assignment. Next, the Segmenter places the template in the gesture space adapting both reciprocally if necessary. This triggers two other processes. The first one is the assignment of body parts to gesture depending on the representation technique in the template, gesture space coordinates, and the availability of body parts for gesturing. The second submodule is involved in planning how these body parts reach the gesture space and retract from there, which roughly forms the phrase structure of gesture (i.e., segmentation, see Section 3.4.1.1). The generation of the phrase structure is informed by the processes in the Grammatical Encoder. The initialisation and termination of gestural movement (within the bounds of a G-phrase) are synchronised with intermediate phrase breaks encoded in the prosodic structure in line with the results of the present study. The Segmenter also receives information about which intermediate phrases are assigned with the target IS unit to figure out its exact span for synchronisation purposes. Note that the Segmenter already has the knowledge of which IS unit is targeted for gesture synchronisation through the sketch input (the categorical sensitivity is achieved at the Action Generator level). Once the intermediate phrase breaks corresponding to the target IS unit are retrieved from the Grammatical Encoder, the Segmenter sets the duration of the template to match the IS unit. The overall duration of the template which carries the gestural meaning makes up the apical area introduced in the present study. This way, the corresponding boundaries of IS units and apical areas achieve synchronisation. Note that at this stage, there is no stroke/post-hold (or preparation/pre-hold) distinction. The Segmenter only plans pre-template movement (i.e., gestural movement that brings the body part to the onset of the template), the template, and post-template movement (i.e., the gestural movement that departs the body part from the offset of the template). The EIH considers pre-holds and post-holds as means of synchronisation which are encoded in the next submodule in the Action Formulator.

In brief, the EIH establishes the information flow between the Segmenter and the Grammatical Encoder by directly linking them (see Figure 7.7). The coordination of these modules results in macro level synchronisation as observed in the present study. After all these processes, the final product of the Segmenter is the template structure, which is ready to be encoded by the *Form Encoder*.

The Form Encoder receives the template structure, and the form pointer in the template triggers the retrieval of the relevant action scheme from the

Action Lexicon. Then, the Form Encoder maps the specific gesture morphology in the action scheme (i.e., motor instructions) onto the template structure constrained by the gestural meaning. This process makes up the lower level form-meaning mapping of gestures. The process precisely specifies the phrase structure - it transforms pre-template, template, and post-template movements into gesture phases (G-phases) drawing their boundaries.

Pre-template movements are transformed into preparation and pre-hold phases. In the preparation phase, the gesturing body part is carried to the apical area onset. In line with the synchronisation process achieved in the Segmenter, the apical area onsets must be synchronised with the IS unit onset. If the action scheme causes the body part to reach the apical area onset earlier than the IS unit onset, the preparation phase pauses, and the body part waits for the IS unit onset. This wait time before the apical area onset constitutes the pre-hold phase. Synchronisation is also imposed on the apical area offsets (i.e., the endpoint of the template) and IS unit offsets. If the action scheme completes the meaningful gestural movement in the template, i.e., stroke, before the IS unit offset, the body part pauses until the IS unit offset is generated, creating a post-hold. The achievement of IS-apical area synchronisation triggers the retraction phase where the gesturing part returns to a rest position.

The Form Encoder also generates dynamically prominent instances (i.e., apexes) within the stroke, in line with its kinetic goals in the action scheme. There can be multiple apexes in the stroke, but the final one signals that the kinetic goal is achieved marking the end of the stroke. The Form Encoder synchronises apex generation according to the input it receives from the Phonological Encoder. Synchronisation is constrained first by the prosodic phrasing. The input from the Phonological Encoder sets the domain of apex synchronisation as the prosodic word suggesting two alternatives (in Turkish): (1) the onsets of prosodic words (2) stressed syllables, both of which are marked with tonal events. Which of these prosodic events apexes are synchronised with is then constrained by the prominence structure which nominates stressed syllables marked with pitch accents for apex synchronisation. In their absence, apexes are synchronised with the onsets of prosodic words as the only other candidate in the domain of synchronisation. Note that the synchronisation of prominences in this manner stems from the communicative intention which needs to be highlighted verbally and gesturally in harmony. Overall, these processes achieve micro level synchronisation.

The EIH allows feedback from the Form Encoder to the Segmenter. The feedback can be utilised to request other templates or as a repair strategy in case the Form Encoder runs into trouble. After all these processes, the final product of the Form Encoder and the Action Formulator is called the motor plan. The motor plan is passed to the *Motor Control*. This module controls the muscular system and other related systems in order to execute the motor plan. The output of the Motor Control is overt gesture/action.

To summarise, the EIH extends the IH model of speech and gesture by introducing a detailed gesture formulation stage, which is used to integrate synchronisation into the model. The synchronisation of speech and gesture is established by linking the processes of gesture formulation to the processes of prosody and information structure formulation, which are also detailed in the model. The EIH needs to be further developed and tested, especially in terms of gesture formulation (e.g., emblem and action formulation). However, the descriptions in the model are consistent with the results of the present study. Overall, the model offers valuable insights into gesture production and the role therein of prosody and information structure.

7.3 Conclusion and Future Directions

This thesis has investigated the synchronisation of gesture and speech in Turkish. It has associated prosody with gesture as a driver of gesture-speech synchronisation. The results have shown that gesture and prosody synchronisation is not only limited to the synchronisation of prominences (i.e., apex and pitch accent), but that boundary marking events can also be synchronised with apexes (i.e., apex and L). This synchronisation has been found to be governed by the prosodic hierarchy which assigns a prosodic domain to synchronisation ensuring that apexes can only be synchronised with the members of that domain. This has revealed for the first time that the prosodic structure is a constraint on synchronisation in addition to prominence.

The results relating to apex synchronisation also contribute to discussion of accentlessness in Turkish. It has been observed that apexes are not synchronised with phrase-final pitch rises when the phrases are deemed accentless, which contradicts the view that the phrase-final pitch rises have a double function marking both a pitch accent and the end of the phrase. Instead, apex synchronisation has been observed to operate as it does in the absence of prominence. Overall, this has shown that gesture, due to its strong relationship with prosody, can be used to supplement phonological investigations - it can help decide phonological representations in other cases and other languages.

The thesis is also one of few studies that has investigated and reported synchronisation between gesture phrases and prosodic phrases. This has shown that gesture-prosody synchronisation persists between units at different levels in their hierarchical organisation, and therefore, gesture and prosody are more connected than previously assumed.

This thesis is the first to suggest and systematically investigate the possibility that gesture is also synchronised with information structure. It has been observed that the semantic functions of gestures are sensitive to information structural categories, and that gesture phrases, through prosody, are synchronised with these categories. This means that in addition to prosody, gesture too is a cue to information structure. Traditionally, only prosody has been considered as a driver of gesture-speech synchronisation. This thesis has shown evidence that information structure also constrains gesture production.

This thesis has highlighted that speech and gesture production models have hardly accounted for synchronisation. It has argued that a full account of a unified speech-gesture production system must represent synchronisation relationships in addition to other conceptual relationships between these modalities. To address this, the thesis has proposed a model, the Extended Interface Hypothesis, as an extension of the Interface Hypothesis. The model integrates synchronisation by establishing links between the relevant stages of speech and gesture production in line with the results of this thesis. This model significantly contributes to our understanding of gesture production as it is the first model to explain a synchronisation mechanism in detail.

More studies are needed to extend the analysis of synchronisation to other languages with different characteristics. Languages can differ in their prosodic systems, how those systems mark IS, and how IS is marked in general. It is possible that all of these could affect gesture synchronisation. Moreover, the claims of this thesis have been mostly about co-speech gestures. More research is needed to explain whether the optionality of speech in emblems and actions have an effect of synchronisation.

It is hoped that this thesis will encourage further studies on gesture-speech synchronisation to adopt a more holistic approach, taking into account linguistic research on prosody and information structure. As seen in this thesis, such comprehensive analyses can reveal systematic synchronisation behaviours between structures in gesture and speech. The thesis also encourages the use of statistically testable synchronisation criteria which result in more transparent synchronisation findings and enable cross-linguistic comparison.

References

Aboudan, R., & Beattie, G. (1996). Cross-cultural similarities in gestures: The deep relationship between gestures and speech which transcends language barriers. *Semiotica*, *111*(3-4), 269–294.

Acredolo, L., & Goodwyn, S. (1988). Symbolic gesturing in normal infants. *Child Development*, 450–466.

Alexanderson, S., House, D., & Beskow, J. (2013). Aspects of co-occurring syllables and head nods in spontaneous dialogue. In S. Ouni, F. Berthommier, & A. Jesse (Eds.), *Proceedings of 12th International Conference on Auditory-Visual Speech Processing* (pp. 169–172). Annecy, France.

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.

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture*, *10*(1), 3–28.

Alibali, M. W., Kita, S., Bigelow, L. J., Wolfman, C. M., & Klein, S. M. (2001). Gesture plays a role in thinking for speaking. In C. Cavé, I. Guaiẗella, & S. Santi (Eds.), *Proceedings of International Oralité et Gestualité Colloquium* (pp. 407–410). Paris, France: L'Harmattan.

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science*, *22*(9), 1138–1144.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3-4), 273–287.

Altmann, H. (2006). *The perception and production of second language stress: A cross-linguistic experimental study* (Unpublished doctoral dissertation). University of Delaware.

Altorfer, A., Jossen, S., Würmle, O., Käsermann, M.-L., Foppa, K., & Zimmermann, H. (2000). Measurement and meaning of head movements in everyday face-to-face communicative interaction. *Behavior Research Methods, Instruments, & Computers*, *32*(1), 17–32.

Ambrazaitis, G., & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, *95*, 100–113.

Ambrazaitis, G., Svensson Lundmark, M., & House, D. (2015). Head beats and eyebrow movements as a function of phonological prominence levels and word accents in Stockholm Swedish news broadcasts. In *Proceedings of the 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing* (p. 42-42). Vienna, Austria.

Arnhold, A. (2014). Prosodic structure and focus realization in West Greenlandic. *Prosodic Typology II*, 216–251.

Arnhold, A., & Kyröläinen, A.-J. (2017). Modelling the interplay of multiple cues in prosodic focus marking. *Laboratory Phonology*, *8*(1), 1–25.

Arvaniti, A., Ladd, D., & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, *26*(1), 3–25.

Atterer, M., & Ladd, D. (2004). On the phonetics and phonology of segmental anchoring of F0: evidence from German. *Journal of Phonetics*, *32*(2), 177–197.

Azar, Z., Backus, A., & Özyürek, A. (2019). General-and language-specific factors influence reference tracking in speech and gesture in discourse. *Discourse Processes*, *56*(7), 553–574.

Azar, Z., & Özyürek, A. (2015). Discourse management: Reference tracking in speech and gesture in Turkish narratives. *Dutch Journal of Applied Linguistics*, *4*(2), 222–240.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415–419.

Barry, W., & Andreeva, B. (2001). Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association*, *31*(1), 51–66.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Baumann, S. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In G. Elordieta & P. P (Eds.), (Vol. 25, pp. 119–162). Berlin: Mouton De Gruyter.

Baumann, S., Grice, M., & Steindamm, S. (2006). Prosodic marking of focus domains-categorical or gradient. In H. M. Rüdiger Hoffman and (Ed.), *Proceedings of the 3rd International Conference on Speech Prosody* (pp. 301–304). Dresden: TUD Press.

Baumann, S., & Kügler, F. (2015). Prosody and information status in typological perspective. *Lingua: International Review of General Linguistics*, *165*(B), 179–182.

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, *70*, 20–38.

Bavelas, J., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, *19*(2), 163–194.

Bavelas, J., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, *15*(4), 469–489.

Bavelas, J., Gerwing, J., & Healing, S. (2014). Effect of dialogue on demonstrations: Direct quotations, facial portrayals, hand gestures, and figurative references. *Discourse Processes*, *51*(8), 619–655.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520.

Bavelas, J., & Healing, S. (2013). Reconciling the effects of mutual visibility on gesturing: A review. *Gesture*, *13*(1), 63–92.

Bavelas, J., Kenwood, C., Johnson, T., & Phillips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, *2*(1), 1–17.

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, *18*(4), 438–462.

Beattie, G., & Shovelton, H. (2002). An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal of Psychology*, *93*(2), 179–192.

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling (version 3). *The OSU Research Foundation*, 1–30.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255–309.

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

Beskow, J., Granström, B., & House, D. (2006). Focal accent and facial movements in expressive speech. In *Proceedings of Fonetik* (pp. 9–12). Lund, Sweden.

Birdwhistell, R. L. (1952). *Introduction to kinesics: An annotation system for analysis of body motion and gesture.* Michigan: University of Michigan Press.

Bloomfield, L. (1933). *Language.* New York: Holt.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127.

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer* [Computer Software]. Retrieved from `http://www.praat.org/` (version 6.0.56)

Bolinger, D. (1983). Intonation and gesture. *American Speech*, *58*(2), 156–174.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7-9), 1044–1098.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*, *109*(2), 274–280.

Büring, D. (2003). On D-trees, beans, and B-accents. *Linguistics and Philosophy*, *26*(5), 511–545.

Butterworth, B., & Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In R.N.Campbell & P. Smith (Eds.), *Recent Advances in the Psychology of Language: Formal and Experimental Approaches* (pp. 347–360). Plenum: New York.

Caldognetto, E. M., Poggi, I., Cosi, P., Cavicchio, F., & Merola, G. (2004). Multimodal score: An ANVIL based annotation scheme for multimodal audio-video analysis. In J. C. Martin, E. D. Os, P. Kühnlein, L. Boves, P. Paggio, & R. Catizone (Eds.), *Proceedings of Language Resources and Evaluation - Workshop on Multimodal Corpora* (Vol. 25, pp. 29–33). Lisbon, Portugal.

Calhoun, S. (2007). *Information structure and the prosodic structure of English: A probabilistic relationship* (Unpublished doctoral dissertation). University of Edinburgh.

Calhoun, S. (2010a). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, 1–42.

Calhoun, S. (2010b). How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, *25*(7-9), 1099–1140.

Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, *47*(1), 14–27.

Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., ... Pelachaud, C. (1994). Modeling the interaction between speech and gesture. *Technical Reports (CIS)*, 341–351.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and Topic*, 27–55.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, *37*(6), 871–887.

Clark, H. (1996). *Using language.* Cambridge: Cambridge University Press.

Clark, H., & Haviland, S. (1977). Comprehension and the given-new contract. discourse production and comprehension. *Discourse Processes: Advances in Research and Theory*, *1*, 1–40.

Cohen, A. A. (1977). The communicative functions of hand illustrators. *Journal of Communication*, *27*(4), 54–63.

Cohen, A. A., & Harrison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, *28*(2), 276–279.

Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2), 1–31.

Condon, W. S. (1976). An analysis of behavioral organization. *Sign Language Studies*(13), 285–318.

Debreslioska, S., & Gullberg, M. (2019). Discourse reference is bimodal: how information status in speech interacts with presence and viewpoint of gestures. *Discourse Processes*, *56*(1), 41–60.

Debreslioska, S., Özyürek, A., Gullberg, M., & Perniss, P. (2013). Gestural viewpoint signals referent accessibility. *Discourse Processes*, *50*(7), 431–456.

De Ruiter, J. P. (2000). The production of gesture and speech. *Language and Gesture*, *2*, 284–311.

De Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, *8*(2), 124–127.

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

232–248.

de Ruiter, J. P., & de Beer, C. (2013). A critical evaluation of models of gesture and speech production for understanding gesture in aphasia. *Aphasiology*, *27*(9), 1015–1030.

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255–1269.

Dohen, M., & Loevenbruck, H. (2004). Pre-focal rephrasing, focal enhancement and postfocal deaccentuation in French. In *Proceedings of the 8th International Conference on Spoken Language Processing* (p. 1313-1316). Jeju, Korea.

Domahs, U., Genc, S., Knaus, J., Wiese, R., & Kabak, B. (2013). Processing (un-) predictable word stress: ERP evidence from Turkish. *Language and Cognitive Processes*, *28*(3), 335–354.

Downing, L. J., & Pompino-Marschall, B. (2013). The focus prosody of Chichewa and the stress-focus constraint: a response to Samek-Lodovici, 2005. *Natural Language & Linguistic Theory*, *31*(3), 647–681.

Ebert, C., Evert, S., & Wilmes, K. (2011). Focus marking via gestures. In *Proceedings of Sinn und Bedeutung* (Vol. 15, pp. 193–208).

ELAN. (2019). [Computer Software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from `https://archive.mpi.nl/tla/elan` (version 5.8)

Emmorey, K., & Casey, S. (2001). Gesture, thought and spatial language. *Gesture*, *1*(1), 35–50.

Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swerts, M., & Prieto, P. (2017). The timing of head movements: The role of prosodic heads and edges. *Journal of the Acoustical Society of America*, *141*(6), 4727–4739.

Esteve-Gibert, N., Loevenbruck, H., Dohen, M., & D'imperio, M. (2019).

Pre-schoolers use head gestures rather than duration or pitch range to signal narrow focus in French. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences.* Melbourne, Australia.

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

Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, *57*, 301–316.

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In *Proceedings of Language Resources and Evaluation - Workshop on Multimodal Corpora* (pp. 86–91). Valletta, Malta.

Ferré, G. (2014). A multimodal approach to markedness in spoken French. *Speech Communication*, *57*, 268–282.

Ferreira, F., & Swets, B. (2002). How incremental is language production? evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57–84.

Féry, C. (2001). Focus and phrasing in French. In C. Féry & W. Sternefeld (Eds.), *Audiatur vox sapientiae. a festschrift for arnim von stechow* (pp. 153–181). Berlin: Akademie-Verlag.

Féry, C., & Krifka, M. (2008). Information structure: Notional distinctions, ways of expression. *Unity and Diversity of Languages*, 123–136.

Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, *36*(4), 680–703.

Feyereisen, P., & De Lannoy, J.-D. (1991). *Gestures and speech: Psychological investigations.* Cambridge: Cambridge University Press.

Fowler, C. A. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech*, *31*(4), 307–319.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression.* Los Angeles, CA: Sage Publications.

Frota, S., & Prieto, P. (2015). *Intonation in Romance.* Oxford: Oxford University Press.

Fung, H. S. H., & Mok, P. P. K. (2018). Temporal coordination between focus prosody and pointing gestures in Cantonese. *Journal of Phonetics*, *71*, 113–125.

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451.

Ganushchak, L. Y., & Chen, Y. (2016). Incrementality in planning of speech during speaking and reading aloud: Evidence from eye-tracking. *Frontiers in Psychology*, *7*, 33.

Genzel, S., Ishihara, S., & Surányi, B. (2015). The prosodic expression of focus, contrast and givenness: A production study of Hungarian. *Lingua*, *165*, 183–204.

Gerwing, J. (2003). *The effect of immediate communicative function on the physical form of conversational hand gestures* (Unpublished doctoral dissertation). University of Victoria, Canada.

Gerwing, J., & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, *4*(2), 157–195.

Göksel, A., & Özsoy, S. (2000). Is there a focus position in Turkish. In A. Göksel & C. Kerslake (Eds.), *Proceedings of studies on turkish and turkic languages (turkologica 4)* (pp. 219–228). Wiesbaden: Harrassowitz Verlag.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*(11), 419–429.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: using the hand to read the mind. *Psychological Review*, *100*(2), 279–297.

Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where Language, Culture, and Cognition Meet* (pp. 85–107). Mahwah, NJ: Erlbaum.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522.

Goldman-Eisler, F. (1967). Sequential temporal patterns and cognitive processes in speech. *Language and Speech*, *10*(2), 122–132.

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., . . . Stoel, R. (2007). Information structure. *Interdisciplinary Studies on Information Structure*, *7*, 147–187.

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

Gullberg, M. (2006). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. *Language Learning*, *56*(1), 155–196.

Güneş, G. (2013). On the role of prosodic constituency in Turkish. In U. Özge (Ed.), *Proceedings of Workshop on Altaic Formal Linguistics* (Vol. 8, pp. 115–128). Cambridge: MITWPL.

Güneş, G. (2015). *Deriving prosodic structures* (Unpublished doctoral dissertation). University of Groningen.

Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents.* Boston: De Gruyter Mouton.

Gussenhoven, C. (2004). *The phonology of tone and intonation.* Cambridge: Cambridge University Press.

Gut, U., & Milde, J.-T. (2003). Annotation and analysis of conversational gestures in the TASX environment. *KI*, *17*(4), 34.

Harrell Jr, F. E. (2019). rms: Regression modeling strategies [Computer software manual]. (R package version 5.1-3.1)

Hartigan, J. A., Hartigan, P. M., et al. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84.

Hilliard, C., & Cook, S. W. (2016). Bridging gaps in common ground: Speakers design their gestures for their listeners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 91–103.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, *79*, 1–17.

Holler, J., & Bavelas, J. (2017). Multi-modal communication of common ground. In R. B. Church, M. W. Alibali, & S. Kelly (Eds.), *Why gesture? how the hands function in speaking, thinking, and communicating* (pp. 213–240). Amsterdam, The Netherlands: Benjamins.

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, *3*(2), 127–154.

Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, *26*(1), 4–27.

Holler, J., Turner, K., & Varcianna, T. (2013). It's on the tip of my fingers: Co-speech gestures during lexical retrieval in different social contexts. *Language and Cognitive Processes*, *28*(10), 1509–1518.

Holler, J., Tutton, M., & Wilkin, K. (2011). Co-speech gestures in the process of meaning coordination. In *Proceedings of Gesture and Speech in Interaction (GESPIN)* (pp. 5–7). Bielefeld, Germany.

Holler, J., & Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences the representation of semantic information in speech and gesture in a narrative task. *Language and Cognitive Processes*, *24*, 267–289.

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics*, *43*(14), 3522–3536.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Hostetter, A. B., & Alibali, M. W. (2019). Gesture as simulated action:

Revisiting the framework. *Psychonomic Bulletin & Review*, *26*(3), 721–752.

House, D., Alexanderson, S., & Beskow, J. (2015). On the temporal domain of co-speech gestures: syllable, phrase or talk spurt? In *Proceedings of Fonetik 2015* (Vol. 10, pp. 63–68).

Ipek, C. (2011). Phonetic realization of focus with no on-focus pitch range expansion in Turkish. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 140–143). Hong Kong.

Ipek, C., & Jun, S.-A. (2013). Towards a model of intonational phonology of Turkish: Neutral intonation. In *Proceedings of Meetings on Acoustics* (Vol. 19, p. 060230). Montreal, Canada: Acoustical Society of America.

Iverson, J. M., & Goldin-Meadow, S. (1997). What's communication got to do with it? Gesture in children blind from birth. *Developmental Psychology*, *33*(3), 453–467.

Iverson, J. M., & Goldin-Meadow, S. (2001). The resilience of gesture in talk: Gesture in blind speakers and listeners. *Developmental Science*, *4*(4), 416–422.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371.

Iverson, J. M., Tencer, H. L., Lany, J., & Goldin-Meadow, S. (2000). The relation between gesture and speech in congenitally blind and sighted language-learners. *Journal of Nonverbal Behavior*, *24*(2), 105–130.

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, *56*(2), 291–303.

Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. In S. Ishihara, M. Schmitz, & A. Schwarz (Eds.), *Interdisciplinary Studies on Information Structure* (Vol. 3, pp. 199–244). Potsdam: Universitätsverlag.

Jeon, H.-S., & Nolan, F. (2017). Prosodic marking of narrow focus in Seoul Korean. *Laboratory Phonology*, *8*(1), 1–30.

Jun, S.-A., & Kim, H.-S. (2007). VP focus and narrow focus in Korean. In *Proceedings of 16th International Congress of Phonetic Sciences* (pp. 6–10). Saarbrücken, Germany.

Jürgen, S. (1996). How to do things with things: Objects trouvés and symbolization. *Human Studies*, *19*, 365–384.

Kabak, B., & Vogel, I. (2001). The phonological word and stress assignment in Turkish. *Phonology*, *18*(3), 315–360.

Kamali, B. (2011). *Topics at the PF interface of Turkish* (Unpublished doctoral dissertation). Harvard University.

Karpiński, M., Jarmołowicz-Nowikow, E., & Malisz, Z. (2009). Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. *Speech and Language Technology*, *11*, 113–122.

Kendon, A. (1972). Some relationships between body motion and speech. *Studies in Dyadic Communication*, *7*, 177–210.

Kendon, A. (1980). Gesture and speech: two aspects of the process of utterance. In M. R. Key (Ed.), (pp. 207–227). The Hague: Mouton.

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

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, *32*(2), 123–131.

Kipp, M. (2005). *Gesture generation by imitation: From human behavior to computer character animation* (Unpublished doctoral dissertation). Saarland University.

Kipp, M., Neff, M., & Albrecht, I. (2007). An annotation scheme for conversational gestures: How to economically capture timing and form. *Language Resources and Evaluation*, *41*(3-4), 325–339.

Kirchhof, C., & Ruiter, J. (2012). On the audiovisual integration of speech and gesture. In *Proceedings of the 5th Conference of the International Society for Gesture Studies.* Lund, Sweden.

Kirkgoz, Y. (2007). English language teaching in Turkey: Policy changes and their implementations. *Regional Language Centre Journal*, *38*(2),

216–228.

Kita, S. (2000). How representational gestures help speaking. *Language and Gesture*, *1*, 162–185.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167.

Kita, S. (2010). A model of speech-gesture production. In E. Morsella (Ed.), *Expressing Oneself/Expressing One's Self: Communication, Cognition, Language, and Identity.* New York: Psychology Press.

Kita, S. (2014). Production of speech-accompanying gesture. In V. Ferreira, M. Goldrick, & M. Miozzo (Eds.), *Oxford Handbook of Language Production* (pp. 451–459). Oxford: Oxford University Press.

Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes*, *24*(5), 761–775.

Kita, S., & Lausberg, H. (2008). Generation of co-speech gestures based on spatial imagery from the right-hemisphere: Evidence from split-brain patients. *Cortex*, *44*(2), 131–139.

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.

Kita, S., Van Gijn, I., & Van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human-Computer Interaction* (pp. 23–35). Springer.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language*, *66*(1), 143–162.

Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40.

Kopp, S., Bergmann, K., & Wachsmuth, I. (2008). Multimodal communication from multimodal thinking-towards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, *2*(01), 115–136.

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

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: a process. *Language and Gesture*, *2*, 261–291.

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, *31*, 533–552.

Krauss, R. M., & Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. *Gesture, speech, and sign*, *93*.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, *61*(5), 743–754.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243–276.

Kruijff-Korbayová, I., & Steedman, M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, *12*(3), 249–259.

Kügler, F. (2011). *The prosodic expression of focus in typologically unrelated languages.* Postdam: Universität Potsdam, Humanwissenschaftliche Fakultät .

Kügler, F., & Calhoun, S. (in press). Prosodic encoding of information structure: a typological perspective. In C. Gussenhoven & A. Chen (Eds.), *Oxford Handbook of Language Prosody* (chap. 31). Oxford: Oxford University Press.

Kuhlen, A. K., Galati, A., & Brennano, S. E. (2012). Gesturing integrates top-down and bottom-up information: Joint effects of speakers' expectations and addressees' feedback. *Language and Cognition*, *4*(1),

17–41.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Ladd, D. R. (2008). *Intonational phonology.* Cambridge: Cambridge University Press.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841–849.

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

Levelt, W. J. (1989). *Speaking: From intention to articulation.* Cambridge: MIT press.

Levi, S. V. (2005). Acoustic correlates of lexical accent in Turkish. *Journal of the International Phonetic Association*, *35*(1), 73–97.

Levy, E. T., & McNeill, D. (1992). Speech, gesture, and discourse. *Discourse Processes*, *15*(3), 277–301.

Loehr, D. P. (2004). *Gesture and intonation* (Unpublished doctoral dissertation). Georgetown University Washington, DC.

Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld speech and gesture alignment corpus (SAGA) and its applications. *Journal on Multimodal User Interfaces*, *7*(1-2), 5–18.

Martell, C. (2002). Form: An extensible, kinematically-based gesture annotation scheme. In *Proceedings of International Conference on Language Resources and Evaluation.* Las Palmas, Canary Island: European Language Resources Association.

Mayberry, R. I., & Jaques, J. (2000). Gesture production during stuttered speech: insights into the nature of gesture-speech integration. *Language and Gesture*, *2*, 199–218.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using Kaldi. In F. Lacerda, D. House, M. Heldner, & J. Gustafson (Eds.), *Proceedings of Interspeech 2017* (pp. 498–502). Stockholm, Sweden.

McClave, E. Z. (1991). *Intonation and gesture* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

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

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

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

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. *Language and Gesture*, 141–161.

McNeill, D., Quek, F., McCullough, K.-E., Duncan, S. D., Furuyama, N., Bryll, R., & Ansari, R. (2001). Catchments, prosody and discourse. *Gesture*, *1*(1), 9–33.

Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes*, *22*(4), 473–500.

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

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2009). The communicative import of gestures: Evidence from a comparative analysis of human–human and human–machine interactions. *Gesture*, *9*(1), 97–126.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2011). Seeing and being seen: The effects on gesture production. *Journal of Computer-Mediated Communication*, *17*(1), 77–100.

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*(3), 615.

Moutinho, L., & Hutcheson, G. D. (2011). *The SAGE dictionary of quantitative management research.* New Delhi: SAGE.

Myrberg, S., & Riad, T. (2016). On the expression of focus in the metrical grid and in the prosodic hierarchy. In C. Fry & S. Ishihara (Eds.), *Oxford Handbook of Information Structure.* Oxford University Press.

Novack, M. A., & Goldin-Meadow, S. (2017). Gesture as representational action: A paper about function. *Psychonomic Bulletin & Review*, *24*(3), 652–665.

Özçalışkan, Ş., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, *96*(3), 101–113.

Özge, U., & Bozsahin, C. (2010). Intonation in the grammar of Turkish. *Lingua*, *120*(1), 132–175.

Özyürek, A. (2002). Speech-gesture synchrony in typologically different languages and second language acquisition. In B. Skarabela, S. Fish, & J. Do (Eds.), *Proceedings of the 27th Boston University Conference on Language Development* (pp. 500–509). Somerville, MA: Cascadilla Press.

Özyürek, A., Kita, S., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2008). Development of cross-linguistic variation in speech and gesture: Motion events in English and Turkish. *Developmental Psychology*, *44*(4), 1040.

Peeters, D., Chu, M., Holler, J., Hagoort, P., & Özyürek, A. (2015). Electrophysiological and kinematic correlates of communicative intent in the planning and production of pointing gestures and speech. *Journal of Cognitive Neuroscience*, *27*(12), 2352–2368.

Perniss, P., & Özyürek, A. (2015). Visible cohesion: A comparison of reference tracking in sign, speech, and co-speech gesture. *Topics in Cognitive Science*, *7*(1), 36–60.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English into-*

nation (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Pine, K. J., Gurney, D. J., & Fletcher, B. (2010). The semantic specificity hypothesis: When gestures do not depend upon the presence of a listener. *Journal of Nonverbal Behavior*, *34*(3), 169–178.

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

Prieto, P., & Torreira, F. (2007). The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics*, *35*(4), 473–500.

Prieto, P., Van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in spanish. *Journal of Phonetics*, *23*(4), 429–451.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Retrieved from `https://www.R-project.org/`

Rimé, B. (1982). The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology*, *12*(2), 113–129.

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *56*(6), 1507–1521.

Röhr, C. T., & Baumann, S. (2011). Decoding information status by type and position of accent in German. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 1706–1709). Hong Kong.

Rohrer, P. L., Prieto, P., & Delais-Roussarie, E. (2019). Beat gestures and prosodic domain marking in French. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1500–1504). Melbourne, Australia.

Roustan, B., & Dohen, M. (2010). Gesture and speech coordination: The influence of the relationship between manual gesture and speech. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the 11th Annual Conference of the International Speech Communication Association.* Makuhari, Japan.

Rusiewicz, H. L. (2010). *The role of prosodic stress and speech perturbation on the temporal synchronization of speech and deictic gestures* (Unpublished doctoral dissertation). University of Pittsburgh.

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of prosody and position on the timing of deictic gestures. *Journal of Speech, Language, and Hearing Research*, *56*(2), 458-473.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1979). A simplest systematics for the organization of turn taking for conversation. *Language*, *50*, 696–735.

Salem, M., Kopp, S., Wachsmuth, I., & Joublin, F. (2009). Towards meaningful robot gesture. In H. Ritter, G. Sagerer, & M. Buss (Eds.), *Human Centered Robot Systems* (pp. 173–182). Berlin: Springer.

Schegloff, E. (1984). On some gestures' relation to speech. In J. Atkinson & J. Heritage (Eds.), *Structures of Social Action: Studies in Conversational Analysis* (pp. 266–296).

Schubotz, L., Holler, J., & Özyürek, A. (2015). Age-related differences in multi-modal audience design: Young, but not old speakers, adapt speech and gestures to their addressee's knowledge. In G. Ferre & M. Tutton (Eds.), *Proceedings of the 4th Gesture and Speech in Interaction Conference.* Nantes, France.

Selkirk, E. (1980). On prosodic structure and its relation to syntactic structure. In T. Fretheim (Ed.), *Nordic prosody II.* Trondheim: TAPIR.

Selkirk, E. (1984). *Phonology and syntax: The relationship between sound and structure.* Cambridge: MIT Press.

Sezer, E. (1981). On non-final stress in Turkish. *Journal of Turkish Studies*, *5*, 61-69.

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

Shattuck-Hufnagel, S., Ren, A., Mathew, M., Yuen, I., Demuth, K., et al. (2016). Non-referential gestures in adult and child speech: Are they prosodic? In *Proceedings of the 8th International Conference on Speech Prosody* (pp. 836–839). Boston, MA.

Shattuck-Hufnagel, S., Ren, P. L., & Tauscher, E. (2010). Are torso movements during speech timed with intonational phrases? In *Proceedings from the 8th International Conference on Speech Prosody.* Chicago, IL.

Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., & Renwick, M. (2007). A method for studying the time alignment of gestures and prosody in American English: Hits and pitch accents in academic-lecture-style speech. In A. Esposito, M. Bratanic, E. Keller, & M. Marinaro (Eds.), *NATO security through science series E human and societal dynamics* (Vol. 18). Washington, DC: IOS PRESS.

Skopeteas, S., Fiedler, I., Hellmuth, S., Schwarz, A., Stoel, R., Fanselow, G., . . . Krifka, M. (2006). *Questionnaire on information structure (QUIS): Reference manual* (Vol. 4). Universitätsverlag Potsdam.

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive science*, *33*(1), 115–125.

Starkey, D., & Fiske, D. W. (1977). *Face-to-face interaction: Research, methods, and theory.* New York: Lawrence Erlbaum Associates.

Steedman, M. (2014). The surface-compositional semantics of English intonation. *Language*, *90*, 2–57.

Trippel, T., Gibbon, D., Thies, A., Milde, J.-T., Looks, K., Hell, B., & Gut, U. (2004). CoGesT: a formal transcription system for conversational gesture. In *Proceedings of Language Resources and Evaluation.* Lisbon, Portugal.

Tuite, K. (1993). The production of gesture. *Semiotica*, *93*(1-2), 83–106.

Vallduví, E. (2016). Information structure. In M. Aloni & P. Dekker (Eds.), *The cambridge handbook of formal semantics* (p. 728-755). Cambridge University Press.

Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, *34*(3), 459–520.

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, *44*(3), 145–174.

Van Welbergen, H., Reidsma, D., & Kopp, S. (2012). An incremental multimodal realizer for behavior co-articulation and coordination. In *Proceedings of International Conference on Intelligent Virtual Agents* (pp. 175–188). Berlin: Springer.

Wagner, A. (2008). *A comprehensive model of intonation for application in speech synthesis* (Unpublished doctoral dissertation). Adam Mickiewicz University, Poznan, Poland.

Wagner, P., & Bryhadyr, N. (2017). Mutual visibility and information structure enhance synchrony between speech and co-speech movements. *Journal of Multimodal Communication Studies*, *4*(1-2), 69-74.

Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 423.

Wang, B., & Xu, Y. (2011). Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. *Journal of Phonetics*, *39*(4), 595–611.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs.L+H*. *Cognitive Science*, *32*(7), 1232–1244.

Wheeldon, L. (2013). Producing spoken sentences: The scope of incremental planning. In S. Fuchs, M. Weirich, D. Pape, & P. Perrier (Eds.), *Speech Production and Perception: Speech Planning and Dynamics* (pp. 97–118).

Wlodarczak, M., Buschmeier, H., Malisz, Z., Kopp, S., & Wagner, P. (2012). Listener head gestures and verbal feedback expressions in a distraction task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialogue.* Stevenson, WA: Skamania Lodge.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, *55*(4), 179–203.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, *27*(1), 55–105.

Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. In *Proceedings of From Sound to Sense* (Vol. 50, pp. 97–102). Cambridge: MIT.

Yoshioka, K. (2008). Gesture and information structure in first and second language. *Gesture*, *8*(2), 236–255.

Zimmermann, M., & Féry, C. (2010). *Information structure: Theoretical, typological, and experimental perspectives.* New York: Oxford University Press.