This paper was presented at the 6th Gesture and Speech in Interaction Conference that was held at Paderborn University, Germany from September 11-13, 2019.

To cite this paper:

Turk, O. (2019). Does gestural hierarchy align in time with prosodic hierarchy? Another modality to consider: Information structure. In: Grimminger, A. (Ed.): *Proceedings of the 6th Gesture and Speech in Interaction – GESPIN 6* (pp. 87-92). Paderborn: Universitaetsbibliothek Paderborn. doi: 10.17619/UNIPB/1-811

Does gestural hierarchy align in time with prosodic hierarchy? Another modality to consider: Information structure

Olcay Turk

Victoria University of Wellington, School of Linguistics and Applied Language Studies, New Zealand

olcay.turk@vuw.ac.nz

Abstract

This study investigates the coordination of gesture with prosody and information structure in Turkish. It has long been known that gesture has a hierarchical structure like prosody. It is also known that gesture is coordinated with prosody on a prominence-related micro level, but less is known about whether this coordination persists at higher levels in the hierarchies. Even less is known about a possible timing relationship to a modality that is also signalled by prosody – information structure. 3 hours of natural speech data was acquired from the narrations of four participants. The study tests the temporal coordination of gesture phrases with multiple levels of phrases within the prosodic hierarchy as well as with information structural units (e.g., topic/focus) that informs the prosodic phrasing. The results show that the hierarchy of alignment is preserved and gesture phrases align with the corresponding prosodic phrases. Information structure units and gesture phrases do not show perfect alignment, but there was a systematic overlap where complete gesture phrases contained the information structure units. Gesture phrase medial stroke + post-hold combinations provided a better anchor for alignment. Overall, the findings confirm multiple levels of alignment between hierarchical structures of gesture and prosody as well as providing empirical evidence for the claim that gesture is informed by information structure in addition to traditional semantic, pragmatic and phonological modalities.

1. Introduction

Speech and gesture have a close relationship in daily human communication; however, the exact nature of their temporal coordination has not yet been fully uncovered. McNeill (1992) suggested three rules that govern the coordination between these modalities: the semantic, pragmatic and phonological synchronization rules. In the light of these, there have been a number of studies investigating the temporal coordination linking prosody to gesture (for an overview, see Wagner, Malisz, and Kopp, 2014) and these studies agree that prominences in prosody and gesture are temporally coordinated. Studies on timing relations have concentrated on prominence-related atomic landmarks at the lowest level within continuous streams of prosody and gesture, but is gesture coordinated with prosody at higher levels and if so what are these larger units that coordinate with gesture?

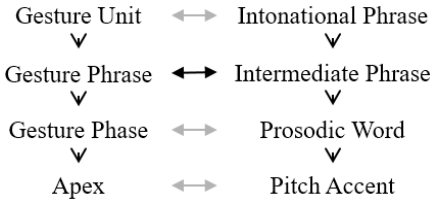

Figure 1. Mapping of gestural and prosodic hierarchies.

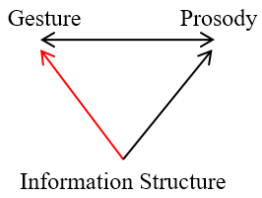

Figure 2. Three-way coordination in production.

Prosodic phrases as described by the current standard phonological framework, Autosegmental-Metrical (AM) model (Ladd, 2008), share structural similarities with gestural structure. That is, they both consist of hierarchically-organized units (see Figure 1) based around an obligatory prominent event, i.e., stroke and nucleus. AM model defines at least two levels of phrasing nested within each other. The terms used differ for each; this study uses intermediate phrase (ip) and

intonational phrase (IP). IPs and ips (a constituent of IPs) are defined based on the degree of juncture/break felt after the phrase offset (greater in IPs) and language specific pitch contours. An ip consists of at least one prosodic word and an IP consists of at least one ip. An ip roughly corresponds to a phrasal syntactic constituent but an IP to a sentence (see Figure 3 for an example of nested phrases).

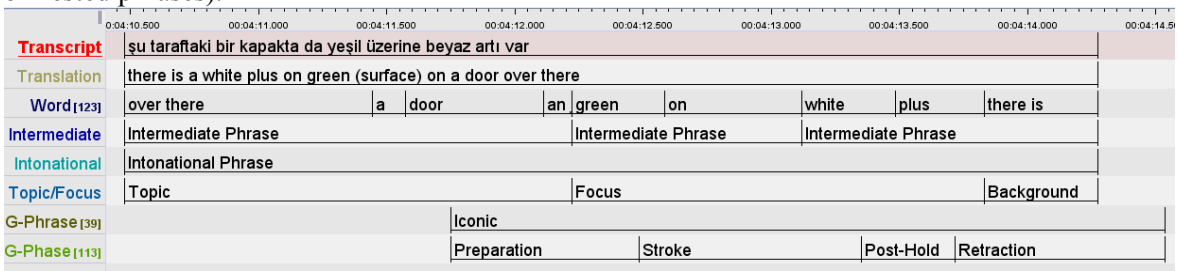

Figure 3. An annotation example showing how different phrases can be mapped onto each other. Prosodic and information structural unit boundaries are marked at the orthographic word boundaries.

Only a few studies have investigated the temporal coordination of gesture and prosody using this phrasing structure (or models similar to AM). For English, Loehr (2004) found that single gesture phrases (GPs) are typically coordinated with single ips, and it was often the case that there were multiple GPs within the span of a single ip. In those cases, their boundaries were sensitive to each other, meaning that GP boundaries occurred within the ip. Unlike Loehr, Ferré (2010) found that in French, GPs overlap with ips, that is, GPs start before their relevant IPs, and end after them. For Polish, Karpiński, Jarmołowicz-Nowikow and Malisz (2009) showed that ips are not temporally coordinated with GPs. A similar investigation of Turkish for such alignment is interesting because of its prosodic structure. In Turkish, prosodic words (see Figure 1) often form their own ips, i.e., there is often only one prosodic word in an ip (see Ipek and Jun, 2013; Kamali, 2011); therefore, they can have a relatively short duration. This duration may potentially be too short for any coordination with the GPs, leading to a different coordination pattern.

The difference in the results of the previous studies may imply that the temporal coordination shows variation depending on the language investigated. Another implication may be that the coordination of gesture with prosody at higher phrasal levels is regulated by another modality, which naturally has linguistic interaction with the prosody of speech. From a gestural point of view, McNeill (1992) and McNeill and Duncan (2000) argue that speech and gesture stem from the same minimal idea units (i.e., growth points) which aim 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: 220). These explanations for the origin of gesture have a lot in common with topic/focus in information structure (IS). IS describes the prominence and organization of information in relation to a discourse, which operates in 3 dimensions: information status, topic/focus, and contrast (Götze et al, 2007). Only topic/focus is investigated in the present study. Topic is the part of an utterance that relates it to previous discourse by setting a frame or by informing what the utterance is about ("on a door over there" in Figure 3), and focus (i.e., new information focus) is the part that carries the discourse forward by introducing new information ("there is a white plus on green" in Figure 3). IS is a relevant modality for gesture alignment also due to its relationship with prosody. Prosody is one of the principal cues to IS for many languages including Turkish (Özge and Bozşsahin, 2010). Topic/focus has been shown to be associated with prosodic features. For instance, topic/focus status decides which pitch accent type a prosodic unit gets; focal area of an utterance includes the prosodically most prominent unit; and more importantly for the present study, prosodic phrasing is sensitive to topic/focus boundaries (Özge and Bozşsahin, 2010; Steedman, 2000).

As shown in prosody-gesture coordination studies above, the coordination of the prosodic and gestural hierarchies seems not to be perfect at the phrasal level. If GPs can span multiple ips or IPs, then this may be linked to potentially larger structures governing alignment, such as topics and foci which can contain multiple prosodic phrases (see Figure 3). A temporal coordination between focus and gesture was assumed before in 3D interactive animation modelling but there was no empirical evidence of such a relationship (see Cassell et al., 1994). To the author's knowledge, the only study that investigates the temporal coordination of IS units with GPs is Ebert, Evert and Wilmes's (2011)

study. Using data in German, they checked whether "focus phrases" are coordinated with GPs. Interestingly, they treated the end of the stroke as the offset of the GP and excluded post-hold and retraction phrases claiming they are semantically empty or "they seem to have a different status as the other phases of a GP" (p. 7) following Loehr's study. They found that GPs start on average 310 milliseconds (ms) earlier than focus phrases but the offsets were not coordinated at all.

The few studies which investigated temporal coordination at phrasal level show different results. The present study aims to contribute to this body of research by investigating a language with particular prosodic structure which can lead to variation in the coordination patterns from those previously observed with other languages. The study looks for the prosodic phrase defined within AM model that is temporally best coordinated with GPs, using Turkish natural speech data. Based on the findings of previous research, this study tests the hypothesis that the domain of coordination for GPs is either the ip or the IP as defined in the AM model. The study postulates that because of the short duration of the ips in Turkish, the alignment between GPs and ips will not be perfect but GPs will display a form of coordination with ips as the most likely candidate in the prosodic hierarchy (see Figure 1). The study also explores a potential coordination between topic/focus areas and GPs by checking whether focus areas as well as topic areas start and end around the same time as GPs. If this is true, then it would introduce information structure as another aspect that governs the coordination of gesture and speech in addition to the traditional semantic, pragmatic and phonological aspects (see Figure 2).

2. Methods

The participants were 4 (2 male, 2 female) 18-25 year-olds who are monolingual native speakers of Turkish. One male confederate listener with the same profile as the participants was also employed. The stimuli consisted of 10 video clips (15-40 secs) where real life actors performed basic daily activities (e.g., passing a book to another) each telling a different story. The participants were shown a video and were asked to recount what they had seen to the confederate listener. The confederate functions to present a communicative target to the participant in order to make the task more meaningful. The confederate could talk and nod freely to reinforce communication but his gestures were not included in the analysis.

3 hours of narrations were video recorded at 60fps. Declarative utterances that contained no speech errors and were accompanied by uninterrupted gestures were randomly sampled for annotation. The annotation of gestures was done in ELAN (Lausberg and Sloetjes, 2009) based on the guidelines in McNeill (1992). The present study considered the offset of the final gesture phase within a GP as the offset, regardless of it being the offset of the stroke or the retraction. Only imagistic gestures (i.e., deictic, iconic and metaphoric) were included in the analysis as only these can bear the same semantic content as speech. The annotation of prosody and IS was done in Praat (Boersma and Weenink, 2019). The annotation of prosody followed Tones and Breaks Indices guidelines where the boundaries between prosodic phrases are defined based on intonation patterns, and breaks or sense of juncture felt at the edge of the prosodic phrases. The annotation scheme for prosody was developed based on the earlier studies on Turkish (Ipek, 2013; Kamali, 2011). The annotation of topic/focus was followed Götze et al. (2007), with the addition of the category "background" for the chunks of utterances that do not qualify as topic or focus (these are left out of the annotation in their scheme). In the data, the total duration of gesture annotation was 20 mins which included 589 GPs. Within this duration 1363 ips and 675 IPs were also annotated. For IS units, the numbers were: 387 topics, 540 foci, and 133 backgrounds. The study tests coordination based on the distance between the nearest relevant annotations of units regardless of their semantic alignment (e.g., nearest ip offset time - GP offset time = offset distance). There is no set number in the literature explaining how near these annotations should be in order to be considered aligned. This study uses the average syllable duration, 160 ms. The cases where an IP included only one ip were excluded from analysis. This study looks for the most suitable prosodic phrase for coordination and such coincidence of boundaries of IPs and ips does not serve this purpose as a possible alignment can be attributed to both the IPs and the ips. At every step of the analysis, the effect of the type of IS unit (topic/focus), gesture type, and ip type (pre-,post-, nuclear ips) on the onset/offsets distances was tested but left out of this paper due to space restrictions.

3. Results

3.1. Alignment with intermediate phrases

Figure 4 shows the distribution of onset/offset distances of GPs from the nearest ip onset/offsets. The negative values on the x-axis show the instances where ip onsets/offsets precede those of GPs. On average, GPs start 70 ms earlier and end 150 ms earlier than ips. A TOST (two one sided t-tests) equivalence test (Lakens, 2017) was used for the statistical analysis. The test checked whether observed time differences (i.e., distances) between GP onsets/offsets and those of ips are statistically equivalent to zero, being the perfect alignment condition. This is done by testing whether the 95% confidence intervals of the mean distance falls within the set equivalence bounds of -160ms and 160 ms. The equivalence test was significant for onsets ($t_{Upper}(513)=-5.75, p < .001$; $t_{Lower}(513)=14.1, p < .001$) and for offsets ($t_{Upper}(487)=-10.4\ p < .001$; $t_{Lower}(487)=9.85, p < .001$) for all participants. Overall, it can be concluded that GP onsets/offsets co-occur in time with those of ips.

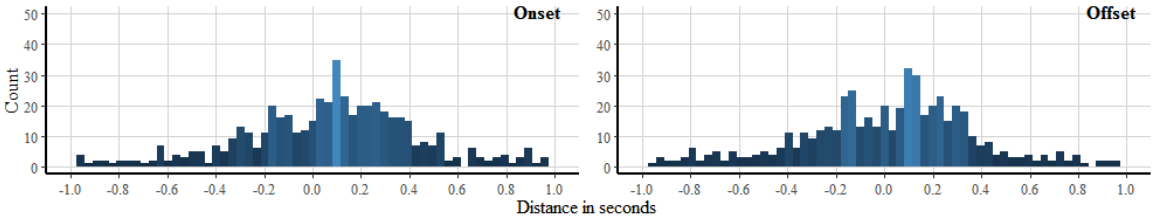

Figure 4. The coordination of ip onsets/offsets with GPs onset/offsets.

3.2. Alignment with intonational phrases

Figure 5 shows the distribution of onset/offset distances of GPs from the nearest IP onset/offsets. The distribution was spread more widely than ips with no apparent peaks observed. In addition, approximately 23% of IP onsets (n=160) and 27% of IPs offsets (n=186) were more than 1s away from the nearest GP onset/offset. Therefore, no further analysis was done and it was concluded that GPs are not coordinated with IPs in Turkish.

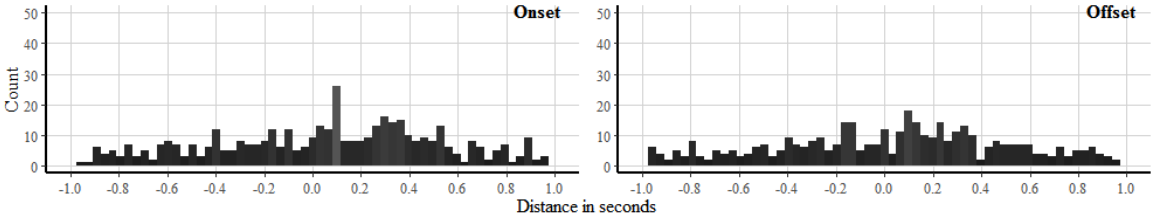

Figure 5. The coordination of IP onsets/offsets with GPs onset/offsets.

3.3. Alignment with topic/focus structure

Confirming the prediction of the growth point theory, GPs tended to mostly co-occur with focus (68%, n=340), followed by topic (27%, n=136) and background (5%, n=27). Figure 6 shows the distribution of onset/offset distances of GPs from the nearest IS unit onset/offsets, regardless of IS unit type. There is a clear compact peak for the onsets (m=413ms, sd=373ms) showing that GPs precede their relevant IS unit by about a word duration on average (390 ms). There is also a minor peak observed for offsets; however, the distribution spreads away from the peak towards the negative values with higher deviation (*m*=-196ms, *sd*=676ms). The equivalence test results for all participants were non-significant both for onsets ($t_{Upper}(499)=15.12$, $p=1.0$; $t_{Lower}(499)=34.4$, $p < .001$) as the upper bound (t_U) was crossed; and for offsets ($t_{Upper}(499)=-11.8\ p < .001$; $t_{Lower}(499)=-1.17$, $p = 0.88$) as the lower bound (t_L) was crossed. This shows that the onset/offset distances were statistically different from zero, therefore IS unit onsets/offsets do not co-occur with GP onsets/offsets. However, the presence of a clear peak may imply a systematic shift for the onsets. Therefore, another equivalence test was applied—this time centering the alignment check on the mean word duration (390 ms) instead of zero in order to match the distribution's peak (i.e., the distances within +/-160 ms from 390 ms are considered aligned). The results were significant ($t_U(499)=-8.19, p < .001$; $t_L(499)=11.0, p < .001$), confirming that the distribution around the peak

was tight enough to consider that there is a displaced alignment (390 ms) between GP onsets and IS unit onsets.

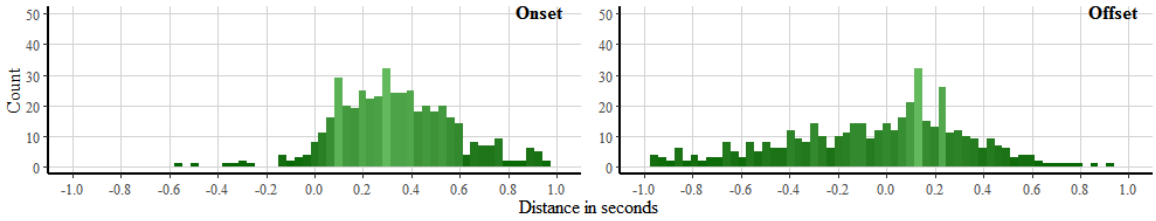

Figure 6. The coordination of topics/foci onsets/offsets with GPs onset/offsets.

3.4. Apical area

The results above indicate that GPs contain topics/foci by starting early and ending later. However, there were peaks observed for both onsets and offsets distances of IS-GP alignment. In addition, the mean of distances for onsets was approximately a phase duration (m=479ms). For offsets, although the standard deviation was high, there was a negative mean (m=-196ms) with 10% of the matches occurring outside of -1s. These may be interpreted as presence of a systematic shift (at least for onsets) in that topics/foci may align with units inside the GPs. As an attempt to find a more refined alignment pattern, the IS-GP alignment was further checked by changing the GP onset/offset. As the core of the GP, the stroke's onset was taken as the GP onset. The offset of the stroke or, if present, the offset of the post-hold was taken as the GP offset. This meant that preparation and retraction phases was ignored for the alignment. This GP central combination, i.e., stroke + (post-hold), contains the meaningful core of the GP and the dynamically most prominent target of the stroke that has been shown to be coordinated with prosodic prominence, the apex (Loehr, 2004). By definition, the post-hold is an apex frozen in time because for most strokes the apex (i.e., the target) is at end of the stroke, which makes the post-hold not as semantically empty as the retraction (cf. Ebert et al., 2011). This combination of phases will be referred as the apical area (AA).

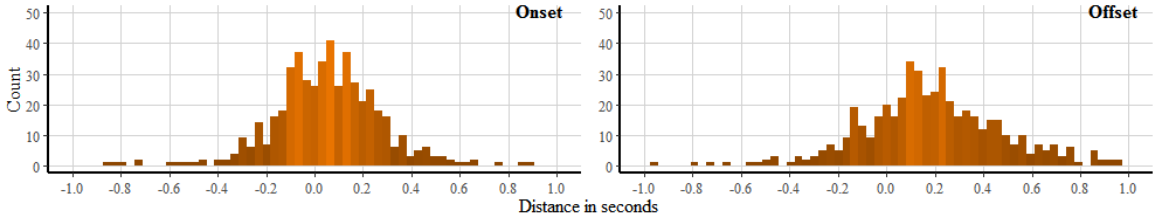

Figure 7. The coordination of topics/foci onsets/offsets with apical area onset/offsets.

Figure 7 shows the distribution of onset/offset distances of AAs as from the nearest IS unit onset/offsets. For the onsets, there was a clear peak with the mean centred very close to zero (m=60ms, sd=551ms). For the offsets, the leftward spread disappeared and the peak appeared more compact (m=196ms, sd=406ms). The equivalence test results were significant for the onsets, showing that the distances were between the set bounds and not statistically different from zero ($t_{Upper}(504)=-4.01, < .001$; $t_{Lower}(504)=9.05, p < .001$). However, the results were non-significant for the offsets as the confidence interval crossed the upper bound by approximately 60 ms, ($t_{Upper}(504)=1.98, p = 0.976$; $t_{Lower}(504)=19.7, p < .001$). These results were consistent for 3 out of 4 participants. However, since there was a clear peak in the distribution, another equivalence test was applied centring the alignment on average syllable duration (160 ms) to account for a shift. The results were significant ($t_{Upper}(504)=-8.87, p < .001$; $t_{Lower}(504)=10.8, p < .001$) for all participants, confirming that the distribution around the peak was tight enough to consider that there is a slightly displaced alignment between AA offsets and IS unit offsets where IS unit offsets end about a syllable duration later than AA's.

4. Discussion

The results show that the ip is the most suitable candidate for coordination with the GP in the prosodic hierarchy of Turkish. The coordination at this level is manifested by the co-occurrence of boundaries, as the durational differences between phrases affect a complete one-to-one alignment. Although more research is required, it seems that the prosodic structural constraints (e.g., the duration of phrases) affect the coordination, which implies a possible variation in the coordination patterns depending on the language investigated. One important note here is that no shift in the alignment hierarchy was observed, in that GPs did not go a level up in the prosodic hierarchy and align with IPs when ips are not suitable for a complete alignment. Instead, the ip-GP boundaries remained temporally sensitive to each other, regardless of how many ips take place between the GP onset and the offset. This way, the hierarchy of alignment was preserved. GPs freely spanning over multiple ips hints at potentially larger structures governing the coordination. IS units are ideal targets for GPs because (1) they typically contain multiple ips in Turkish following their linear ordering. That is, sentence initial topics typically contain multiple pre-nuclear ips; focus areas contain the nuclear ip and pre-nuclear ip(s) (i.e., predicate), and backgrounds contain post-nuclear ip(s). (2) IS units have a shorter duration than IPs. Typically, a combination of topic+focus+background makes up an IP. (3) IS units provide the new and newsworthy information that can be highlighted. The results presented here support the growth point theory as the GPs tended to co-occur with focus over the other IS unit types and the boundaries of these units were temporally coordinated. It is possible to talk to about a gesture-IS coordination at GP level in that there is a displaced alignment between complete units. The study also shows that IS units tightly align with meaningful, well-defined units (AAs) within the GP. Overall, this research contributes to showing hierarchical relationships between speech and gesture at multiple levels (see Figure 2) and concludes that IS could be another level that links gesture and speech in addition to the ones included in McNeillian synchronization rules.

References

Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.48, retrieved 17 February 2019 from http://www.praat.org/

Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., & Achorn, B. (1994). Modeling the interaction between speech and gesture. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In *Language Resources and Evaluation Conference (LREC)*. Workshop on Multimodal Corpora, Malta.

Götze, M., Cornelia, E., Hinterwimmer, S., Fiedler, I., Petrova, S., Schwarz, A., Skopeteas, S., Stoel, R. & Weskott, T. (2007). Information structure. In S. Dipper, M. Götze & S. Skopeteas (Eds.), *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, 147–187. Potsdam: Universitätsverlag Potsdam.

Ipek, C., & Jun, S.-A. (2013). Towards a model of intonational phonology of Turkish: neutral intonation. In *Proceedings of Meetings on Acoustics (POMA), 19*, 060230-069238.

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.

Ladd, D. R. (2008). *Intonational phonology*. 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, Instruments, & Computers, 41*(3), 841-849. doi:10.3758/BRM.41.3.591.

Loehr, D. (2004). *Gesture and Intonation.* (Unpublished doctoral dissertation). Georgetown University, Washington D.C.

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

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

Özge, U., & Bozsahin, C. (2010). Intonation in the grammar of Turkish. *Lingua, 120*(1), 132-175.

Steedman, M. (2000). Information structure and syntax-phonology interface. *Linguistic Inquiry, 31*(4), 641-689.

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