Multimodal marking of information structure: gesture-prosody alignment across languages

Eva Lehečková (Charles University, Prague) —
Jakub Jehlička (Charles University, Prague) —
Magdalena Králová Zíková (Charles University, Prague)

ABSTRACT
In this paper, we first review the existing evidence of gesture-prosody alignment in information structure marking, focusing on specific gestural patterns that were observed to co-occur with various information structure constructions. Then we complement the evidence with the results of a corpus-based study of gesture-speech alignment in Czech. Analyzing a sample of 80 minutes of personal narratives by 16 speakers collected from a Czech multimodal corpus, we observed that by far the most frequent information structure units accompanied by gestures were foci. In line with previous research, we observed that pitch and intensity peaks lag behind the gesture stroke onset (on average by 300 ms). We also provide new evidence for a systematic variation in the duration of the temporal shift related to the marking of discourse contrast.

KEYWORDS
Gesture, prosody, information structure, multimodality, gesture-speech integration

DOI
https://doi.org/10.14712/18059635.2022.1.2

1 INTRODUCTION

Information structure may be marked by a wide range of linguistic means: lexical (e.g. topic and focus particles), syntactic (word order variation or cleft constructions) or prosodic (sentence stress). Additionally, close attention has been paid to the role of co-speech gestures (Kendon 2004; McNeill 2005; Streeck 2009) in information structure marking. Systematic alignment of gesture phrases and speech was observed in a number of languages, characterized by a temporal shift between the apex of gesture phrase (defined as the movement peak of the gesture stroke phase) and the F0 peak within the corresponding intonational unit (pitch accent) (Loehr 2004; Karpiński et al. 2009; Ferré 2010; Leonard and Cummins 2011; Shattuck-Hufnagel and Ren 2018, among others).

The existence of the link between prosody, discourse marking and gesture has long been acknowledged. In one of his pioneering studies, Adam Kendon (1972), conducting a microanalysis of an English conversation, noticed that changes in gesticulation mark the boundaries of discourse units as well as prosodic units. More specifically, in his studies of the Neapolitan gestures, Kendon (1995) observed that a certain type of gesture, namely, the famous "grappolo" handshape, often accompanies the topical part of the discourse. However, rather than specific handshapes, it is

the phase structure of the gestural movement that has attracted most of the attention with respect to information structure marking.

In this paper, we focus on the temporal alignment of the salient phase of hand gestures and markers of intonational prominence — pitch and intensity peaks. The paper is organized as follows: first, we discuss the phasal structure of gestural movement and its linkage with prosody as evidenced in current experimental studies on this topic. Then we review crosslinguistic evidence of systematic patterning of prosody, gesture and information structure. Finally, we present our corpus-based study of gesture-speech alignment in relation to information structure in Czech, a first attempt to describe multimodal marking of information structure in this language.

2 GESTURE-SPEECH INTEGRATION

The basic division of the phases of (manual) gestural movement (introduced by Kendon 1980 and later revisited by Kita 1990) carves up the "ideal" progress of the hand from the resting position to the peak of the movement and back to the resting position into three units that constitute the so-called gesture phrase: (i) *preparation*, in which the hand moves to the main position, the handshape is developed; (ii) *stroke* — the main phase of the gesture, its "meaningful" part, salient in terms of movement intensity and velocity, an obligatory element of a gesture (with pre-stroke and post-stroke *hold* phases as optional elements); and (iii) *retraction* — recovery to the resting position.

Figure 1 illustrates the three main phases:

FIGURE 1: Phasal structure of a gesture phrase.

According to Kendon, a *stroke* is an obligatory phase of a *gesture phrase* (1980) — a single gesture — while other phases do not necessarily need to be realized in the flow of gesture production (e.g., strokes are repeated without visible preparation or retraction in between).

The stroke, as the most salient phase, has been of particular interest to gesture researchers focusing on gesture-speech integration. Although the relationship between gesture execution and suprasegmental features of speech has been noted before (Dobrogaev 1931; Birdwhistell 1970; Bolinger 1983), it was only with the advent of digital video that linguists have been able to assess the exact locus of alignment

between the structure of a gesture phrase and the structure of the intonational phrase.

Thus, across spoken languages, the apex of the gestural stroke was found to tend to be temporally aligned with the pitch accent. This was first evidenced in English (Leonard and Cummins 2011; Loehr 2004), later also in French (Ferré 2010), Polish (Karpiński et al. 2009), German (Ebert et al. 2011) and other languages. However, the alignment between pitch and gestural peaks is often not perfect, with some strokes culminating shortly before the intonational peak. This temporal shift has been linked to specific discourse-marking strategies and will be discussed in more detail in the subsequent section.

The question that arises is whether gestures — especially so-called *beats* that are often assumed not to carry any semantic content related to speech at lexical or phrasal level (McNeill and Levy 1982: 285)[1] — are therefore subordinated to speech, serving as a kind of rhythm-maintenance tool, only highlighting the prosodic contour.

Biau and Soto-Faraco's (2013) ERP[2] study showed that in subjects perceiving a naturalistic discourse, words accompanied by beat gestures induce brain responses associated with early stages of auditory processing and speech recognition.[3] A subsequent study (Biau et al. 2018) revealed, in addition to these responses, an effect related to syntactic processing when the stimuli contained syntactic ambiguities — similarly to a previous study by Holle et al. (2012). However, ERP effects of prosody and beat gestures during semantic processing at the lexical level[4] do not seem to be in accord (Zhang et al. 2021). In their ERP study, Dimitrova et al. (2016) focused on the perception of gestures and information structure (focal words). This study provided evidence for gestural marking of sentence focus, as it revealed specific ERP spikes when beats occurred in a non-focal position; this suggests "increased computation costs needed to arrive at a coherent interpretation of the message when beat gesture emphasizes non-focused information, which should not be highlighted" (p. 1266).

The neurolinguistic evidence thus suggests a more complicated picture of gesture-speech integration than just a subordinate role of visual modality. Rather, there seems to be a dynamic bidirectional influence between the two modalities.

1 Originating in McNeill and Levy's paper (1982) and later elaborated, the standard semiotic-functional typology of gestures includes *iconic* gestures — where the relationship between the gesture's formal features and the associated verbal expression is iconic; *metaphoric* — where the iconic mapping is not established *directly* between the form and verbal semantics, but via metaphoric extension; *deictic* — with a deictically referring function (pointing gestures, typically); and *beats* — rhythmic, baton-like movements often associated with (if anything) only the prosodic and/or discourse structure of the associated speech. Rather than discrete *types*, however, the above classification should be understood in terms of *dimensions* that may be realized simultaneously in natural discourse.

2 An encephalography-based method of measuring brain potentials associated with specific impulses. The ERP responses (EEG wave components) are labeled as negative (N) or positive (P) changes of electric potential against a baseline with a time signature (in milliseconds after the impulse event).

3 N100 and P200 waves, respectively.

4 N400 component.

Further light on the synchronization between gestures and speech has been shed by a recent line of research focusing on the interaction between respiration, vocalization and upper limb movement (Pouw, Harrison and Dixon 2020; Pouw, Harrison, Esteve-Gibert et al. 2020). Respiration is responsible for the modulation of intensity as well as, to a limited extent, pitch. Upper limb movement directly affects respiration by creating muscular impulses reaching lungs and other parts of the respiratory system (Pouw, Harrison and Dixon 2020). A gesture production experiment (Pouw, Harrison, Esteve-Gibert et al. 2020) provided robust evidence that the amount of "energy" invested into gesture execution (in terms of velocity and arm extension) correlates with intensity, while the results for pitch require further investigation. These studies suggest that respiration might provide the biomechanical link between gesture and prosody, explaining the bidirectionality in the gesture-speech relationship reported by previous studies. This was the case in the studies of cross-modal effects on gesture perception (e.g. gesture priming leading to a subjective change of intensity perception (Krahmer and Swerts 2007)), or the "manual McGurk effect" — perception of (actually absent) syllable stress induced by the presence of a gesture (Bosker and Peeters 2020). The fact that the two modalities are so closely interlinked has led to a view of *audiovisual prosody* (Swerts and Krahmer 2008), in which gestures and intonation as well as other prosodic means work in synergy for discourse emphasis. Such a multimodal view of prosody involves not only hand gestures but also other visual cues, such as lip movement (Dohen et al. 2009), eye movements (Ito and Speer 2008), eyebrow movement (Kim et al. 2014) or head movements (Esteve-Gibert et al. 2017). While we subscribe to this view, in this paper, we will limit our discussion to *manual gestures*.

3 MULTIMODAL MARKING OF INFORMATION STRUCTURE ACROSS LANGUAGES

A relatively modest number of studies focused on the three-way interaction between information structure, gestures, and prosody.[5] In general, these studies address three aspects of this interaction: (i) the locus and scope of the alignment between gestural, intonational and discourse units; (ii) description of intonation contours associated with specific gestures or information structure categories; (iii) the degree to which different information structure categories attract gestures. Below, we review the evidence gathered by studies on languages that have been explored so far with regard to gesture-intonation-information structure alignment in speech production — English, German and Turkish.

The first investigation of the synchronization between gestures, intonational structure and information status was carried out by Daniel Loehr (Loehr 2004; 2012), who analysed language production of American English (AmE) speakers. Based on a relatively small sample (four subjects, about three minutes of videos capturing free conversations), Loehr primarily aimed at describing the temporal

5 As for prosody, only intonation has been taken into account in these studies.

alignment between gestural and intonational units at various levels. His main finding in this regard was that in AmE, the apex of a gestural stroke, i.e. the maximal point of the movement both in terms of outward extension of the hand as well as its acceleration within the stroke phase, is aligned with the pitch accent within the stressed syllable. At a higher level, he also found that the gesture phrases tended to align with the *intermediate phrases*.[6] As for information structure, Loehr did not provide a systematic analysis, although he observed some tendencies of co-occurrence of specific tones (described in terms of ToBI[7] annotation (Silverman et al. 1992), types of information status and basic gesture types. For instance, he noticed that with *contrastive focus*, a prominent L+H* pitch accent may occur with a rapid change of gestural movement.

A more systematic inquiry into the relation between gestures, ToBI pitch contours and information structure categories in AmE was carried out by Im and Baumann (2020). Instead of focusing on temporal alignment, they only took into account the association between categorial variables. They found the strongest association of gestures with prominent pitch accents (H* and L+H*) and *contrastive* information status (in general, there was a high co-occurrence of gestures with foci). The results of this study, however, should be treated cautiously due to possible idiosyncrasy of the data, based on the production of a single speaker giving a rehearsed TED talk.

Ebert, Evert and Wilmes (2011) investigated the temporal alignment of gesture, pitch and sentence focus in German multimodal corpus data. The material they analyzed was sampled from the SaGa Corpus (Lücking et al. 2010), which contains recordings of an experimental setting in which speakers give spatial instructions. The authors focused not on the apex of gesture strokes, but on the beginning of the stroke phase (to be discussed below). Pitch accent was again annotated based on the ToBI criteria. Sentence focus was coded as either *new-information* focus or *contrastive* focus[8] (Götze et al. 2007). Ebert et al. found that the stroke phases of gestures accompanying the focal part of a sentence begin on average 360 ms before the corresponding pitch accent within the focus domain, which is in line with Loehr's observation about the alignment between the gesture apex and pitch accent, or a slight precedence of the former.[9] However, there was a clear difference between new-information and

6 The notion of "intermediate phrase" was proposed by Beckmann and Pierrehumbert (1986) to label an intonation unit characterized by the presence of a pitch accent but, unlike a full intonational phrase, the absence of a clear boundary tone. A single intonational phrase may thus consist of multiple intermediate phrases.

7 ToBI = Tones and Break Indices. L stands for a tone with a relatively low pitch, H stands for a tone with a relatively high pitch. Asterisk marks the position of a pitch accent.

8 According to Götze et al. 2007, the question word in the sentence *Who is reading a book?* (p. 176) would be an example of a new-information focus, whereas in the example *We do not export but import goods* (p. 179), the two verbs represent two contrastive foci. Contrastive focus is here defined as "that element of the sentence that evokes a notion of contrast to (an element of) another utterance" (p. 178).

9 In his 1972 study, Adam Kendon was the first to report this asynchrony between the production of gestures and the associated verbal expressions.

contrastive foci: in the latter case, the delay was greater (770 ms vs. 310 ms). Yet, due to the limited sample size, the authors refrained from drawing any generalization about the type of focus marking.

In his integrative study, Türk (2020) analyzed narratives elicited from four speakers of Turkish to investigate the alignment between gestural, intonational and information structure units. Prosody and gesture units were approached at two levels. Prosody was annotated at the level of pitch contour using ToBI labels (customized for Turkish), as well as at the level of intonational phrase. Gesture strokes (and other phases) were coded as well as the apices within the strokes. In addition, gesture type (iconic, deictic, metaphoric) was also coded. Information structure coding was based on the scheme proposed by Götze et al. (2007), which was adapted so as to apply the notion of *contrastiveness* not only to foci but to topics as well.

At the phrasal level, Türk found that gesture phrases are aligned with intermediate phrases or, in the case of gesture phrases spanning multiple intermediate phrases, with the boundaries of these groupings. As for information structure, the main finding was that gesture phrases tend to align mostly with focal domains, but there was also an association with gesture type — deictic gestures were related to topicality and contrastiveness marking. In other words, pointing is often used to refer to discourse entities that were already established as well as in order "to exhaust/differentiate the alternatives for elements that existed in the previous discourse" (Türk 2020: 263).

4 THE PRESENT STUDY

In the research reported in the present paper, we conducted a quantitative analysis of the association between temporal-alignment patterns of gesture phrases and intonation phrases on the one hand and the expression of information structure on the other, as attested in spontaneous Czech discourse. Following particularly the study on German by Ebert al. (2011), we have focused on temporal alignment between the beginning of the gesture stroke, its corresponding prosodic (Fo and intensity) peaks and the information structure category.

4.1 CORPUS

The material was sampled from the multimodal corpus CZICO (CZech Interactional COrpus) that is being developed at Charles University in Prague (https://epocc.ff.cuni.cz/czico/). It comprises recordings of individuals and groups in interaction — during conversations and collaborative problem solving. Our material comes from the part of the corpus containing conversations between the administrator and the subject on a pre-selected topic (either personal experience with *maturita* exams (high-school A-levels) or a recollection of the events of the Velvet Revolution of 1989). These conversations were the middle part of the recording session, so the speakers were already accustomed to the recording and produced gestures in a spontaneous and relaxed manner. The speakers were seated on a sofa and wore a lapel microphone. The scene was captured by two HD cameras. We used the recordings of 16 speak-

ers (ten female, six male; mean age 41.6 years[10]) for our analysis, selecting the initial 5 minutes of each session (80 minutes of video in total).

4.2 CODING

Three independent coders identified all instances of co-speech gestures and corresponding intonation units in the material (968) and annotated them for a) gesture stroke onsets, b) prosodic peaks and c) information structure categories:[11]

a) **Gesture strokes**. Since identifying apices requires the assistance of motion-detection technology, we identified stroke phrase onsets instead. In contrast to apices, the beginning of the movement can be distinguished by the human eye relatively reliably (Kita et al. 1998). The gestures were annotated in ELAN (Wittenburg et al. 2006).

   In this study, we do not focus on semiotic-functional aspects of gestures related to information structure marking, given their potential multifunctionality (Kok et al. 2016). Lumping the gestures together into discrete categories would be too reductionist. Since disentangling their functional-semiotic multidimensionality would exceed the scope of this study, we take all kinds of co-speech gestures into account.

b) **Prosody**. To each gesture phrase, a corresponding F0 and intensity peak were first attributed — the coders listened to the audio track while watching the video and identified the position of intonational prominence, then coded F0 and intensity peaks in Praat (Boersma 2001). The reason for investigating intensity in addition to pitch contour is that the pitch-intensity misalignment has been reported to have a number of specific discourse-pragmatic functions (Ward 2019; 2018). Yet, intensity has so far passed unnoticed by gesture researchers. In the present study, we therefore decided to explore the possible association between the misalignment and the gestural marking of information structure as well.

   The gesture-onset tier from ELAN was subsequently imported to Praat so that the temporal shift between F0/intensity peaks and the onset of the gesture stroke could be extrapolated, yielding two additional tiers with the values of F0 and intensity misalignment.

c) **Information structure.** The information status of the emphasized lexical unit or phrase was annotated using two concepts. The first was the *contextual boundedness* category, derived from the Prague Dependency Treebank annotation guidelines (Mikulová et al. 2005). Contextual boundedness involves two distinctions: given/known vs. non-given/new information. Given information appears in the previous discourse or can be inferred from it, whereas the non-given cannot. Contex-

10 The CZICO corpus contains recordings of speakers from three cohorts: 18–29 years, 30–55 years and 55+. In the sample for the present study, the three cohorts are represented proportionally (6:5:5).

11 The dataset and R scripts are available at <https://osf.io/xq5dh/>.

tually bounded units correspond to *topics*, contextually unbounded instances are referred to as *foci*.

The second aspect of information structure reflected in our annotation is *contrastiveness*: instances coded as contrastive involve an *implicit* or *explicit* profiling of a choice among alternatives (Chafe 1976; Van Valin 2005).

In this respect, we divert from the previous studies, in which *contrastiveness* was defined in terms of lexically encoded contrast (see Götze et al. 2007). We adopt a cognitive perspective, building upon a cognitive process of contrasting (Lambrecht 1994: 291) between contextually known alternatives. The contrastive construal may be either explicit (as in the example of contrastive focus cited in footnote no. 8) or implicit (as in the example (2) below).[12]

The bidimensional annotation of information structure categories can be combined, yielding four possible values: *contrastive topic*, *non-contrastive topic*, *contrastive focus* and *new-information focus*. Examples (1) and (2) illustrate the coding of information structure. Words containing the intonational emphasis associated with a gestural stroke are provided with an information structure annotation in the third line. The still pictures (Figures 2–11) capture the apices of the gesture strokes.

(1) SUBJ05:

Já	*jsem*	*maturova-l-a*	*na*	*hotelov-ý*	*škol-e,*	*tak-že*
1SG	COP	take.exam-PST-F.SG	on	hotel-GEN.SG	school-LOC.SG	so-that

jsem	*mě-l-a*	*jako-by*	*praktick-ou*	*část*	*maturit-y,*	*což*
COP	have-PST-SG.F	as-if	practical-ACC.SG.F	part.ACC.SG	exam-GEN.SG	which

se	*sestáva-l-o*	*z*	*to-ho,*	*že*	*jsem*	*napsa-l-a*
REFL	consist-PST-N.SG	of	DEM-GEN.N.SG	that	COP	write-PST-F.SG

prostě	*takov-ou*	*seminární*	*prác-i*	*na*	*téma*	*školní*
simply	such-ACC.F.SG	seminar	work-ACC.SG	on	topic.ACC.SG	school
		N.FOCUS (Figure 2)				

stravování.	*A*	*protože*	*tehdy*	*stejně*	*jako*	*nyní*
catering	and	because	then	same	as	now

psaní	*prac-í*	*není*	*moj-e*	*nejsilnější*	*stránka,*	*tak*
writing	work-GEN.PL	not.be.PRS.3SG	my.F.SG	strongest	side	so

jsem	*za*	*t-u*	*ee*	*jako*	*tadytudlect-u*	*prác-i*
COP	for	DEM-ACC.F.SG	[noise]	like	DEM.ACC.F.SG	work-ACC.SG
					C.TOPIC (Figure 3)	

12 NB that in spontaneous interactions, the implicit construal of contrast may be inferred not only from the linguistic context but also from non-linguistic (social) cues, context and from the background knowledge shared by the participants.

ee	ta	maturitní	– známka	na	vysvědčení	by-l-a
[noise]	DEM	final.exam	grade	on	exam.report.LOC.SG	be-PST-3F.SG
		C.TOPIC (Figure 4)				
za	tři.					
for	three.ACC					
	N.FOCUS (Figure 5)					

(SUBJ05: *I did my A-levels at a hospitality school, so I had to pass a practical part of the exam, which involved writing a kind of essay that I wrote about school meals. And because writing wasn't, and still isn't, my forte, for this particular essay I got a C on my A-level certificate.*

FIGURE 2

FIGURE 3

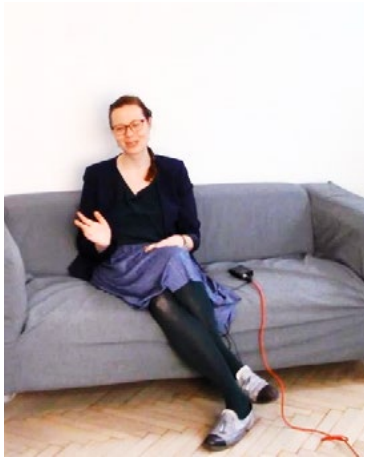

FIGURE 4

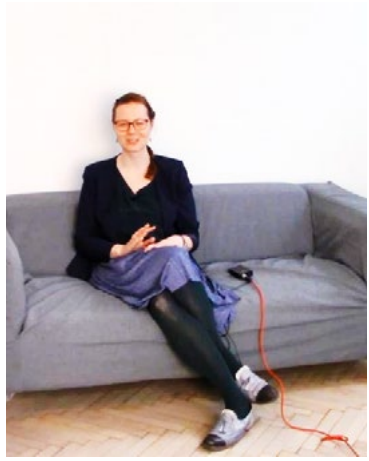

FIGURE 5

(2) SUBJ01:

Vzhledem	*k*	*to-mu,*	*že*	*jsem*	*maturova-l-a*	*z*
with.respect	towards	it-DAT	that	COP	take.exam-PST-F.SG	from

předmět-ů,	*který*	*mě*	*bavi-l-y,*	*který*	*mi*	*š-l-y,*
subject-GEN.PL	which	1SG.GEN	entertain-PST-PL	which	1SG.DAT	go-PST-PL

tak	*to*	*by-l-o*	*poměrně*	*v*	*klid-u.*
so	it	be-PST-N.SG	relatively	in	ease-LOC.SG
					N.FOCUS (Figure 6).

ADMIN:

Z	*če-ho*	*jste*	*maturova-l-a?*
from	what-GEN.SG	COP	take.exam-PST-F.SG

SUB01:

Maturova-l-a	*jsem*	*z*	*literatur-y,*	*z*	*češtin-y,*	*z*
take.exam-PST-F.SG	COP	from	literature-GEN.SG	from	Czech-GEN.SG	from
			C.FOCUS (Figure 7)		C.FOCUS (Figure 8)	

angličtin-y,	*z*	*němčin-y*	*a*	*ze*	*základ-ů*
English-GEN.SG	from	German-GEN.SG	and	from	base-GEN.PL
C.FOCUS (Figure 9)		C.FOCUS (Figure 10)			C.FOCUS (Figure 11)

společensk-ých	*věd*
social-GEN.F.PL	science.GEN.PL

(SUBJ01: *Given that I took only the subjects I liked and was good at, it was easy*. ADMIN: *And what did you take then?* SUBJ01: *I took literature, Czech, English, German and social sciences*)

FIGURE 6

FIGURE 7

FIGURE 8

FIGURE 9

FIGURE 10

FIGURE 11

4.3 RESULTS

The mean delay of Fo peak following gesture stroke onset ("pitch shift") was 311 ms (standard deviation = 261 ms), the mean delay of intensity peak was 333 ms (standard deviation = 258 ms). The histogram below (Figure 12) shows the homogeneous distribution of the shift values in our data.

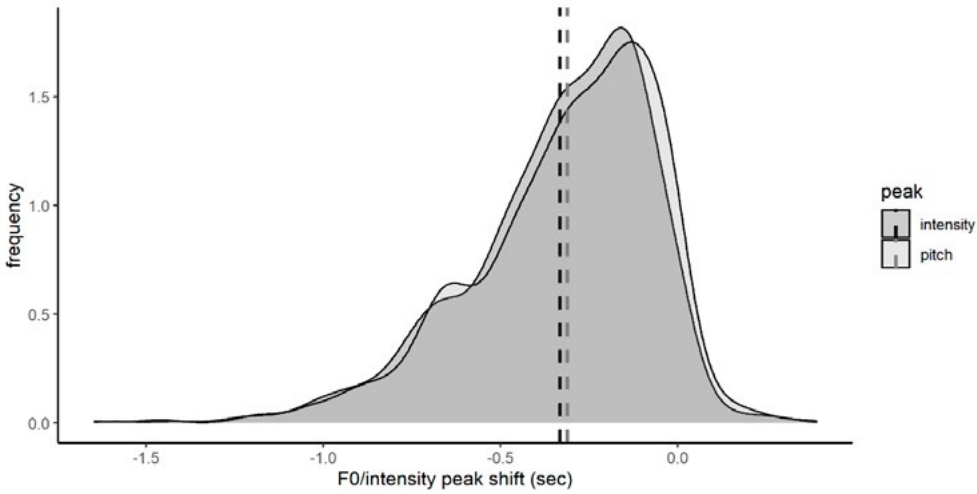

FIGURE 12: Frequency plot for temporal shift values. Lighter grey area = shift between Fo peak and gesture stroke onset (in seconds); darker grey area = shift between intensity peak and gesture stroke onset (in seconds). The vertical dashed lines represent the mean shift (grey = Fo, black = intensity).

Although the mean values of the temporal shift for the two types of prosodic peaks are basically the same, the pitch and intensity peaks were not always perfectly aligned — the mean absolute interval between the two peaks was 47 ms (standard deviation = 67 ms). Most (76%) of the intensity peaks followed after the F0 peak (by 45 ms on average), only 6% were perfectly aligned and a minority (18%) of intensity peaks preceded pitch peaks (average 69 ms).

Table 1 shows the distribution of the information structure categories as well as average values of temporal shift (between gesture and onset respective prosodic peak) for F0 and intensity and the average interval between the two.

category (relative frequency)		**F0 shift (ms)**	**intensity shift (ms)**	**interval (ms)**
topic	120 (0.12)	285	306	44
contrastive topic	54 (0.06)	344	357	55
new-information focus	636 (0.66)	302	327	46
contrastive focus	158 (0.16)	355	367	48

TABLE 1: Distribution of information structure categories, mean shifts (ms) between gesture onsets and F0 and intensity peaks, mean (absolute) interval between F0 and intensity peaks.

More than 80% of all instances represented the focal constructions (n = 754), most of which were non-contrastive foci (n = 636). This is not surprising; a high proportion of focus-accompanying gestures was expected. Considering the misalignment between pitch peaks and gesture onsets, we can see that the difference is between contrastive and non-contrastive contexts: non-contrastive peaks had the average F0 delay of 299 ms (topic = 285 ms, new-information focus = 302 ms), whereas the contrastive peaks trailed the gestures on average by 352 ms (contrastive topic = 344 ms, contrastive focus = 355 ms). Figure 13 visualises the differences between the four information structure categories. The difference between contrastive and non-contrastive instances is apparent.

To see whether this difference is statistically significant, the association between the pitch temporal misalignment (dependent variable) and the information structure category (fixed effects of *context boundedness* and *contrastiveness*) was investigated in R (R Core Team, 2021) with the help of a linear model with random intercepts (*speaker*) using the lmer() function from the *lme4* package (Bates et al. 2015).

Contrastiveness proved to be a significant predictor of the temporal shift (average shift increase = 43 ms, standard error = 20 ms, $p = 0.031$). The linear regression reveals a slight but systematic increase of the shift between the gesture and the corresponding pitch in contrastive contexts, regardless of whether the information is new or given.

The average intensity shift (Table 1) appears to follow a similar pattern as the pitch shift, with contrastive instances exhibiting a slightly greater offset from the beginning of the gesture stroke. Pitch and intensity are highly correlated in our data ($r = 0.95$, $p < 0.001$). However, a mixed-effect linear model with intensity shift as a de-

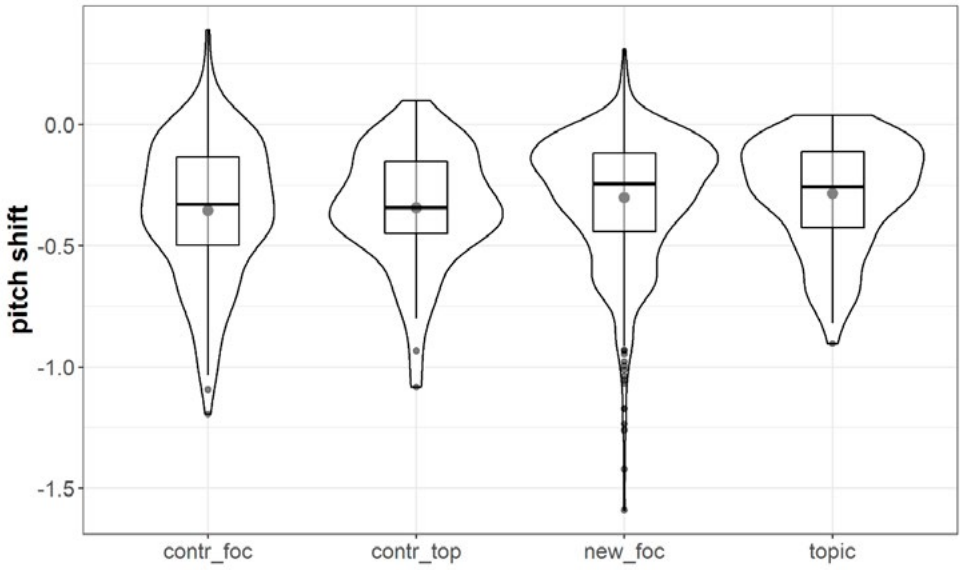

FIGURE 13: Temporal shift (in seconds) between the onset of gesture stroke and the corresponding pitch peak.
Grey dots = mean shift, boxes = interquartile range (IQR), horizontal lines = 1.5×IQR (datapoints beyond this range are considered outliers (small grey dots), thick horizontal lines = median values, curved outlines show the probability density.

pendent variable, and contrastiveness and context-boundedness as predictors (in interaction) did not produce any significant effects. This might be due to a higher degree of inter-speaker variance.

We have already mentioned that in the majority of cases, the Fo peak was closer to the gesture than the intensity peak. This gives rise to the question whether there is any systematicity in the distribution of the perfectly aligned and pitch-preceding intensity peaks with respect to a particular information structure category.

Let us inspect the distribution of the (relative) shift between Fo and intensity peaks in Table 2 and Figure 14:

	Relative shift (ms, intensity — pitch)	**pitch > intensity**	**perfect alignment**	**intensity > pitch**
topic	–20	0.73	0.07	0.20
contrastive topic	–13	0.71	0.05	0.24
focus	–25	0.77	0.06	0.17
contrastive focus	–12	0.78	0.04	0.19

TABLE 2: Mean relative shift (ms) and proportion of instances where the pitch peak precedes the intensity peak, instances of perfect alignment and instances where the intensity peak precedes the pitch peak.

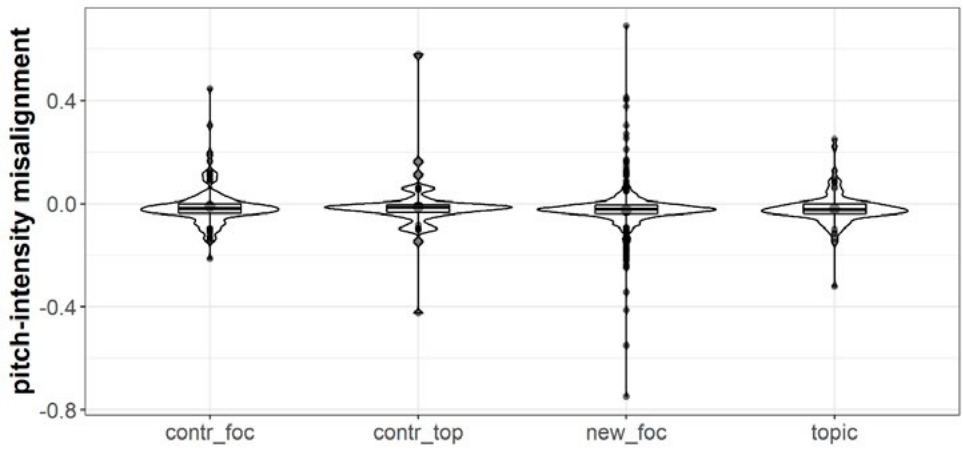

FIGURE 14: Relationship between pitch-intensity misalignment (in seconds) and the information structure categories.
Grey dots = mean misalignment, boxes = interquartile range (IQR), horizontal lines = 1.5×IQR (datapoints beyond this range are considered outliers (small grey dots)), thick horizontal lines = median values, curved outlines show the probability density.

We can see that the pitch-intensity misalignment is distributed uniformly across the information structure categories. The mixed-effect linear model did not reveal any significant effect of the information structure parameters neither, alone nor in interaction.

In sum, the quantitative analysis revealed an association between the pitch peak — stroke onset temporal shift and the information structure. Based on previous studies, we can assume that the gesture stroke *apex* tends to be better aligned with pitch accent in the cases of new information focus marking and non-contrastive topic. The greater shift in the case of contrastive focus and topic may be interpreted as pitch accent lagging behind the gestural apex. Discourse contrast in spoken Czech thus may be marked by misalignment between gesture and intonation.

5 DISCUSSION AND CONCLUSION

Converging crosslinguistic evidence suggests that at the level of gesture phrase, the apex of the gesture's stroke tends to be aligned with the pitch accent (Fo peak) of the stressed syllable of the gesture's "lexical affiliate" (Schegloff 1985). In our study of Czech spontaneous conversations, the gesture strokes exhibited almost the same manner of alignment with pitch accents as was reported in German (Evert et al. 2012). Also, in line with the German study, we observed that in the instances of discourse contrastiveness, pitch accents tend to follow after gesture strokes with a greater delay compared to the general pattern. Moreover, we show that this effect applies not only to the contrastive focus but to contrastive topics as well.

Our study represents the first survey of multimodal marking of information structure in Czech. As such, it provides rather general first-look observations, as we employed relatively coarse-grained measures that did not allow for a closer examination of the gesture-prosody-discourse entanglement. Further investigation of our corpus data has to consider possible alignment not only to F0 peaks but also phrase boundaries or low-pitch tones. However, the approach taken in this study is a good starting point. Also, as pointed out by Türk (2020), it is not necessarily only the apex of a gesture stroke that should be mapped to prosodic markers. For instance, we should focus on whether — and if so, in what contexts — *post-stroke holds* follow strokes that are out of phase with pitch contour.

In addition to pitch contour, the prosodic component of *intensity* was taken into account in this study. Although the quantitative analysis did not reveal any systematic association between intensity and gesture with respect to the four information structure categories, we observed some functional clusters in misaligned contexts containing recurrent gestures or gestural features. These included the away-body movement accompanying expressions of distancing or negation (Bressem and Müller, 2017), or list and enumeration constructions (Inbar 2018) involving finger-counting gestures.[13] A more in-depth qualitative analysis is needed to explore these and other potential multimodal patterns involving pitch-intensity misalignment (and to compare the clusters with those that Ward (2019) links to misalignment in AmE).

The temporal shift between the gesture and the prosodic emphasis can be described as a discourse/pragmatic feature only with regard to other parameters of a multimodal construction as understood in Multimodal Construction Grammar (Schoonjans 2017; Zima and Bergs 2017). Our exploratory study aimed at providing initial insights into temporal alignment between gesture strokes and prosody in Czech spontaneous conversations. The remaining parts of the equation include:

— A closer look at *contrastiveness*: we adopted a broad (cognitive) notion of contrastiveness, while some of the previous studies took a narrower perspective, focusing only on the explicit marking of contrast. Combining these two views in further research may shed more light on this rather complex category. Also, attention should be paid to constructional idiosyncrasies; we saw that from the broader perspective, a number of diverse constructional patterns are hidden behind the category of contrastiveness (see below).
— *Semantics of the referent*: what is the distribution of gestures and prosodic patterns with respect to whether the referent is an animate or inanimate object, abstract concept or event?
— *Frequency*: the delay between the gesture and the lexical affiliate may be caused by slower lexical retrieval (McNeill 1992). This issue can be addressed by factoring in the frequencies of the target words/constructions.
— *Semantic-functional dimensions of the gesture*: not every gesture in our sample can be simply labeled as a "beat" — frequently, gestures aggregate several functions, all of which need to be accounted for.

13 See example (2) and Figures 6–11.

— *Discourse-pragmatic functions*: an important issue is the pragmatic meaning of the multimodal construction in questions (e.g., inquiry, repair, suggesting, enumeration, etc.) — to capture this, thorough qualitative description must first be carried out.
— *Interactional and situational aspects*: how are the patterns in question distributed with respect to the turn-sequential structure of the conversation? Does the same pattern echo across participants?
— *Complex profiles of constructions*: under what conditions (taking into account all of the above and other parameters) does a multimodal construction with the same prosodic contours occur with and without gesture?
— *Non-manual gestures*: head movement, eyebrow movement, shrugging etc.

As for the general information structure categories, it seems that focus marking belongs among the basic functions of co-speech gestures in discourse. Further inquiry should thus be aimed at the circumstances of gestural marking of the topical part of an utterance.

The above list of outstanding issues is, of course, not comprehensive. The points mainly set an agenda for subsequent corpus studies with a larger sample from the CZICO corpus. Another crucial question is whether the gesture-prosody misalignment associated here with *contrastiveness* is actually perceived as such on the part of the receiver. This can only be resolved by subsequent testing in a behavioral experiment.

REFERENCES

Bates, D., M. Mächler, B. Bolker and S. Walker (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67/1. DOI: https://doi.org/10.18637/jss.v067.i01.

Beckman, M. E. and J. B. Pierrehumbert (1986) Intonational structure in Japanese and English. *Phonology Yearbook* 3, 255–309. DOI: https://doi.org/10.1017/S095267570000066X.

Biau, E., L. A. Fromont and S. Soto-Faraco (2018) Beat Gestures and Syntactic Parsing: An ERP Study: Beat Gestures and Syntactic Parsing. *Language Learning* 68, 102–126. DOI: https://doi.org/10.1111/lang.12257.

Biau, E. and S. Soto-Faraco (2013) Beat gestures modulate auditory integration in speech perception. *Brain and Language* 124/2, 143–152. DOI: https://doi.org/10.1016/j.bandl.2012.10.008

Birdwhistell, R. L. (1970) *Kinesics and context: Essays on body motion communication*. Philadelphia: University of Pennsylvania Press.

Boersma, P. (2001) Praat, a system for doing phonetics by computer. *Glot International*, 5/9–10, 341–345.

Bolinger, D. L. (1983) Gesture and intonation. *American Speech* 58/2, 156–174.

Bosker, H. R. and D. Peeters (2020) *Beat gestures influence which speech sounds you hear* [Preprint]. Neuroscience. DOI: https://doi.org/10.1101/2020.07.13.200543.

Bressem, J. and C. Müller (2017) The "Negative-Assessment-Construction" — A multimodal pattern based on a recurrent gesture? *Linguistics Vanguard* 3/s1. DOI: https://doi.org/10.1515/lingvan-2016-0053.

Chafe, W. L. (1976) Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In: Li, C. N. (ed) *Subject and Topic*, 25–55. New York: Academic Press.

Dimitrova, D., M. Chu, L. Wang, A. Özyürek and P. Hagoort (2016) Beat that Word: How Listeners Integrate Beat Gesture and Focus in Multimodal Speech Discourse. *Journal of*

Cognitive Neuroscience 28/9, 1255–1269. DOI: https://doi.org/10.1162/jocn_a_00963.

Dobrogaev, S. M. (1931) Učenie o reflekse v problemach jazykovedenija. *Jazykovedenie i Materializm* 2, 105–173.

Dohen, M., H. Loevenbruck and H. Hill (2009) Recognizing prosody from the lips: Is It Possible to Extract Prosodic Focus from Lip Features? In: Wee-Chung Liew, A. and S. Wang (eds) *Visual Speech Recognition: Lip Segmentation and Mapping*, 416–438. Medical Information Science Reference.

Ebert, C., S. Evert and K. Wilmes (2011) Focus marking via gestures. In: Reich, I., E. Horch, and D. Pauly (eds) *Proceedings of Sinn and Bedeutung 15*, 193–208. Saarbrücken: Saarland University Press.

Esteve-Gibert, N., J. Borràs-Comes, E. Asor, M. Swerts and P. Prieto (2017) The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical Society of America* 141/6, 4727–4739. DOI: https://doi.org/10.1121/1.4986649.

Ferré, G. (2010) Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French. *Workshop on Multimodal Corpora*, 86–91.

Götze, M., T. Weskott, C. Endriss, I. Fiedler, S. Hinterwimmer, S. Petrova, A. Schwarz, S. Skopeteas and R. Stoel (2007) Information structure. In: Dipper, S., M. Götze and S. Skopeteas (eds) *Interdisciplinary Studies on Information Structure*, 147–187. Potsdam: Universitätsverlag Potsdam.

Holle, H., C. Obermeier, M. Schmidt-Kassow, A. D. Friederici, J. Ward and T. C. Gunter (2012) Gesture Facilitates the Syntactic Analysis of Speech. *Frontiers in Psychology*, 3. DOI: https://doi.org/10.3389/fpsyg.2012.00074.

Im, S. and S. Baumann (2020) Probabilistic relation between co-speech gestures, pitch accents and information status. *Proceedings of the Linguistic Society of America* 5/1, 685–697. DOI: https://doi.org/10.3765/plsa.v5i1.4755.

Inbar, A. (2018) *List Construction as a Multimodal Phenomenon: Syntax, Prosody, and Gestures*. 51st Annual Meeting of the Societas Linguistica Europaea (SLE), Tallinn.

Ito, K. and S. R. Speer (2008) Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language* 58/2, 541–73.

Karpiński, M., E. Jarmołowicz-Nowikow and Z. Malisz (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. An analysis of an example. In: Siegman, A. W. and B. Pope (eds) *Studies in Dyadic Communication*, 177–210. Elmsford, NY: Pergamon Press.

Kendon, A. (1980) Gesticulation and speech: Two aspects of the process of utterance. In: Key, M. R. (ed) *The Relationship of Verbal and Nonverbal Communication*, 207–227. Berlin: Mouton.

Kendon, A. (1995) Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics* 23/3, 247–279. DOI: https://doi.org/10.1016/0378-2166(94)00037-F.

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

Kim, J., E. Cvejic and C. Davis (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication* 57, 317–330. DOI: https://doi.org/10.1016/j.specom.2013.06.003.

Kita, S. (1990) *The Temporal Relationship between Gesture and Speech: A Study of Japanese-English Bilinguals* [Master's thesis]. Chicago: University of Chicago.

Kita, S., I. van Gijn and H. van der Hulst (1998) Movement phases in signs and co-speech gestures, and their transcription by human coders. In: Wachsmuth, I. and M. Fröhlich (eds) *Gesture and Sign Language in Human-Computer Interaction* Vol. 1371, 23–35. Berlin-Heidelberg: Springer. DOI: https://doi.org/10.1007/BFb0052986.

Kok, K. I., K. Bergmann, A. Cienki and S. Kopp (2016) Mapping out the multifunctionality of

speakers' gestures. *Gesture* 15/1, 37–59. DOI: https://doi.org/10.1075/gest.15.1.02kok.

Krahmer, E. and M. Swerts (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. DOI: https://doi.org/10.1016/j.jml.2007.06.005.

Lambrecht, K. (1994) *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Leonard, T. and F. Cummins (2011) The temporal relation between beat gestures and speech. *Language and Cognitive Processes* 26/10, 1457–1471. DOI: https://doi.org/10.1080/01690965.2010.500218.

Loehr, D. P. (2004) *Gesture and intonation* [Doctoral dissertation]. Georgetown University.

Loehr, D. P. (2012) Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3/1, 71–89. DOI: https://doi.org/10.1515/lp-2012-0006.

Lücking, A., K. Bergmann, F. Hahn, S. Kopp and H. Rieser (2010) The Bielefeld Speech and Gesture Alignment Corpus (SaGA). *Proceedings of the LREC 2010 Workshop "Multimodal Corpora — Advances in Capturing, Coding and Analyzing Multimodality"*, 92–98. DOI: https://doi.org/10.13140/2.1.4216.1922.

McNeill, D. (1992) *Hand and Mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2005) *Gesture and Thought*. Chicago, IL: University of Chicago Press.

McNeill, D. and E. T. Levy (1982). Conceptual Representations in Language Activity and Gesture. In: Jarvella, R. J. and W. Klein (eds) *Speech, Place, and Action. Studies in Deixis and Related Topics*, 271–295. London: John Wiley and Sons.

Mikulová, M., A. Bémová, J. Hajič, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá and Z. Žabokrtský (2005) *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual.* (Technical Report TR-2006-30). Prague: Charles University.

Pouw, W., S. J. Harrison and J. A. Dixon (2020) Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General* 149/2, 391–404. DOI: https://doi.org/10.1037/xge0000646.

Pouw, W., S. J. Harrison, N. Esteve-Gibert and J. A. Dixon (2020) Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, 148/3, 1231–1247. DOI: https://doi.org/10.1121/10.0001730.

R Core Team (2021) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Schegloff, E. A. (1985) On some gestures' relation to talk. In: Atkinson, J. M. (ed) *Structures of Social Action*, 266–296. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511665868.018.

Schoonjans, S. (2017) Multimodal Construction Grammar issues are Construction Grammar issues. *Linguistics Vanguard* 3/s1. DOI: https://doi.org/10.1515/lingvan-2016-0050.

Shattuck-Hufnagel, S. and A. Ren (2018) The Prosodic Characteristics of Non-referential Co-speech Gestures in a Sample of Academic-Lecture-Style Speech. *Frontiers in Psychology* 9, 1514. DOI: https://doi.org/10.3389/fpsyg.2018.01514.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg (1992) TOBI: A standard for labeling English prosody. *Proceedings of ICSLP 1992*, 867–870.

Streeck, J. (2009) *Gesturecraft: The manu-facture of meaning*. Amsterdam: John Benjamins.

Swerts, M. and E. Krahmer, E. (2008) Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36/2, 219–238. DOI: https://doi.org/10.1016/j.wocn.2007.05.001.

Türk, O. (2020) *Gesture, Prosody and Information Structure Synchronisation in Turkish* [Doctoral

dissertation]. Victoria University of Wellington.
Van Valin, R. D. (2005) *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511610578.
Ward, N. (2018) A Corpus-Based Exploration of the Functions of Disaligned Pitch Peaks in American English Dialog. *9th International Conference on Speech Prosody 2018*, 349–353. DOI: https://doi.org/10.21437/SpeechProsody.2018-71.
Ward, N. (2019) *The prosodic patterns of English conversation*. Cambridge: Cambridge University Press.
Wittenburg, P., H. Brugman, A. Russel, A. Klassmann and H. Sloetjes (2006) ELAN: a Professional Framework for Multimodality Research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.
Zhang, Y., D. Frassinelli, J. Tuomainen, J. I. Skipper, and C. Vigliocco (2021) More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, 288/1955, 20210500. DOI: https://doi.org/10.1098/rspb.2021.0500.
Zima, E. and A. Bergs (2017) Multimodality and construction grammar. *Linguistics Vanguard* 3/s1, 20161006. DOI: https://doi.org/10.1515/lingvan-2016-1006.

Acknowledgements
This work was supported by the European Regional Development Fund project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (reg. no.: CZ.02.1.01/0.0/0.0/16_019/0000734).
The authors thank Klára Matiasovitsová for her assistance with the annotation and Martin Sedláček for proofreading the manuscript.

ABBREVIATIONS:

Grammatical glosses

1	first person
3	third person
ACC	accusative
COP	copula
DEM	demonstrative
DAT	dative
F	feminine
GEN	genitive
M	masculine
N	neuter
PL	plural
PRS	present
PST	past
REFL	reflexive
SG	singular

Information structure categories

C.FOCUS	contrastive focus
C.TOPIC	contrastive topic
N.FOCUS	new-information focus

Eva Lehečková
Institute of Czech Language and Theory of Communication
Faculty of Arts, Charles University
nám. Jana Palacha 2, Praha 1, Czechia 116 38
ORCID ID: 0000-0002-3064-6414
eva.leheckova@ff.cuni.cz

Jakub Jehlička
Department of Linguistics
Faculty of Arts, Charles University
Nám. Jana Palacha 2, 116 38 Prague
ORCID ID: 0000-0001-8176-6873
jakub.jehlicka@ff.cuni.cz

Magdalena Králová Zíková
Department of Linguistics
Faculty of Arts, Charles University
Nám. Jana Palacha 2, 116 38 Prague
ORCID ID: 0000-0002-8046-597X
magdalena.zikova@ff.cuni.cz