

GESPIN 2019

11 - 13 September



Proceedings

6th Gesture and Speech
in Interaction Conference



PADERBORN UNIVERSITY
The University for the Information Society

To cite this version:

Angela Grimminger (Ed.). Proceedings of the 6th Gesture and Speech in Interaction – GESPIN 6, Sept 11-13, 2019, Paderborn: Universitaetsbibliothek Paderborn.

Contents

FOREWORD.....	IV
COMMITTEES.....	V
PLENARY SPEAKERS.....	1
Pointing to words: How gesture provides a helping hand to language development across different learners <i>Şeyda Özçalışkan.....</i>	2
Prosody: Cross-modal Interactions of Form and Function <i>Petra Wagner.....</i>	3
Understanding human behavior using virtual humans: lessons learned and upcoming challenges <i>Alexis Heloir.....</i>	4
Enacting prosody in the classroom: How the prosody in our hands helps us learn pronunciation in a second language <i>Pilar Prieto.....</i>	5
PAPERS.....	7
Smiling for negotiating topic transitions in French conversation <i>Mary Amoyal and Béatrice Priego-Valverde.....</i>	9
Gesture and speech coordination to frame utterances as humorous <i>Marta Buján.....</i>	15
Encouraging gesture use in a narration task increases speakers' gesture rate, gesture salience and the production of representational gestures <i>Alice Cravotta, Pilar Prieto, and M. Grazia Busà.....</i>	21
An integrative platform to capture the orchestration of gesture and speech <i>Christelle Dodane, Dominique Boutet, Ivana Didirkova, Fabrice Hirsch, Slim Ouni, and Aliyah Morgenstern.....</i>	27
Gesture / speech alignment in weather reports <i>Gaëlle Ferré.....</i>	34
Gesture-speech coordination in expression of motion: How far to zoom in to observe semantic synchrony? <i>Katerina Fibigerova and Michèle Guidetti.....</i>	39
The timing of pointing-speech combinations in typically developing and language-delayed toddlers <i>Angela Griminger.....</i>	44
Gestural training benefits L2 phoneme acquisition: Findings from a production and perception perspective <i>Marieke Hoetjes, Lieke van Maastricht, and Lisette van der Heijden.....</i>	50
Synchronization of (dis)fluent speech and gesture: A multimodal approach to (dis)fluency <i>Loulou Kosmala, Maria Candea, and Aliyah Morgenstern.....</i>	56
Children's viewpoint: Iconic co-speech gestures and their relation to linguistic structure across two communicative genres <i>Ulrich Mertens, Olga Abramov, Anne Németh, Friederike Kern, Stefan Kopp, and Katharina J. Rohlfing.....</i>	62
Acoustic specification of upper limb movement in voicing <i>Wim Pouw, Alexandra Paxton, Steven J. Harrison, and James A. Dixon.....</i>	68
Quantifying gesture-speech synchrony <i>Wim Pouw and James A. Dixon.....</i>	75
Embodied reciprocity in conversational argumentation: Soliciting and giving reasons with Palm Up Open Hand gestures <i>Nora Schönfelder and Vivien Heller.....</i>	81
Does gestural hierarchy align in time with prosodic hierarchy? Another modality to consider: Information structure <i>Olcay Turk.....</i>	87
Hand gestures and pitch contours and their distribution at possible speaker change locations: a first investigation <i>Margaret Zellers, Jan Gorisch, David House, and Benno Peters.....</i>	93

Foreword

The uniqueness of previous GESPIN meetings arises from examining gestures in tight coordination with speech (including its lexical, syntactical as well as prosodical properties). The 6th edition of the Gesture and Speech in Interaction (GESPIN) was held in Paderborn, Germany. For this meeting, we focused on the heterogeneity of this coordination. Topics and related questions were:

- Development of gesture-speech coordination: Can general principles of development be identified? Are there sensitive periods and developmental stages?
- Individual differences in coordinating speech and gestures: Are there developmental differences beyond infancy/childhood? How do various population groups (elder people, people with autism spectrum disorders, people with cognitive or language impairments) coordinate gesture and speech?
- Benefits of multimodal coordination for learning in individuals and in a variety of settings
- Computational models dealing with heterogenous data and/or generating behavior that differs across, for example, situation or addressees
- Cross-cultural differences on gesture-speech coordination: Is development following a universal path that is culturally shaped? How do cultural groups differ in how they coordinate gestures and speech?
- Heterogeneity across situations and contexts: Do situations differ due to familiarity with the environment or interlocutors? Do communicative genres require specific types of coordination?

The four keynote speakers addressed the focus of this 6th meeting from different perspectives:

Prof. Dr. Seyda Özçalışkan (Georgia State University) who focusses her research on children's earliest linguistic abilities and on the question whether gesture constitutes a robust aspect of the language learning process. In her talk, she pursues the idea of differences or delays in speech becoming first evident in gesture across different learners.

Prof. Dr. Petra Wagner (Bielefeld University) working on the relationship between prosodic expression in speech and gesture, and currently studying the impact of information structure and visibility between interlocutors on the cross-modal link in prosodic expression. By extending her focus to co-speech movements that are not considered as gestures in the traditional understanding, her talk contributes to the concept of embodiment of communication.

Dr. Alexis Heloir's (Université Polytechnique des Hauts de France) research centers around the question of how virtual agents contribute to understanding human behavior. By pointing to his transdisciplinary collaborations, he will address leading design principles of an agent creation and control framework called YALLAH.

Prof. Dr. Pilar Prieto (ICREA-Universitat Pompeu Fabra, Barcelona, Catalunya) with a strong research interest on the benefits of gesture in the second language classroom, especially with respect to embodied rhythmic movements that might have an effect on L2 pronunciation. In her talk, she proposes a multimodal approach in general and embodied prosodic trainings in specific that are essential to understanding L2 speech learning.

We were pleased that the talks and posters at this year's GESPIN meeting, just as at previous meetings, were presented by international contributors from a variety of countries, and based on them, we were able to offer different formats (workshops, data session) for sharing our research, experiences and ideas.

We also thank all the reviewers who engaged in the selection process this year, and we hope that in the future, GESPIN will remain a community of scholars devoted to the coordination of gestures and speech. Finally, we would like to acknowledge the work of the people at Paderborn University who helped a lot with organizing this event: Sabine Hendriks, and the student volunteers, Camilla Crawshaw, Lisa Enns, Monique Koke, Eileen Sygalla, and Jennifer Truhn.

Katharina J. Rohlfing, Angela Grimminger & Ulrich Mertens

Committees

Local Organizing Committee

Katharina J. Rohlfing
Angela Grimminger
Ulrich Mertens

Reviewers

Olga Abramov (Bielefeld University, Germany)
Manuel Bohn (Stanford University, USA)
Silvia Bonacchi (Warsaw University, Poland)
Jana Bressemer (Technische Universität Chemnitz, Germany)
Hendrik Buschmeier (Bielefeld University, Germany)
Nina Capone Singleton (Seton Hall University, USA)
Alan Cienki (Vrije Universiteit Amsterdam, The Netherlands)
Jean-Marc Colletta (University of Grenoble, France)
Nevena Dimitrova (University of Applied Sciences and Arts of Western Switzerland)
Gaëlle Ferré (University of Nantes, France)
Tilbe Göksun (Koç University, Turkey)
Angela Grimminger (Paderborn University, Germany)
Michelle Guidetti (CNRS-CLLE-University Toulouse 2, France)
Marianne Gullberg (Lund University, Sweden)
Silva H. Ladewig (Europa-Universität Viadrina, Germany)
Vivien Heller (University of Wuppertal)
Konrad Juszcyk (Adam Mickiewicz University, Poland)
Maciej Karpiński (Adam Mickiewicz University, Poland)
Friederike Kern (Bielefeld University, Germany)
Sotaro Kita (University of Warwick, UK)
Carina Lüke (Paderborn University, Germany)
Zofia Malisz (Royal Institute of Technology in Stockholm, Sweden)
Iris Nomikou (University of Plymouth)
Seyda Özçalışkan (Georgia State University, USA)
Asli Özyürek (Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands)
Karola Pitsch (University of Duisburg-Essen, Germany)
Wim Pouw (University of Connecticut, USA)
Katharina Rohlfing (Paderborn University, Germany)
Gale Stam (National Louis University, USA)
Lauren Stites (Georgia State University, USA)
Susanne Vogt (Hochschule Fresenius, Germany)
Paul Vogt (Tilburg University, The Netherlands)
Sławomir Wacewicz (Nicolaus Copernicus University, Poland)
Petra Wagner (Bielefeld University, Germany)
Przemysław Żywiczyński (Nicolaus Copernicus University, Poland)

Plenary Speakers

Pointing to words: How gesture provides a helping hand to language development across different learners

Şeyda Özçalışkan

Department of Psychology, Georgia State University, USA

seyda@gsu.edu

Children communicate using gestures before they speak, and continue to use gesture along with speech even after they begin to produce their first words. Does gesturing merely precede talking, or is it itself relevant to the language-learning process? If gesturing not only precedes language, but also reflects knowledge relevant to the developmental process responsible for language, then the differences and/or delays in speech should become first evident in gesture across different learners. I approach this question by examining early gesture and speech production of children with different developmental profiles—including children with autism, Down syndrome and typical development, who show unique strengths or weaknesses in their early gesture production. I ask whether early gesture predicts later speech across different learners, and if so, what underlies the link between early gesture and later spoken language development.

Prosody: Cross-modal Interactions of Form and Function

Petra Wagner

Bielefeld University, Germany
petra.wagner@uni-bielefeld.de

The strong link between prosodic expression in speech and gesture has been established through ample empirical evidence that prosodic prominence, prosodic phrasing as well as floor management is expressed and processed in a multi-modal fashion. However, the exact functional and formal relationship between the different modalities is still not well understood:

First, we lack knowledge about which aspects of prosodic expression, namely signal features such as pitch movements or rather structural features such as information structure, are actually reproduced across modalities. To shed light on this issue, I will present a recent series of analyses (Wagner et al., in press), where this question was tackled with a newly developed methodological approach: Listeners were asked to "reproduce" acoustically presented utterances in a drumming task. The results revealed that the patterns of drumming intensities closely resemble patterns of perceptual prominence gathered with established annotation schemes, and involving both prosodic experts and naive listeners. However, as unveiled by a Random Forest Analysis, the gestural reproductions of prosody were driven comparatively stronger by signal cues than by linguistic structure. Also, we found different strategies for the gestural interpretations of prosodic patterns: a largely signal-driven, and a more integrative strategy.

Second, we do not know much about the flexibility or stability of the cross-modal link in prosodic expression. While speech economy models predict a flexible relationship depending on communicative demands and cross-modal compensation, a strong view of cross-modal congruence predicts a stable cross-modal co-ordination. To explore this issue, I will present a series of recent studies of semi-spontaneous, task-oriented interaction (Wagner and Bryhadyr, 2017; Wagner et al., 2019a; 2019b; in prep.) aiming at a better understanding of the impact of (1) information structure and (2) visibility between interlocutors on the cross-modal link in prosodic expression. Our results once again confirm a strong cross-modal temporal co-ordination. Furthermore, we detected a systematic modulation of this co-ordination as a function of communicative demands: In important or unpredictable contexts, co-speech movements occur later and align tightly with corresponding pitch peaks if interlocutors can see each others' hands. Also, a lack of facial visibility between interlocutors leads to a significantly earlier production of corresponding co-speech movements. In summary, our results show that co-speech movements in general can express a rich set of signal and structural cues inherent in speech prosody, and that the degree of temporal co-ordination between speech and co-speech movements is a function of communicative needs. As a side result, we found that cross-modal prosodic link also extends to co-speech movements such as drumming or manual moves on a game board, which are not gestures in the traditional understanding of the term.

References

- Wagner, P. and N. Bryhadyr (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, P., Cwiek, A., and B. Samlowski (in press). Exploiting the speech-gesture link to capture fine-grained prominence impressions and listening strategies. *Journal of Phonetics*.
- Wagner P., Bryhadyr, N., Schröer, M., and B. Ludusan (2019a). Does information-structural acoustic prosody change under different visibility conditions? *In: Proceedings of the International Congress of Phonetic Sciences 2019, Melbourne, Australia*.
- Wagner P., Bryhadyr N., and M. Schröer (2019). Pitch Accent Trajectories across Different Conditions of Visibility and Information Structure - Evidence from Spontaneous Dyadic Interaction. *Proceedings of Interspeech 2019, Graz, Austria*.
- Wagner, P. et al. (in prep.). The temporal coordination between speech prosody and co-speech movements as a function of communicative needs.

Understanding human behavior using virtual humans: lessons learned and upcoming challenges

Alexis Heloir

Université Polytechnique des Hauts de France, France

Alexis.Heloir@univ-valenciennes.fr

Animated Virtual Characters exhibit many desirable aspects for who wants to understand human behavior and language. Like their human counterpart, they can display a broad palette of multimodal stimuli, these stimuli can, however, be very precisely altered, fired at exact timestamps, or triggered by specific reactions of a subject taking part in an experiment.

Unfortunately, the integration of virtual characters into a full-fledged experiment setup requires a concentration of many diverse and specific skills which are often out of the reach of the team crafting the experiment. A lightweight, modular, well documented, and easy to deploy agent toolkit is still needed.

This talk starts depicting a series of trans-disciplinary collaborations which lead to the design and implementation of interactive virtual humans in experimental setups involving human counterparts. Each experiment was able to shed an original light on specific aspects of human language or behavior.

This talk later focuses on the lessons learned during these trans-disciplinary collaborations and how we could infer from them the leading design principles of a new agent creation and control framework called YALLAH.

YALLAH stands for Yet Another Low-Level Avatar Handler. It is a framework supporting the creation of real-time interactive virtual humans by non-experts. After a quick overview of YALLAH's features, documentation, and ongoing projects using YALLAH, the talk will conclude by a discussion on how YALLAH could help the community understanding the coordination of gesture with speech.

Enacting prosody in the classroom: How the prosody in our hands helps us learn pronunciation in a second language

Pilar Prieto

ICREA-Universitat Pompeu Fabra, Barcelona, Catalunya

pilar.prieto@upf.edu

When we speak, we use rhythmic hand gestures which are coordinated with prominent parts of speech (e.g., beat gestures). In this talk I will discuss several experiments carried out in our research group that deal with how beat gestures and other embodied rhythmic movements facilitate the learning of second language pronunciation. Even though most of the research on the benefits of gesture in the second language classroom has focused on the effects of representational gestures (e.g., for the acquisition of vocabulary), little is known about the potential beneficial effects of embodied rhythmic movements on the learning of L2 pronunciation. A set of experiments will be presented. Experiments 1 and 2 will assess the potential benefits of observing and performing beat gestures on L2 pronunciation learning with intermediate Catalan learners of English. Experiments 3 and 4 will assess the benefits of hand-clapping on L2 pronunciation learning at initial stages of L2 acquisition of French by Catalan and Chinese native speakers. Widening the scope of this investigation, Experiments 5 and 6 will focus on the positive effects of using melodic and singing trainings for pronunciation learning. Based on the positive findings from these experiments, I will conclude that a multimodal approach is essential to understanding L2 speech learning. I will suggest that not only rhythmic trainings with beat gestures or hand-clapping procedures can act as scaffolding mechanisms for L2 speech production but also melodic trainings based on pitch mimicry and singing. Importantly, both types of embodied prosodic trainings could be successfully applied to language teaching and language treatment contexts.

Papers

Smiling for negotiating topic transitions in French conversation

Mary Amoyal and Béatrice Priego-Valverde

Laboratoire Parole et Langage, CNRS, Aix-Marseille Université, France
mary.amoyal@univ-amu.fr, beatrice.priego-valverde@univ-amu.fr

Abstract

This study focuses on participants' smiling behavior as a resource for negotiating topic transitions in French conversations. The smile will be analyzed as a resource during topic transitions: through its intensities and its development. This study will show that the speaker's smiling dynamic contributes to initiating a transition and that the hearer tends to synchronize his/her smile with the speaker to ratify it.

Index terms: smile, topic transition, conversation, alignment.

1. Introduction

In line with previous work considering the smile as an “interactive gesture” (Bavelas & Gerwing, 2007), smile will be apprehended here as a facial gesture that conveys interactive functions. While it has been mostly analyzed in a binary way (presence/absence), it will be investigated through 5 degrees of intensity, from neutral (0) to laughter (4) (Gironzetti, Attardo & Pickering, 2016). Such an approach will lead us to investigate the way it evolves during a conversation, highlighting the fact that its significance lies not only in its mere presence but also in the way it decreases or increases. Consequently, smile will be investigated in the present study as a resource whose presence and coordination allow participants in a conversation to negotiate topic transitions. Topic transitions are “conversational moves” (Riou, 2015) necessitating negotiations between the participants to be accepted and developed as the next subject under discussion, i.e., “what a portion of the interaction is about” (Berthoud & Mondada, 1995). Following Tannen (1984), a topic transition is considered as such only when the proposed topic is developed by the participants. Several works have pointed out that topic transitions are initiated with various “thematization markers” (De Fornel, 1988; Porhiel, 2005). Among various kinds of markers, smile has been investigated during emotional transition (Kaukomaa, Peräkylä, & Ruusuvuori, 2013). Furthermore, two strong moments are distinguished in the topic transition: the “initiation” (Maynard, 1980), i.e., the proposition of the topic by the speaker (S) and the “ratification” (Riou, 2015) i.e. the approval of the proposition by the hearer (H). In line with previous studies on conversations viewed as collaborative (Sidnell & Stivers, 2012) and as a “joint activity” (Clark, 1996), this study focuses on smiling as a resource for negotiating topic transitions. The question underlying this study is: how does the smile impact the success of a topic transition? Two hypotheses are proposed: (1) while initiating a transition, S displays a different smile intensity according to the presence or absence of verbal markers; (2) while ratifying the transition, H aligns his/her smile with the S's smile. In this exploratory study based on 2 conversations, these hypotheses will be tested using a mixed methodological approach linking quantitative methods used in Corpus Linguistics and qualitative analysis in line with Conversational Analysis and Interactional Linguistics frameworks (Couper-Kuhlen & Selting, 2001).

2. Methodology

2.1. Corpus and participants

This study is based on “Cheese!” (Priego-Valverde, Bigi, Attardo, Pickering, & Gironzetti, 2018) an audio and video corpus recorded in 2016. This corpus is composed of 11 dyadic interactions (around 15 minutes each) between two native French speakers and students at the university. None of them knew the real purpose of the experiment nor did they receive any compensation for their participation. All signed a written consent form. Both mixed and non-mixed dyads were created

without any gender requirement. This present study is based on two interactions of this corpus: JSCL, two 3rd year female students, and MAPC, respectively being 2nd year male and female student.

2.2. Experimental setting

Participants were seated face-to-face in a soundproof room. Two cameras were positioned behind their back and pointed at the other participant's face. Both were fitted with a micro headset, optimally positioned so as not to hide the mouth while preserving the acoustic signal. Each participant was asked to read a text (a canned joke). After the reading part, participants had 15 minutes to discuss as freely as they wished. Our analyses are focused on the conversational part.

2.3. Annotations






2.3.1. IPU parsing and transcription of speech signal

Our selected corpus had been annotated at two levels using SPPAS software (Bigi, 2015). The speech signal was automatically parsed into Inter-Pausal Units (IPUs), i.e., fragments of speech separated by 200 ms breaks. Then, the speech signal was transcribed manually according to the Enriched Spelling Transcription (Bertrand, et al., 2008).

2.3.2. Smiling annotations

Smiles were annotated, relying on the "Smiling Intensity Scale" (SIS) (Gironzetti, Attardo, & Pickering, 2016). The SIS measures the smile intensity gradually from 0 (neutral face) to 4 (laughter), based on Action Units (AUs) detailed by the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). Below the 5 levels of smile intensity are presented by pictures of our corpus:

Table 1
Smiling Intensity Scale (Gironzetti, Attardo & Pickering, 2016)

				
0 - No smile	1 - Closed mouth smile	2 - Open mouth smile	3 - Wide open mouth smile	4 - Laughing smile

According to this scale, manual annotations of smile were performed with ELAN software on each participant (Brugman & Russel, 2004). Each interaction was divided into 400 ms intervals, as this is considered the time necessary to produce or perceive a complex gesture such as smiling (Sanders, 1998; Heerey & Crossley, 2013). Then, each interval was assigned a smile intensity: 2610 smile intensities were annotated in MAPC and 2475 in JSCL. This method allows us to analyze the evolution of each participant's smile (increase/decrease) in a very precise way.

2.3.3. Inter annotator agreement

A counter-coding was carried out on both interactions to validate the reliability of these annotations and the relative objectivity of the scale used. We then calculated Cohen's Kappa (Cohen, 1960), a statistical measure used to compare the annotations of two judges. Both inter-annotator agreement rates were qualified as excellent: 0.87 for MAPC and 0.89 for JSCL.

2.3.4. Topic transition

In line with Riou's methodology (2015), the identification of the transitions was conducted in 5 steps. Below, we illustrate our methodology with a chronological table where each step of annotations is illustrated with an example from our corpus. After having talked about the text that they have read, S asked "what would you like to talk about" and H answers "the semantic course".

Table 2
Methodology to identify topics transitions

Steps	1. Topic under discussion	2. Transition initiation	3. Ratification	4. Themmatization markers	5. Type of markers
Indicator	Key words	Identifying the frontier between topic 1 & topic 2	YES or NO	YES or NO	Verbal or Non verbal
Examples	“semantic course”	“then, what would you like to talk about?”	YES	YES	Verbal “then”

Such methodology leads to an analysis of the complete transition, from its initiation by S to its ratification by H. As a result, 28 transitions were extracted from our corpus.

3. Quantitative results

After having identified the topic transitions (Table 2) present in the conversations of our corpus, we analyzed smiles in these specific moments: S’s smile while s/he initiates a transition, and H’s smile when s/he ratifies the transition.

3.1. Topic transitions

28 topic transitions were identified in the two interactions: 12 in MAPC and 16 in JSCL—on average, one transition per minute. As for topic transitions, the results show that the S tends to initiate a transition more often with than without a verbal marker: 20 transitions were initiated with verbal markers. This trend has led us to investigate the role of S’s smile in these two types of transition, and correlatively, H’s smile when a transition has been proposed.

3.2. Participants’ smiles

During **both entire conversations**, participants smile for more than a third of the time: **39.5%** on average (36% in MAPC and 43% in JSCL). This result is consistent with previous studies, such as (Cosnier, 1987; Bavelas & Gerwing, 2007). More interestingly, comparing the presence of smile in the whole conversation with smile during transitions (from their initiation to their ratification) shows that smile is predominant while the participants are making a transition. Indeed, participants smile during **78.13%** (on average) of the time spent doing transitions. This interesting result shows that **smiling appears even more during topic transitions** than in the rest of the conversation and that smile could have a specific role during topic transition.

In more details, concerning the **initiation phase** of a transition, the results show that there are many more transitions initiated with than without a smile (18 against 10). More precisely, in MAPC 7 transitions are preceded by a smile (out of 12); in JSCL 11 transitions are preceded by a smile (out of 16). These results show that S is more likely to smile when s/he initiates a transition (on average in 63.54% of the initiations). As for the 18 transitions initiated with a smile, S’s smiles during transitions were systematically compared with his/her smiling behavior (increase vs. decrease) before and after the initiation. Two types of evolution were observed:

- S increased their smile in 9 transitions’ initiations: 5 in MAPC and 4 in JSCL.
- S decreased their smile in 17 transitions’ initiations: 5 in MAPC and 12 in JSCL.

In other words, S systematically change the intensity of their smile when they initiated a transition (in 93% of the corpus initiations).

As shown in the figure below, the types of smile shift (increase/decrease) were analyzed according to the type of transition (with/without a verbal marker).

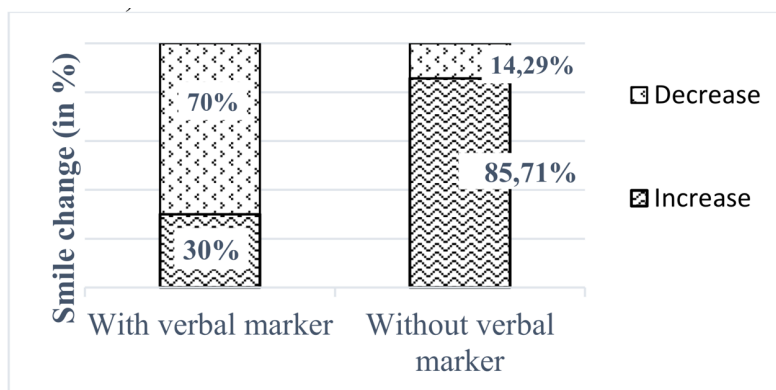


Figure 1. Proportion of S's smile shift according to the presence/absence of verbal marker in topic transition.

The figure above shows a relationship between the presence/absence of verbal markers and the S's smile shift and reveals two results:

- When S initiates a topic transition with a verbal marker, s/he reduces his/her level of smile in 82% of the cases: **a reduction of smile is more likely used when the transition is initiated with a verbal marker.**
- When S does not use any verbal marker to initiate a transition, s/he displays a stronger smile in 80% of the cases: **an increased smile is more likely used when the transition is not initiated with any verbal marker.** These two results show that S's smiling behavior is linked to the presence or absence of verbal markers. Moreover, they highlight the complementarity of smile and verbal resources when S initiates a transition. This tends to confirm our first hypothesis according to which smile change is linked with the way S initiates a transition. Consequently, they suggest that smile, like verbal markers, may be a linguistic resource for sequential organization of the transitions.

As well as S's smiling behavior, **H's smiling behavior** was investigated when s/he ratify a topic transition. Concerning **the ratification phase**, our data show that many more transitions were ratified with than without a smile (19 against 9). This result shows that H is more likely to smile when s/he ratifies a transition (on average in 66.67% of the cases). Nevertheless, the analysis of the two conversations reveals contrasted results, 7 of the 12 topic transitions in MAPC were ratified with a smile, and 12 against 16 in JSCL. This could be explained by the difference in the topics addressed by the participants (in preparation). Then, applying the same classification of smile change (increase/decrease) to H, his/her smiling behavior was compared with S's smiling behavior:

- When S initiates a transition, decreasing his/her smile, H also decreases his/her smile in **83.33%** of every decreasing case. This trend is noticeable in 66.67% of the decreasing smile ratification of MAPC and in every cases of JSCL. This different distribution of smile decrease alignment can be explained by the fact that there are more transitions initiated with a verbal marker in JSCL than in MAPC, thus there are more transitions initiated with a smile decrease in JSCL.
- When S initiates a transition, increasing his/her smile, H also increases his/her smile in **87.5%** of every case. This trend is noticeable in every increasing smile ratification of MAPC and in 75% of JSCL.
- Combining the two interactions reveals that when the transition is accepted by H, both participants of each interaction operate a smile alignment in **85.42%** of the cases.

This result shows that not only participants tend to reciprocate their smiles (Capella, 1997; Hess & Bourgeois, 2010), but they also synchronize their smiling development. Such a result confirms our second hypothesis according to which H aligns his/her smile when a transition is ratified.

interactional patterns highlighted here. Further investigations are currently being conducted. First, we have noticed that the topics identified in the 2 conversations were various (i.e. the soundproof room, the participants' studies, their friends, their romantic relationship); it would be interesting to analyze the impact of **topic type and duration** on smile development during transition. Secondly, some of these topics are deeply related to the participants' **common ground** (in preparation); here again it would be interesting to analyze the impact of common ground on smiling during transitions.

Acknowledgment

We give special thanks to the Centre d'Experimentation de la Parole (CEP), the shared experimental platform for the collection of data, at LPL.

Transcription conventions:

Truncated words: smi-smile

Initials for names: N for Name

Speech in overlap: Underlines words

Smile intensities are aligned with the audio files, one line for each participant.

References

- Bavelas, J. B., & Gerwing, J. (2007). Conversational hand gestures and facial displays in face-to-face dialogue. *Social communication*, 283-308.
- Berthoud, A.-C., & Mondada, L. (1995). Traitement du topic, processus énonciatifs et séquences conversationnelle. *Cahiers de linguistique Française*, 17, 205-228.
- Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, 49(3), 1-30.
- Bigi, B. (2015). SPPAS - Multilingual Approaches to the Automatic Annotation of Speech. *The Phonetician International Society of Phonetic Sciences*, 111(2), 54-69.
- Brugman, H., & Russel, A. (2004). Annotating Multi-media/Multi-modal Resources with ELAN. *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.
- Capella, J. (1997). Journal of Personality and Social Psychology. *Behavioral and judged coordination in adult informal social interactions: Vocal and kinesic indicators*, 72, 119-131.
- Clark, H. (1996). *Using language*. Cambridge: University Press.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.
- Cosnier, J., & Kerbrat-Orecchioni, C. (1987). *Décrire la conversation*. Lyon: Presses universitaires de Lyon.
- Couper-Kuhlen, E. & Selting, M. (2001). *Studies in Interactional Linguistics*. Amsterdam/Philadelphia: John Benjamin Publishing.
- De Fornel, M. (1988). Constructions disloquées, mouvement thématique et organisation préférentielle dans la conversation. *Langue Française* (78), 101-123.
- Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.
- Gironzetti, E., Attardo, S., & Pickering, L. (2016). Smiling, gaze, and humor in conversation. In L. In Ruiz-Gurillo, *Metapragmatics of Humor: Current research trends* (Vol. 14, p. 235). Amsterdam/Philadelphia: John Benjamins Publishing.
- Heerey, E. A., & Crossley, H. M. (2013). Predictive and reactive mechanisms in smile reciprocity. *Psychological science*, 24(8), 1446-1455.
- Hess, U. & Bourgeois, P. (2010). You smile-I smile: Emotion expression in social interaction. *Biological psychology*, 84, 514-520.
- Kaukomaa, T., Peräkylä, A., & Ruusuvuori, J. (2013). Turn-opening smiles: Facial expression constructing emotional transition in conversation. *Journal of Pragmatics*, 55, 21-42.
- Maynard, D. W. (1980). Placement of topic changes in conversation. *Semiotica*, 30(3-4), 263-290.
- Porhiel, S. (2005). Les marqueurs de thématisation: des thèmes phrastiques et textuels. *Travaux de linguistique* (2), 55-84.
- Priego-Valverde, B., Bigi, B., Attardo, S., Pickering, L., & Gironzetti, E. (2018). Is smiling during humor so obvious? A cross-cultural comparison of smiling behavior in humorous sequences in American English and French interactions. *Intercultural Pragmatics*, 15(4), 563-591.
- Riou, M. (2015). A methodology for the identification of topic transitions in interaction. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 16.
- Sanders, A. F. (1998). *Elements of human performance: Reaction processes and attention in human skill*. Mahwah, NJ: Erlbaum.
- Sidnell, J., & Stivers, T. (2012). *The handbook of conversation analysis*. John Wiley & Sons.
- Tannen, D. (1984). *Conversational Style: Analyzing Talk among Friends*. Norwood: Ablex.

Gesture and speech coordination to frame utterances as humorous

Marta Buján

Universidad de Valladolid, Spain

marta.bujan.navarro@gmail.com

Abstract

The purpose of this paper is to present a multimodal study conducted on spontaneous humorous communication, in order to determine whether the pragmatic and discourse use and function of gestures and prosody differ from non-humorous communication. A sample of 14 interviews from *The Late Show with Stephen Colbert* was collected. Only interviewee's speech was analysed to ensure it was not scripted. Utterances were identified as humorous using laughter in the audience as the main criterion. The videos were annotated in ELAN for humour type, gestures (face and head movements), and prosody. The prosodic analysis was done in Praat to look into contrast between humorous and non-humorous utterances in terms of F0 and intensity. No multimodal cues specific to humour were found. The use and function of gestures in humorous utterances bear out previous studies on non-humorous communication.

1. Introduction

Humour is arguably one of the most complex instances of communication, both in terms of production and comprehension (Veale, Brône, & Feyaerts, 2015). Various studies have been conducted to look into consistent multimodal cues of humour, i.e. whether certain gestures, face expressions, head movements, changes in gaze, intonation or prosody patterns invariably associated with humour exist (Pickering et al., 2009; Attardo, Pickering, & Baker, 2011; Urios-Aparisi & Wagner, 2011; Attardo, Pickering, Lomotey, & Menjo, 2013; etc.). Many studies have been conducted on the markers of irony or sarcasm, with conflicting results (Rockwell, 2000; Attardo, Eisterhold, Hay, & Poggi, 2003; Bryant, 2010; Attardo, Pickering, & Baker, 2011; Attardo, Wagner, and Urios-Aparisi, 2011; Tabacaru, 2014, etc.). Fewer studies exist on non-ironical humour, and fewer still focus on spontaneous humour (Archakis & Tsakona, 2005; Attardo, Pickering & Baker, 2011; Feyaerts, 2013, etc.). Nevertheless, the focus on spontaneous —non-scripted— communication is relevant, as humour is based on familiarity (Flamson, Bryant, & Barret, 2011). Given that posed humour needs to reach a wide audience, it may be delivered in an exaggerated manner, and resort to different multimodal resources from those employed in naturally-occurring, non-scripted humorous utterances (Rockwell, 2000; Urios-Aparisi & Wagner, 2011). A survey of the literature shows that most studies have found no consistent multimodal cues of humorous speech, as compared to serious discourse. This is a counterintuitive notion, especially given the abundance of studies claiming that irony, for instance, is associated with certain intonation patterns (Rockwell, 2000; Attardo et al., 2003; Cheang & Pell, 2009; González-Fuente, Escandell-Vidal, & Prieto, 2015; etc.). These studies have yielded a wide range of often conflicting results, whereby irony is associated with flat (Haiman, 1998), rising intonation (Schaffer, 1982), higher (Rockwell, 2000) and lower pitch (Haiman, 1998; Anolli, Ciceri, & Infantino, 2000), heavy exaggerated pitch (Adachi, 1996) and relatively monotonous intonation (Haiman, 1998), etc. Attardo et al. (2003) claimed that there is no such thing as an ironic intonation, but rather that pitch and changes in prosody are just contrastive markers. Regarding gestures, Attardo, Wagner, and Urios-Aparisi (2011) compiled different ironical gestural cues appearing in the literature (Muecke, 1978; Attardo et al., 2003). Tabacaru and Lemmens (2014) argued that raised eyebrows are gestural triggers prompting the hearer to take the utterance as humorous, ironic, or sarcastic. According to González-Fuente et al. (2015), prosody and gesture are just pragmatic facilitators. For these authors, prosody and gestures, therefore, are used as tools to reduce the cognitive effort required from the hearer to interpret the ironic nature of the utterance (Yus, 2003, 2016).

This paper presents a study conducted to gain an insight into how humour is conveyed in face to face interaction. I look into how certain prosodic features, gestures, and speech interplay in the production of non-scripted humorous utterances in English to determine if functions and uses specific to humour can be found, as opposed to non-humorous communication.

2. Methods of data acquisition, annotation and analysis

2.1. Sample

The sample analysed includes 14 interviews from The Late Show with Stephen Colbert (Hoskin, 2015). Only utterances by interviewees have been analysed, avoiding mostly pre-scripted or rehearsed host's speech. The fully spontaneous nature of the interviewees' speech could be questioned, as most of them are people used to speaking in public and may therefore be seen as merely acting out their public persona during the show. Having said that, the aim of the research conducted for this study aimed at confronting non-scripted humorous utterances to those taken from sitcoms, TV shows or stand-up comedies in previous literature. Hence, the sample can at least be considered semi-spontaneous to the extent that it has not been previously scripted.

Each interview was analysed in a different ELAN file. Prosodic features for each selected utterance were studied separately in Praat. The sample contains 103.83 minutes of interviews, out of which 109 humorous utterances were found. For each humorous utterance, annotations on five parameters were made: a) transcription of the utterances selected, b) main construal mechanism underlying humour, c) type of humour involved, d) gestures made in the humorous utterances, e) prosodic analysis (pitch and intensity).

Following Bryant (2010), three different kind of utterances were identified with regards to the prosodic analysis: a) Humorous utterances: For the sake of objectivity, utterances were considered humorous when the audience reacted to them laughing, in order to avoid bias based on the coder's interpretation of humour and following standard practice in the literature (Morreal, 1983; Attardo, Pickering, & Baker, 2011; Archakis & Tsakona, 2005; Flamson et al., 2011; Tabacaru, 2014; Bryant & Gibbs, 2015). b) Baseline and pre-base utterances were also selected to measure prosodic contrast between humorous and non-humorous instances. Baseline utterances were those said immediately before humorous utterances, whereas pre-base were those immediately preceding baseline utterances. A control analysis could thus also be performed comparing non-humorous utterances (pre-base / baseline).

Mean pitch (F0 in Hz) and mean intensity (in dB) were obtained for each utterance. Then, all data was recorded in SPSS in order to estimate the standard deviation (SD) in mean intensity and mean pitch, for each type of utterance per interview, as a proxy measure of variability and prosodic contrast (Purandare & Litman, 2006; Bryant, 2010). SD values per type of utterance were compared within speakers through t-tests (independent variables) to determine whether there was a statistically significant difference in SD, which would lead to conclude that prosodic contrast in F0 and intensity was in turn significant. No statistically significant differences in SD values for F0 and intensity were found in the sample ($p=0.05$). Consequently, no prosodic contrast has been identified between humorous and non-humorous utterances, when it comes to F0 and intensity SD values. Admittedly, the setting and casual tone of the programme, prone to humour, would not require humour be made particularly salient through prosodic cues.

2.2. Multimodal analysis and discussion

In this section, a token of the multimodal analysis performed on the sample is included, on the basis of the most frequent combinations of humour types and gestures. The purpose of the analysis is to delve into the pragmatic and discursive use and functions of co-speech gestures in spontaneous humorous utterances to see whether differences exist with regards to non-humorous communication (Hadar et al., 1985; Poggi & Pelachaud, 1998; McClave, 2000; Kendon, 2002; Lee & Marsella, 2010; Kousidis, Malisz, Wagner, Schlangen, & Ladewig, 2013; Ishi, Ishiguro, & Hagita, 2014; Tabacaru, 2014; etc.). Only face gestures and head movements were annotated, as there was not always a clear view of the hands and the rest of the body. Data on gestures was cross-referenced with both humour types and construal mechanisms identified in the sample. No consistent

correlation patterns emerged beyond what could be expected due to the frequency of occurrence of each type of gesture, construal of humour in the sample.

Head movements and raised eyebrows have been found to serve as beats (Hadar et al., 1984; Pelachaud, Badler, & Steedman, 1996; Krahmer & Swerts, 2007; Guaiatella et al., 2009; Flecha-García, 2010; Tabacaru, 2014), that is, non-representational gestures used to punctuate speech (Kendon, 1980; McNeill, 1992). Head nods are considered to generally signal agreement (Lee & Marsella, 2010), whereas head shakes are associated with explicit or implicit negative statement (Kendon, 2002). Face gestures have been assigned various communicative functions in the literature (Poggi & Pelachaud, 1998): (1) affective display (Ekman & Friesen, 1971); (2) syntactic function, when facial expressions punctuate questions, emphasis, intonational accents, pauses, etc. (Poggi & Pelachaud, 1998); (3) dialogic function (C. Goodwin, 1980); (4) referring function (Ekman, 1979); (5) attitude display, when face gestures express the speaker’s attitude towards the interlocutor (Poggi & Pelachaud, 1998).

In example (1) we find an instance of raised eyebrows and parody. Raised eyebrows have been associated with the notion of surprise, as attention-getting devices (Guaitella et al., 2009), as tools to alert the hearer about important upcoming bits of information (Kim, Cvejic, and Davis, 2014), as underliners contributing to information structure (Flecha-García, 2010) or as gestural triggers to signal that an utterance must be interpreted as humorous. Furthermore, eyebrows have been found to strongly correlate with prosody (Flecha-García, 2010).

Stephen Colbert is interviewing Alec Baldwin, and he brings up a letter that Alec Baldwin received from President Nixon after Baldwin had lost the election for president of his school at George Washington University. Alec Baldwin then takes the letter to read what he considers to be the best part of it.

- (1) Alec Baldwin: You know what the greatest part of this thing is? It’s that he writes: “From our mutual friend Mark Weinberg I’ve learnt of the disappointing results, as far as you are concerned”.



From our mutual friend..



...as far as you are concerned.

Figure 1. Raised eyebrows in example (1).

In this example, Alec Baldwin resorts to parody to delimit the part of the letter that he finds most interesting, as conveying the lack of tact by Nixon, or simply the fact that he did not feel sorry for Alec Baldwin’s defeat. In order to do so, the actor mimics precisely those Nixon’s words, as opposed to the first excerpt from the letter that he reads normally. The parody is shown by a change in voice quality, a significant lower pitch, a palm-up gesture, head shake, and raised eyebrows. Eyebrows are raised over the entire remark “as far as you are concerned”. As I see it, in this particular instance the use of the raised eyebrows could be twofold. On the one hand, to frame the chunk of letter that Alec Baldwin considers more significant, arguably because it is the most telling part about Nixon’s attitude towards his defeat, or because he feels it showed lack of empathy. On the other, the raised eyebrows could also be associated to the expression of surprise felt by Alec Baldwin on reading that part of the letter.

Most examples boast a combination of gestures co-occurring with speech. It is the coordination between modalities which ultimately serves to convey a message. For example, in Alec Baldwin’s interview, upon taking his seat, right after being welcome by the host and by the audience with a very big round of applause, he thanks the audience and stresses what nice people they are. Then he utters: “It’s nice and chilly in here”, which elicits a bout of laughter in the audience. I posit that humour in this utterance arises from a clash in expectations about what he was supposed to say, e.g.

“it’s nice, what a warm welcome”, etc., and the fact that he actually states that it is “chilly”. Furthermore, chilly is reinforced with higher pitch, a nod and a smile.

An illustration of the importance of the interplay between gestures and speech to grasp the meaning of an utterance can be found in example (2) below, where Daniel Kaluuya is teasing Stephen Colbert by mimicking a previous remark made by the host. The humorous nature of (2) can only be understood in the context of the interview knowing what Stephen Colbert had said first, why it had been picked up by Daniel Kaluuya to mock the host, i.e. because it showcased a certain awkwardness due to racial differences, being aware that racial issues was the main topic in the film they are discussing, starred by Daniel Kaluuya. Crucially, only by seeing and listening to Daniel Kaluuya’s speech and multimodal behaviour—mimicking gestures, smile, etc.—, can the humorous intent be fully apprehended.

(2) Daniel Kaluuya: It’s like... What would I say... If I was white... What would I...?



Figure 2. Daniel Kaluuya mocking Stephen Colbert.

3. Conclusion

As briefly pointed in the examples above, the outcome of the analysis leads to conclude that the use and functions of co-speech gestures and prosody in semi-spontaneous humorous utterances in the sample is the same as in non-humorous communication.

One possible explanation of the absence of markers in humour as opposed to irony may be that prosodic cues are used only as metalanguage showing affect, that is, the position and feelings of the speaker with regards to the utterance. In the case of humour, it can be argued that there is no such detachment between the speaker and the humorous text. Both sarcastic/ironical and humorous utterances are manipulated by the speaker, but in two distinct ways. Sarcastic/ironical utterances are manipulated to show what the speaker thinks about the utterance. Humorous speech, on the other hand, is manipulated to mislead the hearer to a false interpretation to be subsequently proved wrong in order to achieve the humorous effect (Tabacaru, 2014).

Another explanation put forward to account for the difference between ironic and non-ironic humour in terms of multimodal framing associates the lack of markers to signal humour with an in-group expression of bonding on the part of the speaker, as relying on the common ground assumed to be shared with the interlocutors, and necessary for humour to be comprehended, thus demonstrating the affinity between participants (Tabacaru, 2014). Interestingly, Flamson et al. (2011) argued that as humour comprehension is influenced by context, the more background information is shared by the participants in the interaction, the less marking would be necessary for humour to be interpreted. In other words, the larger the intended audience of the humorous utterance, the more salient this humour will need to be made in order to ensure it is successfully conveyed (Attardo et al., 2003).

In light of the above, there seems to be no consistent markers of humour. Instead, prosodic and gestural cues, not specific to humour utterances, are sometimes used to communicate humour more effectively. The patterns and salience of the indices involved will eventually depend on the pragmatic context in which humour is conveyed.

References

- Anolli, L., Ciceri, R., & Infantino, M.G. (2000). Irony as a game of implicitness: Acoustic profiles of ironic communication. *Journal of Psycholinguistic Research*, 29(3), 275-311.
- Archakis, A., & Tsakona, V. (2005). Analyzing conversational data in GTVH terms: A new approach to the issue of identity construction via humour. *Humour: International Journal of Humour Research*, 18(1), 41-68.
- Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, 16(2), 243-260.
- Attardo S., Pickering, L. & Baker, A. (2011). Prosodic and multimodal markers of humor in conversation in S. Attardo, M. Wagner, & E. Urios-Aparisi (Eds.), *Prosody and Humor* (pp. 224-247). Amsterdam / Philadelphia: John Benjamins Publishing.
- Attardo, S., Wagner, M., & Urios-Aparisi, E. (2011). Prosody and humor. *Pragmatics & Cognition*, 19(2), 189-201.
- Attardo, S., Pickering, L., Lomotey, F., & Menjo, S. (2013). Multimodality in Conversational Humor. *Review of Cognitive Linguistics*, 11(2), 402-416.
- Bryant, G. (2010). Prosodic Contrast in Ironic Speech. *Discourse Processes*, 47 (7), 545-566.
- Cheang, H. S., & Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America*, 126(3), 1394-1405.
- Ekman, P. (1979). About brows—emotional and conversational signals. In von Cranach, M. K. Foppa, W. Lepenies, & D. Ploog, (Eds.), *Human Ethology*, pp. 169-248. Cambridge: Cambridge University Press.
- Ekman, P., Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), pp. 124 – 129.
- ELAN (Version 5.0.0-beta) [Computer software]. (2017, April 18). Nijmegen: Max Planck Institute for Psycholinguistics.
- Feyaerts, K. (2013). Tackling the complexity of spontaneous humorous interaction: An integrated classroom-modeled corpus approach. In L. Ruiz Gurillo & M. B. Alvarado Ortega (Eds.), *Pragmatics & Beyond New Series* (Vol. 231, pp. 243-268). Amsterdam: John Benjamins Publishing.
- Flamson, T., Bryant, G., & Barret, H. (2011). Prosody in spontaneous humour. *Pragmatics & Cognition*, 19(2), 189-201.
- Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, 52(6), 542-554.
- González-Fuente, S., Escandell-Vidal, V., & Prieto, P. (2015). Gestural codas pave the way to the understanding of verbal irony. *Journal of Pragmatics*, 90, 26-47.
- Guaitella, I., Santi, S., Lagrue, B., & Cave, C. (2009). Are Eyebrow Movements Linked to Voice Variations and Turn-taking in Dialogue? An Experimental Investigation. *Language and Speech*, 52(2/3), 207-222.
- Hadar, U., Steiner, T.J., & Clifford Rose, F. (1985). Head movement during listening turns in conversation. In *Journal of Nonverbal Behavior*, 9, pp. 214 – 228.
- Hadar, U., Steiner, T.J., & Grant, E.C. (1984). The timing of shifts of head postures during conversation. In *Human Movement Science*, 3, pp. 237 – 245.
- Haiman, J. (1998). *Talk is cheap: Sarcasm, alienation, and the evolution of language*. Oxford: Oxford University Press..
- Hoskin, J. (director), & Colbert, S., Spyra, J., Stack, B., & Dinello, P. (writers). (2015-). *The Late Show with Stephen Colbert* [Television show]. Retrieved from <https://www.youtube.com/channel/UCMtFAi84ehTSYSE9XoHefig>
- Ishi, C. T., Ishiguro, H., & Hagita, N. (2014). Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, 57, 233-243.
- Kendon, A. (2002). Some uses of the headshake. *Gesture* 2(2), 147-182.
- Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In M. R. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207-227). The Hague: Mouton and Co.
- Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, 57, 317-330.
- Kousidis, S., Malisz, Z., Wagner, P., & Schlangen, D. (2013). Exploring annotation of head gesture forms in spontaneous human interaction. *TiGeR 2013, Tilburg Gesture Research Meeting*. Retrieved from https://pub.uni-bielefeld.de/download/2567303/2606548/tiger2013_KousidisEtAl.pdf
- 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.
- Lee, J., & Marsella, S. (2010). Predicting speaker head nods and the effects of affective information. *IEEE Transactions on Multimedia*, 12(6), 552-562.
- McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32, 855-878.
- McNeill, D. (1992) *Hand and mind: what gestures reveal about thought*. Chicago & London: University of Chicago Press.
- Morreal, J. (1983). *Taking Laughter Seriously*. New York: State University of New York Press.
- Muecke, D. (1978). Irony markers. *Poetics*, 7(4), 363-375.
- Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, 20, 1-46.
- Pickering, L., Corduas, M., Eisterhold, J., Seifried, B., Eggleston, A., & Attardo, S. (2009). Prosodic markers of saliency in humorous narratives. *Discourse Processes*, 46(6), 517-540.
- Poggi, I. & Pelachaud, C. (1998). Performative faces. *Speech Communication* (26), 5-21.
- Poggi, I., D’Errico, F., & Vincze, L. (2010). Types of Nods. The Polysemy of a Social Signal. In LREC. Retrieved from http://www.academia.edu/download/4683379/596_paper.pdf
- Purandare, A., & Litman, D. (2006). Humor: prosody analysis and automatic recognition for F* R* I* E* N* D* S. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 208-215. Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1610107>
- Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research*, 29(5), 483-495.

- Tabacaru, S. (2014). *Humorous implications and meanings: a multi-modal approach to sarcasm in interactional humor* (Unpublished doctoral dissertation). Université Charles de Gaulle - Lille 3. Retrieved from <http://www.theses.fr/2014LIL30015>
- Tabacaru, S., & Lemmens, M. (2014). Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding. *The European Journal of Humour Research*, 2(2), 11–31.
- Urios-Aparisi, E., & Wagner, M. (2011). Prosody of humor in *Sex and the City*. *Pragmatics & Cognition*, 19(3), 507-529.
- Veale, T. Brône, G., Feytaerts, K. (2015). Humour as the *killer app* of language. In G. Brône, K. Feytaerts, & T. Veale (Eds.), *Cognitive Linguistics and Humour Research* (pp. 1-12). Berlin/Boston: Walter de Gruyter.
- Yus, F. (2003). Humor and the search for relevance. *Journal of Pragmatics*, 35(9), 1295-1331.
- Yus, F. (2016). *Humour and Relevance*. Amsterdam: John Benjamins Publishing.

Encouraging gesture use in a narration task increases speakers' gesture rate, gesture salience and the production of representational gestures

Alice Cravotta¹, Pilar Prieto^{2,3}, and M. Grazia Busà¹

¹Università degli Studi di Padova, Italy; Dipartimento di Studi Linguistici e Letterari (DiSLL)

²Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Catalunya, Spain

³Universitat Pompeu Fabra, Barcelona, Catalunya, Spain; Departament de Traducció i Ciències del Llenguatge

alice.cravotta@phd.unipd.it, pilar.prieto@upf.edu, mariagrazia.busa@unipd.it

Abstract

Previous work has shown the positive effect of encouraging gestures in performing various tasks; in these studies, the participants generally appeared to gesture more when explicitly asked to do it. However, little attention has been paid to whether encouraging gestures also affects other gesture features, i.e., gesture type and salience. In this paper we explore this issue. Twenty native Italian speakers described the content of short comic strips to a listener in 2 conditions: Non-Encouraging gestures (N); Encouraging gestures (E). Co-speech gestures were manually coded and classified according to gesture type (Representational vs. Non-Representational) and gesture salience (Salient vs Non-Salient). The results show that instructing speakers to gesture led to an increase in gesture rate, in gesture salience, and in the number of representational gestures. By contrast, in the non-encouraging condition the rate of Non-Salient gestures was significantly higher, but no difference was found for Non-Representational gestures.

1. Introduction

Researchers in previous studies have explicitly instructed participants to gesture in order to explore the effects of encouraging the use of gesture on activities such as problem solving (Beilock & Goldin-Meadow, 2010; Chu & Kita, 2011), learning math (Broaders, Cook, Mitchell, & Goldin-Meadow, 2007), second language pronunciation (Baills, Suárez-González, González-Fuente, & Prieto, 2019; Llanes-Coromina, Prieto, & Rohrer, 2018), speech fluency and narrative abilities (Vilà-Giménez & Prieto, 2018). These studies have shown that gestures have a beneficial role in thinking, learning, remembering, and speaking. As well, they have shown that instructing participants to gesture generally causes an increase in the participants' gesture rate. Nonetheless, to our knowledge, the only study that has directly focused on the impact of encouraging speakers to use gestures on the way they gesture across genres is Parrill, Cabot, Kent, Chen, & Payneau, (2016). The study compared the differences in gesture rate and gesture type of participants that had been and had not been explicitly instructed to gesture while performing three different discourse tasks (i.e., quasi-conversation, spatial problem solving, and narration). In the study, the instruction to gesture did not change gesture rate or gesture type across the different discourse tasks, suggesting that instructing speakers to gesture will not always work (in the sense that it might not lead them to produce more gestures); at the same time, the instruction does not seem to impact on the type of gestures produced. In sum, the study appears to be in contrast with previous findings, mentioned above, that show that the instruction to gesture should at least contribute to increasing gesture rate. Thus, the issue needs to be further explored.

Gesture production may be influenced by a combination of other factors. For instance, it has been shown that gesture rate, together with gesture type and gesture physical forms (size, salience), can change and be adapted depending on (1) the shared knowledge between interlocutors (Gerwing & Bavelas, 2004; Holler & Wilkin, 2009); (2) the interlocutors' (mutual) visibility (Bavelas, Gerwing, Sutton, & Prevost, 2008; Bavelas, Kenwood, Johnson, & Phillips, 2002); (3) the addressee's feedback (e.g., gesture rate lowers when addressees are less attentive (Jacobs & Garnham, 2007)). Moreover, individual differences in gesture production in terms of rate, type and

physical properties largely depend on the individuals' cognitive abilities, personality traits, cultural and gender differences (Briton & Hall, 1995; Chu, Meyer, Foulkes, & Kita, 2014; Goksun, Goldin-Meadow, Newcombe, & Shipley, 2013; Hostetter & Hopkins, 2002; Hostetter & Potthoff, 2012; Kita, 2009; Nicoladis, Nagpal, Marentette, & Hauer, 2018; O'Carroll, Nicoladis, & Smithson, 2015). These studies suggest that gesture rate, type and salience are key aspects of how gestures are produced, intended and interpreted in the wild. Thus, it seems that instructing participants to gesture can increase their gesture rate, as well as have a more general impact on gesture types and salience. This is interesting from a methodological perspective and deserves further attention: in fact, when setting up an experiment or data collection that requires explicitly asking participants to gesture while speaking, it might be important to assess the possible impact of the instruction to gesture on factors such as gesture salience and type.

To our knowledge, no studies have addressed the question of how encouraging speakers to gesture might affect gesture space or gesture salience. Also, how encouraging speakers to gesture affects gesture rate remains unclear. Thus, the present study will empirically assess, in a narration task, the effects of explicitly asking speakers to gesture on their (1) gesture rate, (2) gesture type, and (3) gesture salience.

2. Methods

The present study used a narration task in which the participants had to watch and describe a set of comic strips in two different conditions: Non-Encouraging (N), in which the participants were given no instructions regarding how to gesture; and Encouraging (E), in which the participants were encouraged to use gestures while telling the story. The experiment has a within-subject design (with a within subject factor: Condition).

2.1. Participants and Materials

Twenty female native speakers of Italian (age $M = 24.2$; $SD = 2.9$) participated in the experiment. They were all female and from the Veneto region (this was done to possibly control for potential gender and cultural differences in gesture production). Sixteen 4-scene comic strips adapted from Simon's Cat by Simon Tofield were used for the narration task (see Figure 1 for an example). The comic strips were carefully selected and adapted so that they were considered equivalent in terms of complexity and length (4-scene narrations). Simon's Cat comic strips do not contain text but feature a variety of characters and show many motion events. The idea was that this characteristic of the selected comic strips would make participants describe the events and spatial relations using gestures.

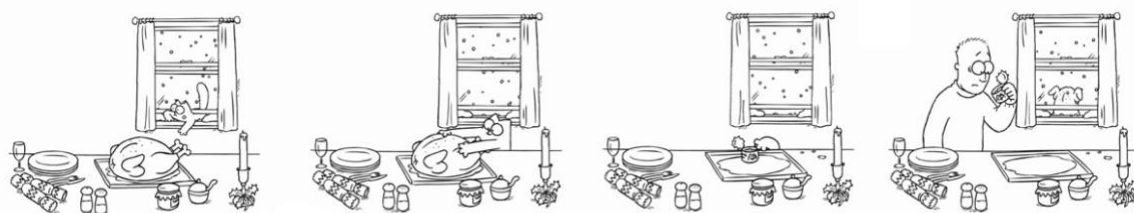


Figure 1. Example of a 4-scene comic strip used for the experiment (from Simon's Cat, by Simon Tofield, reproduced with permission).

2.2. Procedure

The participants were tested individually in a quiet room. Each session was recorded with a HD video camera (JVC GZ-HD7E Everio) connected to a MIPRO wireless head-mounted microphone recorded via a Zoom R16 digital audio mixer. The camera was set in front of the participant (at 2.50 m distance) recording her upper body and face. The participant sat on an office armchair and interacted with a confederate listener that sat in front of her at a distance of 1.50 m. A second video camera was placed in front of the listener and recorded the listener's upper body and face during the whole session. The experimenter (first author) sat at the participant's side. Both the participant and the listener were given written instructions as to how to perform the task (see below). Each participant was introduced to the confederate listener as if he was a fellow participant who did not know the stories in advance. This was done to avoid potential effects of common ground (Holler &

Wilkin, 2009), as well as to give ecological validity to the narration task (the participants would explain the story clearly and fully to their “fellow participant” as he was dependent on them to complete his part of the comprehension task). The confederate listener was instructed to provide basic backchannel and feedback cues to the speaker while listening to the stories as it was shown that speakers’ gesture can be adapted depending on the addressee’s feedback (e.g., gesture rate is lower when addressees are less attentive, Jacobs & Garnham, 2007).

Each participant had to retell a total of 16 stories. The experiment was preceded by 2 familiarization trials so that participants could get acquainted with the task and the setting. Each trial consisted of a three-step sequence: (1) the participant examined a four-scene comic strip to learn the story it depicted; (2) the comic strip was then concealed and the participant told the story to the listener; (3) the listener was then given four cards, each showing one scene of the comic, and had to reconstruct it by putting the four images in the correct order based on the speaker’s story.

The participants had to retell the first half of the comic strips set in the N condition and the second half in the E condition (the order of the comic strips was counter-balanced across conditions). The order of the two conditions was kept the same (N, E) for all the participants (as in Parrill et al. (2016), since we believed that telling participants to “come back” to a N condition after having encouraged them to gesture would lead to carryover effects between E and N). In the E condition the participants were given the following instructions (translated from Italian): “Tell each story and use hand gestures to help you do so”. The written instructions were kept visible in the E condition to remind the participants about the task. The experiment lasted approximately 30 minutes. Audio-visual recordings of a total of 200 short narratives were obtained (20 participants × 10 target trials) lasting a total of 81.2 minutes (39.1 minutes in the N condition and 42.1 in the E condition).

2.3. Gesture annotation

Any instances of co-speech gestures were identified and manually coded with the software ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) by the first author. The annotation criteria consisted in counting any gestural strokes (i.e., the most effortful part of the gesture that usually constitutes its semantic unit, e.g., two hands shaping together a rounded table; Kendon, 2004; McNeill, 1992), and to exclude any non-gestural movement like self-adaptors (e.g., scratching, touching one’s hair). The speakers produced a total of 2396 gestures (1015 in N and 1381 in E). Gesture rate was calculated per every story told as the number of gestures produced per story relative to the number of spoken words in the narrative (Gestures/words*100).

To check whether instructing speakers to gesture also changes the type of gestures performed, the gestures performed were distinguished between Representational (R) vs Non-Representational (NR) gestures. Representational gestures are those gestures that represent semantic information via form, (handshape), trajectory, or location. They can be distinguished from non-Representational gestures which include those that do not primarily serve to depict information and do not refer to a clear referent but which primarily have pragmatic and interactive functions (e.g., presentational, interactive, epistemic; Kendon, 2004; Bavelas, Chovil, Coates, & Roe, 1995; Cooperrider, Abner, & Goldin-Meadow, 2018). Representational gesture rate per story told was computed relatively to the number of words per story (representational gestures/words*100). The same was done for Non-Representational gesture rate.

Furthermore, to assess whether instructing speakers to gesture also changes gesture salience, each stroke was further classified depending on where it was performed (in fact, gestures performed at different height, and span are different in terms of communicativeness and salience; Bavelas et al., 2008; Mol, Krahmer, Maes, & Swerts, 2009; Streeck, 1994). Salience classification was done by using McNeill (1992)’s representation of the gesture space, which is divided into sectors delimited by concentric squares. For the present coding, a simplified 2-sectors version of it was used (as illustrated in Figure 2):

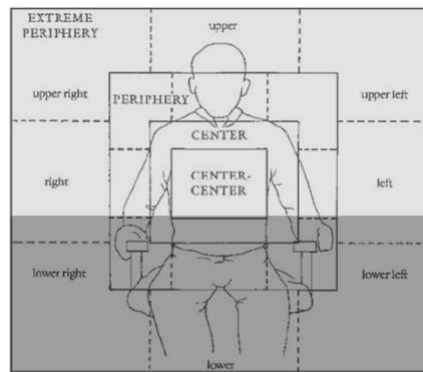


Figure 2. Gesture space. Adapted from McNeill, (1992) with the addition of two shades of gray that highlight the gesture areas of interest for the present study.

When the gesture stroke was produced in a more central, higher and visually prominent area (Streeck, 1994) of the gesture space (the lighter grey area), the gesture was coded as salient, whereas, when the gesture stroke was produced in a less visually prominent area (the lower darker sector), it was coded as non-salient. Salient Gesture (S) rate was computed per every story told as the number of salient gestures produced per story relative to the number of spoken words in the narrative (Salient gesture/words*100). The same was done for Non-Salient (NS) gesture rate.

The effect of gesture encouragement (within-subjects factor) on gesture behaviour was tested via 5 Linear Mixed Effects Models (henceforth LMEMs; R function *lmer* in *lme4* package; see Bates, Mächler, Bolker, & Walker, 2014). Each model included one of the following 5 dependent variables: Gesture (G) rate, Representational (R) gesture rate, Non-Representational (NR) gesture rate, Salient (S) gesture rate, Non-Salient (NS) gesture rate; and had *Condition* (N, E) as a fixed effect and both *Story* and *Participant* as random intercepts. P-values are obtained by likelihood ratio tests of the full model against the model without the fixed effect of interest (i.e., *Condition*).

3. Results

The instruction to gesture had effects on gesture rate, on gesture type and salience, as shown in Table 1 and in Figure 3. The boxplots in Figure 3 represent the different rates per gesture category per condition.

As shown in Table 1, Gesture rate was higher in the E condition (est.=4.134, S.E =0.708, $p < .001$). Regarding the effect on the type of gestures, the rate of Representational Gestures was higher

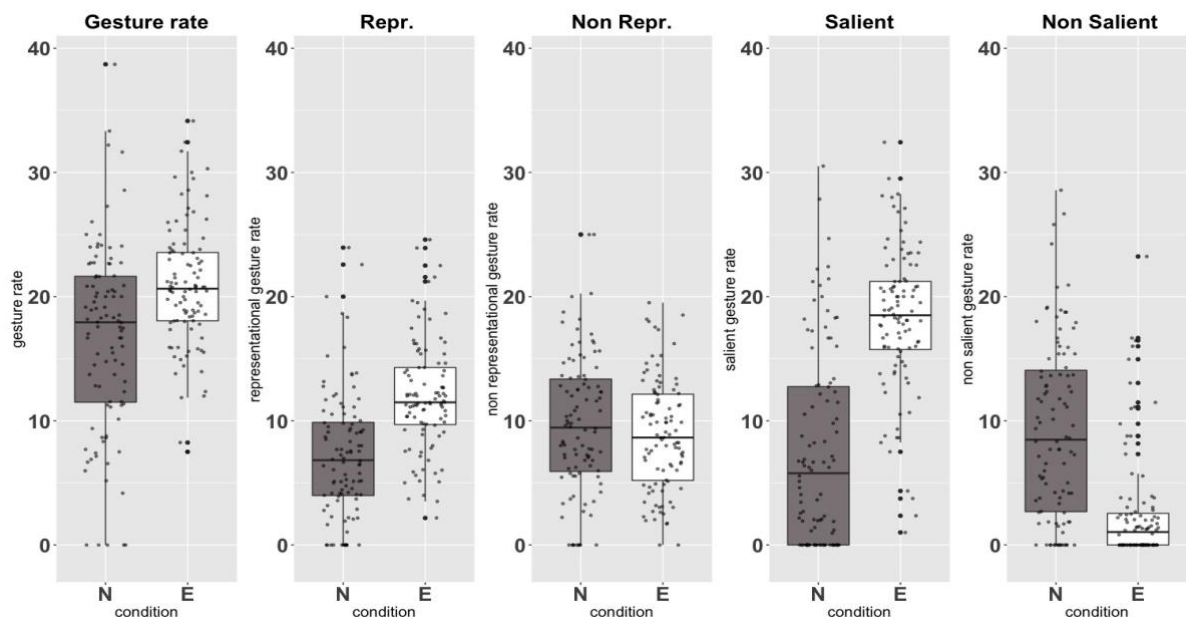


Figure 3. Boxplots representing Gesture rate, Representational, Non-Representational, Salient and non-Salient gesture rates in the two conditions, Non-Encouraging and Encouraging (N, E).

in the E condition (est. = 4.776, S.E =0.586, $p < .001$), while for Non-Representational gesture rate there was no significant difference between the two conditions.

Moreover, there was an effect of *Condition* on Salient gesture rate (est. =10.723, S.E =0.794, $p < .001$) that was found to increase in the E condition. The same applies, in the opposite direction, for Non-Salient gesture rate which is lower in the E condition than in the N condition (est. = - 6.589, S.E =0.65, $p < .001$).

The results show that the instruction to gesture (a) leads speakers to use more gestures; (b) leads to an increase of representational gestures; (c) makes speakers gesture in a higher and more salient gesture space. The latter, to our knowledge, had not been directly investigated before.

Table 1

LMEMs for the effects of Condition on the five measures of gesture rate (per 100 words)

Variable	Estimate	S. E.	C.I		t	Chisq	p
			Lower	Higher			
G rate	4.134	0.708	2.742	5.526	5.838	31.217	<.001
R gesture rate	4.776	0.586	3.624	5.929	8.149	56.306	<.001
NR gesture rate	-0.784	0.588	-1.94	0.371	-1.335	1.781	.182
S gesture rate	10.723	0.794	9.162	12.283	13.51	125.57	<.001
NS gesture rate	-6.589	0.65	-7.868	-5.311	-10.13	80.71	<.001

G: gesture; **R:** Representational; **NR:** Non-Representational; **S:** Salient; **NS:** Non-Salient; **Note:** Models: R function lmer in lme4 package (Bates, Mächler, Bolker, & Walker, 2014). Each model included Condition (N, E) as a fixed effect and both Story and Participant as random intercepts. N. of obs: 200; Groups: participants, 20 | Story, 10. C.I.: Lower 2,5%; Higher 97,5% (R package: *confint*). Levels “N” (baseline) and “E” were recoded by contrasts (i.e., 0 was in between each level, instead of being equal to N).

4. Conclusion

The aim of this study was to assess whether instruction to gesture can increase gesture rate as well as impact on gesture features such as gesture type and salience. The results show that in the gesture encouraged condition participants gestured more and in a higher gesture space. Also, they made more representational gestures than in the non-encouraging condition. These findings suggest that encouraging gesture in a speaking task can drive to effects other than the mere increase in gesture rate. It might be the case that encouraging the use of gestures leads speakers to automatically produce gestures that are more communicative and intended for the listener (Bavelas et al., 2008; Mol, Krahmer, Maes, & Swerts, 2009; Streeck, 1994) (e.g., produced in a higher more visible gesture space). Also, it might well be that explicit instructions on gesture can trigger an unconscious interpretation by speakers to use transparently iconic gestures, leading to an increase in representational gestures rate. It could also be that a narrative task itself is more likely to elicit more iconics compared with other speech tasks and this is worth further investigation. From a theoretical perspective, our results open a number of questions related to how the instruction to gesture have an impact in the process of speech planning and production.

The present study suggests that encouraging speakers to gesture in an experimental setting can effectively lead them to produce more gestures; this can limit the presence of speakers that provide no data, or just help achieving the goal of having the speakers gesture more. However, it should be considered that the prompt can lead speakers to make use of more iconics than they would naturally do and make use of space differently. This information, depending on the scope of the study, can be relevant when setting up an experiment using the instruction to gesture.

References

- Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and Producing Pitch Gestures Facilitates the Learning of Mandarin Chinese Tones and Words. *Studies in Second Language Acquisition*, 41, 33-58. <https://doi.org/10.1017/s0272263118000074>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.
- Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin*, 21(4), 394–405. <https://doi.org/10.1177/0146167295214010>

- 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., Kenwood, C., Johnson, T., & Phillips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, *2*(1), 1–17.
- Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture Changes Thought by Grounding It in Action. *Psychological Science*, *21*(11), 1605–1610. <https://doi.org/10.1177/0956797610385353>
- Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles*, *32*(1–2), 79–90.
- Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making Children Gesture Brings Out Implicit Knowledge and Leads to Learning. *Journal of Experimental Psychology: General*, *136*(4), 539–550. <https://doi.org/10.1037/0096-3445.136.4.539>
- Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, *140*(1), 102–116. <https://doi.org/10.1037/a0021790>
- Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694.
- Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2010). Gesturing makes memories that last. *Journal of Memory and Language*, *63*(4), 465–475.
- Cooperrider, K., Abner, N., & Goldin-Meadow, S. (2018). The Palm-Up Puzzle: Meanings and Origins of a Widespread Form in Gesture and Sign. *Frontiers in Communication*, *3*(June), 1–16. <https://doi.org/10.3389/fcomm.2018.00023>
- Gerwing, J., & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, *4*(2), 157–195. <https://doi.org/10.1075/gest.4.2.04ger>
- Goksun, T., Goldin-Meadow, S., Newcombe, N. S., & Shipley, T. (2013). Individual differences in mental rotation: What does gesture tell us? *Cognitive Processing*, *14*(2), 153–162. <https://doi.org/10.1007/s10339-013-0549-1>. Individual
- Holler, J., & Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, *24*(2), 267–289. <https://doi.org/10.1080/01690960802095545>
- Hostetter, A. B., & Hopkins, W. D. (2002). The effect of thought structure on the production of lexical movements. *Brain and Language*, *82*(1), 22–29.
- Hostetter, A. B., & Potthoff, A. L. (2012). Effects of personality and social situation on representational gesture production. *Gesture*, *12*(1), 62–83. <https://doi.org/10.1075/gest.12.1.04hos>
- Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, *56*(2), 291–303. <https://doi.org/10.1016/j.jml.2006.07.011>
- Kendon, A. (2004). *Gesture: visible action as utterance*. Cambridge University Press.
- Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. <https://doi.org/10.1080/01690960802586188>
- Llanes-Coromina, J., Prieto, P., & Rohrer, P. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. *9th International Conference on Speech Prosody 2018*, (June), 498–502. <https://doi.org/10.21437/SpeechProsody.2018-101>
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- Mol, L., Kraemer, 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. <https://doi.org/10.1075/gest.9.1.04mol>
- Nicoladis, E., Nagpal, J., Marentette, P., & Hauer, B. (2018). Gesture frequency is linked to story-telling style: evidence from bilinguals. *Language and Cognition*, *10*(4), 641–664. <https://doi.org/10.1017/langcog.2018.25>
- O'Carroll, S., Nicoladis, E., & Smithson, L. (2015). The effect of extroversion on communication: Evidence from an interlocutor visibility manipulation. *Speech Communication*, *69*, 1–8. <https://doi.org/10.1016/J.SPECOM.2015.01.005>
- Parrill, F., Cabot, J., Kent, H., Chen, K., & Payneau, A. (2016). Do people gesture more when instructed to? *Gesture*, *15*(3), 357–371.
- Streeck, J. (1994). Gesture as Communication II: The Audience as Co-Author. *Research on Language and Social Interaction*, *27*(3), 239–267. <https://doi.org/10.1207/s15327973rlsi2703>
- Vilà-Giménez, I., & Prieto, P. (2018). Encouraging children to produce rhythmic beat gestures leads to better narrative discourse performances. *9th International Conference on Speech Prosody 2018*, June, 704–708. <https://doi.org/10.21437/SpeechProsody.2018-143>
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. *Proceedings of LREC, 2006*.

An integrative platform to capture the orchestration of gesture and speech

Christelle Dodane¹, Dominique Boutet², Ivana Didirkova³, Fabrice Hirsch¹,
Slim Ouni⁴, and Aliyah Morgenstern⁵

¹Université Paul Valéry, France; ²Université de Rouen, France; ³Université Paris 8, France;

⁴Université de Lorraine, France; ⁵Sorbonne Nouvelle-Paris 3, France

Christelle.Dodane@univ-Montp3.fr; Dominique.Boutet@univ-rouen.fr;

Ivana.Didirkova@univ-paris8.fr; Fabrice.Hirsch@univ-Montp3.fr; Slim.Ouni@loria.fr;

Aliyah.Morgenstern@sorbonne-nouvelle.fr

Abstract

A number of studies have highlighted the coordination of gesture and intonation (Bolinger, 1983; Darwin, 1872; Cruttenden, 1997; Balog & Brentari, 2008; Roustan & Dohen, 2010) but the technological set-ups have been insufficient to couple the acoustic and gestural data with sufficient detail. In this paper, we present the MODALISA platform which enables language specialists to integrate gesture, intonation, speech production and content. The methods of data acquisition, annotation and analysis are detailed. The preliminary results of our pilot study illustrate strong correlations between gestures and intonation when they are simultaneously performed by the speaker. The correlations are particularly strong for proximal segments. Our aim is to expand those results and analyse typical and atypical populations across the lifespan.

1. Introduction

According to Bolinger (1983: 157), “*we READ intonation the same way we read gestures*”. In parallel with Darwin’s observations about gestures (1872), intonation is iconic in the sense that the meaning of upward and downward movements is related to attitudes and indirectly to metaphorical associations with tension, incompleteness and their opposites. Intonation has its own “*symbolizing power thanks to a primitive drive mechanism that raises pitch as tension rises and lowers it as tension falls*” (Bolinger, 1983: 156). It is part of our body movements which are more or less automatically concomitant to our state and our emotions. Bolinger highlights that gestures are coupled with intonation and display the same ascending and descending movements. Gesture and intonation may not systematically be produced together, but when they are, they are synchronized and co-expressive. Their synchrony does not necessarily mean that they work in unison, but rather that the parallel movements are coupled while the non-parallel movements are not.

Adult speakers coordinate their gestural behaviors and intonation when they speak, both in terms of time and direction: downward/forward movements are typically produced with descending contours and upward/backward movements with ascending contours (Bolinger, 1983; Cruttenden, 1997). Balog and Brentari (2008) observed the same type of synchronization in children aged 12 to 24 months and showed that children coordinate their verbal and non-verbal behaviors at the temporal and directional levels as early as the first word period, in order to be better understood by those around them. In their study, gesture coding was done by hand by observers who used a video in slow-motion and they had to indicate whether there was synchronization with the intonation or not. Using motion capture (OptiTrak recordings) on ten speakers, Roustan and Dohen (2010) showed that the prosodic focus attracts the manual gestures (pointing, beat and control gestures), pointing gestures being the most synchronized gestures (mainly between the apex of the pointing gesture and articulatory vocalic targets). Moreover, it has been shown that hand, head and eyebrow movements are aligned with pitch accents in speech and that this contributes to the production and perception of prosodic prominence (Ambrazaitis & House, 2017). All these studies indicate that it is crucial to work on the synchronization of prosody and gestural behaviors, in adults as well as children.

In order to achieve that goal, the MODALISA team has planned to create a multimodal platform that will make it possible to analyze prosody and gesture together. Indeed, to our

knowledge, there is no adequate instrument that makes it easy to measure gestures and prosody together. The objective of the MODALISA¹ project is thus to create an integrative procedure with the existing tools, that would make it possible to align the acoustic data with the gestural data. Instead of manual coding, we aim to use automatic extractions of the different gestural components (movements of the hands, forearms and arms) using several motion capture systems.

The original contribution of our project is that we use the gestural data complemented with articulatory and respiratory data obtained with other devices (laryngograph, articulograph, abdominal belt). This set-up allows us to create a truly multimodal platform for the simultaneous study of speech and gesture. It gives us access to objective, accurate and reliable data that will allow us to develop a large number of studies on speech and gesture. This paper presents our pilot study with the integrative system, our methodological procedure, preliminary results and perspectives.

2. Methods of data acquisition, annotation and analysis

For our pilot study, we implemented and tested the whole multimodal procedure on one participant.

2.1. Participant

A 33 years old French typical right-handed male speaker (MO1) was recorded in the premises of the LORIA laboratory in Nancy. The speaker had previously filled out a document asking for his consent indicating the different steps of the recordings and the equipment used.

2.2. Experimental paradigm

MO1 was recorded during a narrative task, in an experimental situation, inspired by McNeill's protocol (1992). Several sequences from a cartoon of the series *Tweety and Sylvester* (1949, Warner Brothers) were presented to him. After viewing each sequence, MO1 had to narrate it immediately to an interlocutor. MO1 was filmed throughout the duration of the task. We cut the cartoon into 5 sequences, including the "strike" sequence frequently exploited by the gesture community and which was chosen for this study in order to present the processing chain used to study the synchronization of gestures and prosody. For this short paper, we will focus on the acoustic data and on the gestural data exported from the IMU (Inertial Measurement Units, see just below).

2.3. Recording procedures

MO1 was recorded with two different motion capture devices (mocap). The first device (see Figure 1, left) consists of an electromagnetic articulograph (EMA) to record the movements of both hands and speech articulators (lips, tongue, jaw), a microphone to capture the acoustic signal and a video camera placed facing the speaker to film the entire scene.



Figure 1. Speaker (MO1) recorded with two different mocap systems: the electromagnetic articulograph (EMA, on the left) and the IMU suit (in the middle) with the visualization of the latter in Axis Neuron (on the right).

The device is completed by a laryngograph, which records the activity of vocal fold vibrations and a breathing belt recording the subject's abdominal movements. The EMA is normally used to study the movements of the main articulators of speech, i.e. the lips, tongue and jaw. The different movements are recorded every 5 ms using sensors placed on these different articulators. An

¹ Project funded by a grant awarded by the CNRS as part of the "Challenge Instrumentation aux Limites" call for projects in 2017 (Coordination: Christelle Dodane).

2.5. Synchronisation of the recordings

The data retrieved from the three sources - video, audio, and motion capture (mocap) – had to be synchronized with each other since the recordings did not start at the same time. Synchronization was performed under ELAN in which we can integrate the audio, video and mocap sources (with a beep or manual clap at the beginning).

2.6. Sampling frequencies

In addition to synchronization, the frequency of each of these recordings is not the same, it is even different within the same audio source. Indeed, the sampling frequency of the pitch is 10 ms. In concrete terms, this means that the gap increases as time goes by. The sampling rate of the images in the video is 40 ms. The timespan in transcripts under ELAN or Praat is variable and can be done to the nearest millisecond. The timespan for the mocap (Inertial Movement Unit) is 16.5 ms. Four different frequencies coexist in the data, in increasing order: a millisecond for transcription under Praat and / or ELAN, 10 ms for the pitch, 16.5 ms for the mocap and finally the timespan of the video is 40 ms. Video serves us primarily as a visual synchronization element, the data are not processed on this visual basis. In any case, we quickly found gaps in the data. These gaps increase progressively, and vary according to the type of data. It was therefore necessary to re-sample continuously in order to calibrate and coordinate the data without creating false data.

2.7. Resampling method

As we wanted to avoid to create false data, by using interpolation for example, the principle of resampling consisted in aligning the data associated with a short timespan (10 ms) from the existing data associated with a longer span in frequency. Thus, the first four temporal values of the mocap (i / 0 ms, ii / 16.5 ms, iii / 33 ms, iv / 46.5 ms), were aligned with the pitch data associated with the first six values (i / 0 ms, iii / 20 ms, iv / 30 ms, vi / 50 ms). Step by step, every 16.5 ms, the mocap data were compared with the pitch data that corresponded to the closest temporal values. When the matching by resampling of the pitch and mocap data were done, we needed to compare this re-alignment of the data with the Praat transcripts. Each unit (word, syllable, phoneme) has a beginning and an end. These intervals do not correspond to a fixed timespan, they depend entirely on what has been produced by the speaker. The temporal values of word boundaries can be corrected based on the closest values in the mocap output (values are inferior to 8.25 ms, ie 16.5 / 2).

3. Results

Table 1 summarizes our main results for pitch. The speaker has approximately the same speech rate in both tasks. He has a larger speech range and a lower mean pitch with the IMU suit.

Table 1

Measures based on the Prosogram application for the extract in which the motion capture was used along with the EMA and the IMU suit

Mocap	Speech rate	Pitch range	Mean Pitch	Max pitch	Speech time	Phonation time	Pause time
EMA	9,09 syll/sec.	8,1 1/2 tons	154 Hz	196 Hz	110,84 sec.	8,91 sec. (8,04 %)	101,93 sec. (91,96%)
Noitom Suit	9,2 (6,23) syll/sec.	9,2 1/2 tons	129 Hz	169 Hz	71,85 sec.	55,52 sec. (77,28%)	16,33 sec. (22,72 %)

All the studies that have so far explored the relationship between prosody and gestures have followed the positions and movements of the hand according to an absolute frame of reference. Among the sets of devices used in this study, the EMA falls under this type of absolute reference framework. In order to record the co-verbal gestures, these sensors are placed on the hands only, providing data on the position and movement of the hands in a unique and absolute reference frame being located in the recording room. Note that the hands may be submitted to a movement from higher up (arms, shoulders) without having moved on their own, i.e. the consequence of a movement of the arm is measurable on the hand. With a device like the EMA, one cannot detect and analyze the movements of the other segments nor the movement of the hand itself. Thus, to find out what the movements of all the segments of the upper limb are like, we can use the data from the IMU. The IMU enables us to situate the gestures in as many intrinsic reference frames as

there are segments: the position and the movement of each segment are given with respect to the adjacent and proximal segment. Thus, the movements of the arm are calculated according to the shoulder, those of the forearm, according to the arm, those of the hand are determined relative to the forearm. It is therefore possible to measure which segment is moving and by which angle in the three dimensions of each one's own space. The results of the relations between prosody and gesture in these intrinsic frames of reference are presented below. We can thus follow the evolution of the pitch and its possible impact on one of the 8 degrees of freedom of the upper limb, distributed from shoulder to hand. To our knowledge, these links have never been made. A correlation (Bravais-Pearson) was established between the rising ($N = 63$; mean time = 90.40 ms) and descending ($N = 111$; mean time = 92.74 ms) ranges of F0 and the degrees of freedom of the three segments (arms, forearm and hand) of the right upper limb and shoulder for the IMU recording. The linear correlation coefficients range between 1 and -1. Notice that there is a strong affinity between two sets of variables when their value is between 0.8 and 1 or between -0.8 and -1. As we get closer to the value 0, the series are less, if at all, correlated. These results are presented in the two tables below.

Table 2
Percentages of the number of correlation coefficients per range of 0.2 between rising fundamental frequencies and each degree of freedom of the right upper limb. The set of gestural possibilities are defined by degrees of freedom from shoulders to hands included

F0↗	Add/Abd Shoulder & F0↗	Rot Ext/Int Arm & F0↗	Exten/Flex Arm & F0↗	Add/Abd Arm 1 F0↗	Supi/Pro Forearm & F0↗	Exten/Flex Forearm & F0↗	Add/Abd Hand & F0↗	Exten/Flex Hand & F0↗
% corr. coef. $1 > x > 0,8$ or $-0,8 > x > -1$	55,56%	88,89%	82,54%	75,81%	74,60%	80,95%	68,25%	73,02%
% corr. coef. $0,8 > x > 0,6$ or $-0,6 > x > -0,8$	17,46%	3,17%	9,52%	8,06%	9,52%	6,35%	7,94%	9,52%
% corr. coef. $0,6 > x > 0,4$ or $-0,4 > x > -0,6$	6,35%	6,35%	1,59%	3,23%	6,35%	7,94%	9,52%	11,11%
% corr. coef. $0,4 > x > 0,2$ or $-0,2 > x > -0,4$	12,70%	1,59%	6,35%	8,06%	9,52%	3,97%	6,35%	3,17%
% corr. coef. $0,2 > x > -0,2$	7,94%	0,00%	0,00%	4,84%	0,00%	0,79%	7,94%	3,17%

Note: The set of degrees of freedom are defined in the position of the Vitruvian man (man standing with his palms facing forward, circumscribed in a circle, illustrated by Leonardo da Vinci). Abduction / adduction is a degree of freedom that moves a segment or a shoulder away or closer to the bust, in a frontal plane. The extension / flexion makes the segment pass behind or in front of the frontal plane, still in this general reference position of the Vitruvian body. The outer / inner rotation and the supination / pronation are degrees of freedom that turn the segment on itself, arm for the first, forearm for the second.

Table 3
Percentages of the number of correlation coefficients per 0.2 range between descending fundamental frequencies and each degree of freedom of the right upper limb

F0↘	Add/Abd Shoulder & F0↘	Rot Ext/Int Arm & F0↘	Exten/Flex Arm & F0↘	Add/Abd Arm 1 F0↘	Supi/Pro Forearm & F0↘	Exten/Flex Forearm & F0↘	Add/Abd Hand & F0↘	Exten/Flex Hand & F0↘
% corr. coef. $1 > x > 0,8$ or $-0,8 > x > -1$	50,45%	78,18%	72,97%	73,87%	69,37%	75,68%	60,36%	70,27%
% corr. coef. $0,8 > x > 0,6$ or $-0,6 > x > -0,8$	18,02%	12,73%	13,51%	14,41%	16,22%	14,86%	19,82%	15,32%
% corr. coef. $0,6 > x > 0,4$ or $-0,4 > x > -0,6$	16,22%	5,45%	5,41%	6,31%	4,50%	3,60%	6,31%	6,31%
% corr. coef. $0,4 > x > 0,2$ or $-0,2 > x > -0,4$	11,71%	1,82%	5,41%	4,50%	6,31%	3,15%	6,31%	4,50%
% corr. coef. $0,2 > x > -0,2$	3,60%	1,82%	2,70%	0,90%	3,60%	2,70%	7,21%	3,60%

Note: The negative values of the correlation coefficients appear when for the same F0 slope, the pole of the correlated degree of freedom is of an opposite sign. Thus, when the correlation coefficient is negative, for a descending F0, then for example for the arm, its movement corresponds to a flexion (forward or upward). For a positive value, always with descending F0, the arm will have an extension movement (backward or downward).

Tables 2 and 3 indicate that the correlations between the variations of F0 and the degrees of freedom are very strong (at least 60% of the cases higher than a coefficient that is equal to or higher than $\|0.8\|$). It means that the rate of change of joint angles for every segment is correlated with intonation. Moreover, these correlations distributed over all segments of the upper limb, are particularly important for the arm and decrease globally as we take the movement of the forearm and hand into consideration. Even if the high correlation rate remains present for these latter segments, we notice a decrease in the co-variation with the pitch for the distal segments, in particular with a shift towards lower values (between 0.6 and 0.4). In other words, the further one gets away from the bust, in terms of segments (and not of distance), the less powerful this co-variation becomes. We don't know yet whether this anisotropy is structural or if it comes from a temporal shift due to the time needed for the movement of the arm to propagate towards the hand. In favor of this last hypothesis, the average duration of the variations of F0 is about 90 ms when the average duration of the gestures is about 150 ms. A gesture that begins from a proximal segment, cannot have fully developed over all the segments by the time the rise or fall of the fundamental frequency is reached. These questions explain a) the common structuration between prosody and gesture b) their synchronization c) the management of various temporalities.

4. Perspectives

The MODALISA project has reached its technological goal as we have now created a multimodal, multidevice platform in order to collect data on both speech and gesture as well as a methodology to process and analyze the multimodal data. The pilot study we presented in this paper indicates strong correlations between gestures and intonation when they are simultaneously performed by the speaker. The correlations are particularly strong for proximal segments. It would thus be particularly important to analyze head gestures (as advised by Bolinger, 1983). The advantages of the MODALISA platform are that we use MOCAP systems with different frames of reference (absolute (EMA)/intrinsic (IMU) and that we have the possibility to integrate articulatory gestures (EMA), respiratory movements (respiratory belt), vibratory movements of the larynx (laryngograph) with prosody and gesture. Otherwise, we aim to adapt the IMU suit to children's physiological constraints. We can also coordinate the various exported data with our annotations of the video-data on ELAN. The platform is used in various projects by our team to study how prosody and gesture synchronize across the life-span in typical and atypical populations. Our goal is to capture whether integration of polysemiotic resources is quantitatively or qualitatively different in children as their motoric, cognitive and linguistic skills develop and in adults as they reach old age. This instrumental device will give us access to objective, accurate and reliable data that will enable us to develop a large number of studies on child language acquisition, on adult speech, on typical as well as pathological speech. For example, it is crucial to determine how gestures (hands, forearms and arms) interact with articulatory and respiratory levels during episodes of disfluencies produced by stuttering speakers as well as fluent speakers (work in progress). Such multimodal integration of speech is innovative because it allows us to interconnect levels that had not been studied together so far. A better understanding of the way these different levels interact will contribute to a better view of the constraints of speech production with a multiparametric approach.

Acknowledgments

The authors would like to thank Marjorie Bosqué and Cwiosna Roques for their annotation work on the data files.

References

- Ambrazaitis, G. & House, D. (2017). Acoustic features of multimodal prominences: Do visual beat gestures affect pitch accent realization? In Ouni, S., Davis, C., Jesse, A. & Beskow, J. (eds). *Proceedings of the 14th International Conference of Auditory-Visual Speech Processing (AVSP2017)*, KTH.
- Balog, H. and Brentari, D. (2008). The relationship between early gestures and intonation. *First Language*, 28, 141-163
- Boersma, P., Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.15) [Computer Program]. Consulté le 25.01.2017 de PRAAT: <http://www.praat.org/>
- Bolinger, D. L. (1983). Intonation and gesture. *American Speech*, 58, 156-174.
- Cruttenden, A. (1997). *Intonation* (second edition). Cambridge: Cambridge University Press.
- Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*, London: John Murray.
- Goldman, J.P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat Proceedings of InterSpeech, September 2011, Firenze, Italy.

- Lacheret, A., Kahane, S., Beliaou, J., Dister, A., Gerdes, K. et al. (2014). Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. *4e Congrès Mondial de Linguistique Française*, Jul 2014, Berlin, Allemagne. 8, pp.2675-2689.
- Mertens, P. (to appear early 2019). The Prosogram model for pitch stylization and its application in intonation transcription. In Barnes, J.A. & Shattuck-Hufnagel, S. (eds.). *Prosodic Theory and Practice*. Cambridge: MIT Press.
- Roustan, B. & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: temporal coordination and effects on the acoustic and articulatory correlates of focus. *5th International Conference on Speech Prosody (Speech Prosody 2010)*, May 2010, Chicago, United States, 100110:1-4, 2010.
- Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Gesture / speech alignment in weather reports

Gaëlle Ferré

Nantes Université & CNRS UMR 6310 – LLING, Nantes, France

Gaëlle.Ferre@univ-nantes.fr

Abstract

This paper presents a follow-up study of previous work conducted on pointing gestures and their alignment with speech in weather reports (Ferré & Brisson, 2015, Ferré, 2019). Yet, whereas the previous studies concentrated on the expression of viewpoint and how gestures function in association with other semiotic resources, the present study focuses in more detail on the timing relationships between the different modes in speech and the apparent absence of synchronicity in some gesture/speech constructions in French weather reports. What is proposed here is a theoretical analysis rather than a quantitative one in which it will be shown that in order to account for this apparent misalignment of modalities (a) the inclusion of other semiotic modes in the annotation scheme may be useful for the description of specific corpora like weather reports, and (b) temporal graphs that include gesture targets can offer a good representation of the temporal relationships between gesture and other domains involved in communication acts.

1. The challenge of multimodality

Multimodality implies that the meaning making process constantly involves several semiotic resources (Adami, 2017). Oral communication in face-to-face interactions, as an instance of meaning making, involves not only language, but also gesture, posture, facial expression and other bodily behaviors such as proxemics and attitudes. Spoken interactions also typically occur within a physical environment of which certain elements can be integrated in communication acts, as shown by Goodwin (1994, 2007) and Streeck (1996), and form their own semiotic system. Each semiotic system involved in communication acts has its own systemic affordances and material constraints so that what is communicable in speech may not be so easy to communicate in a visual mode (like gesture or graphic representation) and vice versa. This is the reason why Discourse Analysis should not focus on only one modality even if some modalities can be predominant in certain social practices, as in the type of media that is the object of study in the present paper.

In some ideal world, any speech act would contain at least one syntactically complete and grammatically correct clause made of words themselves formed with distinct morphemes. The clause would be bounded by clearly identifiable prosodic boundaries and would be uttered with an intonation contour that would be congruent with the speech act accomplished verbally (whether it be a statement, a question, or any other type). The syntactic clause could also be accompanied by a gesture whose onset and offset would precisely match the syntactic and prosodic boundaries. This gesture would in turn be composed of different phases that would also match the lexical or morphemic boundaries in speech. Lastly, the information conveyed by gesture and prosody would be congruent with the semantics of the clause and its constituents.

This ideal communication act is indeed found in spontaneous interactions, but as anyone who has worked before on naturally occurring interactions knows, misalignments also occur and this therefore makes the issue of the temporal alignment of information units in the different modalities and their conjoint analysis a central one in any multimodal study of video corpora. Considering this issue, the challenge of multimodal discourse analysis, i.e. the study of relationships between different modalities in discourse and the way each modality participates in meaning-making processes, consists in annotating data in linked but nevertheless different semiotic modes that do not always share the same temporal structure and in revealing the interactions between them in as systematic a treatment as possible.

2. Temporal alignment of modalities

Generally speaking, the vast majority of Intonation Phrases temporally coincide with syntactic clauses in speech (Barth-Weingarten, 2016), although this depends a lot on the degree of improvisation and informality of interactions. In this respect, weather reports are well rehearsed types of media, based on scripted material, which means that speech delivery is very fluent. This type of media also takes the form of monologues and prosody is therefore not used as a turn-management device as can be the case in dialogues, in which speakers sometimes purposely avoid to make syntactic and prosodic boundaries coincide not to lose their speech turn.

As far as gestures are concerned, it has been observed that gesture production is linked to the syntactic structure of the speech it accompanies depending on the language of the speaker: Kita and Ozyürek (2003) noted that if some information is typically given in the form of two syntactic clauses in a language, speakers tend to express this same information with two different gestures, whereas when the language enables speakers to express the information in a single syntactic clause, then speakers tend to produce a single gesture to accompany their verbal expression. In terms of discourse structure, McNeill (1992) also observed in a narrative task that speakers tend to produce one gesture per narrative clause which means that a gesture in this case participates in the expression of one idea unit.

Yet, there are differences between gestures and prosody. The major difference between the visual and vocal semiotic modes lies in the fact that whereas one cannot speak without any prosody at all, hand gestures are not required to accompany every piece of verbal information, which means that manual gestures are perhaps a bit more independent from speech than prosody. It also means that not every syntactic clause is accompanied by a gesture.

Speech and gesture may also differ in respect to their temporal structure and this has an impact on their alignment with each other. Shattuck-Hufnagel and Ren (2018) signal that studies concerned with gesture/speech synchronization present contradictory results: while some scholars found a (fairly) good alignment between gesture and speech (Loehr, 2004; Chui, 2005, for instance), others found that some gesture types tend to be produced in anticipation of speech (Schegloff, 1984; Leonard and Cummins, 2009; Ferré, 2010). Shattuck-Hufnagel and Ren (ibid.) however note that the different observations made in this respect may be explained by the fact that scholars were working on different languages and considering different gesture types or even base their observations on different gestural landmarks (gesture apex, whole stroke or even whole gesture phrase) and with different time windows. For McNeill – although the author doesn't specify how precisely he measured this figure – (2005, p. 32) “the stroke is synchronous with co-expressive speech about 90 percent of the time (...). When strokes are asynchronous, they precede the speech to which they link semantically”, i.e. their *lexical affiliate* (Kipp et al., 2007).

3. Semantic gesture/speech mismatches in weather reports

While working on pointing gestures in weather reports both in English and French, we observed a difference between the two languages in terms of gesture/speech alignment. In French, mismatches were found slightly more often than in English between some pointing gestures and the locations pointed at on the map shown in a background screen. Whereas in French weather reports, 9 % of the pointing gestures towards a location on the screen showed clear misalignment with the location mentioned in speech – and therefore fit well with the description provided by McNeill (ibid.) quoted in the previous section – the English corpus showed a lower misalignment rate of 4 %. Although the corpus is extremely limited in size and the difference may not be significant, we may still wonder if we can really talk of gesture/speech mismatches in these cases in French and why the two languages tend to function differently in this respect.

Figure 1 below presents a sequence in a French weather report, in which some gestures do not align with what is referred to in speech. As he begins a new description in (a), the forecaster mentions the ‘Val de Garonne’ and points to this particular location on the map of France that is shown in the background screen. He then initiates a second move but the Intonation Phrase shown in (b) does not contain any spatial reference. Yet, the forecaster anticipates on the next Intonation Phrase and already points to the Pyrenees. In (c) where the Pyrenees are mentioned in speech, he anticipates again in gesture on the next Intonation Phrase and moves his hand directly to the Alps so that as he finishes the word ‘Pyrénées’, his hand is now fully pointing at the Alps on the map.

He catches up in the last Intonation Phrase shown in (d) and points to the Mediterranean as he utters ‘the Alps and the Mediterranean’ packaged in a single Intonation Phrase. The misalignment between speech and gesture is so large that although the apex of each gesture is aligned with the stressed syllable of each Intonation Phrase and the gestures could then be considered as respecting the gesture-speech alignment rules observed by Loehr (2004), there is a mismatch of more than 200 ms in semantic content between what is referred to in speech and what is pointed at in the background screen for two gestures in the sequence.

In sum, the example illustrated in Figure 1 shows that whereas the first and last gestures in the sequence align their apex with the right locations in speech, the second gesture does not align with any spatial location in speech and the third one points at the Alps on the map when the Pyrenees are mentioned in speech. The last gesture produced by the forecaster aligns with one of two locations mentioned in speech. It starts as the Alps are being mentioned but its apex coincides rightly with the mention of the Mediterranean. The gestures in (b) and (c) are then clearly misaligned with their lexical affiliates and the gesture/speech constructions in these two cases seem rather ill-formed, unless one considers that the different elements or *objects* forming the construction can be analyzed in terms of their respective properties and the relationships they entertain with each other on different planes of discourse (Blache, 2004). These relationships between objects in utterances can be represented in the form of temporal graphs (Bird and Liberman, 1999).


	/de la grisaille brumeuse de nouveau/ <i>Grey mist again</i>	
(a)	/dans le Val de Garonne/ <i>in the Val de Garonne</i>	
(b)		/mais beaucoup de soleil/ <i>but very sunny</i>
(c)	/en allant vers les Pyrénées/ <i>towards the Pyrenees</i>	
(d)		/les Alpes et la Méditerranée/ <i>the Alps and the Mediterranean</i>

Figure 1. Gesture / speech temporal (mis)alignment in French (Prévisions Météo-France, 17 Nov. 2015).

4. Temporal graphs for gesture/speech constructions in weather reports

Bird and Liberman (ibid.) consider that any linguistic domain (prosody, gesture, discourse, syntax, phonology...) comprises a number of *objects* organized in a linear way on a temporal axis, so that a multimodal corpus is composed of different objects with an onset and offset time that can be represented by nodes on a timeline. A weather forecast, as said before, is a type of media based on three major semiotic resources: speech, gesture and a background screen. The aim of pointing gestures in this communication type is to establish a link between the background screen and speech content and to open up focus spaces on that screen for the audience to concentrate on (Grosz and Sidner, 1986). The example presented in the previous section can be represented as in Figure 2. (a) shows a multimodal construction in which an Intonation Phrase made of a single syntactic Prepositional Phrase includes a lexical reference to a spatial location. The phrase is accompanied

by a pointing gesture towards a congruent location on the map shown in the background screen. In (b) the syntactic phrase uttered in an Intonation Phrase is also accompanied by a point towards a location on the map, but the gesture-map construction links to the location expressed verbally in (c). The gesture that accompanies the Intonation Phrase here in turn matches a location expressed verbally in (d). This last Intonation Phrase groups two syntactic Noun Phrases that refer to two different spatial locations, but the gesture produced during the utterance of this phrase targets the last spatial location mentioned in speech.

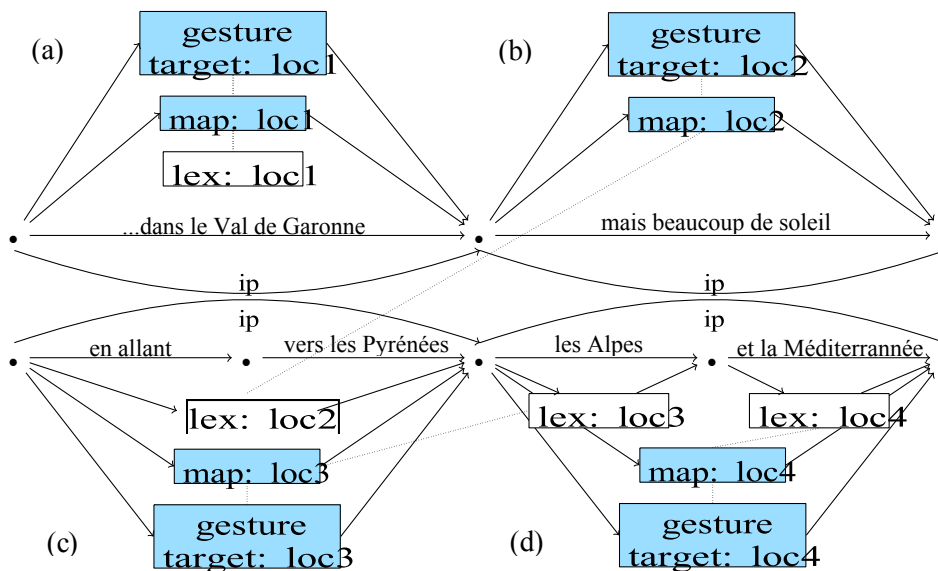


Figure 2. Graph showing dependency relations between syntax on the timeline, prosody, gesture and a visual map (ip = Intonation Phrase, lex = lexical information, loc = location).

5. Conclusion

As was shown in this paper, multimodal constructions may well be composed of objects temporally synchronized with each other as in Figure 2 (a), where the syntactic, semantic, prosodic and gestural domains are all congruent with one another and are besides perfectly coupled with the communication environment (a location on a map, for instance, in the case of weather reports). They may however also be partly synchronous with one another as in Figure 2 (b), (c) and (d): whereas (b) comprises a single syntactic phrase uttered in an Intonation Phrase, (c) and (d) both comprise two syntactic phrases packaged in single Intonation Phrases. Besides, if the gestures produced in these three constructions are nicely aligned with Intonation Phrases, their targets in (b) and (c) are not synchronized with the corresponding spatial locations in speech. This means that multimodal constructions are not always based on the semantics of speech, but rather on the way the information is packaged into prosodic units.

Lastly, although the corpus on which this theoretical paper is based is quite limited in size thus precluding any generalization, it appeared that pointing gestures were more frequently misaligned with referential spatial locations – as they tended to anticipate the lexical reference more often – in French weather reports than in English ones. This might be due to the different information structure of the two languages: whereas spoken English is very similar to written English considering word order, there is a large difference between written and spoken French in terms of information structure, with a tendency to place focused elements at the beginning of a sentence in spoken French. The semantically misaligned gestures in weather reports, that open up focus spaces on a map, may be considered to be following the information structure of oral French (which could be the reason why they tend to align with intonation rather than syntactic phrases), whereas the verbal information, based on scripted material, rather follows the information of written French which could explain the fact that pointing gestures in weather reports anticipate more frequently on speech in this language.

References

- Adami, E. (2017). Multimodality. In García, O., Flores, N., and Spotti, M., editors, *The Oxford Handbook of Language and Society*, pages 451–472. Oxford University Press, Oxford.
- Barth-Weingarten, D. (2016). *Intonation Units Revisited. Cesuras in Talk-in-Interaction*. Amsterdam, Philadelphia: John Benjamins.
- Bird, S. and Liberman, M. (1999). A Formal Framework for Linguistic Annotation. In *Technical Report MS-CIS-99-01*, pages 1–48, University of Pennsylvania.
- Blache, P. (2004). Property Grammars: A Fully Constraint-Based Theory. In Christiansen, H., Rossen Skadhauge, P., and Villadsen, J., editors, *Constraint Solving and Language Processing*, pages 1–16. Springer, Berlin.
- Chui, K. (2005). Temporal Patterning of Speech and Iconic Gestures in Conversational Discourse. *Journal of Pragmatics*, 37:871–887.
- Ferré, G. (2010). Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French. In Kipp, M., Martin, J.-C., Paggio, P., and Heylen, D., editors, *LREC: Workshop on Multimodal Corpora*, pages 86–91, Valetta, Malta. ELRA.
- Ferré, G. (2019). Time Reference in Weather Reports. The Contribution of Gesture in French and English. In Galhano, I., Galvão, E., and Cruz dos Santos, A., editors, *Recent Perspectives on Gesture and Multimodality*, pages 31-40. Cambridge Scholars Publishing Ltd, Cambridge.
- Ferré, G. and Brisson, Q. (2015). “This Area of Rain will Stick South in the Far North”. Pointing and Deixis in Weather Reports. In *Proceedings of GESPIN 4*, pages 101–106, Nantes, France.
- Goodwin, C. (1994). Professional Vision. *American Anthropologist*, 96(3):606–633.
- Goodwin, C. (2007). Environmentally Coupled Gestures. In Duncan, S., Cassell, J., and Levy, E., editors, *Gesture and the Dynamic Dimensions of Language*, pages 195–212. Amsterdam, Philadelphia: John Benjamins.
- Grosz, B. J. and Sidner, C. L. (1986). Attention, Intention, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.
- Kipp, M., Neff, M., and Albrecht, I. (2007). An Annotation Scheme for Conversational Gestures: How to Economically Capture Timing and Form. *Language Resources and Evaluation*, 41:325–339.
- Kita, S. and Ozyürek, 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*, 48:16–32.
- Leonard, T. and Cummins, F. (2009). Temporal Alignment of Gesture and Speech. In *Proceedings of Gespin*, pages 1–6, Poznan, Poland.
- Loehr, D. P. (2004). *Gesture and Intonation*. PhD thesis, Georgetown University, Georgetown.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, London.
- McNeill, D. (2005). *Gesture and Thought*. University of Chicago Press, Chicago, London.
- Schegloff, E. A. (1984). On some Gestures’ Relation to Talk. In Maxwell Atkinson, J. and Heritage, J., editors, *Structures of Social Action. Studies in Conversation Analysis*, pages 266–296. Cambridge University Press, New York.
- Shattuck-Hufnagel, S. and Ren, A. (2018). The Prosodic Characteristics of Non-referential Co-speech Gestures in a Sample of Academic-Lecture-Style Speech. *Frontiers in Psychology*, 9:1–13.
- Streeck, J. (1996). How to Do Things with Things: Objets Trouvés and Symbolization. *Human Studies*, 19:365–384.

Gesture-speech coordination in expression of motion: How far to zoom in to observe semantic synchrony?

Katerina Fibigerova and Michèle Guidetti

CNRS-CLLE-University Toulouse 2, France

katerina.fibigerova@univ-tlse2.fr, michele.guidetti@univ-tlse2.fr

Abstract

The present paper contributes to the discussion about coordination between gesture and speech from the semantic and morpho-syntactical perspective. What information is conveyed in co-speech gesture and how that information relates to the content of the co-occurring segment of speech? Does temporal synchronicity imply semantic synchronicity?

We tackled these questions in the context of description of motion events, in terms of combinations of a specific path (e.g. upward, downward, crossing) and a specific manner (e.g. walking, running, flying). We asked whether gesture depicts the same element(s) of motion that speech does and to ensure variability of verbal content we adopted the comparative method involving French and Czech speakers, two languages offering different patterns for expression of motion path and manner.

This paper/talk presents our most recent results that extend our previous studies in this field. After having observed gesture-speech semantic synchrony at the level of ‘gesture-proposition’ and ‘gesture-word’, it was time to zoom into individual words and explore the ‘gesture-morpheme’ level.

1. Introduction and theory

1.1. Gesture-speech synchrony

From the theoretical perspective, we are interested in the type of gesture that is produced during speech (see ‘gesticulation’ in McNeill 1992). Co-speech gestures are defined as hand and body actions that carry semantic content and co-construct meaning in conjunction with speech (Kendon 2004).

There is evidence the appearance of gestural movements during speech is not fortuitous and that both modalities are semantically synchronized (Kendon 2004; Kita 2000; McNeill 1992). Synchronicity has different aspects. One of them is ‘semantic co-expressivity’ that describes the fact gesture and speech produced simultaneously share the same reference (i.e. they relate the same thing).

Another aspect is ‘semantic redundancy’ which indicates the possibility that gesture and speech produced simultaneously also provide the same pieces of information about the shared reference (i.e. they relate to the same thing depicting the same characteristics of it). Since in a multimodal expression, information is distributed and expressed via different modalities/tools/channels which are different and complementary, a bimodal gesture-speech expression benefits jointly from gesture’s iconic and holistic qualities and the abstract and analytico-syntactical properties of speech (McNeill 1992, 2005). As a result, some elements of an idea might be expressed in one modality rather than in the other one (see ‘information packaging’ in Kita 2000).

1.2. Motion in different languages

Motion is understood here as deliberate ‘change of placement’ (Aurnague 2011) or ‘translocation’ (Levinson & Wilkins 2006). In the traditional conceptual analysis (Talmy 1985, 2000), motion includes several components: figure that is moving (e.g. a dog), path or direction of motion (e.g. across something), manner or how motion is executed (e.g. by running), and finally ground or the reference point (e.g. the street).

There is evidence that languages vary in lexicalization of information about path and manner (Talmy 1985, 2000), which directly impacts the way speakers of different languages talk about motion (e.g. Slobin 2000, 2004).

In ‘verb-framed’ languages, path is encoded in the main verb while manner is typically added in gerunds:

1) French:	<i>descendre</i>	<i>l’escalier</i>	<i>en courant</i>
	to.descend	the stairs	by running

In ‘satellite-framed’ languages, path is encoded in verb satellites while manner is carried in the verb root:

2) English:	to run down the stairs
-------------	------------------------

It seems ‘verb+gerund’ combinations are more complex syntactical constructions—that ask for more robust cognitive treatment—than ‘verb+satellite’ combinations, which manifests in the fact that speakers of verb-framed languages tend to mention only the path of motion, omitting the manner, while speakers of satellite-framed languages systematically indicate both elements.

1.3. Motion expressed in speech and gesture

When the interest in gesture-speech semantic relation meets the study of cross-linguistic variability in motion expression, the general question often raised up is that of the impact of the latter on the former: When path/manner/both is expressed in speech, is it also present in co-occurring gesture? Are gesture and speech about motion path and manner semantically redundant across languages?

Gullberg, Hendriks and Hickmann (2008) observed gesture-speech relation at ‘sentence’ level where the considered verbal units co-occurring with speech were either a simple sentence or a complex sentence with subordinate clause. As French native speakers in their study predominantly verbalized and simultaneously gestured about path, it seemed that gesture and speech about motion were semantically redundant (schematically ‘path in speech as well as path in gesture’). However, since no comparison to other typologically different language was made, it would be problematic to generalize this conclusion.

A comparative study between English and Turkish speakers was conducted by Özyürek, Kita, Allen, Brown, Furman, and Ishizuka (2008). However, the aim of that study was to explore organization of information inside complex verbal and gestural expressions in which both path and manner were indicated. For this purpose, redundant speech-gesture descriptions of type ‘path+manner in speech as well as path+manner in gesture’ were elicited in speakers of both languages as a preliminary condition or first step for subsequent observations rather than spontaneously produced by speakers themselves. This is why this particular study does not really fit into our topic.

When Hickmann, Hendriks and Gullberg (2011) extended their previous study (Gullberg, et al. 2008) by comparing French natives to English ones, they concluded that semantic redundancy was impacted by the type of language as well as by the size of verbal unit considered as simultaneous with gesture or, in other words, by the level of observation chosen for analysis of gesture-speech relation. While French speakers produced predominantly redundant gesture and speech (typically ‘path in speech as well as path in gesture’), in English speakers, the situation was more complex. At ‘gesture – whole surrounding proposition’ level, English gesture and speech were semantically mostly non-redundant (typically ‘path+manner in speech but path alone in gesture’), which resulted into a statistically significant difference between both language groups. This being said, English gesture and speech turned into mostly redundant at ‘gesture – aligned proposition segment’ so that the effect of language totally disappeared. The question remained whether the impact of level/unit of observation on semantic relations between gesture and speech was specific to English or it was a characteristic of satellite-framed languages in general.

For this reason, Fibigerova et al. (Fibigerova 2012; Fibigerova, Guidetti, & Sulova 2012; Fibigerova & Guidetti 2018) replicated the study by Hickmann et al. replacing English with Czech. Their French-to-Czech comparison observed first at ‘gesture – whole surrounding proposition’ level generated results similar to Hickmann et al. The effect of language type manifested through redundancy in French participants (most frequently ‘path in speech as well as path in gesture’) that contrasted with non-redundancy in Czech group (predominantly ‘path+manner in speech but path alone in gesture’). Then, a second analysis was conducted at ‘gesture – aligned proposition segment’ level. This time, although the proportion of redundant gesture-speech couples increased in comparison to the previously explored level of analysis, the difference between Czech and French speakers was still significant. In conclusion, this last result reported by Fibigerova et al. contrasted

with the result brought by Hickmann et al. showing the absence of difference between English and French speakers at this level of analysis. Both studies considered together seem to reveal deeper intra-typological differences inside satellite-framed languages.

In spite of this conclusion, a doubt remains concerning the stated semantic non-redundancy between Czech gesture and speech about motion and this is why we decided to further explore this point.

2. Present study

2.1. Question and hypothesis

Gullberg, Hendriks and Hickmann (2008) as well as Hickmann, Hendriks and Gullberg (2011) define the segment of proposition considered as simultaneous with gesture in terms of ‘word(s)’ that are exactly aligned with main gesture stroke. In English, ‘gesture-word’ level allows to consider the Examples 3 and 4 as cases of redundancy between gesture and speech:

- 3) A bear climbed up the tree.
+ path gesture aligned with ‘up’
- 4) A bear climbed up the tree.
+ manner gesture aligned with ‘climbed’

In Czech language, the situation is more complex. ‘Gesture-word’ level does help to increase the proportion of redundancy in situations illustrated by Examples 5 and 6 – that are nevertheless less frequent – but not in situations illustrated by Example 7 – that are indeed predominant:

- 5) *Medvěd* *vyšplhal* *nahoru* *na strom.*
bear up.climbed upwards on tree
+ path gesture aligned with ‘*nahoru*’
- 6) *Medvěd* *šplhal* *nahoru* *na strom.*
bear was.climbing upwards on tree
+ manner gesture aligned with ‘*šplhal*’
- 7) *Medvěd* *vyšplhal* *na strom.*
bear up.climbed on tree
+ path gesture aligned with ‘*vyšplhal*’

At this point, we wonder whether the proportion of redundancy would increase even more if we could deal with situations illustrated by Example 7. For this purpose, we introduce a third level of observation of semantic relation between gesture and speech that we call ‘gesture-morpheme’ level. Without any aspiration to explore every single gesture-morpheme combination, we use the term of ‘morpheme’ only as tool that will help us to ‘separate’ verbal prefix from verbal root so that we could consider them as two different units, each encoding a single element of motion (either path or manner). Thus, we would be able to identify the finest cases of semantic redundancy as shown in Example 8:

- 8) *Medvěd* *vyšplhal* *na strom.*
bear up.climbed on tree
+ path gesture aligned with ‘*vy-*’

Our aim is to see whether the difference between Czech and French speakers—after being strong and significant at ‘gesture-proposition’ level and less strong but still significant at ‘gesture-word’ level—will finally disappear at ‘gesture-morpheme’ level. Will we obtain the same effect as Hickmann et al. did in their English-French study after having zoomed in a lower level of observation? We formulate the following hypothesis: the tighter/stricter definition of temporal alignment between gesture and speech, the more semantic synchrony between the two modalities emerges.

2.2. Methodology

Our study is based on 24 French and 24 Czech native monolingual speakers, all young adults (20-35 years old), mainly students, living in their respective countries. The data were collected during individual sessions of watching and narrating short video clips showing 50 different motion events

represented as combinations of specific path and manner. The video clips were several seconds long animated stories with the same structure (a character arrives, realizes the target motion and leaves) that were created especially for different motion event studies (Allen et al. 2007; Fibigerova 2012; Fibigerova, Guidetti, Šulová 2012; Hickmann 2006). The filmed narrations were then transcribed and annotated using ELAN Linguistic Annotator (10% of our data was annotated by two independent coders).

Firstly, we selected all sentences related to each target motion and all motion related iconic gestures produced during those sentences. Secondly, each gesture was coded according to which element of motion it expressed: a) path alone, b) manner alone, c) both. Thirdly, we identified the segments of speech – one or more morphemes – that were exactly aligned with the main gesture strokes (i.e. the meaningful part of a hand/body movement). Fourthly, each speech segment simultaneous with gesture was coded according to which element of motion it expressed: a) path alone, b) manner alone, c) both. Finally, each gesture-speech couple was annotated as ‘semantically redundant’ if one of the three situations held: 1) path alone in speech as well as in gesture, 2) manner alone in speech as well as in gesture, 3) both path and manner in speech as well as in gesture. Otherwise, the couple was labeled ‘non-redundant’ (e.g. both path and manner in speech but path alone in gesture).

2.3. Results

After coding, we proceeded to the comparison of the mean proportions of semantically redundant ‘gesture-morpheme’ couples produced in both language groups (see Figure 1). The frequency of redundant cases was much higher in French group ($M = .73$, $SD = .121$) than in Czech group ($M = .26$, $SD = .222$). Since our data were asymmetrically distributed, we used non-parametrical Mann-Whitney U test that confirmed the significance of the found difference ($z = -5.217$, $p < .000$).

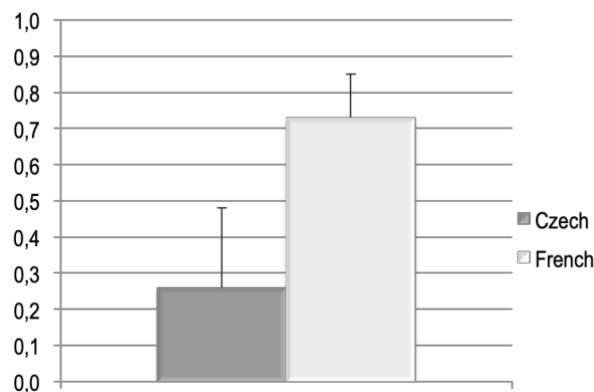


Figure 1. Mean proportions of semantically redundant ‘gesture-morpheme’ couples.

3. Discussion and conclusion

Contrary to our expectations, we did not obtain similar results to those reported by Hickmann, Hendriks and Gullberg (2011). The difference between Czech and French speakers remains significant even at ‘gesture-morpheme’ level. Our hypothesis, according to which the tighter/stricter definition of temporal alignment between gesture and speech, the more semantic synchrony between the two modalities emerges, has been only partially confirmed. In Czech (and perhaps in Slavic languages?), semantic redundancy increases slightly at ‘gesture-word’ level but it decreases again with ‘gesture-morpheme’ level.

As speakers of satellite-framed languages gesture mostly about path alone, we were the most interested in whether path gestures will be synchronized with path satellites. In spite of very similar situations in English and Czech, the fact that English particles are independent elements placed after the verb while Czech prefixes are bound morphemes placed in front of the verb might be the origin of the different findings reported for these two languages (see Dewell 2011 for analyses of ‘prefixed verbs’ vs. ‘particle verbs’ in German).

When ‘gesture-speech-mind’ unity is concerned, the literature typically mentions two semantic levels: 1) gesture stroke and simultaneous word(s) to express a concept or/and 2) gesture stroke and simultaneous proposition to express an idea (see ‘growth point’ McNeill 1992, 2005). A bound

morpheme might not be able to function as a sufficiently solid semantic unit so that gesture could be semantically synchronized with one it, independently of the rest of the word. This might also be related to different cognitive processing of information at morpheme vs. word level (see e.g. Giraudo & Voga 2013 for discussion about the place of prefixes inside mental lexicon).

To sum up, we are going back to our very first question. How far to zoom in to observe semantic synchrony in multimodal expression of motion? First of all, we confirm that simultaneously produced gesture and speech are co-expressive, i.e. both of them refer to a given motion. Semantic redundancy between verbal and co-verbal modality—in terms of whether they depict the same conceptual elements of motion (path and/or manner)—however varies with lexico-syntactical specificities of a given language as well as with level of observation/analysis. In French—and probably in other verb-framed languages—redundancy is obvious when we compare a given gesture to the proposition that envelops it. To observe redundancy in English—and maybe in Germanic subcategory of satellite-framed languages—it is necessary to compare a given gesture to the word(s) that is/are produced simultaneously with it. Finally, in Czech—and maybe in Slavic subcategory of satellite-framed languages—even zooming into relation between a given gesture and the semantically meaningful part(s) of words that is/are exactly aligned with it, does not lead to observe any more important semantic redundancy. Gesture and speech produced by Czech speakers are predominantly non-redundant, which makes them different not only from the situation in verb-framed languages in general but also from some other satellite-framed languages.

References

- Allen, S., Özyürek, A., Kita, S., Brown, A., Furman, R., Ishizuka, T., & Fujii, M. (2007). Language-specific and universal influences in children's packaging of manner and path: A comparison of English, Japanese, and Turkish. *Cognition*, 102, 16–48.
- Aurnague, M. (2011). How motion verbs are spatial: The spatial foundations of intransitive motion verbs in French. *Linguisticae Investigationes*, 34(1), 1–34.
- Dewell, R. (2011). *The meaning of particle/prefix constructions in German*. Amsterdam & Philadelphia: John Benjamins.
- Fibigerova, K. (2012). The impact of language on development of verbal and gestural expression of motion: Comparison between different-aged Czech and French. Unpublished doctoral dissertation, Toulouse: Université Toulouse 2.
- Fibigerova, K., Guidetti, M., & Šulová, L. (2012). Verbal and gestural expression of motion in French and Czech. In L. Filipović & K. M. Jaszczolt (Eds.), *Space and time across languages and cultures II: Language, culture and cognition* (pp. 251–268). Amsterdam & Philadelphia: John Benjamins.
- Fibigerova, K., & Guidetti, M. (2018). The impact of language on gesture in descriptions of voluntary motion in French and Czech adults and children. *Language, Interaction and Acquisition*, 9(1), 99–134.
- Giraudo, H., & Voga, M. (2013). Prefix units within the mental lexicon. In N. Hathout, F. Montermini, & J. Tseng (Eds.), *Morphology in Toulouse: Selected proceedings of Décembrettes 8* (pp. 61–78), Munchen: Lincom Europa.
- Gullberg, M., Hendrix, H., & Hickmann, M. (2008). Learning to talk and gesture about motion in French. *First Language*, 28(2), 200–236.
- Hickmann, M. (2006). The relativity of motion in first language acquisition. In M. Hickmann & S. Robert (Eds.), *Space in languages: Linguistic systems and cognitive categories* (pp. 281–308). Amsterdam & Philadelphia: John Benjamins.
- Hickmann, M., Hendriks, H., & Gullberg, M. (2011). Developmental perspectives on the expression of motion in speech and gesture: A comparison of French and English. *Language, Interaction and Acquisition*, 2(1), 129–156.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and Gesture* (pp. 162–185). Cambridge: Cambridge University Press.
- Levinson, S. C., & Wilkins, D. (Eds.) (2006). *Grammars of space: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
- McNeill, D. (2005). *Gesture and Thought*. Chicago & London: University of Chicago 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), 1040–1054.
- Slobin, D. I. (2000). Verbalized events: A dynamic approach to linguistic relativity and determinism. In S. Niemeier, & R. Dirven (Eds.), *Evidence for linguistic relativity* (pp. 107–138). Amsterdam & Philadelphia: John Benjamins.
- Slobin, D. I. (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. In S. Stromqvist & L. Verhoeven (Eds.), *Relating events in narrative: Typological and contextual perspectives* (pp. 219–257). New York: Lawrence Erlbaum Associates.
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical form. In T. Shopen (Ed.), *Language typology and syntactic description* (vol. 3): *Grammatical categories and the lexicon* (pp. 57–149). Cambridge: Cambridge University Press.
- Talmy, L. (2000). *Towards a cognitive semantics*. Cambridge, MA: MIT Press.

The timing of pointing-speech combinations in typically developing and language-delayed toddlers

Angela Grimminger

Paderborn University, Germany
angela.grimminger@uni-paderborn.de

Abstract

Research on the development of the gesture–speech integrated system suggests that the temporal alignment becomes closer with progression in linguistic skills. In this study, the multimodal communicative combinations of pointing gestures with speech (vocalizations and first words) in two groups of 18-month-old children with different developmental trajectories in their linguistic development were analyzed: a group of typically developed children and a group of children delayed in language acquisition—as attested retrospectively by a standardized test. Using the reliable paradigm of the decorated room to elicit pointing behavior in children, the analyses focussed on the timing between the two modalities and the temporal distances between gesture and speech onsets. Similar patterns of gesture–speech integration were found for both groups.

1. Introduction

For adult speakers, there is wide consensus that gesture and speech form an integrated communicative system (Kendon, 1980; Kita & Özyürek, 2003; McNeill, 1992). This view is based on the observation that gesture and speech are temporally and semantically synchronized (McNeill, 1992). Synchronization on the semantic level means that both modalities refer to the same idea, either by expressing similar information in gesture and speech (i.e., with one modality being redundant with or complementing aspects to the content of the other, e.g., Iverson & Goldin-Meadow, 2005), or by gesture and speech expressing information that supplement one another. Semantic synchronization has been used to show how children first express themselves multimodally and are increasingly able to combine information within one modality (Iverson & Goldin-Meadow, 2005).

Beside the semantic level, synchronization on the temporal level means that the most prominent part of the gesture, i.e. the stroke, co-occurs with the most prominent part of a speech unit. Following Kendon (1980), a gesture’s execution can be described in different phases: the “preparation phase”, in which the hand leaves its rest position, the “stroke”, and the “retraction”, in which the hand or arm return to the rest position. For adult speakers, it has been reported that the onset of a gesture precedes the onset of speech (Bergmann, Aksu, & Kopp, 2011), while the gestural stroke is temporally closely aligned with the onset of the spoken part of an utterance (with a mean temporal distance of about 128 ms between the stroke onset and the onset of speech; Bergmann et al., 2011).

For children’s developing communication system, it is of question how and when speech and gesture become integrated. This question is addressed by analyzing the age at which gesture and speech are synchronized on the semantic and temporal level respectively (e.g., Butcher & Goldin-Meadow, 2000; Esteve-Gibert & Prieto, 2014; Murillo, Ortega, Otones, Rujas, & Casla, 2018). In this paper, I will focus on the temporal level.

Studies on precursors for the integrated gesture–speech system investigated the co-development of hand movements and movements of the mouth or vocalizations, and found both modalities to be temporally integrated from very early on (e.g., Ejiri & Masataka, 2001; Iverson & Thelen, 1999; Masataka, 2003). However, with increasing linguistic capabilities this temporal relation becomes closer, as Iverson and Thelen (1999) report for children between 16 and 18 months of age compared to younger children. Similar results were obtained for communicative gestures, of which deictic gestures are the most frequently used type observed in young children (Bates, 1976; Capone & McGregor, 2004). Butcher and Goldin-Meadow (2000) reported that children temporally synchronize verbal utterances and their communicative gestures, mainly deictic gestures at this age,

not until they started to produce their gestures together with meaningful words, as opposed to gestures produced with speech sounds. Note that in this study, temporal integration was operationalized as a complete overlap of the verbal part with the gestural stroke. Esteve-Gibert and Prieto (2014) also included partial overlaps in their analyses of a longitudinal study and were able to shed light onto developmental changes in the temporal coordination of the earliest gesture–speech combinations in a fine-grained way using different measures. This way, the authors revealed support for a closer temporal relation between communicative gestures and speech with increasing linguistic abilities: At 11 months of age, when infants are at the babbling stage, they already combine about 40% of their (mainly deictic) gestures with verbal utterances; critically, once infants produce their first words, the majority of gestures are produced together with speech. Further, and similar to the temporal coordination reported for adult speakers, in this study, the infants’ gesture onset preceded the onset of the verbal utterance. Interestingly, the temporal distance between those two measures appeared to be more adult-like in the single-word period compared to the babbling period. Analyses of the temporal distance between the gestural stroke and the onset of the spoken part of the utterances revealed very small differences between the two measures, thus showing that gestures and words were almost simultaneously produced. In sum, the two studies mentioned above show that children start to temporally integrate speech and gestures at an early age, and this temporal relation becomes closer aligned with progress in linguistic skills (see also Murillo et al., 2018).

Many studies have shown a strong and positive relation of infants’ use of deictic gestures and their subsequent lexical and syntactic development (e.g., Beuker, Rommelse, Donders, & Buitelaar, 2013; Colonnese, Stams, Koster, & Nool, 2010; Rowe & Goldin-Meadow, 2009; Rowe, Özçalışkan, & Goldin-Meadow, 2008), not only in children whose language develops typically but also in clinical populations (Brady, Marquis, Fleming, & McLean, 2004; Lüke, Grimminger, Rohlfsing, Liszkowski, & Ritterfeld, 2017; Özçalışkan, Adamson, & Dimitrova, 2016; Stolt et al., 2014). Beyond these results, recently the focus seems to have moved from considering the frequency of pointing alone to considering pointing–speech combinations, which might be an even better predictor of lexical skills (Igalada, Bosch, & Prieto, 2015; Murillo & Belinchón, 2012; Wu & Gros-Louis, 2014) and advances in syntactic development (e.g., Fasolo & D’Odorico, 2012; Rowe & Goldin-Meadow, 2009).

Up to date, however, studies addressing the use of pointing–speech combinations in children who show different rates in their early language development are missing. Given the results that the frequency of pointing–speech combinations predicts later linguistic skills, and that the temporal integration becomes closer with increasing linguistic skills, for this study, I hypothesize that the timing of pointing–speech combinations of LD children will differ from that of children with whose linguistic skills are lower. More specifically, I assume that the temporal distance between the onset of the gesture stroke and the onset of the spoken part of the utterance is greater in children with LD. A second hypothesis is that the combinations of pointing gestures with words are temporally closer aligned than pointing gestures with vocalizations. These two hypotheses are examined in children at the age of 18 months, at which period typically developing children start to produce two-word-utterances in speech (cf. Klann-Delius, 2016).

2. Methods

2.1. Participants

14 German-learning children were drawn from a larger sample of 34 families participating in a longitudinal study between 12 and 30 months of age (Grimminger, 2017). Within the whole sample, eight children were identified as being language delayed (LD) at 24 months of age (see below). One of the children with LD was excluded from the analyses here because the parents did not give consent for further analyses. The final sample thus consisted of seven children with LD (2 girls, 5 boys). Seven children with typical language development (TD) were matched for gender.

2.2. Setting and procedure

To elicit spontaneous pointing and verbal utterances, the infants and one of their caregivers (85 % mothers) were observed in a semi-naturalistic setting within a laboratory room that was selectively decorated with 16 interesting objects, pictures (Liszkowski & Tomasello, 2011), and events (e.g., sudden onset of a water fountain). Because we started the longitudinal study when the children were

12 months old, at 18 months of age, they had been in this decorated room with a caregiver several times before. At every session, caregivers were instructed to engage with their children while carrying them for 6 minutes and looking at the objects presented in the room without touching any of them. The data were videorecorded using four cameras from different angles of the room.

2.3. Assessment of language development

To assess the children's language development at 24 months of age, a German standardized language test was administered (*Sprachentwicklungstest für zweijährige Kinder – SETK-2*) [test of language acquisition for two-year-old children] (Grimm, 2000). In accordance with other authors (Heilmann, Ellis Weismer, Evans, & Hollar, 2005; Sachse & von Suchodoletz, 2008), a two-year-old child was defined as being language delayed if she or he scored 1½ standard deviations below the mean (i.e., T-score of ≤ 35) in at least one of the four subtests of the SETK-2 and one standard deviation below the mean in at least one additional subtest (i.e., T-score of < 40).

2.4. Coding

All verbal utterances and gestures of the children were coded using ELAN (Sloetjes & Wittenburg, 2008). Pointing gestures, defined as the extension of the index finger or the whole hand towards an object or location, were the majority of gestures used, and thus, other gesture types will be omitted for the analyses here. For each pointing gesture that was accompanied by a verbal utterance the gesture onset, i.e. the beginning of the preparation phase, and the stroke was coded in order to analyze the timing between the onset of a verbal utterance and the onset of a gesture and its stroke. The gestural stroke was coded as the interval in which the arm and/or index finger were maximally extended (see Esteve-Gibert & Prieto, 2014). The pointing–speech combinations were assigned to one of the following categories: (a) pointing+vocalization; (b) pointing+protoword (German “da”): these combinations were coded as an extra category (see also Liszkowski & Tomasello, 2011), because they mark a transition phase to semantically more complex forms of pointing–speech combinations that contain a word (Clark, 1978); (c) pointing+word. Pointing+two-word utterances were observed only on few occasions, and were therefore excluded from the statistical analyses of temporal distance.

Pointing gestures and verbal utterances that did not at least partially overlap were not considered. If a pointing gesture was accompanied by more than one vocalization, the vocalization closest to the gestural stroke was considered. As an additional measure of temporal coordination, it was coded if the gestural stroke overlapped with the verbal part, either fully, partially or not.

3. Results

First, in most cases of pointing–speech combinations (371 cases in total), the gesture onset preceded the onset of the verbal part (91.4%), and children in both language development groups, TD versus LD, were as likely to do so: TD children, $M=92.5\%$ ($SD=5.6$), children with LD, $M=84.8\%$ ($SD=16.5$), $p > .05$ (Mann-Whitney). Second, in the majority of cases of pointing–speech combinations, the onset of the pointing gesture's stroke followed the onset of speech (73.8%). Again, no differences between both language development groups were found, TD children, $M=70.1\%$ ($SD=2.3$), children with LD, $M=77.4\%$ ($SD=14.7$), $p > .05$ (Mann-Whitney). These results are consistent with previous research and show that at 18 months of age, TD children and children with LD are comparable in their overall pattern of how pointing gestures and speech are aligned. However, some of the children with LD mainly used either vocalizations together with their pointing gestures or protowords, whereas TD children used all forms of combinations. This observation is reflected in the results showing significant group differences in the number of pointing+1-word utterance, $Z= -3.27$, $p= .001$, and pointing+two-word utterances, $Z= -2.25$, $p= .05$, but not in the number of pointing+vocalizations and pointing+protowords (both $p > .05$, Fig. 1).

While above, the onsets of the two modalities were compared to each other, in the following analyses their temporal distance is considered. Following the finding that temporal integration becomes closer with increasing linguistic skills, it was hypothesized that group differences can be found in the temporal distance between gesture and speech. Thus, the mean temporal distances in

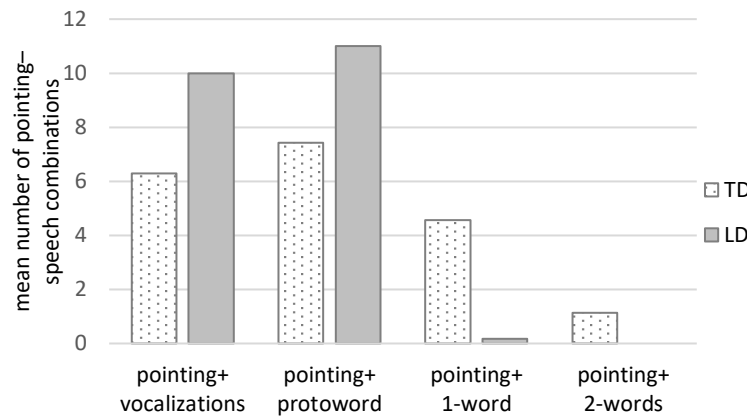


Figure 1. Comparison of the two groups in their use of different types of pointing–speech combinations.

milliseconds between (1) gesture onset and speech onset, and (2) between speech onset and gestural stroke onset were calculated (see Esteve-Gibert & Prieto, 2014). For each participant, the mean temporal distances in milliseconds were separately calculated for each category of pointing–speech combinations, and these variables were compared across the two language development groups. Only those cases were included in which the gesture onset preceded the onset of the verbal part. Contrary to the assumption, the groups did not differ significantly, neither in the temporal distance between pointing onset and speech onset, separately for combinations with vocalizations, proto-words, or one-word utterances (each $p > .05$), nor in the temporal distance between speech onset and stroke onset, again separately for each type of combination (each $p > .05$, Table 1 for the descriptive statistics).

Table 1
Descriptive statistics of temporal distances in milliseconds

	TD		LD	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
gesture onset – speech onset				
vocalization	673,68	220,77	948,62	647,03
protowords	568,68	203,89	487,31	188,10
1-word	455,90	290,72	530,00	.
speech onset – stroke onset				
vocalization	225,51	134,67	253,45	94,92
protowords	266,58	107,13	260,95	92,42
1-word	291,10	163,83	434,00	---

A further hypothesis was that the combinations of pointing gestures with words are temporally closer aligned than pointing gestures with vocalizations. Because no group differences were found, the sample was analyzed as a whole to address this hypothesis. As shown above, the children with LD barely used pointing+1-word combinations, and a Friedman’s test was conducted to compare the temporal distances of pointing onset with vocalizations, protowords, and one-word utterances, respectively. This analyses did not yield significant differences, $\chi^2(2) = .67, p > .05$. Because four of the children with LD did not produce any pointing gestures with words, this analysis was performed with $N = 8$. To include more children, we additionally compared the temporal distances of pointing onsets with vocalizations and pointing onsets with protowords using a Wilcoxon test. No differences were found, $Z = -1.65, p > .05, N = 11$. To compare the distance between the onset of vocalizations and protowords, respectively, and the onset of the gesture stroke, we applied a Wilcoxon test. Again, no differences were found, $Z = -1.48, p > .05, N = 12$. We also did not find significant differences between the language development groups when comparing the percentage of pointing–speech combinations in which the gesture stroke did not overlap with the spoken part.

4. Discussion

For adult speakers, gesture and speech form an integrated system. Research on gesture–speech integration from a developmental perspective focusses on how and when this integration is achieved, and has impressively shown that infants quite early start to temporally align their manual and vocal activities (e.g., Ejiri & Masataka, 2001; Iverson & Thelen, 1999; Masataka, 2003). However, with increasing linguistic skills, this integrated system becomes more adult-like (Esteve-Gibert & Prieto, 2014; Murillo et al., 2018). Therefore, in this study, I investigated the temporal integration of children’s early vocalizations, protowords and words with their gestures in two groups of children that retrospectively differed in their language developmental paths to receive insights into the development of the system. I hypothesized that the timing of pointing–speech combinations of children with LD will differ from that of TD children who I assumed to show a closer temporal alignment of the two modalities. In addition, it was hypothesized that the combinations of pointing gestures with words are temporally closer aligned than pointing gestures with vocalizations. No group differences were found in how children synchronize their pointing gestures with vocal behaviors. Instead, the results for both groups of children are consistent with previous research in adults and infants: The gesture onset preceded the onset of the verbal part, and the onset of the pointing gesture’s stroke followed the onset of speech (Bergmann et al., 2011; Esteve-Gibert & Prieto, 2014). This result is intriguing as it suggests a similar processing of early communicative behavior for both, the TD children and children with LD. The only difference that was found pertains to the number of pointing gestures with words being higher in TD children than in children with LD (see Figure 1). Thus, while the TD children at 18 months of age might have a more diverse repertoire of their multimodal communicative means, by using pointing gestures together with different kinds of verbal utterances, the children with LD make greater use of pointing gestures with vocalizations or protowords still. This finding accords with previous findings suggesting that while toward the end of the second year TD children increasingly used words, the pointing gestures of children with LD are accompanied by reduced expressive language (Lüke et al., 2017).

However, some methodological limitations of this study should be stressed. First, the sample size in each group is very small which is due to fact that all children were recruited at 12 months of age for this longitudinal study, and language delay could only be assessed as early as 24 months of age. The results here therefore need further verification. Second, by averaging the temporal distances of gesture onset or stroke onset and speech onset across each subject, the data of the stability in individual patterns were likely stripped: It is possible that the multimodal utterances of children with LD are less consistently synchronized when considering all pointing attempts; vice versa, TD children appear to be rather stable in the way they integrate their vocal behaviors with pointing gestures. To confirm such observations, more fine-grained analyses and methods that are sensitive to patterns on the individual level are necessary. Secondly, even though Lüke et al. (2017) report about early communicative attempts consisting of two types of pointing gestures, namely the index-finger pointing and hand pointing, we excluded pointing gestures performed with the whole-hand from the analyses here, because pointing with the whole hand was shown to be rather negatively related with later language skills. Further investigations taking any gestural form into account might thus look at the way integration with verbal behavior is achieved in whole-hand pointing.

Concerning the debate whether children with LD use their gestures to compensate for their language deficits, our results can be interpreted against the compensation effect, because we found similar integration of pointing with verbal behaviors in both group of children. In addition, as can be viewed from Figure 1, it is rather the group of TD children who is using more pointing–word combinations. With respect to the other forms of speech, no group differences were found. However, it is possible that focusing our analyses on the data point at which children were 18 months old is problematic, since the compensation effect might come into play when children use words rather than protowords and vocalizations. Thus, it can rather be that while the gesture–speech system is developing a similar way in both groups, with increasing linguistic skills, it might start to serve different functions resulting in increasing compensation. This possibility links to a discussion about early communicative attempts being of similar or different nature than conventional use of language (Dore, 1975).

Acknowledgments

This work was supported by the German Research Foundation (DFG; RO 2443/3-1). I thank Juliane Rode for help with annotating the data.

References

- Bates, E. (1976). *Language and context: The acquisition of pragmatics* (Vol. 13). New York: Academic Press.
- Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*.
- Beuker, K. T., Rommelse, N. N. J., Donders, R., & Buitelaar, Jan. K. (2013). Development of early communication skills in the first two years of life. *Infant Behavior and Development, 36*, 71–83.
- Brady, N. C., Marquis, J., Fleming, K., & McLean, L. (2004). Prelinguistic predictors of language growth in children with developmental disabilities. *Journal of Speech, Language, and Hearing Research, 47*, 663.
- Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one-to two-word speech: When hand and mouth come together. *Language and Gesture, 28*, 235–257.
- Capone, N. C., & McGregor, K. K. (2004). Gesture development: A review for clinical and research practices. *Journal of Speech, Language, and Hearing Research, 47*, 173–186.
- Clark, E. V. (1978). From gesture to word: On the natural history of deixis in language acquisition. In J. S. Bruner (Ed.), *Human growth and development* (pp. 85–120). Oxford: Clarendon Press/Oxford University Press.
- Colonnese, C., Stams, G. J. J. M., Koster, I., & Nool, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review, 30*, 352–366.
- Ejiri, K., & Masataka, N. (2001). Co-occurrences of preverbal vocal behavior and motor action in early infancy. *Developmental Science, 4*, 40–48.
- Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication, 57*, 301–316.
- Fasolo, M., & D’Oro, L. (2012). Gesture-plus-word combinations, transitional forms, and language development. *Gesture, 12*, 1–15.
- Grimm, H. (2000). *Sprachentwicklungstest für zweijährige Kinder (SETK-2)*. Göttingen: Hogrefe.
- Grimminger, A. (2017). *Gestische und sprachliche Kommunikation von 12 – 16 Monate alten Kindern und ihren Bezugspersonen in wiederkehrenden, semi-naturalistischen Interaktionen und individuelle Unterschiede in der späteren Sprachentwicklung*. Paderborn: Universitätsbibliothek.
- Heilmann, J., Ellis Weismer, S., Evans, J. L., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology, 14*, 40–51.
- Igualada, A., Bosch, L., & Prieto, P. (2015). Language development at 18 months is related to multimodal communicative strategies at 12 months. *Infant Behavior and Development, 39*, 42–52.
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychol. Sci., 16*, 367–371.
- Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies, 6*, 19–40.
- Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207–227). Mouton: The Hague.
- 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*, 16–32.
- Klann-Delius, G. (2016). *Spracherwerb* (3rd ed.). Stuttgart: J.B. Metzler.
- Lüke, C., Grimminger, A., Rohlfing, K. J., Liszkowski, U., & Ritterfeld, U. (2017). In infants’ hands: Identification of preverbal infants at risk for primary language delay. *Child Development, 88*, 484–492.
- Masataka, N. (2003). From index-finger extension to index-finger pointing: Ontogenesis of pointing in preverbal infants. In S. Kita (Ed.), *Pointing: Where Language, Culture, and Cognition Meet* (pp. 69–84). Mahwah: Lawrence Erlbaum.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago, USA: Chicago University Press.
- Murillo, E., & Belinchón, M. (2012). Gestural-vocal coordination: Longitudinal changes and predictive value on early lexical development. *Gesture, 12*, 16–39.
- Murillo, E., Ortega, C., Otones, A., Rujas, I., & Casla, M. (2018). Changes in the Synchrony of Multimodal Communication in Early Language Development. *Journal of Speech, Language, and Hearing Research, 1–11*.
- Özçalışkan, Ş., Adamson, L. B., & Dimitrova, N. (2016). Early deictic but not other gestures predict later vocabulary in both typical development and autism. *Autism, 20*, 754–763.
- Rowe, M. L., & Goldin-Meadow, S. (2009). Early gesture selectively predicts later language learning. *Developmental Science, 12*, 182–187.
- Rowe, M. L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture’s role in predicting vocabulary development. *First Language, 28*, 182–199.
- Sachse, S., & von Suchodoletz, W. (2008). Early Identification of Language Delay by Direct Language Assessment or Parent Report? *Journal of Developmental & Behavioral Pediatrics, 29*, 34–41.
- Sloetjes, H., & Wittenburg, P. (2008). Annotation by Category: ELAN and ISO DCR. *LREC Proceedings, 816–820*. Marrakesch, Marokko.
- Stolt, S., Mäkilä, A.-M., Matomäki, J., Lehtonen, L., Lapinleimu, H., & Haataja, L. (2014). The development and predictive value of gestures in very-low-birth-weight children: A longitudinal study. *International Journal of Speech-Language Pathology, 16*, 121–131.
- Wu, Z., & Gros-Louis, J. (2014). Infants’ prelinguistic communicative acts and maternal responses: Relations to linguistic development. *First Language, 34*, 72–90.

Gestural training benefits L2 phoneme acquisition: Findings from a production and perception perspective

Marieke Hoetjes, Lieke van Maastricht, and Lisette van der Heijden

Centre for Language Studies, Radboud University, The Netherlands
m.hoetjes@let.ru.nl, l.vanmaastricht@let.ru.nl, ll.vanderheijden@student.ru.nl

Abstract

This paper aims to study whether training with gestures benefits L2 phoneme acquisition from both a production and perception perspective. In the production study, Dutch learners of Spanish received pronunciation training for the phonemes /u/ and /θ/ in one of four conditions: audio-only, audio-visual, audio-visual with pointing gestures, or audio-visual with iconic gestures. Results show that in general, gestural training benefits L2 phoneme acquisition, but different gestures benefit the acquisition of different phonemes, possibly depending on their complexity. The perception study, in which L1 speakers of Spanish judged the L2 Spanish material on accentedness and comprehensibility, corroborate the findings from the production study: Including visual information in training generally lowered the perceived accentedness and increased the perceived comprehensibility of speech, but the type of phoneme matters. Together, these studies suggest that gestural training can benefit L2 phoneme acquisition, yet certain gestures work better for certain phonemes than others.

1. Introduction

It has been established that speech and gesture are closely related (Kendon, 2004; McNeill, 1992), evidenced, for example, by the semantic and temporal coordination of speech and gesture (see e.g. Gullberg, 2006, for an overview). Moreover, previous studies have shown that gesture is relevant in language development, as children produce pointing gestures for objects they do not yet have labels for (e.g., Iverson & Goldin-Meadow, 2005), and these gestures predict the words that are to appear next in their vocabulary. Gestures have also been shown to facilitate learning, both in non-linguistic (Yeo, Wagner Cook, Nathan, Popescu, & Alibali, 2018), and linguistic contexts. For example, gestures have been shown to benefit the acquisition of novel L2 words (Kelly, McDevitt, & Esch, 2009; Tellier, 2008). Regarding L2 phonemic contrasts specifically, prior studies have demonstrated that seeing the speaker benefits the production of phonemes by L2 learners (Hardison, 2003; Hazan, Sennema, Iba, & Faulkner, 2005), yet studies on the role of gestures in the perception of L2 tonal and phonemic contrasts report contrasting findings: Hannah, Wang, Jongman, and Sereno (2017) and Kelly, Bailey, and Hirata (2017) revealed that training with gestures significantly improves the perception of non-native phonemic tones and intonation contours, while work by Hirata, Kelly, and colleagues (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly et al., 2017; Kelly, Hirata, Manansala, & Huang, 2014) revealed no significant improvement in the perception of non-native phonemic vowel length distinctions after gestural training. Kelly et al. (2017) concluded that “gestures help with some – but not all – novel speech sounds in a foreign language” (p. 1).

Thus, while gestures are known to be a common and effective resource in L1 communication, less is known about their possibly beneficial effects in the context of L2 acquisition, especially concerning the educational value of different types of gestures. Previous studies on gestures in L2 pronunciation training have used varying gesture types and hand movements, e.g., beats (e.g. Gluhareva & Prieto, 2017), metaphoric gestures (e.g. Kelly et al., 2014), clapping (e.g. Zhang, Baills, & Prieto, 2018), with varying methods and results, which complicates the process of determining which kinds of gestures facilitate L2 phoneme acquisition and under which conditions. In addition, most studies on the effects of gesture on L2 phoneme acquisition rely only on perception measures, either by analysing on-target phoneme production through L1 perception measures (e.g. Gluhareva & Prieto, 2017), or by testing whether L2 learners can discriminate between different L2 phonemes in a perception task (e.g. Kelly et al., 2017). Therefore, we aim to

determine whether instruction modality affects L2 phoneme acquisition, distinguishing between four training conditions: 1) training in which examples are presented as audio fragments only; 2) training in which examples are presented as video fragments but the trainer does not gesture; 3) training in which examples are presented as video fragments and the trainer produces a pointing gesture towards her mouth when producing the target phoneme; and 4) training in which examples are presented as video fragments and the trainer produces an iconic gesture visualizing the position and/or form of the relevant articulators near her mouth when producing the target phoneme. As measures of successful L2 phoneme acquisition, we analyse both the phonetic characteristics of L2 speech (Study 1) and L1 listeners' ratings of foreign accentedness and comprehensibility (Study 2).

Based on prior research, we hypothesize that adding audio-visual information to language training will be beneficial for phoneme acquisition compared to providing only audio information (Hardison, 2003; Hirata & Kelly, 2010; Wang, Behne, & Jiang, 2008). Given that the use of gestures is helpful in the acquisition of certain segments (Hannah et al., 2017; Kelly et al., 2017), using gestures in the audio-visual training will be more beneficial than not including them. As prior work has not yet compared the effect of different types of gestures, no predictions can be made regarding comparisons between iconic and pointing gestures. Based on Zhang, Baills and Prieto (2018), we predict that the findings for our production and perception measures will be congruent, with possibly a stronger effect of gestural training on perception than on production.

2. Study 1: Production

This study was set up using a pre-test (T1) – training – post-test (T2) design. Fifty-one L1 speakers of Dutch (30 female, mean age 25 years old, range 18-61 years old), who did not speak Spanish, took part in one of four training conditions: audio-only (AO), audio-visual (AV), audio-visual with pointing gestures (AV-P), or audio-visual with iconic gestures (AV-I).

2.1. Materials

We focused on the acquisition of the Spanish phonemes /θ/ and /u/, since their nativelike production by L2 learners is often complicated by two factors: 1) The difference in grapheme-to-phoneme conversion between Dutch and Spanish. The grapheme 'u' should be pronounced as /u/ in Spanish, whereas in Dutch it is generally pronounced as /y/, /ə/, or /ʏ/. Likewise, the grapheme 'z' is pronounced as /θ/ in Spanish, yet as /z/ or /s/ in Dutch. 2) The possible absence of L2 segments in the L1 inventory. While the /u/ exists in the Dutch phoneme inventory, /θ/ does not.

The phonemes were embedded in 16 four-word sentences, which were read aloud by participants at T1 and T2 in one of two randomised orders. Each sentence was presented on a separate PowerPoint slide. Above each sentence, a picture illustrated its meaning. Half of the sentences were experimental items, which had a word containing the target phoneme as the second word of the sentence. The target phoneme always occurred in the first syllable of this two-syllable word (e.g., *La nube es blanca*, *La zeta es verde*). Each of the two target phonemes occurred in four target words. The remaining eight sentences were fillers, which were not currently analysed.

After T1 and before T2, the participants received a short training focusing on the Spanish pronunciation of /θ/ and /u/ (in counterbalanced order). Training consisted of a set of PowerPoint slides on which information was given about how each target phoneme is pronounced in Spanish. Specifically, participants were told that the Spanish pronunciation of both graphemes is different from the Dutch pronunciation of these graphemes, and it was explained which articulatory gestures are necessary for nativelike pronunciation (e.g., “when pronouncing the letter ‘u’ in Spanish, you round your lips”). The training included various examples, produced by an L1 speaker of Spanish; one example was given on the same slide as the written information about the respective target phoneme, and two examples were given on successive slides.

To manipulate the training modality, the visual information that was presented to participants in each condition was varied while the audio (from the L1 speaker seen in the video) was dubbed over all conditions: In the AO condition, participants heard the audio example but did not see the video. In the AV condition, a video of the speaker was shown, but the speaker did not gesture. In the AV-P condition, the speaker pointed towards her mouth while producing the target phoneme. In the AV-I condition, the speaker made an iconic gesture representing the articulatory gesture needed for on-target segment production as she produced the target phoneme. For the /u/, this was

a one-handed gesture indicating the rounding of the lips, and for the /θ/, this was a one-handed gesture indicating that the speaker should push their tongue out between their teeth (see Figure 1).

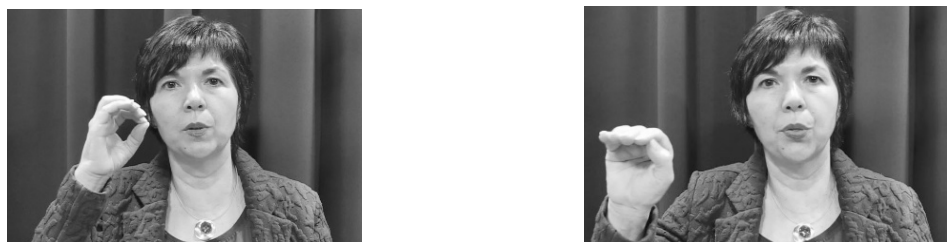


Figure 1. Stills from training video in AV-I condition showing the articulatory gesture needed for /u/ (left) and /θ/ (right).

2.2. Procedure

The experiment took place in Dutch (with the exception of the Spanish sentences) in a soundproof booth. After receiving instructions, participants read out the 16 Spanish sentences into a microphone (T1). After T1, participants completed a language background questionnaire, followed by one of the four types of training. Participants then read a reordered version of the same sentences (T2). The audio produced during T1 and T2 was recorded, and production of the target phonemes was analyzed using Praat (Boersma & Weenink, 2018). The target phonemes were annotated by two phonetically trained coders, distinguishing between nativelike production (i.e., as an L1 speaker of Iberian Spanish would do) and several non-nativelike options (for /θ/ these were /s/, /z/, or ‘other’; for /u/: these were /y/, /ə/, /ʏ/, or ‘other’). For the present analysis, on-target productions were distinguished from non-target productions, collapsing data across the non-target options. There was an overlap in coding of 50%, with a good interrater reliability, $\kappa = .900$, $p < .001$. Annotations for the same sentences were compared between T1 and T2, resulting in 4 options: 1) the participant was able to produce the target phoneme at T1, but not anymore at T2; 2) the participant was not able to produce the target phoneme at either T1 or T2; 3) the participant was able to pronounce the target phoneme at both T1 and T2; or 4) the participant was unable to produce the target phoneme at T1, but able to do so at T2. In the current analysis, we distinguish between progress (i.e., option 4), and no progress (i.e., options 1-3) and conduct chi-square analyses to analyse whether training modality affected target phoneme production.

2.3. Results

The analysis of the results for on-target /u/, i.e., using only those productions coded as option 4, did not reveal a significant association between training condition and progress, $\chi^2(3) = 6.679$, $p = .083$. However, the highest proportion of learning was obtained after the AV-I training, which is substantially higher than the proportion of learning after AO training (see Figure 2). The frequencies of the results coded as options 1-3 showed that in 64.6% of all cases participants already produced the /u/ correctly at T1, and continued to do so at T2 (vs. the 31.3% of all cases coded as option 4).

The analysis for target production of /θ/ showed a significant association between training condition and progress, $\chi^2(3) = 9.155$, $p = .027$. The progress in the AV-P and AV-I conditions differed significantly from the expected values. The analysis revealed that the proportion of cases with progress in the AV-P condition (37%) was significantly higher, and the proportion of cases with progress in the AV-I condition (15%) was significantly lower than expected. In other words, for the acquisition of /θ/, the AV-P condition is particularly helpful but the AV-I condition is particularly harmful (see Figure 2). Interestingly, inspection of the frequencies of the results for /θ/ show that in the majority of all cases (64.5%) participants never learned to produce the /θ/ correctly. This suggests that /θ/ is particularly challenging for L2 learners, in contrast to /u/, which appears to be substantially less challenging.

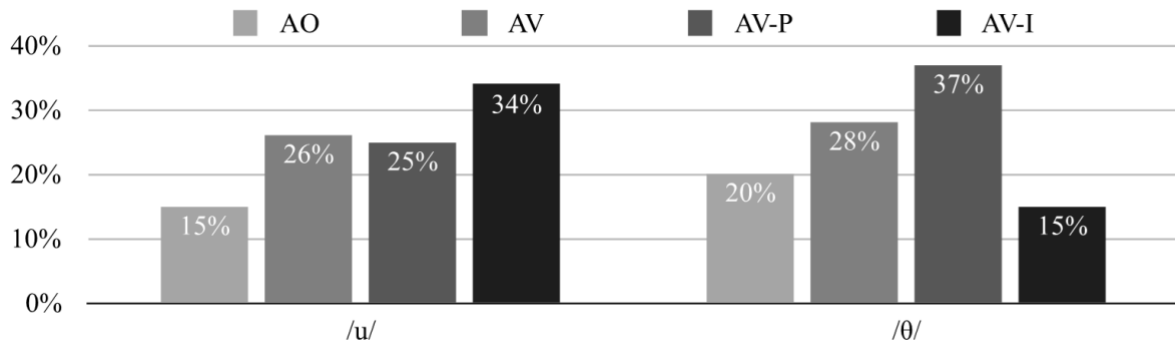


Figure 2. Percentages of /u/ (left) and /θ/ (right) acquisition, separated by training condition.

3. Study 2: Perception

In this within-subjects design, 46 L1 Spanish speakers (19 females, mean age 31 years old, range 19-70 years old) listened to a selection of target words, produced at T1 and at T2 after AV, AV-P, or AV-I training, and rated them on accentedness (21 subjects) or comprehensibility (25 subjects).

3.1. Materials and Procedure

In order to keep the length of the experiment acceptable, we used 8 items (2 with /θ/ and 2 with /u/ from T1 and T2) from 21 speakers (7 randomly chosen speakers from 3 types of training condition: AV, AV-P, and AV-I) from the production study, resulting in 168 items. The AO condition was left out as it represents the least realistic learning context. Accentedness was measured with the statement “This speaker speaks ...”, followed by a 7-point semantic differential anchored by “without a foreign accent - with a strong foreign accent” (based on Jesney, 2004). Comprehensibility was measured with the statement “I find this speaker easy to understand” followed by a 7-point Likert scale anchored by “totally disagree - totally agree” (based on Derwing & Munro, 1997). Before participants rated the items, a brief explanation of either accentedness or comprehensibility was given. The entire experiment took place in Spanish. Subjects were requested to wear headphones.

3.2. Results

Accentedness ratings were transformed to reflect the same direction of effect as comprehensibility ratings, i.e., a higher rating always reflects more nativelike speech. A repeated measures analysis for accentedness with Type of training (4 levels: T1, AV, AV-P, AV-I) and Target sound (2 levels: /u/ or /θ/) as within-subject factors showed a significant main effect of Type of training ($F(3, 72) = 16.17, p < .001, \eta^2 = .40$), no main effect of Target sound ($F(1, 24) < 1, p = .606$), and a significant interaction between Type of training and Target sound ($F(3, 72) = 12.94, p < .001, \eta^2 = .35$). Pairwise comparisons between training conditions within individual phonemes revealed that for words containing /u/, scores for identical items increased significantly between T1 and T2 after AV-P ($p < .001$) and AV-I training ($p < .001$), but not after AV training. For words containing /θ/, scores for identical items increased significantly between T1 and T2 after AV-P training ($p < .001$) but not after AV or AV-I training, implying that L2 speakers only benefitted from AV-P training.

The analysis for comprehensibility revealed a significant main effect of Type of training ($F(2.04, 40.72) = 10.26, p < .001, \eta^2 = .34$), no significant main effect of Target sound ($F(1, 20) < 1, p = .749$), and a significant interaction between Type of training and Target sound ($F(3, 60) = 10.74, p < .001, \eta^2 = .35$). Pairwise comparisons between training conditions within individual phonemes revealed that for words containing /u/, scores were significantly higher at T2 than at T1 after both AV-P ($p = .003$) and AV-I training ($p = .001$), but not after AV training. Conversely, for words containing /θ/, scores for identical items increased significantly between T1 and T2 after AV-P training ($p = .001$) but decreased significantly after AV-I training ($p = .044$). This suggests that, for comprehensibility, L2 speakers always benefitted from AV-P training, but were actually hindered by the AV-I training for /θ/, see Figure 3, which shows the mean ratings of accentedness and comprehensibility for both /u/ and /θ/.

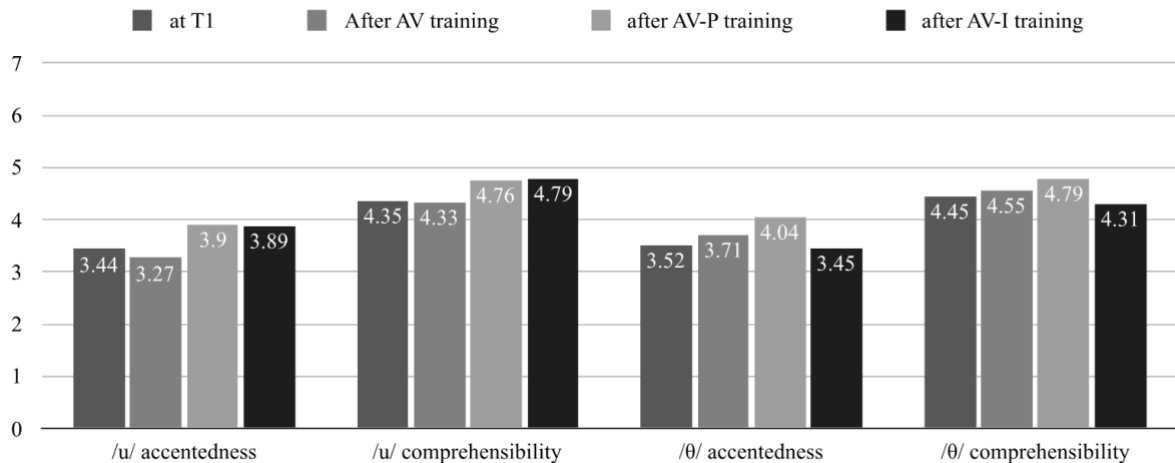


Figure 3. Mean ratings of accentedness (left) and comprehensibility (right) for /u/ and /θ/.

4. General Discussion

The aim of the current studies was to determine whether training with gestures can facilitate L2 phoneme acquisition. Specifically, we aimed to determine whether instruction modality affects L2 phoneme acquisition. We analysed the phonetic characteristics of speech produced by L2 speakers of Spanish (Study 1), and L1 listeners' ratings of accentedness and comprehensibility for a selection of these same items (Study 2). Our first hypothesis was that adding audio-visual information to training would be more beneficial than providing audio only information during training. In addition, we specifically expected that including gestures in the audio-visual training would be more beneficial than not including gestures. Whether there would be differences concerning the facilitatory effects of the different types of gestures was an open question. Moreover, we expected findings from the perception study to confirm those of the production study.

The results from Study 1 showed different effects for the two phonemes under investigation: For /u/, there was no significant association between training condition and progress, even though more cases of native-like phoneme production occurred in all types of audio-visual training in comparison to AO training, and the AV-I training appeared most beneficial. It is important to note though, that /u/ was relatively easy for participants to produce, which was apparent by the fact that in many cases, native-like production already took place at T1. For /θ/, however, results showed that many speakers never acquired a native-like production. For those speakers who did acquire native-like production after training, our results showed that AV-P training was helpful, but AV-I training was harmful. Thus, the findings from Study 1 suggest that in general, including visual information in phoneme training helps, but also that the combination of type of gesture and type of phoneme matters: iconic gestures, which typically provide more semantic information than pointing gestures, seem helpful only when the phoneme is relatively easy to acquire. For a phoneme that learners find difficult to acquire, including an iconic gesture in training is not helpful for acquisition. However, for this difficult phoneme, including a pointing gesture, which mainly served to point the listener's attention to the mouth of the trainer, does facilitate acquisition. This is in line with prior research stating that the use of iconic gestures in training benefits L2 word learning, but only when the cognitive demands of the target words are low (Kelly & Lee, 2012). Similarly, seeing lip movements with speech made it easier for L2 learners to discriminate between phonemes, but adding (here meta-phoric) gestures to audio-visual training actually hindered them (Hirata & Kelly, 2010). The results from Study 2 corroborated the findings from Study 1 in the sense that including gesture in training generally led to speech that was perceived as less accented and more comprehensible. Similarly, in the perception study, this effect differed between the two types of phonemes: Production of /θ/ was perceived as less accented and more comprehensible after AV-P training, but not after AV-I training. For /u/, both types of gesture training led to equal improvement in perception, with lower perceived accentedness and higher perceived comprehensibility.

Taken together, these studies suggest that gestures can benefit L2 phoneme acquisition. Not only can gestures help L2 speakers produce native-like phonemes, but this progress in phoneme production in turn also has a positive effect on speakers' perception with regard to accentedness

and comprehensibility. However, the differing findings for the two phonemes under investigation also indicate that this process cannot easily be generalized and that more research is required comparing the use of different gesture types in more and less challenging acquisition contexts.

Acknowledgments

Many thanks to Núria Domínguez, Judith Peters and Nick Theelen for their help in creating the material, data collection, and data analysis.

References

- Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer (Version 6.0.49) [Computer program]. Retrieved March 2, 2019, from <http://www.praat.org/>
- Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in second language acquisition*, 20, 1-10.
- Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, 21(5), 609-631.
- Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (hommage a Adam Kendon). *International Review of Applied Linguistics*, 44, 103-124.
- Hannah, B., Wang, Y., Jongman, A., & Sereno, J. A. (2017). Cross-modal association between auditory and visual-spatial information in Mandarin tone perception. *The Journal of the Acoustical Society of America*, 140(4), 3225-3225.
- Hardison, D. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, 24(4), 495-522.
- Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English *Speech Communication*, 47, 360-378.
- Hirata, Y., & Kelly, S. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, 53, 298-310.
- Hirata, Y., Kelly, S., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, 57, 2090-2101.
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16(5), 367-371.
- Jesney, K. (2004). The use of global foreign accent rating in studies of L2 acquisition. *Calgary, AB: University of Calgary Language Research Centre Reports*, 1-44.
- Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology*, 3(1), 7. <https://doi.org/10.1525/collabra.76>
- Kelly, S., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, 5(673).
- Kelly, S., & Lee, A. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and cognitive processes*, 27(6), 793-807.
- Kelly, S., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and cognitive processes*, 24, 313-334. doi:10.1080/01690960802365567
- Kendon, A. (2004). *Gesture. Visible action as utterance*. Cambridge: Cambridge University Press.
- McNeill, D. (1992). *Hand and mind. What gestures reveal about thought*. Chicago: University of Chicago Press.
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8(2), 219-235.
- Wang, Y., Behne, D., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *Journal of the Acoustical Society of America*, 124, 1716-1726.
- Yeo, A., Wagner Cook, S., Nathan, M. J., Popescu, V., & Alibali, M. (2018). Instructor gesture improves encoding of mathematical representation. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2723-2728). Austin, TX: Cognitive Science Society.
- Zhang, Y., Baills, F., & Prieto, P. (2018). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from training Chinese adolescents with French words. *Language Teaching Research*.

Synchronization of (dis)fluent speech and gesture: A multimodal approach to (dis)fluency

Loulou Kosmala, Maria Candea, and Aliyah Morgenstern

Sorbonne Nouvelle University, France

{Loulou.kosmala, maria.candea, aliyah.morgenstern}@sorbonne-nouvelle.fr

Abstract

Disfluency is verbally expressed by several markers (filled, unfilled pauses, repetitions, self-repairs, etc). This study is grounded in the functionally ambivalent view of (Dis)fluency following Crible, (2017) and Götz (2013), but with a multimodal and interactional approach. Previous research has shown a coordination between speech and gesture suspension (Gullberg, 2013, 2018; Seyfedinnipur 2006). The aim of our paper is thus to examine how (dis)fluent speech and gestures can be synchronized, and how visual-gestural features can provide a finer understanding of (dis)fluency. Our analyses are conducted on 3 pairs of French and American speakers interacting both in their L1 and their L2. (Dis)fluency markers were annotated according to their multimodal features. Qualitative analyses revealed how the notions of time suspension and planning associated with (dis)fluency were also found in gesture. This strongly supports the idea that (dis)fluency is to be considered a multimodal phenomenon, and its visual cues are essential for a closer examination of its pragmatic functions.

1. Introduction

In spontaneous typical speech, the course of human language can never be largely continuous, as speakers do not know in advance the specific content they are going to deliver, and how they are going to formulate it. They end up producing a number of “disfluent” utterances in the midst of their discourse. Verbal disfluency is usually defined as a temporary suspension of the speech flow (Ferreira & Bailey, 2004) through filled pauses, silence, repetitions, or whole new utterances. Disfluency is not only a vocal phenomenon and it can be signalled through other modalities: (1) facial expressions, (2) head movements, (3) shoulder movements (Jokinen & Alwood, 2010 p.57). Disfluency markers can also be considered as devices used by speakers to achieve fluency. This paper thus focuses on (dis)fluency as an ambivalent process and on its multimodal features.

A number of studies have been conducted on the relation between speech and gesture following Kendon (2004) or McNeill (1985), but less is known about the relationship between gesture and (dis)fluency specifically. Gullberg (2006) points out two opposite functions served by gestures: (1) an interactional function – gestures that can be useful for turn taking regulation, agreement marking, and attention directing; (2) a self-directed function—gestures addressed towards oneself, dealing with the organization of thought. Similarly, a certain duality can be found in disfluency. Two main views emerge from the literature: (1) Disfluency is the result of speech production “problems” linked to a cognitive load, which disrupt the fluidity of utterances (e.g. Bortfeld et al. 2001; Finlayson & Corley, 2012; Schachter, Christenfeld, & Bilous, 1991); (2) (Dis)fluency markers are communicative strategic devices and time-buying tools which serve discourse planning and structuring functions, and therefore restore continuity in speech. (Allwood, Nivre, & Ahlsén, 1990; Crible, Degand, & Gilquin, 2017; Götz, 2013; Kjellmer, 2003; Kosmala & Morgenstern, 2019; Swerts, 1998; Tottie, 2014).

In the first view, disfluencies are seen as mostly self-directed, as speakers are trying to deal with production problems, while in the second one, they are mostly interactional as disfluencies can also positively contribute to the interaction. Therefore, recent approaches to *(dis)fluency* highlight their functional ambivalence (Crible, Dumont, Grosman, & Notarrigo, 2019; Götz, 2013): (dis)fluencies can both show signs of *fluency* (more other-directed, contributing to the interaction) and *disfluency* (self-directed, disrupting speech). In line with this approach, this paper is grounded in a multimodal, interactive, functional approach to language captured in situated discourse, and aims to explore the ambivalence of (dis)fluency markers conveyed in various modalities. We

examine how discourse suspension and planning associated with (dis)fluency markers can also be conveyed in the visual-gestural channel.

Previous studies have demonstrated the importance of a multimodal approach. Seyfeddinipur (2006) investigated the coordination of speech disfluencies and gestures. Her analysis of speech interruptions and gesture phrases in a corpus study indicated that out of 432 speech suspensions, 306 were accompanied by gestures. This suggested that speech disfluency could affect gesture execution as gestures were suspended at the same time as speech. In an earlier study conducted by Seyfeddinipur & Kita (2001), similar results were found, as they concluded that gestures tended to be suspended prior to the production of speech disfluencies. Gaze could also be seen as an indicator of (dis)fluency. Goodwin & Goodwin (1996) found that speakers frequently gazed away from their interlocutor during word search. They explained that gaze withdrawals usually occurred near “perturbations in the talk displaying initiation of a word search” (p.57). Gestures can also be used to compensate for linguistic problems. In L2 acquisition for example, Gullberg (2006: 108) argues that L2 learners do not only need to acquire grammar and vocabulary, but also “appropriate language use in a broader sense in order to be communicatively competent in a new language”. She further suggests that gesture production reflects the planning load. In this perspective, the use of gesture may help L2 learners to keep talking.

Our analysis is conducted on French and American speakers in L1 and L2 productions. Our hypothesis is that the notion of time suspension, which is inherent to speech (dis)fluencies, is also reflected in other modalities; and that the combination of vocal and gestural features can show overt traces of speech processing. The multimodal features of (dis)fluency can thus provide a better understanding of these ambivalent processes.

2. Data, methods, and results

2.1. Data

The data used for our analysis is drawn from the SITAF Corpus (Horgues & Scheuer, 2015) which comprises tandem interactions between French and English native speakers (undergraduate students, aged 18-21) engaged in production tasks in L1-L1 or L1-L2. Our analysis was conducted on 10 video recordings comprising 6 L1-L2 pairings and 4 L1-L1 pairings, and involved 6 subjects: A03, A07, and A13 (American speakers), and F03, F07, and F13 (French speakers). The pairings included either one native speaker and one non-native speaker (tandem condition) or two native speakers (control condition). The participants performed two tasks which involved telling a story and inserting three lies that the partner had to identify (task 1) and discussing a controversial topic and deciding on their degree of agreement (task 2). The tasks were done respectively in their L1 and in their L2. The duration of our selected corpus is approximately 40 minutes. It should be noted that the purpose of this study was not to compare Task 1 and Task 2 nor the tandem and control conditions specifically but rather to focus on the relationship between (dis)fluency and gesture, so the two tasks were grouped together in the results.

2.2. Methods and annotation

The methodology used for our analysis is derived from a previous pilot study (Kosmala & Morgenstern, 2017). In line with Crible 2017 and Crible, Dumont, Grosman, & Notarrigo (2019), the term “sequence” was adopted to refer to the cluster of immediately adjacent (dis)fluency markers which include: (1) filled pause (uh/um), (2) unfilled or silent pause, (3) syllable prolongations, (4) non-lexical repetitions, (5) self-repairs, (6) self-interruptions, and (7) non-lexical sounds, such as tongue clicks, creaky voice. They were coded according to their position in the utterance, their duration (in ms), and their level of complexity (whether they appear isolated or combined, e.g. *filled pause + unfilled pause*). Their accompanying (total overlap) gestural features were also analysed based on the “gestural phrases” taken from Kendon (2004) and Seyfeddinipur (2006). Gestures were also classified into three functional types, adapted from Kendon (2004) and Gullberg (2011): (1) referential gestures (2) deictic gestures, and (3) pragmatic gestures – gestures not related to the content of discourse but on its structure or “breakdown”. 48% of the (dis)fluencies (230 observations) were annotated by a second coder, and received Cohen’s Kappa measure of 0.84

for the gesture phrase, and 0.78 for the gesture type. The video recordings were transcribed and coded using ELAN (Sloetjes & Wittenburg, 2008).

2.3. Quantitative results

A total of 475 (dis)fluent sequences were found (279 in L1, and 196 in L2), along with 164 accompanying gestures (80 in L1, 94 in L2). Results show that all speakers mostly kept their hands in rest position (64% of the time, ($p < 0.05$)). This is consistent with the view that speakers tend not to gesture when they produce (dis)fluencies (Christenfeld, Schachter, & Bilous, 1991) and that gestures occur much more frequently during fluent speech (Graziano & Gullberg, 2013). However, in cases when speakers did produce gestures, they tended to be suspended or interrupted at the same time as speech, (48% of the time overall ($p < 0.05$)). Additionally, there were more gestures co-occurring with disfluent speech in L2 (47%) than in L1 (25%) ($p < 0.05$). This is consistent with the idea that L2 learners produce more gestures in their L2 than in their L1 (Graziano & Gullberg, 2013, 2018). As shown in Table 2, all speakers mostly produced pragmatic gestures during (dis)fluent speech (approx. 70% both in L1 and L2), which stresses the fact that (dis)fluencies do not necessarily relate to the content of the interaction but rather to its structure, or its “breakdown”, in line with Gullberg (2011).

Table 2
Number of completed gestures during (dis)fluent sequences

	L1	L2
Gest. Type	45	50
referential	7	8
deictic	4	5
pragmatic	34	26

Cases of interrupted or held gestures indicate a synchronization between speech suspension and gesture suspension, while cases of completed gestures show a relation between gesture activity and planning activity. This will be analysed in detail in the next section.

Speakers averted their gaze 81% of the time when producing (dis)fluent sequences, with no significant differences between L1 and L2 (85% and 82%), which is consistent with Goodwin & Goodwin (1996). Speakers did not often display salient facial expressions during disfluent speech (14%) but they were more prominent in L2 speech (21%) than in L1 speech (9%). This may confirm that L2 is more cognitively demanding than L1, and that speakers are more likely to produce “thinking” gestures in L2 in order to seek help from their interlocutors (Gullberg, 2011). Due to the limited size of the data, and the fact that gestures rarely accompany (dis)fluencies, this paper will now focus on a few qualitative examples from the corpus, drawing more specifically on the relationship between (dis)fluency and multimodality. The notions of time suspension and planning will be explored in further detail.

3. On the relation between (dis)fluency and multimodality: qualitative examples

3.1. Gesture and Speech Suspension

One characteristic of speech (dis)fluencies is that they embody a delay in speech (Schegloff, 2010, Clark & Fox Tree, 2002), as their very presence causes a suspension in speech. The following examples will show how this same suspension is also conveyed in gesture with *holds*. Let us consider Example (A), which is an utterance taken from the American speaker A13 speaking in his L2 (French)

- (A) Je suppose que c'est important de:e (1650) [//] d'être là pour ton ami.
I suppose it's important to:o (1650) [//] to be there for your friend.

The underlined part shows the complex (dis)fluent sequence, which contains a prolongation, (*de:e*) an unfilled pause (1650 ms) and a self-repair (*de:e [//] d'être*); the total duration of the sequence is 2.164 ms, which represents a fairly long time of suspension. When looking at its gestural manifestation (Fig.1, first picture) we can see that the speaker is holding his left hand in the same

position, and then slowly moves up his right hand until they are both aligned. It is only then that the speaker returns to fluent speech.



Figure 1. Gestural expression in the (dis)fluent sequence.

There seems to be a synchrony between the suspension of speech indicated by the unfilled pause and the suspension of the hand gesture epitomized in the hold gesture; but more interestingly, there seems to be a relation between the complexity of the (dis)fluent sequence, which is composed of several different (dis)fluency markers (an unfilled pause, a prolongation, and a self-repair) and the gestural activity which is a combination of a hold (left hand) and a (right) hand movement. Both the production of the (dis)fluent sequence and the manual gesture are then followed by fluent speech. In example (B), taken from another American speaker (A03) in her L1 (English), the same notion of suspension is found, but this time with a simple (dis)fluent sequence (i.e. no combination).

(B) we:e went afterwards [/] we:e went to his aunt’s house/ which is closer to my house/**um**
his house is further away.

In this case, the (dis)fluent sequence (the filled pause *um*) lasts 465 ms, so is not as vocally perceptible as in (A), but it is clearly visible in her hand gestures (Fig. 1, second picture). Here the speaker uses two deictic gestures, one directed towards her chest, which points to her house, and the other one directed towards her right, which refers to her boyfriend’s house. Between the two descriptions, her hands momentarily return to the same rest position during the short length of the production of the filled pause *um*.

These examples have shown how the retraction, or suspension of a gesture can be synchronized with the production of the (dis)fluency, which corroborates Graziano & Gullberg (2013, 2018)’s findings. This suggests that (dis)fluency is a multimodal phenomenon, as time suspension is conveyed in the two modalities.

3.2. Planning activity

The moment of suspension signaled by (dis)fluencies can also be used for planning purposes; (dis)fluencies can thus be seen as time-buying tools for planning (see Nicholson, 2007; Tottie, 2014). While (dis)fluencies carry no semantic weight, the accompanying gestures can provide visual cues and help understand the pragmatic functions served by the verbal markers. We will be looking at two cases of pragmatic cyclic gestures accompanying the (dis)fluent sequences. Cyclic gestures can be used “in the transition from non-fluent to fluent speech when finding the word/concept” (Ladewig 2011, p.8). The following examples illustrate this point.

(C) Um he was staying at **ou:ur** like dormroom you know there’s like 6 beds in there.

This utterance is taken from Participant A07 when performing the production task in her L1. She is talking about her stay in South Africa in a youth hostel and the people she met there. In this example, she produces a simple (dis)fluent sequence (marked by a prolongation of 448 ms) before retrieving the noun “dormroom”. Figure 2 (first picture) shows that she is producing a cyclic gesture at the same time as the production of the (dis)fluency, and prior to the lexical item to be retrieved. As soon as the target word is found, she gazes back at her interlocutor, and completes her gesture.

The circular movement of the gesture may be an indication that the prolongation serves a word finding function. Therefore, it could be interpreted as a way to facilitate lexical retrieval, as the cyclic movement could refer to the lexical item to be retrieved. Producing the movement may thus help retrieve the word more quickly, following Krauss (1998) and is in synchrony with the prosodic expression (the 448 ms phonemic prolongation).

A similar example is found in F03's multimodal utterance, also performing the production task in her L1 (French) with another French speaker (F07):

(D) F03: Alors personnellement pendant les dernières vacances donc pas celles de Noël mais celles (**0.662**) après le petit trou qu'on a eu.

F07: ouais.

*So personally during the last vacation so not Christmas vacation but (**0.662**) after the short gap we had.*

Here, the speaker produces an unfilled pause before planning a rather long prepositional phrase (“après le petit trou qu'on a eu”/after the short gap we had) which probably refers to reading week at university. However, she does not seem to know (or has perhaps forgotten) how that short break is called, and therefore describes it by using her own words “le petit trou” (the short gap). While she is trying to retrieve the words, she also produces a cyclic gesture for the duration of her pause (Fig. 2, second picture). But as opposed to (C), her gaze is fixed on her interlocutor while she is producing the cyclic gesture. This could indicate that she is seeking help from her interlocutor, but it may also show that the two speakers share common ground; that is, both are students from the same university, so both are aware of what “the short gap we had” refers to. The fact that she is gazing at her interlocutor thus serves an additional interactive function. As a result, her interlocutor answers with the use of verbal backchanneling (“ouais”/yeah), and nods in agreement.



Figure 2. Cyclic gesture during word search.

These examples have shown how cyclic gestures, used in similar word-finding contexts, co-occurring with a (dis)fluency marker may indicate that the speaker is currently planning parts of the utterance, but can also determine whether the planning process was more self-oriented, therefore more DISfluent (in C) or other-oriented, more communicative, contributing to the fluency of the interaction (in D). The multimodal features of (dis)fluencies thus allow for a finer understanding of these ambivalent processes.

4. Conclusion

This study of L1 and L2 speakers of French and English has shown that (dis)fluent speech and gestures can be synchronized, as speech and gesture production were sometimes suspended at the same time. Moreover, the gestural features have proven to be useful indicators of the pragmatic planning functions associated with (dis)fluencies. However, gestures were not frequent with disfluent speech. A comparative analysis of fluent speech will thus be explored in a larger dataset for future studies. The quantitative findings suggested a higher gestural activity in L2 than in L1 during disfluent speech, and a higher number of pragmatic gestures during (dis)fluencies, which supports previous findings (e.g. Graziano & Gullberg, 2013, 2018), but more quantitative and qualitative work needs to be done on those differences. Overall, the findings provide strong support for the idea that (dis)fluency should not only be viewed as a purely verbal and vocal process, but as a multimodal one as well. While vocal (dis)fluency markers are typically non-lexical as they lack propositional content, their co-occurring visual-gestural features can add visual content and richer meanings, thus providing a finer understanding of these ambivalent processes, typical of spontaneous interactions.

References

- Allwood, J., Nivre, J., & Ahlsén, E. (1990). Speech Management—on the Non-written Life of Speech. *Nordic Journal of Linguistics*, 13(1), 3–48.
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44(2), 123–147.
- Christenfeld, N., Schachter, S., & Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, 20(1), 1–10.
- Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73–111.
- Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses. *Languages in Contrast*, 17(1), 69–95.
- Crible, L., Dumont, A., Grosman, I., & Notarrigo, I. (2019). (Dis)fluency across spoken and signed languages: Application of an interoperable annotation scheme. In L. Degand, G. Gilquin, & A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (Corpora and Language in Use-Proceedings 4). Louvain-la-Neuve: Presses universitaires de Louvain.
- Ferreira, F., & Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, 8(5), 231–237.
- Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: an intentional signal from the speaker? | SpringerLink. *Psychonomic Bulletin & Review*, 19(5), 921–928.
- Goodwin, C., & Goodwin, M. H. (1996). Seeing as a situated activity: Formulating planes. In D. Middleton & Y. Engeström (Eds.), *Cognition and Communication at Work*. Cambridge: Cambridge University Press.
- Götz, S. (2013). *Fluency in native and nonnative English speech* (Vol. 53). John Benjamins Publishing.
- Graziano, M., & Gullberg, M. (2013). Gesture production and speech fluency in competent speakers and language learners. *Presentado En TIGER, Tilburg University, Holanda*.
- Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology*, 9, 879.
- Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL-International Review of Applied Linguistics in Language Teaching*, 44(2), 103–124.
- Gullberg, M. (2011). Multilingual multimodality: Communicative difficulties and their solutions in second-language use. *Embodied Interaction: Language and Body in the Material World*, 137–151.
- Horgues, C., & Scheuer, S. (2015). Why some things are better done in tandem. In *Investigating English Pronunciation* (pp. 47–82). Springer.
- Jokinen, K., & Allwood, J. (2010). Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging. In *Culture and computing* (pp. 55–70). Springer.
- Kjellmer, G. (2003). Hesitation. In defence of er and erm. *English Studies*, 84(2), 170–198.
- Kosmala, L., & Morgenstern, A. (2019). Should “uh” and “um” be categorized as markers of disfluency? The use of fillers in a challenging conversational context. In L. Degand, G. Gilquin, & A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (Corpora and Language in Use-Proceedings 4). Louvain-la-Neuve: Presses universitaires de Louvain.
- Kosmala, L., & Morgenstern, A. (2017). A preliminary study of hesitation phenomena in L1 and L2 productions: a multimodal approach. *TMH-QPSR*, 37.
- Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, 7(2), 54–54.
- Ladewig, S. H. (2011). Putting the cyclic gesture on a cognitive basis. *CogniTextes. Revue de l'Association Française de Linguistique Cognitive*, (Volume 6).
- Nicholson, H. B. M. (2007). *Disfluency in dialogue: attention, structure and function*. University of Edinburgh.
- Schachter, S., Christenfeld, N., & Bilous, F. (1991). Speech Disfluency and the Structure of Knowledge. *Journal of Personality and Social Psychology*, 60(3), 362–367.
- Schegloff, E. A. (2010). Some other “uh (m)” s. *Discourse Processes*, 47(2), 130–174.
- Seyfeddinipur, M. (Ed.). (2006). *Disfluency: Interrupting speech and gesture*. MPI-Series in Psycholinguistics.
- Seyfeddinipur, M., & Kita, S. (2001). Gesture as an indicator of early error detection in self-monitoring of speech. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.
- Sloetjes, H., & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.
- Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4), 485–496.
- Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, 21(1), 6–29.

Children’s viewpoint: Iconic co-speech gestures and their relation to linguistic structure across two communicative genres

Ulrich Mertens¹, Olga Abramov², Anne Németh³, Friederike Kern³, Stefan Kopp², and Katharina J. Rohlfing¹

¹Paderborn University, Germany; ²CITEC, Bielefeld University, Germany; ³Bielefeld University, Germany

ulrich.mertens@uni-paderborn.de, olga.abramov@uni-bielefeld.de,
anne.nemeth@uni-bielefeld.de, friederike.kern@uni-bielefeld.de,
skopp@techfak.uni-bielefeld.de, katharina.rohlfing@uni-paderborn.de

Abstract

In this study, two different communicative genres (explanation vs. report) were elicited in 38 German preschool children at the age of 4 years. In one part of the study, explanations of a game were elicited from the child. The game involved spatial movements and figures with various geometrical shapes. In a subsequent part, children reported about a puppet and its odd behaviour to their caregiver. We examined children’s viewpoint in iconic co-speech gestures and related it to the children’s event structures and linguistic structures that differed in terms of transitivity. Our findings suggest that children do not use viewpoints in a unified way—which had been reported from studies with adults. In contrast, our results indicate a great variability in the ways children use viewpoint in iconic co-speech gesture. We found that different communicative genres (explanation vs. report) evoke different viewpoints in gesture, due to their different event structure and linguistic structure. During the genre “explanation”, O-VPT gestures occurred more frequently with intransitive utterances, whereas during the genre “report”, C-VPT gestures occurred more frequently with transitive utterances. Moreover, neither of the events within the communicative genres exclusively evoked one specific viewpoint.

1. Introduction

During communication, gesture provides insights into people’s viewpoints. The viewpoints mainly used in iconic gesture are character viewpoint (C-VPT) and observer viewpoint (O-VPT) (McNeill, 1992). Describing an event with C-VPT gestures, a speaker focuses on the manner of the observed action and takes the first-person perspective of the character (McNeill, 1992). Research suggests that by taking the C-VPT, people conceptualize an event from an internal view and thus gain first-person knowledge (Black, Turner, & Bower, 1979; Demir, Levine, & Goldin-Meadow, 2015; Speer, Reynolds, Swallow, & Zacks, 2009). In contrast, using O-VPT within an iconic gesture, a speaker presents an event from the third-person perspective, focusing on the path of a character’s action (McNeill, 1992). For example, depicting how a character climbed up a ladder, using the O-VPT, the speaker would move her or his hand from below to above, which would represent the whole character and the trajectory of the character. In contrast, taking the C-VPT, a speaker could mimic the movements of the character’s hands, showing how the character actually climbed up the ladder. Why speakers take a specific perspective in gesture is uncertain and remains widely debated in the literature (e.g. Dancygier & Vandelanotte, 2017; McNeill, 1992; Parrill, 2010). McNeill (1992) hypothesizes that maximally salient or newsworthy information evokes C-VPT. He also argues that the centrality of an event and the linguistic structure (transitivity) of an utterance lead to a specific viewpoint. An event can be central or peripheral to a discourse (McNeill, 1992; Parrill, 2010; Stein & Glenn, 1975), and the structure of an event can be transitive (the verb requires a direct object), or intransitive (the verb requires no direct object). While central events of narrations and transitive utterances evoke character viewpoints, peripheral events and intransitive utterances are related to observer viewpoints in gesture (McNeill, 1992). Parrill (2010) conducted a study with adults and confirmed McNeill (1992)’s proposition that C-VPTs occurred with transitive utterances

and O-VPTs with intransitive utterances. However, she did not confirm whether the centrality of an event is related to a specific viewpoint. Furthermore, Parrill (2010) showed that the structure of an event alone is more likely to evoke a specific viewpoint than the accompanying linguistic structure of an utterance.

Events which evoked C-VPTs involved some sort of handling, use of the torso, emotions, or events which are difficult to depict from an O-VPT. Events which evoked O-VPTs included trajectories. Many event structures exclusively evoke one viewpoint in particular. However, Parrill (2010) argued, that C-VPT events like reading a newspaper mostly occur with transitive utterances (“he is reading a newspaper”) while O-VPT events with trajectories usually occur with a subject-verb-prepositional phrase (intransitive).

1.1. Children’s manner of gesturing

To our knowledge, how children’s viewpoint is related to event structures and linguistic structures has not yet been investigated. Children’s manner of gesturing strongly differs from adults, e.g. in terms of object presentation and speech-gesture synchronization (Butcher & Goldin-Meadow, 2000; Heilman, Rothi, & Valenstein, 1982; Overton & Jackson, 1973). Heilman and colleagues (1982) found that children between 3 – 6 years often use body parts to represent an object physically, whereas older children and adults use their hands as hands and therefore construct an imaginary object. The differences between children and adults might be related to children’s gestural system being coupled differently with communicative behavior. In fact, speech-gesture-systems take a while to develop the proper timing (Esteve-Gibert & Prieto, 2014) and meaningful synchrony (Butcher & Goldin-Meadow, 2000). Regarding initial communicative gestures in infants, Esteve-Gibert and Prieto (2014) observed that prosodic features of vocalization and related features of the gesture execution (the gestural stroke) became more closely related to each other temporally once infants began to produce their first words. However, the differences between adults and children in the use of the VPT within gestures were never considered. The study by Demir and colleagues (2015) found that not all children at the age of five used C-VPT during a retelling task. It was observed that children were more likely to tell better structured stories at a later age when they expressed a character’s viewpoint in gesture. It is thus reasonable to argue that children’s speech behavior is related to the way that they gesture viewpoints. Little is known about how children’s speech-gesture system is established and organized, especially across communicative genres (Kern, 2011; Streeck, 2009). Communicative genres differ within their discursive demands on the interactants and require different communicative solutions (Günthner & Knoblauch, 1995; Quasthoff, Heller, & Morek, 2017). Finding patterns in communicative behavior across communicative genres would confirm that particular communicative behaviors are context-specific, or a general phenomenon.

In this study, we examined how children’s viewpoint within iconic co-speech gestures is related to the structure of an event, and to the linguistic structure of an utterance. For this purpose, we compared our findings in two different communicative genres with findings on adults’ use of viewpoints from the existing literature.

2. Method

Preschool children ($n = 38$) from Germany at the age of 4 participated in our study. Together with their caregivers, the children visited the lab. The caregiver played an active role in our setting and interacted with the child during the communicative tasks of explanation and report. During both tasks, the children spoke about events where they either performed actions by themselves, or observed a character performing actions with objects. As the children were not given a time limit for their responses, we have a wide variation in the amount of utterances and gestures children produced during their performance. To make children’s responses comparable, we divided the total number of children’s use of gesture types by children’s total number of intonation phrases.

2.1. Procedure

The main idea behind the study was to offer the children the opportunity to gesture freely. In our pilot studies, we noticed that children felt less comfortable with an experimenter, resulting in the children speaking in short utterances with fewer gestures. We therefore decided to facilitate the caregiver playing an active role as an interaction partner, and thus to elicit communicative behavior

from children which is more natural and more directly reflects their abilities. In both tasks discussed here, the caregiver was not present during a phase in which the child engaged in an activity with an experimenter. After this activity, the caregiver who had been waiting outside came back into the room and initiated a conversation with the child. Note that we did not control for the caregiver's behavior. This is because we are convinced that children at this age are scaffolded within dialogue and more complex communicative genres, which serves as a first step on their way to being able to provide a complex monologic explanation on their own.

During the communicative genre 'explanation', the experimenter and a child played a jigsaw game designed by the authors. It consisted of dice and a 20 cm x 35 cm x 0.2 cm Plexiglas board that was painted to depict a landscape. This landscape reflected a city at night: a church, a house, a moon and a star (see: Figure 1). The shapes on the dice included a: triangle, quadrangle, circle, star, and moon. One side of the dice lacked a shape. Some pieces of the board were cut out, leaving a hole that could be covered by a fitting piece. During the game, the child was allowed to throw the dice six times. If they rolled a particular shape for the first time, the child 'flew' with a small figurine through the corresponding shape on the board. First, the experimenter explained the game to the child, and then they played it together. Afterwards, the caregiver who re-entered the room asked the child what he or she had been doing. After hearing that the child had played a game, the caregiver asked the child to explain the game in order to play it with him or her later. Our analyses only refer to the situation in which the child explains the game. During the communicative genre report, the experimenter acted out a puppet character (a dog) performing some incorrect actions. The puppet dog told the child it was excited to show them how it had learned some everyday actions from humans. For example, the dog showed the child how to eat with a spoon (but held the spoon on the wrong side) or how to drink from a bottle (but drank from the bottom of the bottle) (see: Figure 1). After performing an action, the child was allowed to correct the puppet. After the activity, and after re-entering the room, the caregiver asked the child about what he or she experienced in order to elicit a report by the child. Our analyses refer only to the situation in which the child reported the event to the caregiver.

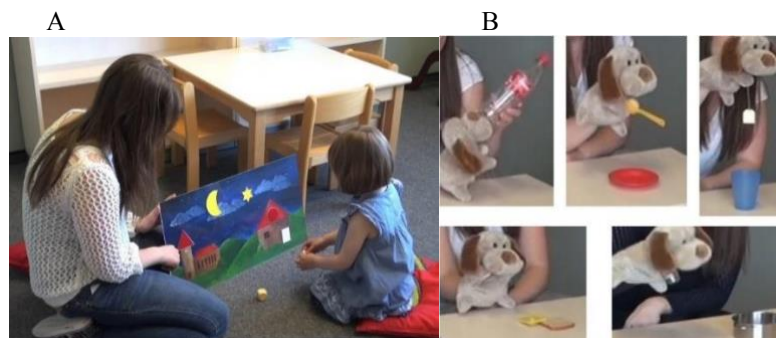


Figure 1. (A) Child plays a self-made jigsaw-puzzle-game with the experimenter (Explanation). (B) Activities the dog performed incorrectly (Report).

2.2. Coding

All verbal productions produced by narrators were transcribed and separated into intonation phrases (Barth-Weingarten, 2016; Halliday, 2015). Each utterance that was accompanied by an iconic gesture was coded regarding its transitivity (transitive: “the dog drinks water”, intransitive: “the figure flies”). All iconic gestures produced by children were coded for viewpoints. Viewpoints were sorted into four categories: character viewpoint (C-VPT); observer viewpoint (O-VPT), dual viewpoint (D-VPT) and no viewpoint (N-VPT). As almost no D-VPT and N-VPT were carried out, these categories were excluded from the analyses. Children's verbal utterances were transcribed into three types of linguistic structure. The linguistic structures were transitive, intransitive and neither. We focused on transitive and intransitive utterances. We measured the agreement between the coders using Cohen's Kappa (Cohen, 1960) for speech and gesture separately. The mean Kappa value for viewpoints is $k = .860$ ($SD = .093$) and for linguistic structure $k = .887$ ($SD = .027$). During the game explanation, the events were 'throwing a dice' and 'performing flying actions with a figure'. Within the genre report, the events were: 'eating properly with a spoon', 'drinking properly from a bottle',

'placing cheese properly on a slice of bread', 'placing a teabag properly into a cup', 'putting salt with a salt shaker properly into a pot'.

3. Results

3.1. Viewpoints with respect to the event structure

During the genre explanation, children used descriptively more O-VPTs ($M = .061$; $SE = .009$) than C-VPTs ($M = .038$; $SE = .010$) but this effect was not significant ($Z = -1.708$; $r = .277$; $p = .088$). In addition, neither the event of throwing a dice nor the event of flying with a figure elicited one particular viewpoint in a dominant manner. During the genre report, children used more C-VPTs ($M = .127$; $SE = .019$) than O-VPTs ($M = .027$; $SE = .006$) reaching significant effects ($Z = -4.469$; $r = .725$; $p = .001$). However, when looking at the whole sample, we found no event that was expressed with one specific viewpoint; rather, events could be expressed with different viewpoints (see Figure 2).

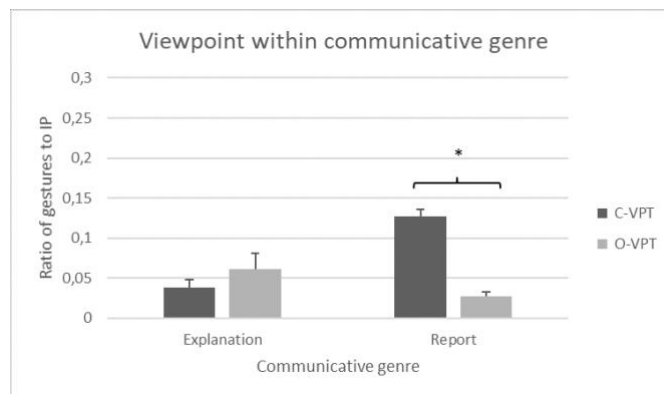


Figure 2. Ratio of gestures to intonation phrases (IP) of each viewpoint type according to communicative genres.

3.2. Linguistic structure

During the genre "explanation", children used descriptively more C-VPTs accompanied by transitive utterances ($SE = .011$; $M = .030$) than C-VPTs accompanied by intransitive utterances ($M = .014$; $SE = .005$), but this effect did not reach significance ($Z = -1.961$; $r = .319$; $p = .051$). Moreover, children used more O-VPTs accompanied by intransitive utterances ($M = .044$; $SE = .010$) than O-VPTs accompanied by transitive utterances ($SE = .005$; $M = .010$); this effect is significant ($Z = -3.043$; $r = .494$; $p = .002$).

During the genre "report", children used significantly more C-VPTs accompanied by transitive utterances ($M = .043$; $SE = .008$) than C-VPTs accompanied by intransitive utterances ($M = .011$; $SE = .003$); this effect reached significance ($Z = -3.264$; $r = .530$; $p = .001$). This means that character viewpoint was synchronized with children's utterances containing verbs with objects. Overall, children used descriptively less O-VPTs accompanied by intransitive utterances ($M = .002$; $SE = .008$) than O-VPTs accompanied transitive utterances ($M = .003$; $SE = .009$); this effect did not reach significance ($Z = -3.588$; $r = .582$; $p = .999$) (see Figure 3).

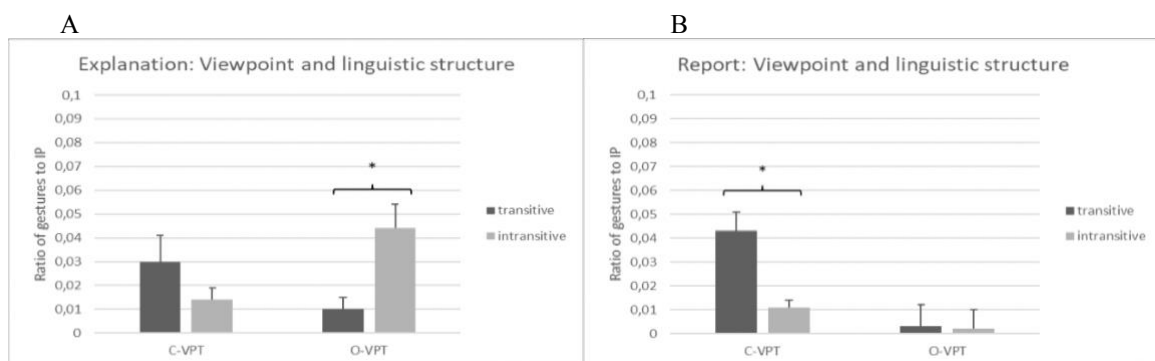


Figure 3. (A) Ratio of gestures to intonation phrases (IP) of each viewpoint type according to linguistic structure in Explanation (A) vs. Report (B).

4. Discussion

The aim of the study was to identify patterns within children's speech-gesture system. Therefore, we examined how children's linguistic structure is related to viewpoint in speech-accompanying iconic gestures across two different communicative genres with different event structures. We found that different communicative genres (explanation vs. report) evoke different viewpoints in gesture. During the genre "explanation", O-VPT gestures were more likely to occur with intransitive utterances, whereas during the genre "report", C-VPTs occurred more frequently with transitive utterances. Moreover, neither of the events evoked one specific viewpoint exclusively.

Parrill (2010) reported that adults tend to use C-VPTs when speaking about events with motoric properties, consisting of character's hands and torso, while visuo-spatial events with trajectories seem to evoke O-VPTs. In this study with adults, many events seemed to exclusively evoke one specific viewpoint in gesture (Parrill, 2010). Regarding the linguistic structure, C-VPTs seemed to occur with transitive, and O-VPTs with intransitive utterances (McNeill, 1992; Parrill, 2010). Our results extend Parrill's (2010) findings on adults' viewpoint in two ways: Firstly, we were able to analyze children's iconic co-speech gestures; secondly, we analyzed children's performance within two different communicative genres. Our findings suggest that the occurrence of children's viewpoint in gesture differs from adults in several ways. Firstly, no event was exclusively performed with one specific viewpoint in gesture. Secondly, specific viewpoints did not occur with a specific kind of linguistic structure. If Parrill's (2010) findings on adults were to be transferred to our findings, we would see some events executed exclusively with one specific viewpoint. However, in our study, the events evoked both C-VPT and O-VPT gestures. For example, using a saltshaker evoked different viewpoints: Some children represented the salt with their fingers and showed how the salt falls (O-VPT). Therefore, children represented the salt physically. Adults would be more likely mimic the holding of a salt shaker and fulfil a shaking action (C-VPT) by representing the salt shaker and the salt imaginarily (Heilman et al., 1982; Parrill, 2010). Another example is that of children showing the cube rolling over the floor, by representing the dice with their hands physically (O-VPT). There are several explanations possible for this effect. One drawn from literature suggests that children's "internal reference", or representation system, might not be fully developed (Heilman et al., 1982). Another explanation, which we favour, is that the iconic practices children use are not fully following conventionalized practices of gesturing (Streeck, 2009). Beside these differences between adults and children, our findings suggest that the events within the communicative genre of a report evoke significantly more C-VPTs than O-VPTs. This is in accordance with Parrill's (2010) findings suggesting that events with motoric properties including actions with character's hands and torso evoke C-VPTs in gesture.

For the communicative genre of an explanation, we found that intransitive utterances are more often accompanied by O-VPT gestures than transitive utterances. However, in contrast to adults, the linguistic structure (transitivity) of children's utterances did not differ when using C-VPTs. In parallel, during the communicative genre of a report, transitive utterances occurred mostly with C-VPTs. However, we found no differences between the linguistic structures that accompanied O-VPT gestures. But this could be because children used very few O-VPT gestures during the explanation phase. Therefore, we need more data to verify these results.

One explanation for why children's viewpoint does not map a specific iconic meaning to a kind of linguistic structure could be the way in which children represent objects. Children at the age of four often represent objects physically, while adults would represent the object imaginarily (Overton & Jackson, 1973), which could be due to the fact that children's iconic practices do not fully follow conventional iconic gestural practices (Streeck, 2009). This form of object representation might result in children's gestures taking an O-VPT, when adults would use a C-VPT gesture. This alternative viewpoint resulting from differences in object representation within children could be one reason why the viewpoint in gestures is not as uniform across the same linguistic structures in children as in adults.

5. Conclusion

Children's use of viewpoints in iconic co-speech gestures differs in many ways from adult usage, and between communicative genres. In addition, when comparing across different communicative genres (explanation vs. report), we do not see a unified manner in which children use their viewpoints to provide specific iconic meaning to a certain kind of linguistic structure. Instead, we found that the relation of children's viewpoint and the linguistic structure underlying children's utterances differs across communicative genres. Therefore, we can conclude that the relationship between children's viewpoints in gesture and linguistic or event structures differs from that in adults. More research is needed on how children's way of representing objects is related to their viewpoint.

Acknowledgments

This work was supported by the German Research Foundation (DFG; KE1627/3-1, KO 3510/4-1, RO 2443/8-1). We thank all the participating children and parents.

References

- Barth-Weingarten, D. (2016). *Intonation units revisited: Cesuras in talk-in-interaction. Studies in language and social interaction: Volume 29*. Amsterdam, Philadelphia: John Benjamins Publishing.
- Black, J. B., Turner, T. J., & Bower, G. H. (1979). Point of view in narrative comprehension, memory, and production. *Journal of Verbal Learning and Verbal Behavior*, 18(2), 187–198.
- Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech: When hand and mouth come together. In D. McNeill (Ed.), *Language, culture and cognition: Vol. 2. Language and gesture* (1st ed., pp. 235–258). Cambridge: Cambridge Univ. Press.
- Dancygier, B., & Vandelanotte, L. (2017). Viewpoint phenomena in multimodal communication. *Cognitive Linguistics*, 28(3), 164.
- Demir, Ö. E., Levine, S. C., & Goldin-Meadow, S. (2015). A tale of two hands: Children's early gesture use in narrative production predicts later narrative structure in speech. *Journal of Child Language*, 42(3), 662–681.
- Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, 57, 301–316.
- Günthner, S., & Knoblauch, H. (1995). Culturally patterned speaking practices - the analysis of communicative genres. *Pragmatics*, 5(1), 1–32.
- Halliday, M. A. K. (2015). *Intonation and grammar* (Reprint 2015). *Janua Linguarum. Series Practica: Vol. 48*. Berlin/Boston: De Gruyter; De Gruyter Mouton.
- Heilman, K. M., Rothi, L. J., & Valenstein, E. (1982). Two forms of ideomotor apraxia. *Neurology*, 32(4), 342.
- Kern. (2011). Der Erwerb kommunikativer Praktiken und Formen am Beispiel des Erzählens und Erklärens. In S. Habscheid (Ed.), *de Gruyter Lexikon. Textsorten, Handlungsmuster, Oberflächen: Linguistische Typologien der Kommunikation* (pp. 231–254). Berlin: De Gruyter.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: The University of Chicago Press.
- Overton, W. F., & Jackson, J. P. (1973). The Representation of Imagined Objects in Action Sequences: A Developmental Study. *Child Development*, 44(2), 309.
- Parrill, F. (2010). Viewpoint in speech–gesture integration: Linguistic structure, discourse structure, and event structure. *Language and Cognitive Processes*, 25(5), 650–668.
- Quasthoff, U., Heller, V., & Morek, M. (2017). On the sequential organization and genre-orientation of discourse units in interaction: An analytic framework. *Discourse Studies*, 19(1), 84–110.
- Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, 20(8), 989–999.
- Stein, N., & Glenn, C. G. (1975). An Analysis of Story Comprehension in Elementary School Children: A Test of a Schema. R. O. Freedle (Ed.), *New directions in discourse processing*. Norwood, NJ: Ablex.
- Streeck, J. (2009). *Gesturecraft* (Vol. 2). Amsterdam: John Benjamins Publishing.

Acoustic specification of upper limb movement in voicing

Wim Pouw^{1,2}, Alexandra Paxton^{1,3}, Steven J. Harrison¹, and James A. Dixon¹

¹Center for the Ecological Study of Perception and Action, University of Connecticut

²Department of Psychology, Education, & Child Studies, Erasmus University Rotterdam

³Department of Psychological Sciences, University of Connecticut Affiliation

wimpouw@uconn.edu, alexandra.paxton@uconn.edu, steven.harrison@uconn.edu,
james.dixon@uconn.edu

Abstract

Hand gestures communicate through the visual information created by movement. Recently, we found that there are also direct biomechanical effects of high-impetus upper limb movement on voice acoustics. Here we explored whether listeners could detect information about movement in the voicing of another person. In this exploratory study, participants listened to a recorded vocalizer who was simultaneously producing low-(wrist movement) or high-(arm movement) impetus movements at three different tempos. Listeners were asked to synchronize their own movement (wrist or arm movement) with the vocalizer. Listeners coupled with the frequency of the vocalizer arm (but not wrist) movements, and showed phase-coupling with vocalizer arm (but not wrist) movements. However, we found that this synchronization occurred regardless of whether the listener was moving their wrist or arm. This study shows that, in principle, there is acoustic specification of arm movements in voicing, but not wrist movements. These results, if replicated, provide novel insight into the possible interpersonal functions of gesture acoustics, which may lie in communicating bodily states.

1. Introduction

A conundrum in gesture studies is that gestures are often recruited by a gesturer who knows full well that gestures will never visually reach the listener. For example, during phone conversations, we do not stop gesturing (Bavelas, Gerwing, Sutton, & Prevost, 2008). Even speakers with congenital blindness gesture to listeners who also are blind since birth (Iverson & Goldin-Meadow, 2001).

Here we explore the possibility that visual information from gesture is but one of its (communicatively meaningful) products. Recently, we have found that upper-limb movements with relatively high physical impetus produce prominent but non-intentional changes in voice quality (Pouw, Harrison, & Dixon, 2018b). Specifically, we found peaks in the fundamental frequency (F0) and the amplitude envelope of continuous phonation of the vowel /a:/ when participants made high-impact movements that recruited the entire arm but not when producing low-impact wrist movements or when standing still (see Figure 1). Such peaks in phonation were observed at the moment at which posturally destabilizing forces of the arm movements were highest and at which the body counteracted such forces by tensioning of the muscles in anticipatory fashion. These results accommodate findings as observed in naturalistic contexts. Namely, sudden increases in speech intensity and fundamental frequency are key properties that define the prosody of speech, and spontaneous co-speech gestures are known to synchronize with such prosodic aspects of speech (Wagner, Malisz, & Kopp, 2014). Scaling up to natural speech, other work has found that infants' babbling becomes more adult-like in voice quality when infants simultaneously and rhythmically move their arms (Ejiri & Masataka, 2001) and that encouraging gesture production during adults' speech production boosted maximum observed F0 and intensity of speech (Cravotta, Busà, & Prieto, 2018).

We could wonder therefore whether there is information in speech acoustics specifying bodily gestures. Note, Hoetjes et al. (2004) found no statistically significant changes in acoustics when participants were (restrained) from gesturing, nor were listeners able to detect whether

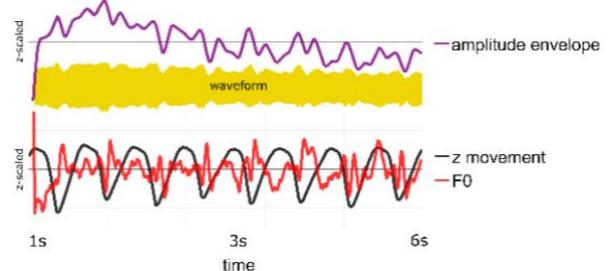


Figure 1. Example motion vis-à-vis acoustics.

someone was gesturing based on listening to their speech. However, we could argue mixed results might be obtained by averaging acoustic metrics over time (cf. Hoetjes et al.) and inferences about acoustics and gesture might be obscured when not taking into account physical impetus of gestures.

Since gestures—especially of the more forcefully beat-like kind (see e.g., <https://osf.io/29h8z/>)—affect voice acoustics we should assess whether listeners can pick up information about a gesturer’s movement. The idea that there is acoustic information that specifies an object or event in the environment is actually non-controversial in ecological psychology of language (Fowler, 1986) and object perception (Carello, Wagman, & Turvey, 2005). Namely, Carol Fowler famously asserted that we do not hear speech sounds that we need to translate into meaningful objects of language perception but that we directly hear the cause of the sound - the articulatory gestures. Evidence for this includes studies on the McGurk effect, in which otherwise ambiguous speech sounds are disambiguated by visually or even manually perceiving the articulatory gesture (Fowler & Dekle, 1991). Furthermore, a line of research in ecological acoustics has shown that properties of objects (e.g., object length, object thickness; relative position) can actually be directly perceived by attuning to acoustic properties of the objects (Carello, Anderson, & Kunkler-Peck, 1998).

The current idea that we can hear bodily gestures is then complementary to these Gibsonian perspectives (Gibson, 2014). However, we are after a direct specification of bodily action in speech acoustics. If such specification exists to some degree, this would open up the investigation into whether bodily gestures’ communicative function lies in part in its direct linkage with speech acoustics. We have a long experimental road ahead before we could conclude that gesture acoustics serve such a communicative role in a manner similar to the visual information created by gesture. Indeed, to date there is simply no evidence that humans can hear gesture (Hoetjes, Krahmer, & Swerts, 2014).

In the current exploratory study two participants were asked to make a wrist or arm motion while listening to a recording from a vocalizer, an original participant from Pouw, Harrison, et al. (2018a). The vocalizer was continuously voicing the vowel /a:/ while making a high-impetus arm motion or a low-impetus wrist motion at slow, medium, or fast movement tempos. Arm motions have a higher physical impetus on the body as larger body parts are involved in the movement, as compared to wrist movements. The current participants’ task was to synchronize their own movement with the movement of the vocalizer, as they perceive it via the acoustics. The current exploratory study served as a basis for a pre-registration of a planned confirmatory study (see <https://osf.io/9843h/>). For a comprehensive follow-up study and description of methods see (Pouw, Paxton, Harrison, & Dixon, under review).

2. Methods and results

2.1. Participants and design

Two female graduate students (ages: 22 and 28) participated in the current exploratory study. The current study entailed a full 2 x 2 x 3 within-subject design: a two-level factor ‘listener movement’ condition (listening while moving wrist vs. listening while moving arm), a two-level factor ‘vocalizer movement’ condition (wrist movement vocalization vs. arm movement vocalization), and a three-level factor ‘tempo’ (slowed down vs. self-paced vs. sped up). Note that slowed-down and sped-up versions were derived from the original self-paced movement vocalizer (see procedure below). Participants performed 12 trials, one for each cell of the design (2 x 3 x 2). Each trial consisted of 5 phonation cycles where the vocalizer took a full breadth and phonated until breadth was almost emptied and phonation could not be steadily maintained (Pouw, Harrison, et al., 2018a, 2018b).

2.2. Stimuli

We extracted two trials collected with a participant in a previous study wherein (henceforth referred to as the ‘vocalizer’). For both trials, the vocalizer continuously voiced the vowel /a:/ (as in ‘cinema’) while making repetitive up-and-down upper-limb movements at a self-paced tempo (around 1.8Hz). The movements were made on the sagittal plane with fingers fully extended, with a higher velocity

in the down-phase so as to have a beat-like movement profile. The vocalizer was instructed to try not to let voicing be affected by the movement.

In the extracted wrist-movement vocalization trial, repetitive wrist movements of the dominant right hand were made with no movement around the elbow joint. For this wrist movement, the elbow joint was kept at a 90-degree angle. The wrist movement vocalization trial (listen to the sound-clip here: <https://osf.io/rvx3c/>) reflected a low-impetus movement relative to second arm-movement vocalization trial (sound-clip available here: <https://osf.io/ymqnu/>). The arm-movement vocalization trial was produced by the participant moving her lower arm around the elbow joint, while keeping the wrist joint locked at 0 degrees.

To construct the stimuli of different tempos, we first looped the original audio track from the vocalizer (i.e., movement with self-directed speed) 5 times, such that there were 5 voicing episodes with intermittent pauses where the vocalizer took a full breath. This self-paced vocalization track serves as our “normal” tempo stimuli. We then created two additional versions of this vocalization track, one that was artificially slowed down by 20% and one that was artificially sped up by 20%. These transformations were done with AVS Audio Editor (Online Media Technologies Ltd.), which allows for tempo transformation while maintaining the original pitch. We made a set of three vocalization tempos for both the wrist-movement vocalization conditions and arm-movement vocalization conditions. Effectively this resulted in 3 tempo conditions (slow down vs. self-paced vs. sped up). The tempo conditions provide additional information whether participants are sensitive to movement-induced rhythm in voicing, which would be evident in lower or higher frequency of listener movement for slowed down or sped up condition (respectively) as compared to self-paced tempo condition.

2.3. Motion-tracking equipment

A Polhemus Liberty sampling at 240Hz was used to record movement of the listener (L). Given that upper-limb motions were primarily defined by vertical motion (in the z-dimension), we will only perform synchrony analyses for this dimension. We smoothed z-position traces with a first-order low-pass Butterworth filter of 33 Hz.

Audio Presentation. Participants wore a Samsung Level On EO-PN900BBEGUS headphone (with noise-cancelling deactivated) with a wired connection to the PC. Volume was set at a comfortable level for the participant. The audio was pre-buffered and then played using a custom C++ script that started the audio at the exact moment that the motion tracker started recording. This ensured that the original vocalizer motion-tracking data is completely synchronized with listener motion-tracking data.

Procedure. Each participant (i.e., listener) was asked to stand upright with their elbow in a 90-degree angle. The experimenters then demonstrated the two movement types that the listener needed to make: one wrist movement and one arm movement. Subsequently the participant was informed that they would repeatedly listen to someone voicing, while the vocalizer had been concurrently moving her upper limbs at different speeds (although no additional information was given about speed manipulations). The listener was then told that they would need to synchronize with the movements of the vocalizer, based solely on hearing her voice. Participants briefly practiced the synchronization task with an arm movement vocalization trial of self-paced tempo, while the listener was asked to move their wrist or arm in synchrony. After practicing, participants performed 12 trials (in randomized order) containing all 2 (listener movement) x 2 (vocalizer movement) x 3 (tempo) conditions wherein they heard 5 voicing episodes before going to the next trial.

2.4. Analyses

Spectral Analyses (FFT). We performed spectral analyses (fast Fourier transform or FFT) using R’s native stats package (function `spectrum`) to assess changes in listeners’ movement frequency as a function of vocalizer tempo condition.

Relative Phase Analyses (Φ). We performed relative-phase analyses using a simple continuous point-wise estimation method (e.g., Zelic, Kim, & Davis, 2015; see also Kelso, Del Colle, & Schoner, 1990). To calculate Φ we used the equation

$$\phi = 2\pi\Delta t/T_v$$

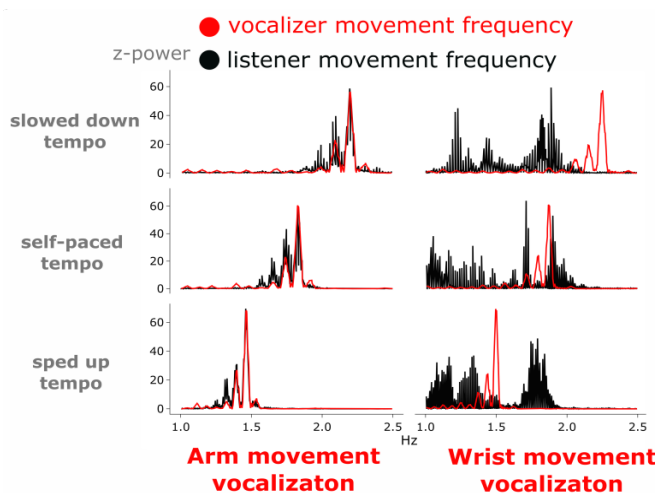
where T_v is the current time interval for the original vocalizer’s maximum vertical extension (i.e., the time between each beat of the vocalizer movement). Δt reflects the asynchrony of the moment

of maximum vertical extension of the listener versus the vocalizer. 2π transforms temporal dispersion into the angular dispersion. We converted the equation's output from radians to degrees such that 0 degrees indicated in-phase coordination, -180 degree indicated anti-phase coordination with listener in the lead, and 180 degrees indicated anti-phase coordination with vocalizer in the lead.

2.5. Overview results

There are two hierarchically organized research questions that need to be answered before concluding that there is (some) acoustic specification of upper-limb movements in phonation. Firstly, can participants attune to the rhythmic tempo of the movement? Secondly, if indeed participants are sensitive to rhythm in phonation, are participants able to attune to the exact phasing of the vocalizers' movement in a 1:1 in-phase fashion? Note, a supplementary figure is available at <https://osf.io/zngb2/> containing an example time series of the listener (participant 1) against the vocalizer for different movement tempos.

Acoustic Specification of Movement Tempo in Phonation (FFT Analyses). Next, we formally assessed for all the data the degree to which participants were attuning to tempo information in the vocalization. Figure 2 shows the mean spectral results for the arm- vs. wrist-movement vocalization conditions for all three tempos (and regardless of which movement listeners were making). Namely, there were no clear effects for when listeners were trying to synchronize while making wrist- versus arm-movements (see additional plot with listener movement conditions here: <https://osf.io/6adm4/>). As shown in Figure 2, we found clear evidence for listener-vocalizer tempo-specific movement coupling when the listener heard the clips in which the vocalizer was making arm movements (i.e., arm-movement vocalizations), but not those trials in which the vocalizer was making wrist movements (wrist-movement vocalizations). Thus, for wrist-movement vocalizations, participants seemed to fail to pick up movement tempo information; while listening to arm-movement vocalizations, participants could both adjust tempo of their own arm movements and wrist movements.



Note. FFT results for all movement frequencies (horizontal axis in hertz [Hz]; vertical axis z -standardized power for that frequency) for both the vocalizer and the listener movement frequencies. The vocalizer wrist and arm movement frequencies show slower or faster frequencies per tempo condition by design (as we artificially manipulated the tempo for these conditions). The natural frequency of the vocalizer (original tempo condition) was about 1.75 Hz, with a slight faster frequency for when the vocalizer was making a wrist movement. Interestingly, in the arm vocalization conditions, there was clear frequency coupling between listeners' movement with that of the vocalizer. This is indicated by the large overlap of spectral peaks in arm vocalization condition.

Figure 2. Spectral results movement frequencies for vocalizer and listener.

For statistical confirmation of the results obtained in Figure 1, we assessed whether listener's dominant movement frequencies were affected by tempo condition (as well as vocalizer movement condition and listener movement condition). That is, we quantified whether tempo conditions predicted dominant frequency—with higher frequencies for fast tempo conditions and lower frequencies for slow tempo conditions, both as compared to the original tempo. To test this, we extracted the frequency with the highest observed power (i.e., dominant frequency) for each trial. Subsequently we performed `nlme` mixed regressions using participant as a random intercept (adding adding random slopes caused the model not to converge), identifying the best model by comparing model fits at increasing levels of complexity.

Compared to a model predicting the overall mean for dominant frequency, entering tempo condition as a predictor for dominant frequency improved the fit of the model (change in $\chi^2 [1] = 10.32, p = .006$). Adding to the previous model, vocalizer movement condition improved the fit of the model further, change in $\chi^2 [1] = 6.95, p = .008$. Adding the interaction between tempo (3 levels)

the model further, change in $\chi^2 [1] = 6.95, p = .008$. Adding the interaction between tempo (3 levels) and vocalizer-movement conditions to the previous model further improved the model, change in $\chi^2 [2] = 7.67, p = .021$. Finally, adding listener movement condition to this previous model did not significantly improve predictions of dominant frequency further (change in $\chi^2 [1] = 2.41, p = .12$).

The best-fit model with vocalizer movement and tempo (and their interaction) was assessed with post-hoc comparisons with the R package `lsmeans` (using Tukey correction for multiple comparisons). Only the arm movement vocalization condition showed tempo scaling of listener movement with that of the vocalizer, and this was only statistically reliable for the contrast between sped-up vs. slowed down tempo condition. All model results are reported in Table 1.

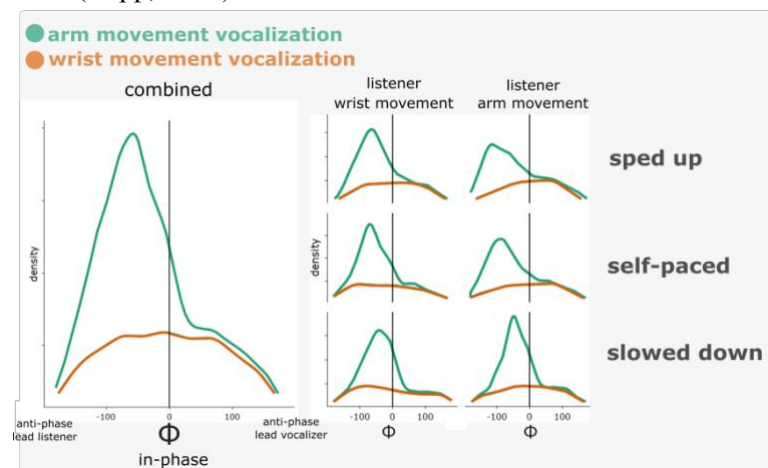
Table 1

Post-hoc comparisons for frequency scaling per tempo and vocalizer movement condition

Arm Movement Vocalization	Difference estimate	t ($df=17$)	p -value (corrected)
Sped up - slowed down tempo	0.71Hz	4.89	<.001
Sped up - self-paced tempo	0.36Hz	2.52	.054
Self-paced - slowed down tempo	0.34 Hz	-2.37	.073
Wrist Movement Vocalization			
Sped up - slowed down tempo	0.181	1.25	0.443
Sped up - self-paced tempo	0.192	1.33	0.402
Self-paced - slowed down tempo	-0.011	-0.077	0.996

Acoustic Specification of Movement Phasing in Phonation (Relative Phase Analyses: $SD \Phi$).

Now that we have established that there is frequency-coupling between listener and vocalizer movement (but only for vocalizer arm movement), we assess whether there is also phase-synchronization. Note that it is possible that participants picked up the rhythmic structure in the voicing while being oblivious about the exact phases of the vocalizer's movement. Figure 3, however, clearly shows that there was indeed listener-vocalizer phase-coupling but only for the vocalizer arm movement. Furthermore, there is not perfect in-phase locking but rather a negative mean asynchrony whereby the listener anticipates the vocalizer with about $\Phi = 50$ degrees. Note that negative mean asynchrony is a very common phenomenon in sensorimotor synchronization tasks (Repp, 2005).



Note. The left panel shows the relative phase distributions for all data combined for the vocalizer wrist movement and the vocalizer arm movement condition. On the right-hand side, data are parsed for each tempo and listener movement conditions. The clear peaked relative phase distributions for the vocalizer arm motions indicating phase-coupling for this condition, but we also saw that listeners tended to anticipate vocalizer movement.

Figure 3. Distributions relative phase listener-vocalizer.

We statistically confirmed the phase-coupling results by computing the standard deviation of Φ per trial performed. If phase-coupling is more pronounced, lower $SD \Phi$ will be observed (i.e., less variable/more stable phase relations around the average relative phase). We assessed this using

nlme mixed regressions (again using participant as the sole random intercept without random slopes, as adding random slopes caused the model to fail to converge).

As compared to a model predicting the overall mean, entering vocalizer movement condition as a predictor for SD Φ led to increased fit of the model (change in $\chi^2 [1] = 30.62, p < .001$). Adding tempo condition as an additional predictor did not further improve prediction of relative-phase (change in $\chi^2 [1] = 1.34, p = 0.51$). Adding listener movement condition as a predictor for relative phase (next to vocalizer movement) also did not improve the previous model (change in $\chi^2 [1] = 0.74, p = 0.378$). Therefore, the resulting best-fit model—which included vocalizer movement condition as the sole fixed effect—revealed that vocalizer arm movement condition had a lower SD Φ of 51 degrees as compared to the vocalizer wrist movement condition, $b = -51.09, t[21] = -7.81, p < .001$. These findings support our hypothesis that listeners synchronized their movement phasing with phase information in the vocalizer acoustics.

3. Discussion

While preliminary (results require replication), the current exploratory study demonstrates that it is—in principle—possible to glean information about bodily movement from voice acoustics alone. We found that listeners demonstrated both frequency-coupling and phase-coupling of their own movements with that of a vocalizer who was moving at different tempos while producing a single vowel sound. As predicted based on the absence of acoustic effects (Pouw, Harrison, et al., 2018a, 2018b), the vocalizer wrist movements (as opposed to vocalizer arm movements) were not reliably detected by the listeners; no evidence was obtained for frequency- or phase-locking in the wrist movement vocalizations. Although it appears that there must be some information about bodily gestures in phonation acoustics, the current exploratory study is unable to determine how pervasive the couplings might be.

The current results suggest that we do not necessarily hear voicing as only voicing: Intriguingly, we can also detect within voicing the bodily states of the voicer on the basis of acoustic-body invariants. The current research therefore directly aligns with the ecological psychology of language (Fowler, 2010) and the acoustic perception of object geometry (Carello et al., 2005). Our findings may extend this research program with the idea that prosodic contrasts in speech are direct informational sources of bodily tensioned states (including hand gestures). The findings further align with research on other animals, who often modulate their vocal activity so as to appear larger (and more intimidating or appealing) in size (Hardus, Lameira, Van Schaik, & Wich, 2009).

References

- 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.
- Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science, 9*(3), 211–214.
- Carello, C., Wagman, J. B., & Turvey, M. T. (2005). Acoustic specification of object property. In J. D. Anderson & B. Fisher Anderson (Eds.), *Moving*, 79–104.
- Cravotta, A., Busà, M. G., & Prieto, P. (2018). Restraining and encouraging the use of hand gestures: Effects on speech. In *9th International Conference on Speech Prosody 2018* (pp. 206–210). ISCA. doi: 10.21437/SpeechProsody.2018-42
- Ejiri, K., & Masataka, N. (2001). Co-occurrences of preverbal vocal behavior and motor action in early infancy. *Developmental Science, 4*(1), 40–48.
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*(1), 3–28.
- Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 816–828.
- Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press.
- Hardus, M. E., Lameira, A. R., Van Schaik, C. P., & Wich, S. A. (2009). Tool use in wild orangutans modifies sound production: A functionally deceptive innovation? *Proceedings. Biological Sciences, 276*(1673), 3689–3694.
- Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication, 57*, 257–267.
- 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.
- Pouw, W., Paxton, A., Harrison, S., & Dixon, J. A. (under review). Social Resonance: Acoustic information about upper limb movement in voicing. <https://psyarxiv.com/ny39e>
- Pouw, W., Harrison, S. A., & Dixon, J. A. (2018a). The physical basis of gesture-speech synchrony: Exploratory study and pre-registration. <https://doi.org/10.31234/osf.io/9fzsv>

- Pouw, W., Harrison, S. J., & Dixon, J. A. (2018b). Gesture-speech physics: The biomechanical basis of gesture-speech synchrony. <https://doi.org/10.31234/osf.io/tgua4>
- Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review*, *12*(6), 969-992.
- Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209-232.

Quantifying gesture-speech synchrony

Wim Pouw^{1,2} and James A. Dixon¹

¹Center for the Ecological Study of Perception and Action, University of Connecticut

²Department of Psychology, Education, & Child Studies, Erasmus University Rotterdam

wimpouw@uconn.edu, james.dixon@uconn.edu

Abstract

Spontaneously occurring speech is often seamlessly accompanied by hand gestures. Detailed observations of video data suggest that speech and gesture are tightly synchronized in time, consistent with a dynamic interplay between body and mind. However, spontaneous gesture-speech synchrony has rarely been objectively quantified beyond analyses of video data, which do not allow for identification of kinematic properties of gestures. Consequently, the point in gesture which is held to couple with speech, the so-called moment of “maximum effort”, has been variably equated with the peak velocity, peak acceleration, peak deceleration, or the onset of the gesture. In the current exploratory report, we provide novel evidence from motion-tracking and acoustic data that peak velocity is closely aligned, and shortly leads, the peak pitch (F0) of speech.

1. Introduction

Humans across all known cultures tend to move their hands during speaking (Kendon, 2004), suggesting a fundamental connection between communicative vocalizations and hand movements (Iverson & Thelen, 1999). There is one fundamental aspect of gesture that is central to its functioning: gestures are performed in synchrony with speech. Without synchrony with speech, gestures would fail to unambiguously point to objects or portray them through depiction, and be meaningless as markers of semantic or emotional salience (Quine, 1968).

Although it is widely accepted that synchrony is fundamental to gesture’s functioning, fine-grained quantification of gesture-speech synchrony as it occurs spontaneously during speaking is currently lacking (see however Danner, Barbosa, Goldstein, 2018; Pouw & Dixon, 2018). There is abundant research showing that the moment of “maximum effort” within a gesture is closely timed with the prosodic contrasts made in speech, but such evidence has varying degrees of objectivity and generalizability. Specifically, the primary evidence is based either on: a) artificial data (e.g., gestures produced by the experimenter, e.g., Leonard & Cummins, 2010), b) pointing gestures that are produced in a repetitive way outside the context of fluid speech (e.g., Rochet-Capellan, Laboissiere, Galvan, & Schwartz, 2008), or c) analyses of video recordings that do not allow for quantification of kinematic properties of gesture production (Loehr, 2012). To be clear, such research has been crucial in the study of gesture-speech synchrony, but also solicits an important next research objective given the technological advancements in current day and age (e.g., Danner et al., 2018; Pouw, Trujillo, Dixon, in press): A fine grained quantification of the synchrony of spontaneous gesture kinematics relative to speech.

For example, the most promising evidence for gesture-speech synchrony relies on methodology involving experimenter judgments of the intensity of gestural hand movements, the “maximum effort” of a gesture (Loehr, 2012; Wagner, Malisz, & Kopp, 2014). The maximum effort is the supposed to be the moment at which there is an energetic peak in the gesture stroke. However, as Wagner and colleagues (2014) conclude, the concept of maximum effort is an ambiguous spatiotemporal marker of a gesture “[the maximum effort is studied] with varying degrees of measurement objectivity and with varying definitions of what counts as an observation of maximum effort. Most definitions evoke a kinesthetic quality of effort or *peak effort* (Kendon, 2004) correlated with abrupt changes in visible movement either as periods of movement acceleration or strokes (Kita, van Gijn, & van der Hulst, 1998), as sudden halts or *hits* (Shattuck-Hufnagel, Veilleux, & Renwick, 2007), or as maximal movement extensions in space called *apexes* (Leonard & Cummins, 2008)” (p. 221, original emphasis).

As such, there is a need for a more fine-grained quantification of spatio-temporal properties of gesture in the form of specific *measurable* energetic peaks (e.g., peak acceleration, peak velocity). Such energetic peaks may provide the much sought after “anchor point” in gesture, the property of gesture that supposedly couples to a property of speech, thus creating synchrony. In the current exploratory data report, we provide preliminary evidence for key objective anchor points to study gesture and speech synchrony *for fluid speech and spontaneous gestures*, and at the conference we will report on a larger scale replication of this study. This should provide a novel quantification of temporal coordination of spontaneous gesture and speech. In addition to fundamental insights about how speech and gesture arise, the applied importance of quantifying synchrony of gesture and speech is immediately evident for the field of psychopathology and speech pathology. Such fields have already attempted to relate measures of gesture-speech synchrony to the diagnosis of certain pathologies (e.g., De Marchena & Eigsti, 2010). Other immediate applications of reliable quantifications of synchrony could one day be found in education (Iani, Cutica, & Bucciarelli, 2017).

2. Current approach

Subjects in the current exploratory study ($N = 4$) retold the narrative of a cartoon they had just watched, a common gesture-elicitation method (McNeill, 2005), which yielded about 230 gesture events. We employed high resolution motion-tracking of the dominant hand (240 Hz) during narration (non-dominant hand was not used for gesturing). From the movement time series, we identified energetic peaks during each gesture event (peak velocity, peak acceleration, peak deceleration), providing an objective measurement of gesture kinematics. Gesture identification was performed using ELAN (Lausberg & Sloetjes, 2009) so as to categorize different gestures, and to define the onset of a gesture based on assistance of hand-movement time series (see method and Crasborn, Sloetjes, Auer, & Wittenburg, 2006). Similar to previous studies (e.g., Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2010), we further extracted pitch ($F0$) from acoustic data so as to identify peaks of pitch within relevant gesture-speech events, which we show is a reliable anchor point for gesture-speech synchrony. Gesture-speech synchrony was quantified by the difference (D) in milliseconds between peak pitch and the relevant gesture anchor points (e.g., gesture onset, peak velocity). In the current study, we focused on three major gesture types, namely beat, iconic, and narrative pointing gestures. This exploratory study will allow us to answer a host of classic questions that have not been quantitatively studied to the current extent, including: What reliable kinematic anchor point in a gesture event is most closely synchronized with peak pitch? How strong is the synchrony between gesture and speech? Do beat, iconic, and pointing gestures differ in gesture-speech synchrony? To what extent are there individual differences in gesture-speech synchrony? For this conference contribution, we will report results of a larger scale study with 50 participants.

3. Method & Results

Four male right-handed graduate students at the University of Connecticut participated in this study (ages = 30, 38, 23, 34). Two participants were native speakers of American English and two were native speakers of Spanish with high proficiency in spoken and written English. In total, we collected movement and speech data from about 15 minutes of narration. Note that this much narration is considerable relative to other comprehensive studies of temporal coordination of gesture and speech, which have been naturally time-constrained because of the time-intensiveness of video analytic annotation (e.g., Loehr, 2012)

3.1. Apparatus

Motion tracking. We used a Polhemus Liberty (Polhemus Corporation, Colchester, VT, USA) with a single motion-sensor collecting 3D position data at 240Hz (~0.13 mm spatial resolution). The motion sensor was attached to the top of the participant’s index finger (at the height of the fingernail). This allowed us to capture arm movements together with movements of the wrists and fingers. We recorded the motion of only one hand to simplify data collection and analysis.

Audio. Instead of using the noisier sound stream of the video camera, we obtained speech data by using a RT20 Audio Technica Cardioid microphone (44.1kHz) which suppresses surrounding noises including any unintended experimenter noise (e.g., coughs).

Motion & audio recording. We used C++ to simultaneously call and write audio and movement data. We modified a C++ script made publicly available by Michael Richardson (Richardson, n.d.) in which we included scripts to enable recording of sound from a microphone (using toolbox SFML for C++ <https://www.sfml-dev.org/>).

Camera. We videotaped participants using Sony Digital HD Camera HDR-XR5504 Recorder, sampling at 29.97 frames per second.

3.2. Procedure

Participants were first equipped with a glove for the dominant hand that allowed us to attach the motion sensor of the Polhemus Liberty via Velcro to the index finger. Then a full clip of Tweety and Sylvester “Canary road” was watched. This cartoon clip is often used in gesture research, which lasts about 350 seconds. Participants were informed beforehand that they would later retell the narrative to the experimenter. The glove was attached prior to watching the video so that the subject could get used to wearing it. After watching the clip, participants were asked to retell the narrative of the cartoon while holding their non-dominant hand in their pocket as the recording equipment was running. No instructions were provided about hand gesturing.

3.3. Data Preparation

Gesture annotation phase. In the annotation phase, the first author transcribed speech and identified gesture events. For the annotation phase, we loaded in the video data, audio data, as well as the time series of the motion tracking into ELAN (Crasborn et al., 2006). ELAN allows the user to visually present the movement time series along with the video data. As such, the emergence of gesture could be identified based on the actual movement data rather than the lower resolution method of identifying movement on the basis of changes in movement per video frame, which can be difficult. As introduced by Crasborn and colleagues (2006), this provides clear advantages over traditional gesture video analysis.

The procedure of marking a gesture in the current dataset was as follows. Gesture onset was identified by spotting a gesture in the video, categorizing it as either a beat, iconic, or pointing gesture (based on gesture categorization guidelines by McNeill, 2005). In cases where the gesture was not of a clear nature, it was categorized as “undefined”; we also categorized “abandoned” gestures that were not completed (see e.g., Kita, Alibali, & Chu, 2017). After having spotted a gesture, the experimenter would go back to beginning of the gesture event and seek out the onset of the gesture (first fluent change from static position), on the basis of the time series of the kinematic data (with the use of x and y axis, and velocity trace). The gesture event was marked as ending at the place where the gesture completed its main stroke, thus not including a possible post-stroke hold, and not including a retraction phase. Excluding these optional end-phases of gesture allowed us to ensure that our peak-finding functions do not pick out possible energetic peaks in the retraction phase (which is generally known not to coordinate meaningfully with speech).

Speech Pitch. We extracted pitch time series of the audio recording using PRAAT with default range suitable for males 75-500 Hz (Boersma, 2001). We matched the sampling rate of pitch with that of the motion tracker (1 sample per 4.16 milliseconds).

Speech content. For exploratory purposes, also using ELAN, speech was transcribed and lexical affiliates of iconic gesture were identified if possible, but not when gestures did not clearly refer to what was mentioned in text.

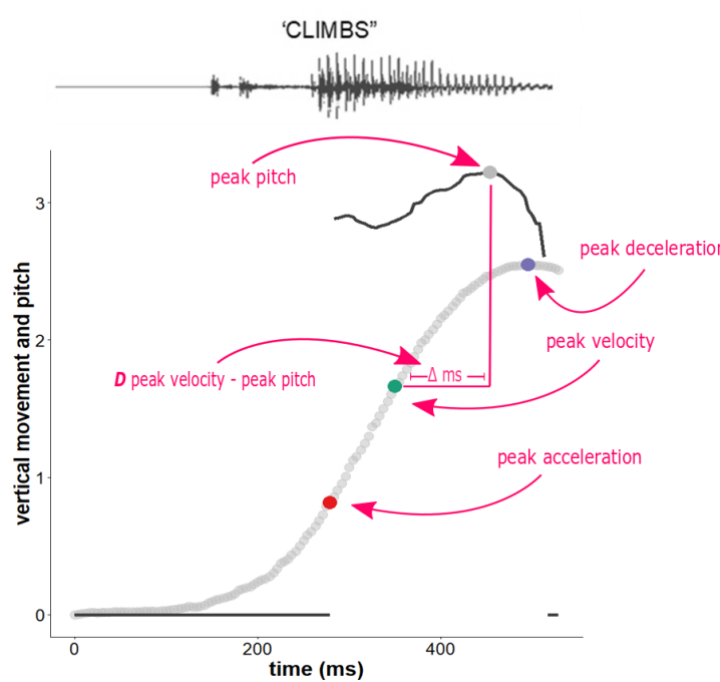
Data aggregation and analysis. We wrote a custom code in R (R core Team 2013) to aggregate the ELAN, PRAAT, and motion tracking data. We interpolated the movement data to match the pitch data with an interpolation function in R (code available on the Open Science Framework; <https://osf.io/5ja6y>). Using a custom-made function, we automatically read in ELAN gesture and speech annotation files so that these events were marked in the movement and pitch time series.

For each gesture event, the peak velocity, peak acceleration, and peak deceleration were extracted by a custom-written function in R. Since our peak-finding function could be sensitive to small but significant jumps in position data due to noise, we applied a low-pass Butterworth filter to the position velocity and accuracy traces with a cut-off of 10Hz (e.g., Leonard & Cummins, 20).

Data Availability. All (raw) data, pitch data (PRAAT), annotation data (ELAN), experiment code (C++), data preparation code (R), & analyses code (R) generated for this exploratory study are publicly available on the Open Science Framework (<https://osf.io/5ja6y>).

3.4. Descriptive Results

A total of 231 gesture events were observed (beat = 152, iconic = 44, pointing = 31, undefined/abandon = 4). Average time for gesture events was 829 ms ($SD = 602$ ms); beat gesture $M = 739$ ($SD = 398$), iconic gesture $M = 947$ ($SD = 789$), pointing gesture $M = 667$ ($SD = 443$). Table A (see here: <https://osf.io/3n79f/>) provides an overview of the production rates of the different gestures, as well as speech rate (spoken words per minute narration). It is important to note that these gesture ratios are very comparable to other studies that have used the same retelling of cartoon procedure (see e.g., McNeill, 2005, p. 42, where a comparable 41% of iconic gestures was found). This serves as evidence that in the current sample the glove and measuring apparatus did not seem to greatly alter spontaneous gesture tendencies.



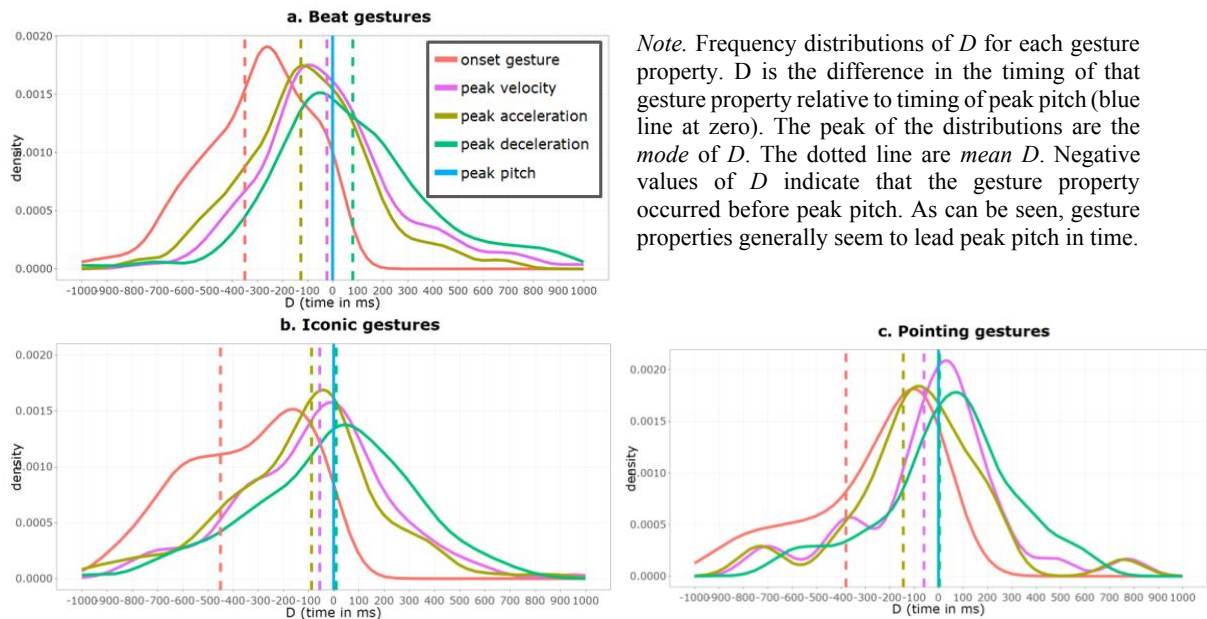
Note. Example of change y-axis position (grey) and pitch track (black) over time (ms; centered and scaled) for the “CLIMBS the wall” gesture. Red dot = peak acceleration, green dot = peak velocity, purple dot = peak deceleration, solid grey dot = peak pitch. These gesture properties were extracted using the custom-written function in R. Further note that speech starts around onset of gesture. We have super imposed the raw sound waveform in red above. The pitch (F0) reflects the vocal fold opening at pronouncing the “I” in “climbs”.

Figure 1. Visual example peak extraction method.

3.5. Gestures and Peak Pitch

We first assessed the temporal relation between speech (peak pitch) and properties of gesture. Table B (see here: <https://osf.io/c7qbm/>) shows the mean difference in milliseconds, D , between the peak pitch and the different kinematic properties of gesture - gesture onset, peak velocity, peak acceleration and peak deceleration, for each gesture type separately. Figure 2 shows the relative frequency distributions of D for these gesture properties relative to peak pitch (which defines the zero point).

A flat distribution curve of D would be an indication of a random occurrence of a kinematic property of gesture with regards to peak pitch. We obtain a clearly non-uniform distribution of D for beat, iconic, and pointing gestures, showing an impressive temporal coupling between gesture and speech prosody. Furthermore, the data show that gesture’s peak velocity, peak acceleration and gesture onset, all lead peak pitch in time (and is followed by peak deceleration). Gesture onset and peak acceleration are clearly not the point at which gestures synchronize with peak pitch. For each gesture property separately (i.e., onset, peak velocity, peak acceleration, peak deceleration), we performed a within-subjects ANOVA to assess differences in D between each gesture type (3 levels: beat vs. iconic vs. pointing gesture events; see Table B <https://osf.io/c7qbm/>). In the current sample, we did not find statistically significant differences between gesture types for D . This suggests that all the gestures types addressed here in this exploratory sample are roughly comparable in the degree to which they synchronize with peak pitch. The Bayesian Analyses further show that the observed data were 3 times or more likely under the null-hypothesis (absence of effect of gesture type) for gesture onset, peak velocity, peak acceleration. However, for peak deceleration we did not find substantial evidence for the null-model, suggesting that peak deceleration may differ in D between gesture types (when tested with larger samples).



Note. Frequency distributions of D for each gesture property. D is the difference in the timing of that gesture property relative to timing of peak pitch (blue line at zero). The peak of the distributions are the *mode* of D . The dotted line are *mean* D . Negative values of D indicate that the gesture property occurred before peak pitch. As can be seen, gesture properties generally seem to lead peak pitch in time.

Figure 2. Distribution of D 's: Gesture properties relative to peak pitch.

A further question that arises is whether there is one particular gesture property that is most closely coordinated with peak pitch in speech. Since we did not find reliable statistical differences in D between gesture types, we collapsed all beat, iconic, and pointing gesture events for the following analyses. With this combined data, we performed a within-subjects ANOVA with gesture property (peak velocity vs. peak acceleration vs. peak deceleration) as a within-subjects variable, and D as the dependent variable.

We found that these gesture properties differed reliably in their D 's, $F(2, 6) = 17.54, p < .001$. Paired post-hoc comparisons (p-values Bonferroni corrected) revealed that peak velocity shortly led peak pitch ($M_D = -39, SD_D = 454, 95\%CI[-90 : 11]$), as compared to peak deceleration which followed peak pitch ($p < .001; M_D = 44, SD_D = 424, 95\%CI[-3 : 92]$). Peak acceleration was furthest from peak pitch ($M_D = -113, SD_D = 494, 95\%CI[-168 : -58]$), and was statistically different from peak velocity and peak deceleration ($ps < .001$). As can be seen, both peak velocity and peak deceleration have 0 in their confidence intervals, suggesting that both closely synchronize with peak pitch, with peak velocity shortly leading (39 ms), and peak deceleration shortly following (44 ms) peak pitch.

4. Discussion

This exploratory study has provided the following preliminary implications with regards to classic questions in gesture research. These implications should be regarded as tentative.

Firstly, gesture-speech synchrony is obviously occurring, as indicated by clear peaks in the distributions of difference in timing (D) between peak pitch and kinematic gesture properties. This synchrony with speech is remarkable given that beat, iconic, and pointing gestures each serve different functions. The current results suggest that regardless of their role in discourse, all gestures tend to emerge in synchrony with speech. However, it is clear from the relatively large standard deviations of D that gesture-speech synchrony is not a 1-1 coupling, suggesting a more loose temporal relation between gesture and speech [Loehr, 2012; McClave, 1994].

Secondly, we have disambiguated gesture's anchor point with speech, by objectively assessing which energetic peak in manual movement most closely aligns with energetic peak pitch. Most clearly, gesture onset, and peak acceleration are not most closely synchronized with peak pitch. For all gestures, peak velocity is closely synchronized with peak pitch (gestures lead speech with 39 milliseconds), but most notably for beat gestures. For iconic and pointing gestures peak deceleration could also be a good anchor point for studying gesture-speech synchronization.

Future research is needed to ensure that our findings can be reproduced in a more comprehensive experiment. Also, we may wonder whether findings can be reproduced when movement of the non-dominant hand can be made as well.

References

- Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glott International* 5 (9/10), 341-345.
- Chu, M., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, 143(4), 1726-1741.
- Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics*, 71, 268–283.
- Crasborn, O., Sloetjes, H., Auer, E., & Wittenburg, P. (2006). Combining video and numeric data in the analysis of sign languages with the ELAN annotation software. In C. Vettori (Ed.), *Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios* (pp. 82-87). Paris: ELRA.
- de Marchena, A., & Eigsti, I. M. (2010). Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency. *Autism Research*, 3(6), 311-322.
- 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(3), 850-864.
- Iani, F., Cutica, I., & Bucciarelli, M. (2017). Timing of gestures: Gestures anticipating or simultaneous with speech as indexes of text comprehension in children and adults. *Cognitive Science*, 41(6), 1549-1566.
- Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, 6(11-12), 19-40.
- Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, 124(3), 245.
- 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*, Proceedings International Gesture Workshop Bielefeld, Germany, September 17-19, 1997.
- Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology*, 8(1), 1-36.
- Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers*, 41(3), 841-849.
- Leonard, T., Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471.
- Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1), 71-89.
- McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, 23(1), 45-66.
- McNeill, D. *Gesture and Thought*. Chicago: University of Chicago press, 2005.
- Quine, W. V. O. (1968). Ontological relativity. *Journal of Philosophy*, 65, 185–212.
- Rochet-Capellan, A., Laboissiere, R., Galvan, A., Schwartz, J. (2008). The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, 51(6), 1507–1521.
- Richardson, M. (n.d.). Retrieved from <http://xkiwilabs.com/software-toolboxes/>
- Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, 57, 283-300.
- Pouw, W., Trujillo, J., & Dixon, J. A. (in press). The quantification of gesture-speech synchrony: A tutorial and validation of multi-modal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*. <https://doi.org/10.31234/osf.io/jm3hk>
- Wagner, P., Malisz, Z., & Kopp, S (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209-232.

Embodied reciprocity in conversational argumentation: Soliciting and giving reasons with Palm Up Open Hand gestures

Nora Schönfelder and Vivien Heller

University of Wuppertal, Germany

schoenfelder@uni-wuppertal.de, vheller@uni-wuppertal.de

Abstract

Based on video-recorded peer interactions, this paper discusses the interactive functions of Palm Up Open Hand gestures in conversational argumentation. Drawing on conversation analysis, we demonstrate that PUOH gestures occur in sequential positions where new perspectives come up for discussion and divergent positions are established; they are thus resources for soliciting or giving reasons. It is argued that as publicly visible resources reciprocal PUOH gestures facilitate the orderly production of contiguous responses and ensure coherence between turns.

1. Introduction

This paper deals with “palm up” gestures, a family of *pragmatic* (Kendon, 2004, Streeck, 2007, Müller, 2004) or *interactive* gestures (Bavelas et al., 1995). According to Streeck, pragmatic gestures enact or embody communicative functions. For instance, they can display what an utterance is designed to do, embody the speaker’s stance towards the utterance or punctuate spoken discourse. Pragmatic gestures can be coupled with interaction units of different scopes, ranging from “turns, turn-construction units, speech acts, and speech act sequences” (Streeck, 2009, p. 179).

In a “Palm Up Open Hand” gesture (Müller, 2004; henceforth referred to as PUOH), the speaker presents the hand palm upwards in the shared space of perception; in this way, s/he enacts the physical act of giving, handing over or offering an object (Streeck, 2007, Müller, 2004). The meaning of these acts is evoked metonymically: the open hand presents an “abstract discursive object” (Müller, 2004, p. 233), for instance an opinion or a proposal, offers it for joint inspection, and invites the co-participants to position themselves towards the perspective offered in the speaker’s open hand. In this way, the discursive object receives a “transitional status” and “cannot be used for anything until the transaction of giving or receiving has been completed” (Kendon, 2004, p. 274). Aside from the presenting and offering function, PUOH gestures can also be used for soliciting responses. Streeck (2007) points out that the way in which the gesture modulates the verbal utterance changes depending on its duration, shifting from offering to expecting to receiving something. This means that the longer the open hand is held at the end of the turn, the stronger the obligation upon the recipient to respond becomes. One function of PUOH as a publicly visible resource is thus to display or amplify a conditional relevance. With this term, Schegloff (1968) refers to the relationship between adjacent turns. Particular sequence-initiating actions, for instance a summons or invitation, impose a normative obligation on the co-participant to perform a type-fitted response. PUOHs are one of the resources that serve to establish a conditional relevance. Past research has exclusively focussed on individual palm up gestures. In our data of conversational argumentations, PUOHs are not only employed in turns that establish an obligation to provide reasons (Quasthoff, Heller & Morek 2017), they are also used in subsequent turns to demonstrate that the conditional relevance has been fulfilled. Taking such reciprocal palm up gestures as our focus, we describe the uses and interactive functions of series of PUOHs in conversational argumentations.

2. Data and analytical approach

2.1. Data

The analysis is based on a corpus of 14 videotaped group discussions (46:19 minutes) of children aged between 7;0 and 13;6. The corpus includes 19 instances of reciprocal PUOHs, i.e. sequences in which at least two subsequent PUOHs were produced by different speakers. Groups of three to five children were asked to deal with a fictitious problem involving a shipwreck. The task was to arrive at a jointly agreed decision on three essential survival items. Since the scenario allowed for a variety of solutions, it was likely that the children's positions would diverge.

The data was transcribed in accordance with GAT 2 conventions (Selting et al., 2011); they include final pitch movements of intonation phrases, the focus accents (noted in capital letters) and multimodal phenomena. To represent PUOH gestures and gaze, still pictures were embedded into the transcripts.

2.2. Analytical Approach

Drawing on multimodal conversation analysis, we analysed reciprocal PUOHs with regard to their sequential embedding in conversational argumentations and their precise position in the emerging turn. Conversational argumentation is approached as a discursive practice with a complex sequential organisation that involves certain conversational “jobs”, i.e. “constituting dissent/ problematizing”, “establishing an obligation to provide reasons”, “providing and challenging reasons”, “closing” and “transition” (Quasthoff, Heller & Morek, 2017, pp. 97-101). Conversational argumentation can either be framed as controversial/persuasive or as consensual/collaborative reasoning (Ehlich, 2014). We therefore examined reciprocal PUOHs in both types of contexts. Regardless of the contextual framing, reciprocal PUOH gestures occur in the jobs *establishing an obligation to provide reasons* and *providing and challenging reasons*. In the following, we present two examples of reciprocal palm up gestures produced when participants provided and challenged reasons.

3. Results

3.1. Reciprocal PUOH gestures in a controversial context

Extract (1) shows Damira and Sila in a controversial moment of their discussion. The participants have already constituted a dissent and established an obligation to provide reasons. The extract starts with Damira pointing to the illustrations of the first aid kit and the mobile phone on the handout, thereby taking up the positions of previous speakers and formulating an interim conclusion (l. 71). She then establishes a fictitious scenario (l. 72) that results in another choice: a knife or matches (l. 76). A series of three PUOH gestures now occurs: the first gesture is produced by Damira when she offers her reasoned position, the second by Sila, who challenges the position, and the third again by Damira when she substantiates her claim. Note that some of the PUOHs contain two or three gesture phrases (Kendon, 2004).

Extract 1: DAM - Damira, SIL - Sila

```





071  DAM  |DAS hier| und          |das hier          | auf jeden fall;=
        |this one and        |this one          | for sure
        |((points at first aid kit))| |((points at phone))|
072  =aber (-) wenn wenn |jetzt jemand KOMMT? | =zum BEispiel,
        but if if       |someone comes   | for example
        |((PUOH + head shake))|

```

```

((...))
076  dann braucht man entweder |das hier oder    |
        then you'll need either |this or         |
        |((points at knife))|
        |<<creaky> DAS;>      |
        |this                  |
        |((points at matches, transforms hand into PUOH))|

```


- 077 SIL |<<doubting> n **MES**se:r,>|
 a knife
 |((PUOH on handout)) |
- 
- 078 |damit man den Tötet (-), |,
 to kill him
 |((PUOH tow. Dam, then downwards))|
- 
- 079 SIL |↓oder WAS;|
 or what
 |((PUOH)) |
- 
- 080 [((holds PUOH))]
 081 DAM [((palm up oh + head shake))]
- 
- 082 damit man keine ANGST bekommt <<p> vielleicht?>
 so you don't get scared maybe

The first PUOH occurs within an argumentative multi-unit turn produced by Damira. It is part of a gesture unit comprising three gesture phrases: a PUOH followed by a deictic gesture that is again transformed into a PUOH. While formulating a hypothetical condition of her fictitious scenario (“if someone comes”), Damira performs a PUOH. In this way, she is indicating to her co-participants that a new idea is being offered. The headshake that conveys an uncertain epistemic stance modulates her position and contextualises it as a proposal. As a consequence of the hypothetical condition, she formulates two new items by pointing to the knife and the matches (l. 76). During the turn-completion, the deictic gesture is transformed into a brief PUOH. At this turn position, the gesture not only hands over the turn to the co-participants, but also invites them to inspect the proposal.

In response to this invitation, Sila produces another argumentative multi-unit turn in which she challenges Damira’s proposal. Within her turn, a reciprocal PUOH gesture with three gesture phrases is performed. Gazing at Damira, Sila brings her open hand towards the illustration of the knife, holds it there and repeats “knife” with an intonation that conveys doubt (l. 77). At the same time, the hand, together with the intonation, also shows that the knife has received a transitional status: it is ‘put on the table’ for further negotiation. The gesture thus contributes to disambiguating reference to a particular object as well as establishing its disputability; as a visible resource, it provides an observable cue to the new epistemic order (Heller, 2018). Sila continues her turn by animating a potential argument in favour of the knife (l. 78: “to kill him?”). The facial shrug and the rising final pitch movement modulate the meaning of the verbal utterance and contextualise it as a rhetorical question that challenges Damira’s proposal. Temporally aligned with the rhetorical question, Sila initiates the second gesture phrase: she lifts the open hand and moves it towards her addressee, who now also gazes at her. She then lowers her hand and brings it to a hold next to the



paper. Temporally aligned with this third gesture phrase is a colloquial tag question (l. 79: “or what;”) which establishes a conditional relevance for a response. In this turn position, the third gesture phrase of the PUOH becomes a “gesture of waiting – a hand waiting, as it were, to be filled with a response” (Streeck 2009, l. 175). It displays the unfulfilled conditional relevance of the turn and solicits a contiguous response. A third reciprocal PUOH gesture occurs in the forefield of Damira’s next turn. Still gazing at each other, both Sila and Damira hold their PUOHs above the table. While Sila’s hand solicits an answer, Damira projects that she is about to deal with the challenge, which is done in the next turn (l. 82: “so you don’t get scared maybe?”). This *pas de deux* of PUOHs embodies reciprocity in dealing with divergent perspectives.

In summary, three reciprocal PUOHs have been performed by different speakers in subsequent multi-unit turns. The first gesture embodied the communicative function of *offering a reasoned position*. At the same time, it served as a *handover and invitation to inspect the proposal*. The reciprocal PUOH by the next speaker embodied the communicative function of *challenging the position*. The final component of this gesture phrase, initiated at turn-completion, selected a next speaker and reinforced the conditional relevance to deal with the challenge. The third reciprocal PUOH was again produced by the first speaker in advance of the next turn, at a moment when the last speaker’s PUOH was still visible. In this position, it *projected the fulfilment of the conditional relevance*, i.e. the production of another argument. In the controversial sequence analysed here, the series of reciprocal PUOHs emerges, due to the fact that a proposal – accompanied by a gesture – is challenged in the next turn. This raises the question as to whether series of reciprocal gestures also occur in consensual contexts in which a challenge is absent.

3.2. Reciprocal PUOH gestures in a consensual context

The second extract shows a moment of consensual reasoning. The speakers have already discussed different proposals. Now, Zaim provides a list of three options (ll. 74-76). The following extract shows the subsequent negotiation of the item ‘tent’.

Extract 2: ZAI – Zaim, CEN – Cennet

- 074 ZAI | =entweder **MES**ser, |
 either knife
 | ((lh: palm up open hand)) |
- 
- 075 | **ZELT**, |
 tent
 | ((rh: deictic PUOH)) |
- 
- 076 | **STREICH**holz; |
 | ((rh: deictic PUOH)) |
 match
- 077 (3.0)
- 078 CEN °°h (-) tja | den **ZELT** lassen wir, = |
 well we keep the tent
 | ((lh: deictic PUOH)) |

one of the proposals within a multi-unit turn (l. 78); the third gesture is produced in addition to the agreement, but now embodies the elaboration of the reason (l. 83). This extract shows that reciprocal PUOHs are also produced in consensual contexts. Unlike the first extract, they co-occur within sequences of co-constructive turns. Both transcripts support the argument that global semantic coherence between turns is facilitated by PUOHs: they embody participants' reciprocity in dealing with an abstract discursive object in both controversial and consensual contexts.

4. Discussion

Embodied reciprocity is not only an interesting phenomenon from the analyst's point of view, but is, first and foremost, a matter for the participants themselves. This is especially the case when participants need to negotiate how divergent positions are to be dealt with. In such situations, the collaborative continuation of talk is potentially at risk. In such conversational environments, PUOHs fulfil important functions with regard to the participants' "working consensus" (Goffman, 1959, pp. 9-10) on the purpose and structure of the activity in progress. As publicly visible resources, reciprocal PUOHs both enact and embody the give-and-take of arguments and the constantly changing epistemic order. A comparison of controversial and consensual argumentative contexts revealed that reciprocal PUOHs were employed as long as the disputability of a position needed to be established and negotiated. Once the participants achieved a consensus, no more reciprocal PUOHs could be observed (extract 2, l. 80). Our analysis shows that reciprocal PUOHs facilitated the orderly production of contiguous responses and ensured coherence between turns.

The present study was based on child interactions. Recent research on PUOHs (Müller, 2004, Streeck, 2007) shows that these gestures also occur in adult conversational argumentation. Whether series of reciprocal PUOHs exist in adult interaction remains the topic for future research.

An interesting question is the emergence of embodied reciprocity in ontogenetic development. Previous studies on PUOHs in narratives (Graziano, 2014) and explanations (Alamillo, Colletta, & Guidetti 2012) indicate that pragmatic gestures are used as early as the age of four; yet the variety of communicative functions, especially their modal use, seems to develop in parallel to other linguistic resources and interactive competences. Future studies should further investigate the development of embodied reciprocity as one (key) component of discourse competence.

References

- Alamillo, A., Colletta, J., & Guidetti, M. (2012). Gesture and language in narratives and explanations. The effects of age and communicative activity on late multimodal discourse development. *Journal of Child Language* 40(3), 511-538.
- Bavelas, J.B., Chovil, N., Lawrie, D.A., & Wade, A. (1992). Interactive gestures. *Discourse Processes* 15(4), 469-489.
- Ehlich, K. (2014): Argumentieren als sprachliche Ressource diskursiven Lernens. In A. Hornung, G. Carobbio, D. Sorrentino, (Eds.): *Diskursive und textuelle Strukturen in der Hochschuldidaktik. Deutsch und Italienisch im Vergleich* (pp. 41-54). Münster: Waxmann.
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Doubleday Anchor.
- Graziano, M. (2014). The development of two pragmatic gestures of the so-called Open Hand Supine family in Italian children. In M. Seyfeddinipur, & M. Gullberg (Eds.), *From Gesture in Conversation to Visible Action as Utterance. Essays in honor of Adam Kendon* (pp. 311-328). Amsterdam, Philadelphia: John Benjamins Publishing.
- Heller, V. (2018). Embodying epistemic responsibility. The interplay of gaze and stance-taking in children's collaborative reasoning. *Research on Children and Social Interaction* 2(2), 262-285.
- Holler, J. (2010). Speaker's Use of Interactive Gestures as Markers of Common Ground. In S. Kopp, & I. Wachsmuth (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction* (pp. 11-22). Berlin: Springer.
- Kendon, A. (2004). *Gesture. Visible action as utterance*. Cambridge, New York: Cambridge University Press.
- Müller, C. (2004). Forms and uses of the Palm Up Open Hand: A case of a gesture family? In C. Müller, & R. Posner (Eds.), *The semantics and pragmatics of everyday gestures* (pp. 233-256). Berlin: Weidler.
- Quasthoff, U., Heller, V., & Morek, M. (2017). On the sequential organization and genre-orientation of discourse units in interaction. An analytic framework. *Discourse Studies* 19(1), 84-110.
- Selting, M. et al. (2011). A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung* 12, 1-51.
- Streeck, J. (2007). Geste und verstreichende Zeit: Innehalten und Bedeutungswandel der „bietenden Hand“. In H. Hausendorf (Ed.), *Gespräch als Prozess. Linguistische Aspekte der Zeitlichkeit verbaler Interaktion* (pp. 157-177). Tübingen: Gunter Narr Verlag.
- Streeck, J. (2009). Forward-Gesturing. *Discourse Processes* 46(2-3), 161-179.

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?

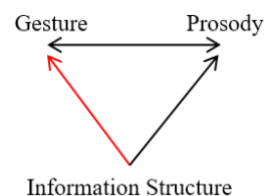
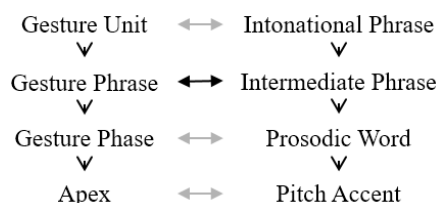


Figure 1. Mapping of gestural and prosodic hierarchies. Figure 2. Three-way coordination in production.

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).

	0:04:10.500	0:04:11.000	0:04:11.500	0:04:12.000	0:04:12.500	0:04:13.000	0:04:13.500	0:04:14.000	0:04:14.500	
Transcript	şu taraftaki bir kapakta da yeşil üzerine beyaz artı var									
Translation	there is a white plus on green (surface) on a door over there									
Word [123]	over	there	a	door	an	green	on	white	plus	there is
Intermediate	Intermediate Phrase				Intermediate Phrase			Intermediate Phrase		
Intonational	Intonational Phrase									
Topic/Focus	Topic				Focus			Background		
G-Phrase [39]					Iconic					
G-Phase [113]				Preparation		Stroke		Post-Hold	Retraction	

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 Bozşahin, 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 Bozşahin, 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.

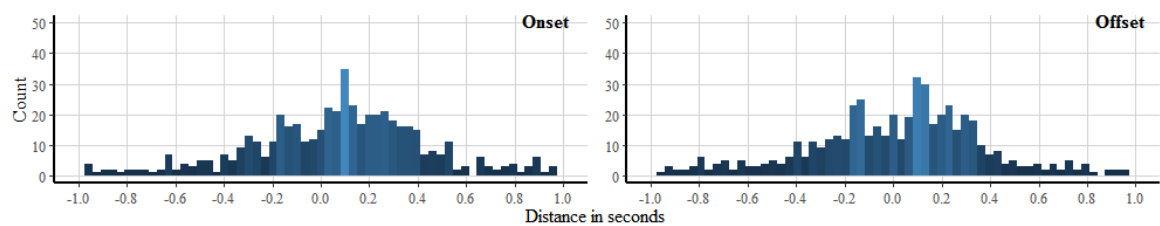


Figure 4. The coordination of ip onsets/offsets with GPs onset/offsets.

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.

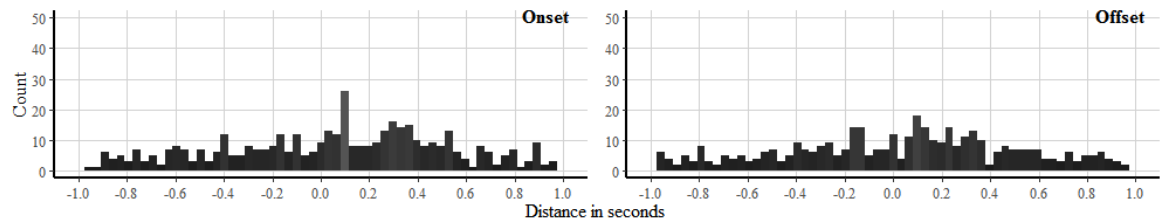


Figure 5. The coordination of IP onsets/offsets with GPs onset/offsets.

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.

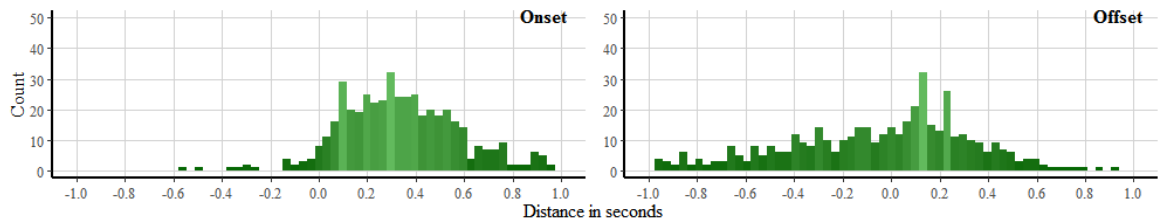


Figure 6. The coordination of topics/foci onsets/offsets with GPs onset/offsets.

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 ($m=479\text{ms}$). For offsets, although the standard deviation was high, there was a negative mean ($m=-196\text{ms}$) 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).

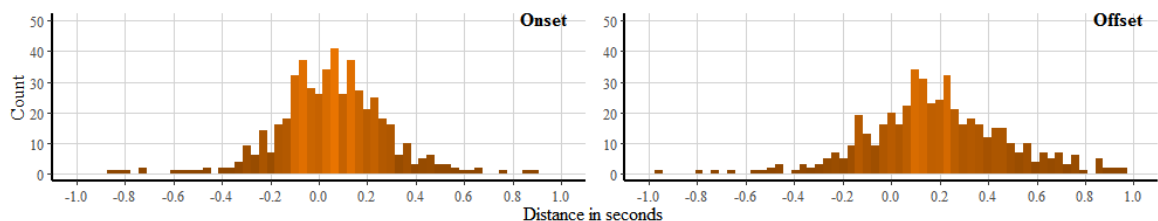


Figure 7. The coordination of topics/foci onsets/offsets with apical area onset/offsets.

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 ($m=60\text{ms}$, $sd=551\text{ms}$). For the offsets, the leftward spread disappeared and the peak appeared more compact ($m=196\text{ms}$, $sd=406\text{ms}$). The equivalence test results were significant for the onsets, showing that the distances were between the set bounds and not statistically different from zero ($t_{\text{Upper}}(504)=-4.01$, $p < .001$; $t_{\text{Lower}}(504)=9.05$, $p < .001$). However, the results were non-significant for the offsets as the confidence interval crossed the upper bound by approximately 60 ms, ($t_{\text{Upper}}(504)=1.98$, $p = 0.976$; $t_{\text{Lower}}(504)=19.7$, $p < .001$). 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 ($t_{\text{Upper}}(504)=-8.87$, $p < .001$; $t_{\text{Lower}}(504)=10.8$, $p < .001$) 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., Jarmolowicz-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.

Hand gestures and pitch contours and their distribution at possible speaker change locations: a first investigation

Margaret Zellers¹, Jan Gorisch², David House³, and Benno Peters¹

¹Institute for Scandinavian Studies, Frisian Studies, and General Linguistics, University of Kiel, Germany; ²Leibniz-Institute for the German Language, Mannheim, Germany;
³KTH Speech, Music & Hearing, Stockholm, Sweden
mzellers@isfas.uni-kiel.de, gorisch@ids-mannheim.de,
davidh@speech.kth.se, peters@ipds.uni-kiel.de

Abstract

Smooth turn-taking in conversation depends in part on speakers being able to communicate their intention to hold or cede the floor. Both prosodic and gestural cues have been shown to be used in this context. We investigate the interplay of pitch movements and hand gestures at locations at which speaker change becomes relevant, comparing their use in German and Swedish. We find that there are some shared functions of prosody and gesture with regard to turn-taking in the two languages, but that these shared functions appear to be mediated by the different phonological demands on pitch in the two languages.

1. Introduction

Everyday conversation, the fundamental context in which spoken language is used, has been demonstrated to have consistent structural features to which conversational participants orient, in particular with regard to turn-taking. Sacks, Schegloff, & Jefferson (1974) report that at Transition Relevance Places—i.e. locations where speaker change may become relevant—new speakers have priority to take up a turn, with the current speaker only continuing if a new speaker does not take the floor. However, in many cases a single chunk of speech may be insufficient for the current speaker to achieve an interactive goal, for example, telling a story, meaning that the current speaker must have a means of holding the floor if s/he is to be able to achieve this goal. Similarly, the speaker may also wish to invite input from an interlocutor, or even to directly cede the floor at a certain point. While in some cases a listener may be able to predict the upcoming conclusion of a current speaker's communicative project, this is by no means the case in every circumstance. Thus, the current speaker must have ways of communicating her/his intention to hold or cede the floor to an interlocutor.

A variety of communicative means have been proposed by which floor-holding and floor-ceding can be achieved in conversation. These can be broadly grouped into the categories of linguistic, phonetic, and gestural means. By linguistic means, we primarily refer to syntactic or semantic completion of an utterance in context. Phonetic means may include such features as pitch variation (choice of contour or size of pitch movements), amplitude variation, and speech rate variation. Gestural means can include body movements of any type, such as those of the eyes, eyebrows, head, and/or hands. The role and interplay of linguistic and phonetic cues at turn boundaries have been widely investigated, suggesting that syntactic/semantic completion is a strong cue to speaker change, while pitch, phonation quality, and duration variation can also contribute as turn-taking cues (Schaffer, 1983; Auer, 1996; Local, Kelly, & Wells, 1986; Koiso, Horiuchi, Tutiya, Ichikawa, & Den, 1998; Gravano & Hirschberg, 2009, 2011; Kane, Yanushevskaya, de Looze, Vaughan, & Ní Chasaide, 2014; Heldner & Włodarczak, 2015; Zellers, 2017, *inter alia*). Similarly, a variety of gestural cues have been shown to impact turn-taking, including gaze direction (Edlund & Beskow, 2007, 2009) and hand movements (Streeck & Hartge, 1992; Mondada, 2007; Sikveland & Ogden, 2012).

Since some aspects of turn-taking signalling involve the linguistic system, it is particularly interesting to make comparisons across languages which show relevant structural differences. In the case of pitch movements, for example, German and Swedish differ in terms of the functional load: German is an intonation language, in which pitch contours bear pragmatic meanings, while in

(most varieties of) Swedish, a lexical pitch accent contrast is also signalled by pitch. Thus we might expect that the availability of pitch/fundamental frequency (f_0) for providing information relevant to turn-taking might be different in these two languages; and indeed, previous studies have given some evidence for language-specific differences (Peters, 2006; Zellers, 2014; Bergmann, 2018), and also indicating that rising contours are relatively infrequent in Swedish (House, 2005).

The larger goal of our research project is to identify points of interaction between prosody and gesture in conversational settings. In the current work, we operationalize prosody as f_0 contours, and gesture as hand gestures, and investigate their relevance in the particular conversational function of turn-taking. In particular, our research questions are as follows:

- Do pitch and hand gestures carry out the same functions with regard to turn-taking?
- Does the relationship between pitch and hand gesture at potential turn boundaries differ in German and Swedish?

2. Data sets, annotation and analysis

2.1. Data sets

In making a cross-linguistic comparison, it would be ideal to have corpora from each language which were collected and annotated using the same methodology. In the current case, comparable recorded data in the two target languages are not available. The selection of the German data was made primarily on the basis of its similarity to the Swedish data; however, as will be seen, the similarity between the datasets is not always straightforward.

2.1.1. Swedish data: Spontal

The Swedish data in our study are taken from the Spontal corpus (Edlund et al., 2010). The Spontal corpus consists of two-party conversations recorded in a laboratory setting. Participants sat facing each other and were each filmed with a video camera directly facing them which captured their body from approximately the lap upwards. Audio recordings were made using both head-mounted microphones and a set of more distant microphones. Additionally, participants wore a set of motion-capture markers, with the goal of being able to automatically process data about their body movements. Some participant pairs knew each other prior to the data collection, while others were strangers.

Our data set consists of five five-minute chunks of conversations from Spontal (portions of 09-06, 09-20, 09-22, 09-28, and 09-36), comprising ten participants in total (8 male, 2 female). We used only the video and audio data, since no motion capture data was available for German.

2.1.2. German data: FOLK

The German data in our study are taken from the FOLK corpus (Schmidt, 2014). FOLK consists of a wide variety of spontaneous and semi-spontaneous speech situations, ranging from informal conversation to televised political debates. Most recordings were made with one video camera and microphone, though there is considerable variation. For the current study, we have excerpted three seven-minute chunks of two-party interactions.

In order to be as similar as possible to the Spontal data, we were particularly interested in two-party, relatively spontaneous interactions in which the participants were face-to-face, and their hand movements were easily visible. Our final selection thus includes two cases of mock job interviews (where a candidate interacts with the interviewer and then receives feedback) and one interview with a specialist on birds-of-prey (portions of FOLK_E_00173, 00174, and 00261). Since the interviewer is the same in both mock interviews, and we do not annotate the interviewer in the birds-of-prey interview since she is holding a microphone the whole time, these data comprise four participants (all male). While FOLK contains some more informal conversational settings, we determined that these were inappropriate for our goals since they either involved more than two participants, or else gesture annotation would have been impossible due to the recording conditions (i.e. hands were not visible or participants were carrying out some task with their hands).

2.1.3. Comparability

While we have endeavoured to use data that is as comparable as possible from the two languages, the conversational settings involved in the Swedish and German data are substantially different, as are the quality of the recordings. Without collecting entirely new, identically structured data, this

will always be a problem. Thus, our cross-linguistic comparisons are also mediated by the differences in interactional setting.

2.2. Annotation

An annotation scheme to address our research questions must minimally meet the following criteria:

- Allow for parallel analysis of prosody and gesture at turn boundaries
- Successful in conversational speech
- Applicable to corpora with different content/structures
- Involve enriched gestural annotations so function can be taken into account

Thus, for the current study, we have annotated the following fields:

- Turn-taking: syntactic/semantic completeness in context; type of turn transition
- Phonetics: final f₀ contour (span measured in semitones)
- Gesture: presence of gesture within 1 sec of offset of speech; gesture phase at offset of speech (following Kendon, 2004)

Syntactic/semantic completion was annotated with reference to the orthographic transcription; that is, without taking into account prosodic features which might additionally signal completion or incompleteness. The phonetic annotations were carried out in Praat (Boersma & Weenink, 2018) without access to the video signal. Conversely, gesture locations and phases were annotated using ELAN (Version 5.4, Max Planck Institute, 2019) without access to the audio signal.

2.2.1. Transition types

The crucial locations for our data were places in conversation where speaker change could become relevant. These locations were defined using two criteria: first, the presence of a silent pause, and second, the potential syntactic/semantic completion in context of the lexical material at that location. If both criteria were met, a location was given a label defining the turn-taking behaviour at that point:

- *Change*: the current speaker produces a complete full turn in declarative form, and then the next speaker launches a full turn
- *Keep*: the current speaker produces a complete full turn in declarative form, and then the same speaker launches an additional full turn
- *Backchannel*: the current speaker produces a complete full turn in declarative form, the other speaker produces a short response token (e.g. ja, mhm), and then the first speaker launches an additional full turn
- *Question*: the current speaker produces a complete full turn with lexical/syntactic interrogative form, and then the next speaker launches a full turn

Unclear cases, including cases where turns ended in tag questions, were not used in the analysis. Locations in which the speakers produced full turns in overlap were also discarded, since if a next speaker launches a turn before the offset of the previous turn, they by definition cannot have been orienting to features occurring at the offset of the prior turn.

2.3. Data extraction and statistical analysis

ELAN annotation files were converted to Praat TextGrids and merged with the phonetic annotations. The data were then automatically extracted using scripts. Substantial difficulty arose during the f₀ extraction, since many speakers in both languages were very creaky or used whisper near the ends of their turns. Time did not allow for a manual correction of all missing f₀ values, but the values that were used were manually checked and a few octave errors were hand-corrected.

2.4. Results

A total of 212 transition locations were identified in the German data, and 286 in the Swedish data. Of these, 98 in the German data and 102 in the Swedish data had hand gestures occurring in the vicinity of the speech offset. However, in only 73 and 38 (respectively) of the locations with hand gesture were f₀ measurements possible. Thus we must be cautious in the interpretation of the f₀ data.

2.4.1. Gesture phases at speech offset

For the 98 (German) and 102 (Swedish) cases where hand gesture occurred in the vicinity of the offset of speech, the distribution of gesture phases is shown in Figure 1. Ongoing gestures of all kinds at the offset of speech were much more frequent at Backchannel and Keep locations than at Changes and Questions in both languages. In terms of gesture phases, gesture strokes co-occurring with the offset of speech only occur at Keep and Backchannel locations, while the other gesture phases can occur at Backchannels, Changes, and Keeps. There is not enough data to draw any conclusions about possible gesture phases at Questions.

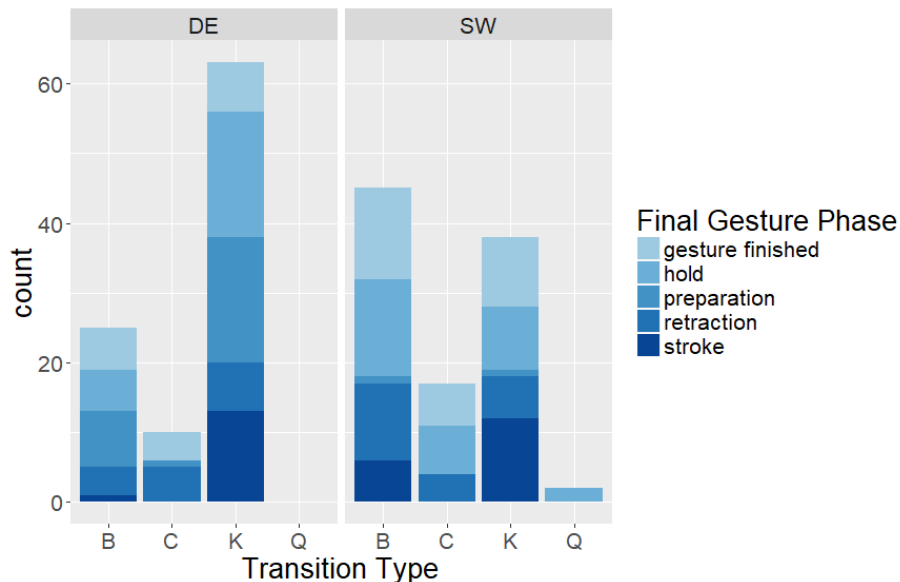


Figure 1. Gesture phase at offset of speech in German (DE) and Swedish (SW) at different transition types (B = Backchannel, C = Change, K = Keep, Q = Question).

2.4.2. Pitch movement and gesture at speech offset

For the transition locations where both pitch measurements and gestures were available, the final f_0 contours were classified as representing either rising, falling, or level pitch, with level contours comprising those which changed less than 1 semitone in either direction between the two measurement points. A comparison of pitch contours at transition locations with and without accompanying gesture is shown in Figure 2. As expected, rising contours are more frequent in German than in Swedish; a chi-square test confirms this distributional difference ($\chi^2(2) = 18.393$, $p < .001$). Rising contours also appear to be more frequent when there is no accompanying hand gesture than when there is an accompanying hand gesture; however, this trend could not be confirmed by a chi-square test, possibly due to the unbalanced data set ($\chi^2(2) = 0.770$, $p = .68$).

3. Discussion

The results presented here reflect a relatively small dataset, and should additionally be interpreted with caution due to the different interactional settings in the two languages tested. However, the results are still suggestive of some patterns of turn-taking signalling in the two languages, and the interplay between pitch and gestural cues in particular.

In both languages, we have observed a much higher proportion of gestures occurring at the offset of speech in cases where the current speaker takes up an additional full turn following the initial turn (i.e. in Keeps and Backchannels) compared to cases where the other speaker takes up the next full turn (i.e. Changes); cf. Figure 1. This is consistent with the report from Duncan (1972) that ending or relaxing a hand gesture is treated as a turn-yielding cue, while ongoing gestures are (speech-)attempt-suppressing cues. In our data, gestures in their stroke phase (i.e. the meaningful portion of the gesture) can apparently only accompany speech offset if the current speaker continues with the next full turn following the silence. It is possible that gesture strokes (or potentially only those which are referential; this remains to be tested) contain some kind of semantic content that is “lexical” enough to be interpreted as the current speaker continuing to speak. In this case, although

there is a silence from an auditory point of view, it would be possible to make the argument that the current speaker has not actually stopped speaking, thus hindering another speaker from taking up a full turn (though, of course, backchannels in overlap with current speech—or ongoing gesture strokes—are permissible).

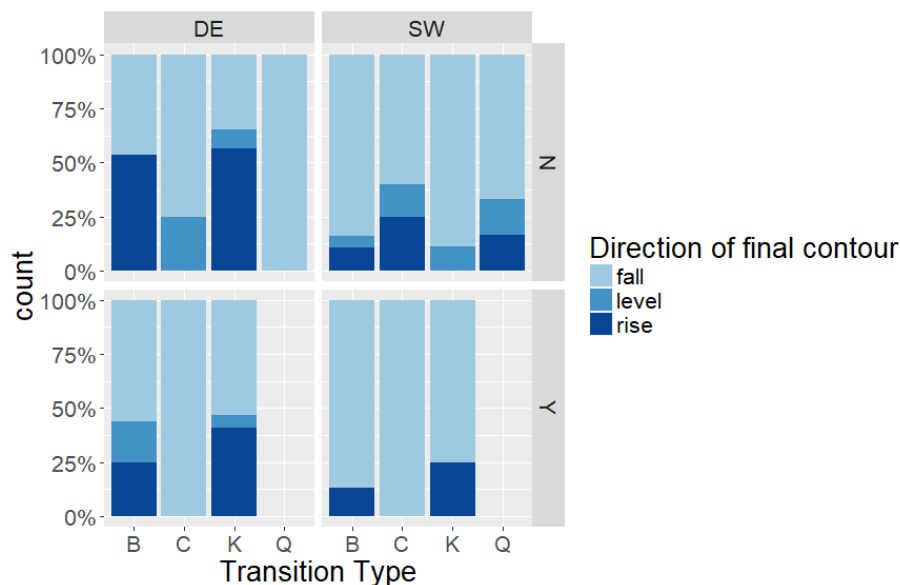


Figure 2. Distribution of final pitch contours in contexts without (N) and with (Y) accompanying hand gesture in the vicinity of speech offset at different transition types (B = Backchannel, C = Change, K = Keep, Q = Question).

Keeping in mind the differences between the communicative situations in the two datasets, cross-linguistic differences begin to emerge when we investigate pitch movements and gestures together. First, as would be expected from previous data, we find that rising contours are overall less frequent in Swedish than in German. This supports the argument that Swedish may have less flexibility to modify pitch in order to provide information about turn-taking. Furthermore, as seen in Figure 2, it appears that in general, turn-final pitch is more variable in both languages in contexts where there is no accompanying hand gesture. This supports the hypothesis that pitch and gesture share a functional load in conversation with regards to turn-taking: when hand gestures are present and can be manipulated in terms of their timing relative to the offset of speech, there is less need to vary pitch to carry the same meanings.

One observation which remains to be explained is the incidence of rising contours in Swedish Changes without accompanying hand gestures. Since all Changes involved turns in declarative, not interrogative form, it is unclear why rising contours appear in these cases. Heldner & Włodarczak (2015) report that final pitch that deviates substantially from a speaker’s midpoint in either direction is associated with floor-release, whereas Zellers (2017) found no relationship between turn-final pitch and speaker change (although Swedish listeners made limited use of pitch variations if duration cues to speaker change were not available). Neither of these studies carried out a phonological analysis of the pitch contours, so further research is needed to clarify the role of pitch here.

Despite the limitations of this study, we have been able to provide preliminary responses to both of our research questions. There appears to be at least some overlap between the functions of pitch and hand gestures with regard to signalling turn-taking in both German and Swedish; however, this relationship appears to be mediated by the different phonological demands on pitch in the two languages, with pitch being overall less flexible in Swedish with regard to turn-taking. Future research will expand the pitch measurements and take into account other phonetic features, while also considering the semantic and pragmatic content of the hand gestures investigated.

Acknowledgments

This work was supported by the German Research Foundation (DFG; GO 3063/1-1, PE 2879/1-1, ZE 1178/1-1), the Swedish Research Council (VR-2017-02140), and the Riksbankens Jubileumsfond (P12-0634:1). We are grateful to Simon Alexanderson, Jonas Beskow, and Jens Edlund for assistance with Spontal, and to Caroline Kleen for supplementary annotation work.

References

- Auer, P. (1996). On the prosody and syntax of turn-continuations. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: interactional studies* (pp. 57-100). Cambridge, UK: Cambridge University Press.
- Bergmann, P. (2018). Prosody in Interaction. *Linguistik Online*, 88(1).
- Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer. Retrieved from <http://www.praat.org/>.
- Duncan, S., Jr. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292.
- Edlund, J., & Beskow, J. (2007). Pushy versus meek - using avatars to influence turn-taking behaviour. In *Proceedings of Interspeech 2007*, Antwerp, Belgium.
- Edlund, J., & Beskow, J. (2009). Mushypeek: a framework for online investigation of audiovisual dialogue phenomena. *Language and Speech*, 52, 351-367.
- Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proceedings of LREC 2010*, Valetta, Malta.
- ELAN (Version 5.4) [Computer software]. (2019). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from <https://tla.mpi.nl/tools/tla-tools/elan/>
- Gravano, A., & Hirschberg, J. (2009). Turn-yielding cues in task-oriented dialogue. In *Proceedings of SIGDIAL 2009*, Queen Mary University of London, UK (pp. 253-261).
- Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25, 601-634.
- Heldner, M., & Włodarczak, M. (2015). Pitch slope and end point as turn-taking cues in Swedish. In *Proceedings of ICPHS*, Glasgow, Scotland (pp. 10-15).
- House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue. *Speech Communication*, 46, 268-283.
- Kane, J., Yanushevskaya, I., de Looze, C., Vaughan, B., & Ni Chasaide, A. (2014). Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions. In *Proceedings of 15th Interspeech*, Singapore (pp. 333-337).
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge, UK: Cambridge University Press.
- Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41, 295-321.
- Local, J., Kelly, J., & Wells, W. H. G. (1986). Towards a phonology for conversation: turn-taking in Tyneside English. *Journal of Linguistics*, 22, 411-437.
- Mondada, L. (2007). Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2), 194-225.
- Peters, B. (2006). *Form und Funktion prosodischer Grenzen im Gespräch* (Doctoral dissertation). Christian-Albrechts Universität zu Kiel.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, 50(4), 696-735.
- Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, 11, 243-257.
- Schmidt, T. (2014). The research and teaching corpus of spoken German—FOLK. In *Proceedings of LREC 2014*, European Language Resources Association (ELRA).
- Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructive units and turns in conversation. *Pragmatics*, 6, 357-388.
- Sikveland, R. O., & Ogden, R. (2012). Holding gestures across turns: moments to generate shared understanding. *Gesture*, 12(2), 166-199.
- Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer & A. di Luzio (Eds.), *The Contextualization of Language* (pp. 135-158). Amsterdam: Benjamins B.V.
- Zellers, M. (2014). Duration and pitch in perception of turn transition by Swedish and English listeners. In Heldner, M. (ed.), *Proceedings of FONETIK 2014*, Stockholm, Sweden, 9-11 June 2014.
- Zellers, M. (2017). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. *Language and Speech*, 60(3), 454-478.