

Book of Abstracts



MMSYM

**2nd international Multimodal
Communication Symposium**

September 25 - 27, 2024



Table of Contents



Words of Welcome	2
Keynotes	4
Committees	5
Practical Information	6
Programm	7
Programm Overview.....	7
Detailed Programm.....	9
September 25th.....	17
Keynote: Julie Hunter.....	17
Session 1: Prosody-Gesture Interaction.....	19
Session 2: Processing of Multimodal Data.....	28
Postersession 1.....	37
Session 3: Multimodality and Development.....	60
September 26th.....	69
Keynote 2: Judith Holler.....	69
Session 4: Embodiment and Arts.....	72
Session 5: Sign Languages	79
Session 6: Methodical Perspectives.....	88
Postersession 2.....	95
Session 7: Phonetic Aspects of Gestures.....	118
September 27th.....	127
Session 8: Semantics / Pragmatics.....	127
Postersession 3.....	136
Keynote 3: Petra Wagner.....	159
Index of Authors.....	161



WELCOME TO MMSYM 2024!



We are delighted to welcome you to the Second International **Multimodal Communication Symposium**, MMSYM 2024, to be held at Goethe University Frankfurt, on September 25-27, 2024. The symposium aims to offer a vibrant, multidisciplinary forum for researchers from diverse fields to explore and discuss multimodality in both human communication and human-computer interaction. The event is organized and generously supported by the *Institute of Linguistics at Goethe University Frankfurt*. We would also like to express our sincere gratitude for the additional financial support provided by Goethe University's *research profile area 'Universality and Diversity'*, in collaboration with the *DFG Priority Program 2392 'Visual Communication (ViCom)'* and the *'Alfons und Gertrud Kassel Stiftung'*.

The Department of Linguistics at Goethe University Frankfurt has a strong research focus on theoretical linguistics, has been highly successful in securing external funding for collaborative and interdisciplinary research initiatives, and in recent years has become known for its innovative, interdisciplinary research focus on multimodality. This year's MMSYM symposium aligns seamlessly with this key research focus, continuing a tradition of multimodality symposia that has developed over the past two decades from Scandinavia to Europe. Last year, this tradition culminated in the first international MMSYM held in Barcelona in April 2023, and we are excited to build on that success with this year's event.

This year, the call for papers has centered on three key research themes of special interest to the prosody-gesture research community. The first (1) is **gesture-speech integration**, with a particular focus on the prosody-gesture link, concretely how gesture and prosody interact in discourse structuring. The second (2) is **formal, automatic and machine-learning approaches**, which are highly relevant for the future of multimodal communication research. The third (3) is **psycholinguistic approaches** in multimodal settings, including multimodality in acquisition as well as the cognitive processing of communication in different modalities.

All three themes will be explored in depth by our esteemed keynote speakers, who have graciously accepted our invitation to present and share their significant research experience with us. In the first keynote, **Dr. Julie Hunter** from the LinaGora Labs in Toulouse talks about *Situated conversation and Conversational Cobots*, examining how embodied agents collaborate with humans to reach improvement in human-robot multimodal conversation. This talk is thus centered around the second main research theme of MMSYM. The second keynote by **Dr. Judith Holler** from the Donders Institute for Brain, Cognition and Behavior in Nijmegen focuses on *Producing and comprehending of multimodal utterances in face-to-face interaction*. Assuming a multimodal language processing model, the talk focuses on the use of hand, head and facial gesture to convey information, and on how these cues are processed to facilitate understanding in conversation. Therefore, this talk is closely connected to the first and third main research themes of the

conference. In the final keynote, **Prof. Petra Wagner** will talk about *The multimodal expressions of (non-)understanding in dyadic explanations*. Based on a board game explanation corpus, she will illustrate the interplay of verbal and non-verbal explanations and listeners' feedback. This talk is therefore closely related to the first research theme of MMSYM.

We are pleased to present the MMSYM 2024 conference program, which revolves around the main research themes of this year's event. Our diverse lineup includes presentations that perfectly complement the keynote addresses and each other, fostering a rich environment for discussion. In addition to the three keynote talks, MMSYM 2024 will feature 30 presentations and 33 posters across the following topics: Prosody-Gesture Interaction, Processing of Multimodal Data, Multimodality and Development, Embodiment and Arts, Sign Languages, Methodological Perspectives, Phonetic Aspects of Gestures, and Semantics/Pragmatics of Gestures. Each presentation has been meticulously selected, being the top-most rated by three blind reviewers. We would like to express our sincere appreciation to the reviewers, the scientific committee, and the local organizing team for their invaluable contributions in shaping the program and making this event possible.

We hope that this exciting program, along with the discussions it sparks, will further strengthen the ties within our research community. Our goal is to foster a welcoming and intellectually stimulating environment where we can openly share our work and ideas. We especially encourage early career researchers to actively engage in these discussions and share their perspectives. To support this, we will prioritize questions from early career researchers during the discussion sessions following the oral presentations.

Finally, thank you for attending the conference and for coming to Frankfurt. We hope you have the opportunity to enjoy German and Hessian culture, food and traditions. During the welcome reception on the first day, you will be treated to classical music by local Frankfurt composers, performed by a string quartet, while enjoying some sparkling wine. We wish you a productive, collaborative, and exciting stay in Frankfurt for MMSYM 2024.

Warm regards,



Frank Kügler

Chair of MMSYM 2024

Professor of Linguistics/Phonology at Goethe University Frankfurt

On behalf of the MMSYM Organizing Committee

Keynotes at MMSYM 2024

We have invited three experts on multimodal communication to present their work on the three main conference themes at MMSYM. We are delighted that they have accepted our invitation and are introducing them here:



Dr. Julie Hunter (*LinaGora Labs Toulouse*)

Julie Hunter develops models of human conversation, including multimodal interactions, from a pragmatic perspective. Working on the development and automatic annotation of audio-visual communication corpora, Julie Hunter's research contributes to the facilitation and efficiency of multimodal communication research at the interface of linguistics and computational science.



Dr. Judith Holler (*Donders Institute Nijmegen*)

Judith Holler's research focuses on how people convey and comprehend messages with the verbal and visual modalities at their disposal in face-to-face interaction and how they use them to structure their interactions. She combines different approaches, including quantitative corpus analyses, behavioural and neurocognitive experimental methods, as well as tools such as mobile eyetracking and virtual reality.



Prof. Dr. Petra Wagner (*Bielefeld University*)

Petra Wagner's research focuses on phonetics, prosody and multimodal prosody, as well as speech synthesis, conversational acts of speech and human-machine interaction. Among other research interests, Petra Wagner is involved in projects tackling gesture-speech coordination, shedding light on multimodality from different perspectives.

Committees

Organizing a conference is a team effort! We are very thankful for everyone involved in the organization of MMSYM. Thanks to Patrizia Paggio and the GeHM network for founding the conference and for support, as well as to Pilar Prieto with her team at UPF for a great conference in Barcelona and for handing over the conference.

Thanks to our student assistants and helping hands Malin, Leah, Leoni, Jule, Natascha, Ai, Jonas and Miles for their help before, during and after the conference. This wouldn't be possible without you!

These are the main organizing committees for MMSYM:

Local Organizing Committee



Frank Kügler



Alina Gregori



Kathryn Barnes



Tina Bögel



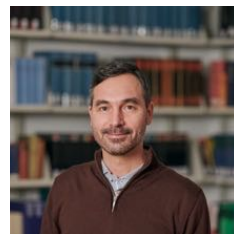
Cornelia Ebert



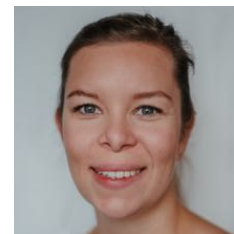
Lennart Fritzsche



Corinna Langer



Andy Lücking



Anna Pressler



Theresa Stender

Scientific Committee

- Gilbert Ambrazaitis (Linnaeus University, Växjö, Sweden)
- Kathryn Barnes (Goethe University Frankfurt, Germany)
- Florence Baills (Universitat de Lleida, Catalonia)
- Cornelia Ebert (Goethe University Frankfurt, Germany)
- Alina Gregori (Goethe University Frankfurt, Germany)
- Frank Kügler (Goethe University Frankfurt, Germany)
- Andy Lücking (Goethe University Frankfurt, Germany)
- Patrizia Paggio (University of Copenhagen, Denmark)
- Wim Pouw (Radboud University Nijmegen, Netherlands)
- Pilar Prieto (ICREA & Universitat Pompeu Fabra, Spain)
- Patrick Louis Rohrer (Radboud University Nijmegen, Netherlands)
- Markus Steinbach (Göttingen University, Germany)

Practical Information for the Conference

This is to provide some practical information to ensure a convenient stay in Frankfurt and at Goethe University!

This is the address of the conference venue, and a map on how to get to all relevant buildings:

**Goethe University Frankfurt
Westend Campus, Casino Building
Norbert-Wollheim-Platz 1
60629 Frankfurt, Germany**

You can reach the university best by subway, using lines U1, U2, U3 or U8 getting off at “Holzhausenstraße”.

Alternatively, you can use busses M36 or 75 getting off at “Uni Campus Westend”.

While the campus is in the north-western part of the city, you can reach the city center quickly! Orient yourself towards “Konstablerwache”, “Hauptwache” (public transport station) or “Alt-Sachsenhausen” (district).



For lunch, there are multiple Mensa options available: “Mensa Casino” (directly below the conference rooms), “Cafeteria Dasein” (5 min walk) or “Cafeteria Hoagascht” (8 min walk). All of them have vegetarian options available. You can find their menus [here](#).

Alternatively, you can also go to the “Rotunde” (IG Farben Building, 3 min walk) or the “Sturm und Drang” (Hörsaalzentrum, 3 min walk) for Sandwiches and Snacks.



MMSYM Program Overview

25.09. – 27.09.2024

Goethe University Frankfurt am Main; Campus Westend Casino Building

Time	Wednesday, 25.09.2024	Thursday, 26.09.2024	Friday, 27.09.2024
08:30 - 40	Registration (08:30 – 17:00)	Registration (08:30 – 17:00)	
08:40 - 50			
08:50 - 09			
09:00 – 10	Opening Ceremony	Keynote 2: Judith Holler	Oral session 8 (4 Slots) Semantics/Pragmatics of gestures
09:10 - 20			
09:20 - 30			
09:30 - 40	Keynote 1: Julie Hunter	Chair: Wim Pouw	Chair: Cornelia Ebert
09:40 - 50			
09:50 - 10			
10:00 - 10	Chair: Andy Lücking	Oral session 4 (3 Slots) Embodiment and Arts	Poster session 3
10:10 - 20			
10:20 - 30			
10:30 - 40	Coffee break	Chair: Patrick Rohrer	Chair: Andy Lücking
10:40 - 50			
10:50 - 11			
11:00 - 10	Oral session 1 (4 Slots) Prosody-Gesture Interaction	Coffee break	Coffee break
11:10 - 20			
11:20 - 30			
11:30 - 40	Chair: Petra Wagner	Oral session 5 (4 Slots) Sign Languages	Chair: Frank Kügler
11:40 - 50			
11:50 - 12			
12:00 - 10	Lunch	Chair: Markus Steinbach	Keynote 3: Petra Wagner
12:10 - 20			
12:20 - 30			
12:30 - 40		Lunch	Chair: Frank Kügler
12:40 - 50			
12:50 - 13			
13:00 - 10			Farewell
13:10 - 20			
13:20 - 30			
13:30 - 40	Oral session 2 (4 Slots) Processing of Multimodal Data		
13:40 - 50			
13:50 - 14			
14:00 - 10		Oral session 6 (3 Slots)	
14:10 - 20			
14:20 - 30			
14:30 - 40			Campus Tour -

14:40 - 50		Methodological Perspectives	Goethe University Campus Westend
14:50 - 15	Chair: Patrizia Paggio		
15:00 - 10	Poster session 1	Chair: Aleksandra Ćwiek	
15:10 - 20			
15:20 - 30		Poster session 2	
15:30 - 40			
15:40 - 50			
15:50 - 16			
16:00 - 10			
16:10 - 20			
16:20 - 30	Chair: Alina Gregori		
16:30 - 40	Coffee break	Chair: Kathryn Barnes	
16:40 - 50		Coffee break	
16:50 - 17	Oral session 3 (4 Slots)		
17:00 - 10	Multimodality and Development		
17:10 - 20		Oral session 7 (4 Slots)	
17:20 - 30			
17:30 - 40		Phonetic aspects of gestures	
17:40 - 50			
17:50 - 18			
18:00 - 10	Chair: Carina Lücke		
18:10 - 20	Short break		
18:20 - 30		Chair: Stefan Baumann	
18:30 - 40	Welcome Reception at IG-Farben building with a performance of a string quartet	Travelling to Sachsenhausen	
18:40 - 50			
18:50 - 19			
19:00 - 10			
19:10 - 20			
19:20 - 30			
19:30 - 40			
19:40 - 50		Conference Dinner at "Homburger Hof"	
19:50 - 20			
20 - open			

A highlight of the **welcome reception** will be the performance of a string quartet who will play a selection of classical music. We are very happy to welcome these artists:

Carolin Grün (Akademie des hr-Sinfonieorchesters) - Violin 1

Mixia Kang (Akademie des hr-Sinfonieorchesters) - Violin 2

Franziska Hügel (Akademie des hr-Sinfonieorchesters) - Viola

Simon Napp (Hochschule für Musik und Darstellende Kunst Frankfurt am Main (HfMDK)) - Violoncello



Detailed program for MMSYM

Sessions and Contributions



Keynote 1 (25.09. 09:30 – 10:30)

Julie Hunter
LinaGora Labs, Toulouse

“Situated Conversation and Conversational Cobots”

Oral session 1 (25.09. 11:00 – 12:20): Prosody-Gesture Interaction

Patrick Louis Rohrer, Ronny Bujok, Lieke van Maastricht & Hans Rutger Bosker
Donders Centre for Cognition, Radboud University, Nijmegen; Max Planck Institute for Psycholinguistics, Nijmegen; Centre for Language Studies, Radboud University, Nijmegen

“The timing of non-referential beat gestures affects lexical stress perception in Spanish regardless of individuals’ working memory capacity”

Massimo Moneglia & Giorgina Cantalini
University of Florence; Civica Scuola Interpreti e Traduttori ‘Altiero Spinelli’, Milan

“Prosodic Synchrony and the Semantic Anchors of Referential Gestures”

Paula G. Sánchez-Ramón, Frank Kügler & Pilar Prieto
Universitat Pompeu Fabra; Goethe University Frankfurt; Institució Catalana de Recerca i Estudis Avançats

“The influence of gesture presence in the prosodic realization of focus types in the Catalan language”

Florence Baills & Stefan Baumann
Universitat de Lleida; University of Cologne

“Gesture, prosody and syntax as markers of information structure in French”

Oral session 2 (25.09. 13:40 – 15:00): Processing of multimodal data

Walter Philip Dych, Karee Garvin & Kathryn Franich
Binghamton University; Harvard University

“A toolkit for automating co-speech gesture data annotation and analysis”

Lena Pagel, Simon Roessig & Doris Mücke
University of Cologne; University of York

“Introducing DiCE: A novel approach to elicit and capture multimodal accommodation via 3D electromagnetic articulography, audio, and video”

Romain Pastureau & Nicola Molinaro
Basque Center on Cognition, Brain and Language (BCBL), San Sebastián; Universidad del País Vasco/Euskal Herriko Unibertsitatea, San Sebastián; Ikerbasque, Basque Foundation for Science

“Krajjat: A Python Toolbox for Analysing Body Movement and Investigating its Relationship with Speech”

Davide Ahmar, Šárka Kadavá & Wim Pouw
Donders Institute for Brain, Cognition and Behavior; Leibniz Centre for General Linguistics

“MOBILE MULTIMODAL LAB: An Open-Source, Low-Cost and Portable Laboratory for the study of Multimodal Human Behavior”

Poster session 1 (25.09. 15:00 – 16:20)

- | | | |
|-------|--|--|
| D1-01 | Alina Naomi Riechmann & Hendrik Buschmeier
Bielefeld University | <i>“Automatic Reconstruction of Dialogue Participants’ Coordinating Gaze Behaviour from Multiple Camera Perspectives”</i> |
| D1-02 | Luca Béres, Ádám Boncz, Péter Nagy & István Winkler
HUN-REN Research Centre for Natural Sciences, Budapest; Budapest University of Technology and Economics, Budapest | <i>“The role of synchronization in face-to-face communication: A dual eye-tracking and motion capture study”</i> |
| D1-03 | Sharice Clough, Beyza Sümer, Kristel de Laat, Annick Tanguay, Sarah Brown-Schmidt, Melissa C. Duff & Aslı Özyürek
MPI for Psycholinguistics; Vanderbilt University Medical Center; University of Amsterdam | <i>“Spatial Narratives from Remote and Recent Memory in Individuals with Alzheimer’s Disease and Healthy Older Adults: A Multimodal and Kinematic Perspective”</i> |
| D1-04 | Stefanie Shattuck-Hufnagel & Ada Ren-Mitchell
MIT RLE Speech Communications Group; MIT Media Lab | <i>“Kinematic gestural evidence for higher-level prosodic constituents in speech”</i> |
| D1-05 | Aleksandra Ćwiek, Šárka Kadavá, Wim Pouw & Susanne Fuchs
Leibniz-Centre General Linguistics; Donders Institute for Brain, Cognition, and Behaviour; University of Göttingen | <i>“The Communicative Consequences of Multimodal Coordination”</i> |
| D1-06 | Schuyler Laparle & Merel Scholman
Tilburg University; Utrecht University; Saarland University | <i>“Signaling discourse relations in multimodal communication”</i> |
| D1-07 | Marion Schulte
Rostock University | <i>“Social meaning and multimodality: The performance of scientific authority”</i> |

- | | | |
|-------|--|---|
| D1-08 | Vera Wolfrum, Carina Lücke & Simone Schaeffner
Julius-Maximilians University
Würzburg | <i>“The influence of linguistic input on the multimodal language processing of primary school children”</i> |
| D1-09 | Stefan Lazarov & Angela Grimminger
Paderborn University | <i>“Verbal signals of understanding do not predict a decrease of gesture deixis”</i> |
| D1-10 | Elena Nicoladis, Anahita Shokrkon & Shiva Zarezadehkheibari
University of British Columbia;
University of Alberta | <i>“Farsi-English bilinguals’ gesture production while telling a story”</i> |
| D1-11 | Kazuki Sekine & Ikuko Nonaka
Waseda University | <i>“Effects of Bowing During Japanese Telephone Conversation on Acoustic Properties”</i> |

Oral session 3 (25.09. 16:50 – 18:10): Multimodality and Development

- | | |
|--|--|
| Sara Coego, Núria Estve-Gibert & Pilar Prieto
Universitat Pompeu Fabra; Universitat Oberta de Catalunya; Institució Catalana de Recerca i Estudis Avançats (ICREA) | <i>“Preschoolers’ use of prosody and gesture in marking focus types”</i> |
| Anita Slonimska, Alessia Giulimondi, Emanuela Campisi & Asli Ozyurek
Max Planck Institute for Psycholinguistics, Nijmegen; Utrecht University; Catania University; Donders Institute for Brain, Cognition and Behavior, Nijmegen | <i>“Simultaneity in iconic two-handed gestures: a communicative strategy for children”</i> |
| Joel Espejo-Álvarez, Júlia Florit-Pons, Claire Lien Luong, Mireia Gómez i Martínez, Alfonso Igualada & Pilar Prieto
Universitat Pompeu Fabra; University of Cork; Universitat Oberta de Catalunya; Institució Catalana de Recerca i Estudis Avançats | <i>“The impact of a multimodal oral narrative intervention on boosting the frequency of use and the quality of children’s non-dominant language”</i> |
| Mariia Pronina, Júlia Florit-Pons, Sara Coego & Pilar Prieto
The University of the Balearic Islands; Universitat Pompeu Fabra; Institució Catalana de Recerca i Estudis Avançats (ICREA) | <i>“Different developmental paths of multimodal imitation in typically and non-typically developing preschool and primary school children”</i> |

Keynote 2 (26.09. 09:00 – 10:00)

Judith Holler

Donders Institute for Brain, Cognition &
Behaviour, Nijmegen

*“Producing and comprehending multimodal utterances
in face-to-face conversation”*

Oral session 4 (26.09. 10:00 – 11:00): Embodiment and Arts

**Lara Pearson, Thomas Nuttall & Wim
Pouw**

Max Planck Institute for Empirical Aesthetics,
Frankfurt; Universitat Pompeu Fabra; Donders
Institute for Brain, Cognition, and Behaviour,
Radboud University, Nijmegen

*“Motif-Gesture Contiguity in Karnatak Vocal
Performance: A Multimodal Computational Analysis”*

**Nasim Mahdinazhad Sardhaei, Marzena
Zygis & Hamid Sharifzadeh**

Leibniz Center for General Linguistics

*“Orofacial signals beyond sight: A study of expressive
faces and whispered voices in German”*

Elena Nicoladis

University of British Columbia

“The effects of familiarity on children’s pantomimes”

Oral session 5 (26.09. 11:30 – 12:50): Sign languages

**Sonja Gipper, Anastasia Bauer, Jana
Hosemann & Tobias-Alexander
Herrmann**

University of Cologne

*“Multimodal feedback in signed and spoken
languages: Evidence for a shared infrastructure of
conversation”*

Marisa Cruz & Sónia Frota

University of Lisbon

*“Four seasons in one head: The prosodic phrasing of
enumerations in Portuguese Sign Language”*

Clara Lombart

University of Namur, NaLTT, LSFB-Lab

*“How visual cues make information units more
prominent in spoken and signed languages: A case
study on French and French Belgian Sign Language
(LSFB)”*

**Anastasia Bauer, Anna Kuder, Marc
Schulder & Job Schepens**

University of Cologne; University of Hamburg

*“The phonetics of addressee’s head nods in signed and
spoken interaction using a computer vision solution”*

Oral session 6 (26.09. 14:20 – 15:20): Methodological Perspectives

Geert Brône, Bert Oben & Julie Janssens
University of Leuven

“Looking together. An eye-tracking corpus of museum visitors’ shared experience and joint attention”

Sam O’Connor Russell & Naomi Harte
Trinity College Dublin, Ireland

“Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction”

Mojenn Schubert
Leibniz-Institute for the German Language,
Mannheim

“Navigating the topical landscape: Pointing at others as an embodied backlinking device in multi-party interaction”

Poster session 2 (26.09. 15:20 – 16:40)

D2-01 **Alexander Henlein, Alexander Mehler & Andy Lücking**
Goethe University Frankfurt, Text
Technology Lab

“Virtually Restricting Modalities in Interactions: Va.Si.Li-Lab for Experimental Multimodal Research”

D2-02 **Han Zhou**
Heidelberg University

“A Theoretical Model for Analyzing Metaphors in Multimodal Communication: Exemplified by Pictorial and Verbo-Pictorial Metaphors in Editorial Cartoons”

D2-03 **Patrizia Paggio, Manex Agirrezabal & Bart Jongejan**
University of Copenhagen; University of
Malta

“Movement entrainment in online meetings”

D2-04 **Alina Gregori & Susanne Fuchs**
Goethe University Frankfurt; Leibniz-
Center for General Linguistics; ILCB
and IMéRA at Aix-Marseille University

“Moving Meetings by Moving Prosody and Gesture”

D2-05 **Marion Bonnet, Cornelia Ebert, Kurt Erbach & Markus Steinbach**
Göttingen University; Goethe University
Frankfurt

“Show me the choice”

D2-06 **Christoph Rühlemann & James Trujillo**
University of Freiburg; University of
Amsterdam

“The effect of gesture expressivity on emotional resonance in storytelling interaction”

D2-07 **Himmet Sarıtaş & Şeyda Özçalışkan**
Georgia State University

“Does gesture play a similar role in the communication of second language learners in face-to-face and online interactions?”

D2-08 **Emanuel Schütt, Merle Weicker & Carolin Dudschig**
University of Tübingen; Goethe
University Frankfurt; University of
Cologne

“Human language comprehenders appear to integrate rapidly gestural and verbal expressions of “yes” and “no”: Evidence from a two-choice response time task”

- D2-09 **Ingrid Vilà-Giménez, Mariia Pronina & Pilar Prieto**
Universitat de Girona; Universitat de les Illes Balears; Institució Catalana de Recerca i Estudis Avançats; Universitat Pompeu Fabra
“Exploring children’s storytelling: The link between narrative abilities, receptive vocabulary and gesture rate in 7- to 9-year-olds”
- D2-10 **Nathalie Frey & Carina Lüke**
University of Würzburg
“Multimodal insights into the lexical development of mono- and multilingual children with SLCN”
- D2-11 **Júlia Florit-Pons, Pilar Prieto, Alfonso Igualada**
Universitat Pompeu Fabra; Institució Catalana de Recerca i Estudis Avançats; Universitat Oberta de Catalunya; Institut Guttmann
“A multimodal narrative intervention for boosting NDD children’s oral narrative skills”

Session 7 (26.09. 17:10 – 18:30): Phonetic aspects of gestures

- Gilbert Ambrazaitis, Margaret Zellers & David House**
Linnaeus University, Växjö; Kiel University; KTH Royal Institute of Technology, Stockholm
“Pitch accent realization as a function of accompanying manual or eyebrow gestures in spontaneous Swedish dialogue”
- Martine Grice, Alexandra Vella, Maria Lialiou, Florence Bails, Aviad Albert, Petra B. Schumacher, Nadia Pelageina & Solveigh Janzen**
University of Cologne; University of Malta; Universitat de Lleida
“Gesture apex coordination with prosodic structure and tonal events in Maltese English”
- Kathryn Franich & Vincent Nwosu**
Harvard University; University of Calgary
“Timing of Co-Speech Gesture in Igbo: Influence of Metrical Prominence and Tonal Melody”
- Helene Springer, Henrik Garde, Frida Splendido & Marianne Gullberg**
Lund University; Lund University Humanities Lab
“Quantifying the visual salience of Swedish vowels: A computer vision approach”

Session 8 (27.09. 09:00 – 10:20): Semantics/Pragmatics of gestures

- Andy Lücking, Alexander Mehler & Alexander Henlein**
Goethe University Frankfurt, Text Technology Lab
“The Gesture--Prosody Link in Multimodal Grammar”

Silva H. Ladewig
University of Göttingen

“From Hand to Discourse: The Stabilization of the Slicing Gesture and its Meta-Pragmatic Function”

Daniel K. E. Reisinger & Marianne Huijsmans
University of British Columbia; University of Alberta

“On the Role of Co-speech Gesture with ʔayʔajuθəm D Elements”

Cornelia Loos & Sophie Repp
University of Hamburg; University of Cologne

“The many ways to mark agreement & rejection: Multimodal polar responses in German”

Poster session 3 (27.09. 10:20 – 11:40)

- | | | |
|-------|---|---|
| D3-01 | Arianna Colombani, Varghese Peter, Quian Yin Mai, Outi Tuomainen, Natalie Boll-Avetisyan, Amanda Saksida & Mridula Sharma
International Doctorate for Experimental Approaches to Language and Brain (IDEALAB); Macquarie University; University of Potsdam; University of the Sunshine Coast, Brisbane; Institute for Maternal and Child Health-IRCCS “Burlo Garofolo”, Trieste | <i>“Cross-situational learning of word-pseudosign pairs in children and adults: a behavioral and event-related potential study”</i> |
| D3-02 | Stéphanie Caët, Loulou Kosmala, Carla Ferran & Marine Laval
Université de Lille; UMR 8163 Savoirs, Textes, Langage; Université Paris Nanterre; EA 370 CREA | <i>“Participation of deaf children with a cochlear implant in family dinner interactions: the role of gesture”</i> |
| D3-03 | Katharina J. Rohlfing, Nils Tolksdorf, Angela Grimmering, Koki Honda & Kazuki Sekine
Paderborn University; Waseda University | <i>“Using social robots for cross-cultural gesture elicitation in children: Psycholinguistic considerations on dialogue design”</i> |
| D3-04 | Maria Graziano, Joost van de Weijer & Marianne Gullberg
Lund University Humanities Lab; Centre for Languages and Literature, Lund University | <i>“Exploring gesture distribution over disfluency markers in competent speakers and language learners”</i> |
| D3-05 | Joanna Wójcicka, Anna Kuder & Justyna Kotowicz
University of Warsaw; University of Cologne; University of Silesia Katowice | <i>“Language Control and Multimodal Behavior in Native Hearing PJM-Polish Bilinguals Using Spoken Polish”</i> |

- | | | |
|-------|---|--|
| D3-06 | Fien Andries, Katharina Meissl & Clarissa de Vries
KU Leuven | <i>“Mocking enactments: a case-study of multimodal stance-stacking”</i> |
| D3-07 | Margaret Zellers, Jan Gorisch & David House
Kiel University; Leibniz-Institut für Deutsche Sprache; Kungliga Tekniska Högskolan | <i>“Referential gestures and the management of turn-taking in conversation”</i> |
| D3-08 | Johannes Heim, Rebecca Woods, Franziska Busche & Sophie Repp
University of Aberdeen; Newcastle University; University of Cologne | <i>“Multimodal profiles of different (negative) question types”</i> |
| D3-09 | Federica Raschellà, Frida Splendido, Nadja Althaus, Marieke Hoetjes & Gilbert Ambrazaitis
Linnaeus University, Växjö; Lund University; University of East Anglia, Norwich; Radboud University, Nijmegen | <i>“Embodied pronunciation training for the Swedish complementary length contrast”</i> |
| D3-10 | Anna Inbar & Yael Maschler
The Academic College Levinsky-Wingate; University of Haifa | <i>“Pointing at the addressee in Hebrew face-to-face interaction”</i> |
| D3-11 | Vivien Lohmer & Friederike Kern
Bielefeld University | <i>“The role of interactive gestures in explanatory interactions”</i> |

Keynote 3 (27.09. 12:10 – 13:10)

Petra Wagner
Bielefeld University

“The multimodal expression of (non-)understanding in dyadic explanations – some lessons learned”



Keynote 1:

Julie Hunter

25.09.2024
9:30-10:30



Situated Conversation and Conversational Cobots

Julie Hunter

LinaGora Labs, Toulouse

jhunter@linagora.com

Embodied agents that need to collaborate with humans in real-time on a task would benefit from being able to acquire skills -- including repeatable skills in the form of reusable programs -- through conversation. Human teachers, however, will in general stop short of fully specifying a set of instructions, and when things are left too open, planning can be difficult. A central function of conversation in such a case is to allow participants to jointly fill in and create -- through a series of actions and complex conversational moves such as questions, answers, elaborations, corrections etc. -- the final content of the instructions or program that the agent needs to follow.

In this talk, I present our efforts to build models of multimodal conversation for collaborative conversational agents such as cobots. In the first part, I introduce the work we have done on the Minecraft Dialogue Corpus, which contains collaborative, task-oriented dialogues of the sort we wish to target. I also describe our model of multimodal task-oriented dialogue trained on this corpus, as well as our approach to predicting and evaluating action sequences based on human instructions. I then show how we are using large language models (LLMs) to retrieve simpler, reusable concepts that can then be exploited to build more complex constructions. Finally, I explain how we build upon this previous work to build a dialogue model for the COCOBOTS corpus, which we have designed to closely reflect an actual industrial use case involving a collaborative robot.

Session 1: Prosody-Gesture Interaction

25.09.2024

11:00-12:20



The timing of non-referential beat gestures affects lexical stress perception in Spanish regardless of individuals' working memory capacity

Patrick Louis Rohrer ¹, Ronny Bujok ², Lieke van Maastricht ³, Hans Rutger Bosker ^{1,2}

Donders Centre for Cognition, Radboud University, Nijmegen, The Netherlands¹

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands²

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands³

`hansrutger.bosker@donders.ru.nl`

In conversation, addressees can make use of both auditory and visual cues to facilitate speech processing. For example, facial and manual gestures have been shown to affect the disambiguation of ambiguous syntactical structures [1], [2], the recognition of speech acts [3] or even the perception of prominence at both the phrasal and the lexical level [4], [5]. Regarding the perception of word-level stress, recent evidence has shown the existence of a *manual McGurk effect*, where the timing of a non-referential “beat” gesture influences spoken word recognition of Dutch lexical stress minimal pairs. Specifically, Dutch listeners were biased to hear lexical stress on syllables that coincided with a beat gesture, regardless of the acoustic cues to stress present in the auditory stream.

Moreover, evidence suggests that there is a relationship between multimodal speech processing and individual cognitive abilities, such as visuospatial working memory (VWM). For example, it has been shown that individuals with a high VWM are more sensitive to the semantic congruency of iconic gestures and their accompanying speech than those with lower VWM, and such effects are not found for phonological working memory (PWM) [6], [7]. Similarly, adults receiving mathematical instruction with gestures tended to learn and transfer their mathematical abilities better when they had a higher VWM [8]. This suggests that integrating gesture with speech for language comprehension taxes VWM capacity, with a higher VWM leading to greater sensitivity to gesture (see [9] for a review).

The current study aims to further our knowledge of the manual McGurk effect in two ways. First, it investigates how it surfaces in Spanish, a language where duration is the primary acoustic cue to lexical stress (and F0 functioning more to mark phrase-level prominence) [10], and where the lexical stress contrast is present in the regular verb conjugation system, representing a highly relevant cue for everyday speech comprehension. Second, it investigates how individuals' PWM and VWM influence the size of the manual McGurk effect.

Acoustic lexical stress continua were created for 18 disyllabic minimal pairs across 7 steps by interpolating syllabic duration, intensity, and the F0 contour (e.g., going from strong-weak [SW] ‘bailo’, *I dance*, with word-initial stress to weak-strong [WS] ‘bailó’, *he/she danced*, with word-final stress). The audio was then superimposed on a video of a face-masked speaker producing a non-referential beat gesture timed to occur on either the first or the second syllable. The audiovisual stimuli were presented to one hundred native Spanish speakers in an online study, where they indicated which of the two words they heard. The participants subsequently carried out Digit Span and Corsi Block Tapping tasks to assess individual PWM and VWM capacities.

The results showed that participants were biased to perceive lexical stress on the syllable that visually co-occurred with a beat gesture, with the effect being stronger in acoustically more ambiguous steps. That is, the same acoustic continuum was perceived as more SW-like if combined with a gesture falling on the first syllable (vs. second syllable; Figure 1). However, neither measure of working memory correlated with individual effect sizes (Figure 2), which suggests that VWM may be less relevant for non-referential gestures than iconic ones, as they do not represent meaning visuospatially but rather temporally. These findings on Spanish corroborate the impact of gestures' timing on prominence perception and spoken word recognition.

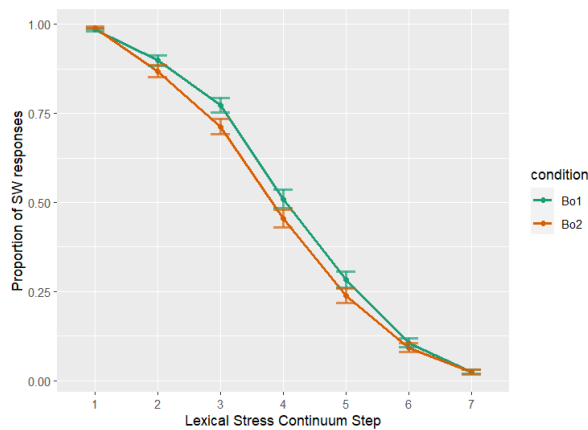


Figure 1: Proportion of ‘strong-weak’ (SW) responses across continuum steps as a function of Beat condition (Bo1 = beat-on-first-syllable, in green; Bo2 = beat-on-second-syllable, in orange)

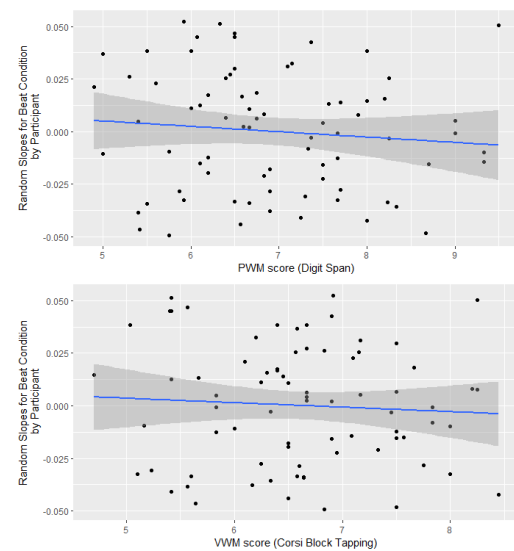


Figure 2: Scatterplot of the effect size (random slopes of the beat condition by participant) against PWM scores (top panel) and VWM scores (bottom panel)

References

- [1] B. Guellai, A. Langus, and M. Nespors, “Prosody in the hands of the speaker,” *Front. Psychol.*, vol. 5, Jul. 2014, doi: 10.3389/fpsyg.2014.00700.
- [2] H. Holle, C. Obermeier, M. Schmidt-Kassow, A. D. Friederici, J. Ward, and T. C. Gunter, “Gesture Facilitates the Syntactic Analysis of Speech,” *Front. Psychol.*, vol. 3, 2012, doi: 10.3389/fpsyg.2012.00074.
- [3] N. Nota, J. P. Trujillo, and J. Holler, “Conversational Eyebrow Frowns Facilitate Question Identification: An Online Study Using Virtual Avatars,” *Cogn. Sci.*, vol. 47, no. 12, p. e13392, 2023, doi: 10.1111/cogs.13392.
- [4] E. Krahmer and M. Swerts, “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception,” *J. Mem. Lang.*, vol. 57, no. 3, pp. 396–414, Oct. 2007, doi: 10.1016/j.jml.2007.06.005.
- [5] H. R. Bosker and D. Peeters, “Beat gestures influence which speech sounds you hear,” *Proc. R. Soc. B Biol. Sci.*, vol. 288, no. 1943, p. 20202419, Jan. 2021, doi: 10.1098/rspb.2020.2419.
- [6] Y. C. Wu and S. Coulson, “Co-speech iconic gestures and visuo-spatial working memory,” *Acta Psychol. (Amst.)*, vol. 153, pp. 39–50, Nov. 2014, doi: 10.1016/j.actpsy.2014.09.002.
- [7] D. Özer and T. Göksun, “Visual-spatial and verbal abilities differentially affect processing of gestural vs. spoken expressions,” *Lang. Cogn. Neurosci.*, vol. 35, no. 7, pp. 896–914, Sep. 2020, doi: 10.1080/23273798.2019.1703016.
- [8] M. Aldugom, K. Fenn, and S. W. Cook, “Gesture during math instruction specifically benefits learners with high visuospatial working memory capacity,” *Cogn. Res. Princ. Implic.*, vol. 5, no. 1, p. 27, Jun. 2020, doi: 10.1186/s41235-020-00215-8.
- [9] E. Congdon, M. Novack, and E. Wakefield, “Exploring Individual Differences: A Case for Measuring Children’s Spontaneous Gesture Production as a Predictor of Learning From Gesture Instruction,” *Top. Cogn. Sci.*, Jan. 2024, doi: 10.1111/tops.12722.
- [10] M. Ortega-Llebaria and P. Prieto, “Acoustic Correlates of Stress in Central Catalan and Castilian Spanish,” *Lang. Speech*, vol. 54, no. 1, pp. 73–97, Mar. 2011, doi: 10.1177/0023830910388014.

Prosodic Synchrony and the Semantic Anchors of Referential Gestures

Massimo Moneglia¹ Giorgia Cantalini²

University of Florence¹ Civica Scuola Interpreti e Traduttori 'Altiero Spinelli', Milan²

Corresponding author massimo.moneglia@unifi.it

In his seminal work, McNeill states three synchrony rules governing co-speech gestures: semantic, pragmatic, and phonological. When speech and gesture co-occur, they are expected to present the same semantic information or perform the same pragmatic function, and the Stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech [1]. Lexical affiliate, however, does not automatically correspond to the co-expressive speech segment. Gestures are 'windows onto thinking' [2] and can refer to the underlying concept rather than to the emerging speech: 'conceptual affiliate', rather than 'lexical affiliate' [3]. Therefore, the notion of 'lexical affiliate' is insufficient to specify the semantic relations that a gesture may find in speech [4].

A substantial body of research on gesture/prosody synchronization has been conducted in the last twenty years, particularly in the Autosegmental frame, starting from the work of Loehr [5], which shows a strong synchronization between prominent pitch accents and strokes' apex. More recently, the importance of prosodic edges has been taken into account, and information structure has also been considered in connection to gesture functions [6] [7] [8].

In the Language into Act Theory perspective [9] [10], the Perceptively Relevant Prosodic Movements (PRMs) [11] characterizing prosodic units (PUs), convey functional values, such as the Illocutionary force, the Topic-function, and the Parentheses-function. PMRs, by definition, constitute the main prosodic prominence in the utterance since they are intentionally performed [11], and functional information can constitute the affiliate of gestures [12].

Assuming this approach, the paper investigates to which extent prosodic synchrony and semantic affiliation overlap and whether the alignment to the prosodic prominence so defined signals the gesture's semantic relationship.

We observed a monologue taken from an informal interview given by an actor on his work belonging to the Cantalini Corpus [13] previously annotated with regard to a) Gesture hierarchy according to standard models [14] [15] (G-Units, G-Phrases, and G-Phases, Preparation, Pre-stroke hold, Stroke, Post-stroke hold, Retraction); b) Prosodic boundaries identifying PUs; c) Information function of PUs, according to the L-AcT tagset [16].

We assumed an operative definition of the semantic relation between a gesture and speech information based on the disclosure of the metaphoric or metonymic representation embodied by the gesture in the given context [17] [18]. The semantic affiliate (*anchor* in our terms) is *the linguistic information allowing the referential interpretation of a co-speech gesture* [19]. By adding to the dataset, the annotation of PRMs characterizing PUs and the linguistic Anchor found for each Stroke, the relation among PRMs, Strokes, and Anchors is investigated.

The paper will present the annotation procedure and the main findings of the research. The 206 prosodic units in the dataset guest 237 PRMs and 132 Strokes. 117 Strokes are referential gestures (deictic, metaphorical, iconic). A significant part of them (47 Strokes) were not anchored in a single lexical entry in the utterance. This set is almost equally divided between Strokes that find their anchor in an information function (Illocution, Topic, Parenthesis) and, very interestingly, strokes expressing a *modal evaluation*, lacking lexical or functional anchors.

The synchrony of strokes with PRMs and the semantic relevance of this synchrony are largely confirmed: 107 strokes are aligned to PRMs that also mark a lexical, functional, or modal value. Asynchronies may occur in coincidence with Catchment phenomena [20] or Post-stroke hold. A genuine asynchrony may arise when the Stroke synchronizes to a specific lexical Anchor not marked by prosody. For instance, in the Topic unit, the PRM is necessarily on the right, while the lexical head of the unit anchoring the Stroke may be on the left.

References

- [1] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992.
- [2] D. McNeill, & S. D. Duncan, "Growth point in thinking-for-speaking". In D. McNeill (ed.), *Language and Gesture* (p. 141-161) Cambridge University Press, Cambridge, 2000.
- [3] C. Kirchhof, "So What's Your Affiliation With Gesture?" in C. Kirchhof, Z. Malisz, P. Wagner, eds. *GeSpIn*. Bielefeld; 2011.
- [4] A. Cienki, "The study of gesture in cognitive linguistics: How it could inform and inspire other research in cognitive science," *WIREs Cognitive Science*, 13(6), e1623. 2022
- [5] D. Loehr, *Gesture and Intonation*. Ph.D dissertation. Washington, DC: Georgetown University, 2004
- [6] G. Cantalini, M. Moneglia, "The annotation of Gesture and Gesture / Prosody synchronization in Multimodal Speech Corpora," *Journal of Speech Science*, Vol 9, pp. 1-24, 2020.
- [7] P. Rohrer, E. Delais-Roussarie, P. Prieto, "Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses," *Lingua* 293, 2023.
- [8] P. Rohrer, "A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system," Ph.D. dissertation, University Pompeu Fabra, Barcelona, 2022.
- [9] E. Cresti, *Corpus di italiano parlato*, Accademia della Crusca, Firenze, 2000.
- [10] E. Cresti, M. Moneglia, "The illocutionary basis of information structure. Language into Act Theory (L-AcT)," in E. Adamou, K. Haude, M. Vanhove (eds.) *Information structure in lesser-described languages: Studies in prosody and syntax*, pp. 359-401, Benjamins, Amsterdam, 2018.
- [11] J. 't Hart, R. Collier, S. Cohen, *A Perceptual Study of Intonation. An Experimental-Phonetic Approach to Speech Melody*, CUP, Cambridge, 1990.
- [12] M. Moneglia, "Gesture / speech synchronization and gesture's scope across Information Unit types", in N. Korotaev, N. Sumbatova (eds.) *Corpus scientiae: Papers in honor of Vera I. Podlesskaya*, pp. 339-365, Buki Vedi, Moscow, 2024.
- [13] G. Cantalini, "Corpus multimodale annotato per lo studio della gestualità co-verbale nel parlato-parlato e nel parlato-recitato," in E. Cresti, M. Moneglia (eds.) *Corpora e Studi Linguistici*, pp. 135-149, Officinaventuno, Milano, 2022.
- [14] S. Kita, I. van Gijn, H. van der Hulst, "Movement phases in signs and co-speech gestures, and their transcription by human coders" in Wachsmuth I, Fröhlich M. (eds), *Gesture and Sign Language in Human-Computer Interaction* (pp. 23-35): Springer, Berlin 1998.
- [15] J. Bresse, SH. Ladewig, C. Müller, "Linguistic Annotation System for Gestures (LASG)", in C. Müller, A. Cienki, E. Fricke, SH. Ladewig, D. McNeill, S. Teßendorf (eds.), *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*, pp. 1098-1125, Vol. 1. De Gruyter-Mouton. 2013.
- [16] M. Moneglia, T. Raso, "Notes on Language into Act Theory (L-AcT)", in T. Raso, H. Mello (eds.), *Spoken Corpora and Linguistic Studies*, pp. 468–495, Benjamins, Amsterdam, 2014
- [17] A. Cienki "Speakers' Gestures and Semantic Analysis," *Cognitive Semantics*, vol. 9, pp. 167–191 2023.
- [18] A. Cienki, C. Müller (eds) *Metaphor and gesture*, Amsterdam, Benjamins 2008.
- [19] A. Kendon, *Gesture: Visible Action as Utterance*. CUP, Cambridge, 2004.
- [20] D. McNeill, F. Quek, K.E. McCullough, S. D. Duncan, N. Furuyama, R. Bryll, N. Furuyama and R. Ansari "Catchments, prosody and discourse," *Gesture*, Volume 1, Issue 1, pp. 9 – 33, 2001.

The influence of gesture presence on the prosodic realization of focus types in the Catalan language

Paula G. Sánchez-Ramón^{1,2}, Frank Kügler², Pilar Prieto^{3,1}

Universitat Pompeu Fabra¹, Goethe University Frankfurt², Institució Catalana de Recerca i Estudis Avançats³

paulaginesa.sanchez@upf.edu

Research has shown that speech and gesture are highly interconnected, and that humans make use of prosodic as well as gestural strategies to convey meaning (e.g. [1]; [2]; [3]). Recent studies have shown that the production of a visual beat (a manual beat gesture, a head nod or a rapid eyebrow movement) affects the acoustic realization of the accompanying speech. For instance, [4] analyzed 10 speakers producing a target sentence in Dutch with different distributions of pitch accents and visual cues and found a longer duration and a lower second formant (F2) in syllables produced with a visual beat, regardless of the presence or position of pitch accents. Also, [5] found that French children's focused words co-occurring with gestures had a longer syllable duration and a wider pitch range compared to focused words produced without gestures. To our knowledge, little is known about the effects of gesture presence in contexts where increasing levels of prosodic prominence are expected due to the expression of different pragmatic domains like distinct focus types. Following the findings by [4] and [5] we expect that the presence of gesture will affect the prosodic realization of focus types, e.g. inducing a higher-than expected duration and pitch range.

The present investigation will assess the effects of gesture presence in a context where different levels of prosodic prominence are driven by pragmatic meaning distinctions, e.g. across focus types ([6], [7]). For prosody, a general increase in prosodic prominence is expected across focus types (e.g., information focus < contrastive focus < corrective focus) in the Catalan language (see [8]; [9]). Our specific aims are the following: (1) to explore which are the prosodic cues in words co-occurring with a gesture involved in the marking of three increasing layers of pragmatic meaning in focus (information focus < contrastive focus < corrective focus); and (2) to explore which are the prosodic cues occurring without an accompanying gesture, involved in the marking of these three increasing layers of pragmatic meaning in focus. Based on [4] and [5], we expect, on the one hand, that gesture presence will produce an augmentation effect in the prosodic cues for focus marking. On the other hand, for the words occurring without gestures, following [8] and [9], we hypothesize that the pragmatically strongest focus types (e.g. contrastive, corrective) will trigger an increase in prosodic prominence, and thus greater duration and pitch range of the target accented syllables.

The method (inspired by [5] and based on [10]) consists of a focus elicitation task in which a total of 35 participants instruct a digital character to take certain objects from a bag. Before the task, participants were asked to use their body to express themselves. The target focused noun phrases (see Figure 1 for an example of the sample dialogue) in the three different focus conditions were elicited in a semi-controlled environment using a set of pictures prompted in the game and by the responses from the digital character. By now, the target focused adjectives (N = 303) produced by 15 participants have been annotated in terms of prosodic prominence levels from no prominence "0" to extra strong prominence "3" (DIMA, [11]). Preliminary results show an expected increase in measures of perceived prosodic prominence across focus types. The Spearman correlation test reveals a significant positive relation between the prosodic prominence ratings and the increased layers of pragmatic meaning ($\rho = 0.29$, $p < 0.0001$) (see Figure 2). Follow-up analyses will include an assessment of the exact acoustic cues that are involved in this prosodic increase of prominence in the whole database, as well as the analysis of the contributing effects of gesture presence and focus type conditions. The complete set of analyses will be presented at the conference.

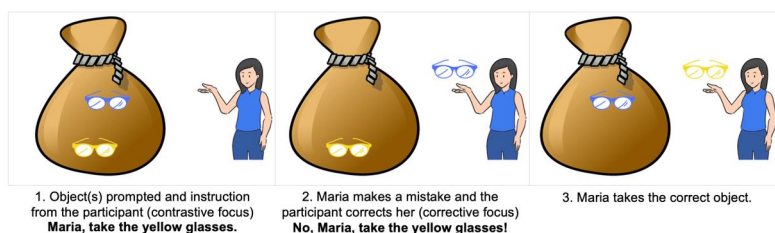


Figure 1: Example trial of the study: contrastive and corrective items. Contrastive item: “Maria, agafa les ulleres [grogues]^F” (Maria, take the [yellow]^F glasses). Corrective item: “No, Maria, agafa les ulleres [grogues]^F!” (No, Maria, take the [yellow]^F glasses!).

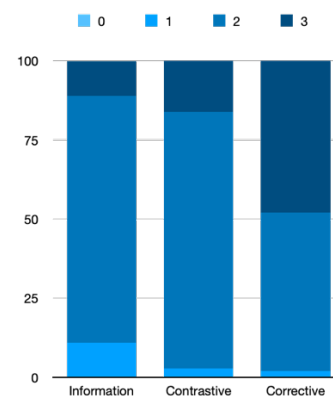


Figure 2: Perceived prosodic prominence ratings (0-3) across focus types.

References

- [1] Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In Key, N. R. (Ed.), *The Relationship of Verbal and Nonverbal Communication*, 25, 207-227. Mouton.
- [2] McNeill, D. (1992). *Hand and mind: what gestures reveal about thought*. University of Chicago Press.
- [3] Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1), 71-89.
- [4] 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.
- [5] Esteve-Gibert, N., Løevenbruck, H., Dohen M. & D'Imperio, M. (2021). Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech. *Developmental Science*, e13154.
- [6] Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3), 243–276.
- [7] Repp, S. (2014). *Contrast: Dissecting an elusive information-structural notion and its role in grammar*. In Caroline Féry & Shinichiro Ishihara (eds.), *OUP Handbook of Information Structure*.
- [8] Vanrell, M. M., Stella, A., Gili-Fivela, B. & Prieto, P. (2013). Prosodic manifestations of the Effort Code in Catalan, Italian and Spanish contrastive focus. *Journal of the International Phonetic Association*, 43, 195–220.
- [9] Estebas-Vilaplana, E. (2009). *The use and realization of accentual focus in Central Catalan with a comparison to English*. Munich: Lincom Europa.
- [10] Gregori, A., Sánchez-Ramón, P., Prieto, P. & Kügler, F. (2023). Prosodic and gestural marking of focus types in Catalan and German. *Proceedings of the 12th International Conference on Speech Prosody*. July 02–05, 2024. University of Leiden, The Netherlands.
- [11] Kügler, F., Baumann, S., and Röhr, C. T. (2022). “Deutsche Intonation, Modellierung und Annotation (DIMA),” *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion*, Tübingen: Narr, 23-54.

Gesture, prosody and syntax as markers of information structure in French

Florence Baills¹ & Stefan Baumann²
Universitat de Lleida¹, Universität zu Köln²
 Florence.baills@udl.cat

Natural languages signal the important elements in speech at various linguistic levels, e.g. by the choice of words and their order [1], prosody [2], and gestures co-occurring with speech [3, 4]. These cues may be used to signal the information structure of a language, for instance, referents which are new in the discourse or constituents that the speaker wants to emphasize - that is, elements which need to be especially noticed by a listener for successful communication. While the syntax-phonology interface has been theorized and exemplified for information structure [5], it remains to be observed through the quantitative analysis of oral speech corpora. Moreover, the role of gesture for IS marking has mostly been examined separately [3], even though there is growing evidence that prosody and gesture work in an integrated manner [4, 6, 7].

In the present study, we aim at giving an exhaustive description of how contrastive focus and information status – the level of givenness or cognitive activation of an expression in discourse – are signaled through head gestures, prosody and syntax by French native speakers in spontaneous speech. French is a language which has been claimed to signal information structure mainly through syntactic transformations such as clefting and dislocations [8], however, there is recent evidence with spontaneous speech data showing that French does mark focus prosodically and sometimes allows given information to be deaccented [9, 10].

Nineteen native French speakers were video-recorded while speaking about their best friend (total phonation time = 19.46 min, mean phonation time = 58 s). They were talking in front of the webcam with the upper part of their body and their face being clearly visible. The sounds were extracted from the video files and orthographically annotated in Praat. In a first step, information status and contrastive focus were annotated following a simplified adaptation of the RefLex system [11]. For prosody, prominence scores were attributed perceptually (DIMA [12]) and independently, a phonetic annotation of the pitch accents was performed [13]. For gestures, only head movements were annotated, as hand gestures were scarce in the corpus. We annotated gestures' strokes and apices following the M3D scheme [14]. In a second step, the annotation of non-canonical (non-SVO) syntactic constructions such as clefts, dislocations, fronting, and passive sentences will be performed following the taxonomy proposed by Brunetti et al. [15].

First results looking at prosodic and gestural marking of information structure show that *contrastive*, *new* and *inferable* information were significantly more prominent and more frequently marked with pitch accents alone or with a combination of pitch accents and head gestures compared to non-contrastive and *given* information (Fig.1). Moreover, head gestures marking referents and focused elements were rarely used without being accompanied by a pitch accent. These results show that even in a language which has been claimed to rely on syntactic cues to express information structure, prosody and head gestures also play a role for distinguishing *new* and *given* expressions. Our findings reinforce the idea of a *multimodal prosody* to mark IS and support previous evidence claiming that the tight synchrony between prosody and gesture arises from a unique multimodal system. Nevertheless, while head gestures generally did not appear on their own, pitch accents certainly did, which may show that pitch accents are the default cue to convey prominence and that head gestures serve as a redundant cue for signaling *new* or important information by adding some extra prominence.

The annotation of the syntactic constructions marking information structure will allow us to present results concerning the types of syntactic structure used to mark *new* vs. *given* information, their frequency of use, and their interplay with gestural and prosodic markers.

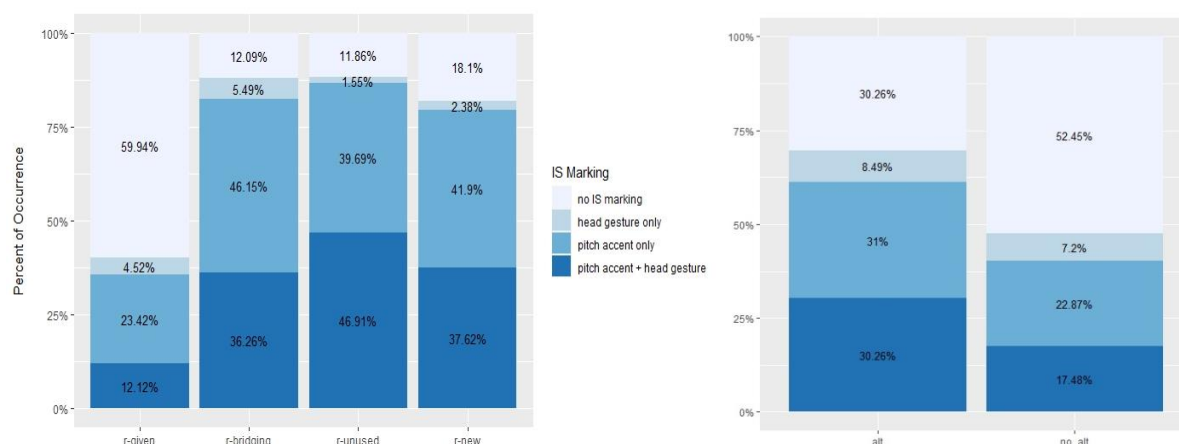


Figure 1: *Prosodic and Gestural Marking of Information Status (left panel) and Contrastive Focus (right panel) in French.*

Note: r-given = given referents, r-bridging = contextually inferable referents, r-unused = new but identifiable referents, r-new = new referents, alt = contrasted expressions, no_alt = non-contrasted

References

- [1] N. Erteschik-Shir, "Information structure: The syntax-discourse interface". Oxford University Press, 2007.
- [2] F. Kügler, & S. Calhoun, S. "Prosodic encoding of information structure: A typological perspective," in C. Gussenhoven & A. Chen (eds.), *The Oxford handbook of language prosody*. Oxford Academic, p. 454–467, 2020. doi: 10.1093/oxfordhb/9780198832232.013.30.
- [3] S. Debreslioska, A. Özyürek, M. Gullberg, & P. Perniss, "Gestural viewpoint signals referent accessibility," *Discourse Processes*, 50, p. 431–456, 2013. doi: 10.1080/0163853X.2013.824286.
- [4] C. Ebert, S. Evert, & K. Wilmes, "Focus marking via gestures," in I. Reich, E. Horch, & D. Pauly (eds.), *Proceedings of Sinn und Bedeutung 15*, p. 193–208, 2011.
- [5] D. Büring, "Syntax, information structure, and prosody," in M. den Dikken (ed.), *The Cambridge Handbook of Generative Syntax*. Cambridge University Press, p. 860–896, 2013.
- [6] S. Im & S. Baumann, "Probabilistic relation between co-speech gestures, pitch accents and information status," *Proceedings of the LSA*, 5, p. 685–697, 2020. doi: 10.3765/plsa.v5i1.4755
- [7] P. Rohrer, "A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system," PhD dissertation, Universität Pompeu Fabra, 2022
- [8] W. Klein, "The information structure of French," in M. Krifka and R. Musan (eds.), *The expression of information structure*. De Gruyter, p. 95–126, 2012.
- [9] C. Féry, "Focus and phrasing in French," in C. Féry & W. Sternefeld (ed.), *Audiatur Vox Sapientiae: A Festschrift for Arnim von Stechow*. Akademie Verlag, p. 153–181, 2001. doi: 10.1515/9783050080116.153.
- [10] M. Dohen & H. Loevenbruck, H. "Pre-focal rephrasing, focal enhancement and postfocal deaccentuation in French," *Proceedings of Interspeech 2004*, p. 785–788, 2004.
- [11] A. Riester & S. Baumann, "The RefLex Scheme – Annotation Guidelines," *SinSpeC, Working Papers of the SFB 732*, vol. 14. University of Stuttgart, 2017.
- [12] F. Kügler, S. Baumann, & C.T. Röhr, "Deutsche Intonation, Modellierung und Annotation (DIMA)," in C. Schwarze & S. Grawunder (eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion*. Narr, p. 23–54, 2022.
- [13] J.I. Hualde & P. Prieto, "Towards an international prosodic alphabet (IPrA)," *Laboratory Phonology*, 7, 5, 2106. doi: 10.5334/labphon.11
- [14] P. Rohrer, U. Tütüncübasi, I. Vilà-Giménez, ... & P. Prieto, "The MultiModal MultiDimensional (M3D) labeling system". doi: 10.17605/OSF.IO/ANKDX
- [15] L. Brunetti, S. Bott, ... & E. Vallduví, "A multilingual annotated corpus for the study of Information Structure 1," *Dritte internationale Konferenz Grammatik und Korpora*, 2009.

Session 2:

Processing of multimodal data

25.09.2024
13:40-15:00



A toolkit for automating co-speech gesture data annotation and analysis

Walter Philip Dych¹, Karee Garvin² and Kathryn Franich²

¹*Binghamton University*, ²*Harvard University*

wdych@binghamton.edu, garvinkaree@gmail.com, kfranich@fas.harvard.edu

Progress in the study of multimodal communication is hindered by a scarcity of tools for automatic coding of co-speech gestures from video data, particularly where naturalistic conversation is concerned. Coding of gestures by hand is time-consuming, and, following best practices, usually requires at least two coders for the establishment of inter-rater reliability [1]. While marker-based motion-capture technologies can be useful for avoiding pitfalls of manual annotation, such systems are often not available for the study of under-documented languages spoken in areas of the world where linguistics labs are not common. Here, we present a set of tools adapted for the automatic coding of co-speech gestures in video data, demonstrating their efficacy in coding video data based on speech samples from several typologically unrelated languages.

Our toolkit includes two workflows: 1) automatic annotation of apexes using manually coded strokes across gesture types; or 2) automatic annotation of movement onset and offset, an interval that comprises the preparation, stroke, hold, and recovery of a gesture [2], in addition to automatic apex annotation, where the apex has been shown to be closely timed to prominence in the speech signal, e.g., stressed syllables [3]–[5]. For both the automatic apex detection and automatic movement boundary detection, the data processing pipeline uses MediaPipe [6] markerless motion capture technology to track the horizontal and vertical movement of the articulators from video inputs to extract keypoints, as shown in Figure 1. The keypoint data is then used to calculate speed and velocity curves, and perform low-distortion signal smoothing using a Savitzky-Golay filter [7]. The apexes can then be automatically annotated as the point of minimum speed within the manually coded stroke boundaries imported from ELAN [8]. This method has been shown to generate apex annotations that closely correspond with manually coded apexes [9].

Alternatively, the processing pipeline allows for automatic movement boundary detection as an alternative to manually coded strokes, tested here the kinematic profile for these out-and-back gestures consists of two dominant peaks in the speed curve. These peaks are detected through peak-prominence and can then be used as an event window to detect and annotate gesture events. The workflow can then annotate either maximum speed or minimum speed within the event window as the apex, depending on the goals of the study. In this demonstration we use minimum speed within the event window, which roughly corresponds with the point of maximum extension, as seen in Figure 2. In a sample of 500 gestures, the movement boundary detection algorithm successfully segmented all gestures identified by human coders. Differences in manual vs automated results typically reflected erroneous inclusion of non-gesture events, e.g., scratch nose, which have a shared kinematic profile with other out-and-back gestures.

We propose a set of next steps for kinematic parsing of addition gestural phases, i.e., preps, stroke, holds, recoveries, and annotation of more complex gestures, such as bimanual cyclic gestures (Figure 3). We conclude by discussing how automated tools for co-speech gesture coding can broaden the range of languages for which efficient multimodal data analysis can be possible.

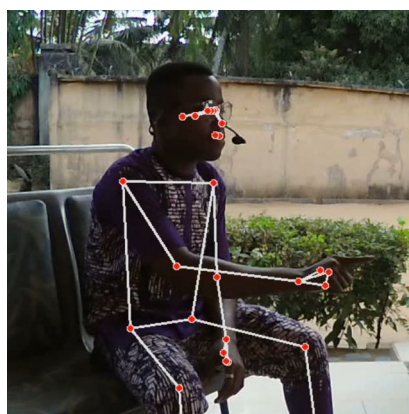


Figure 1: MediaPipe Keypoints

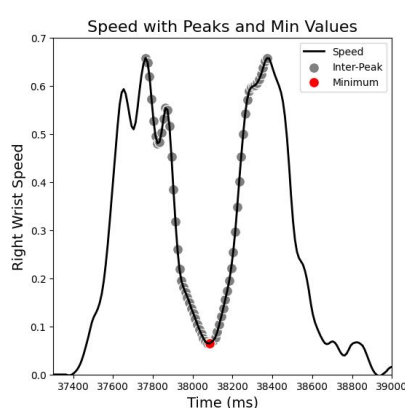


Figure 2: Gesture Event Speed Curve



Figure 3: Cyclic-Beat Gesture

References

- [1] Speech Communication Group, *Scg gesture coding manual*. [Online]. Available: <http://scg.mit.edu/gesture/coding-manual.html>.
- [2] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, DE GRUYTER MOUTON, Dec. 1980, pp. 207–228.
- [3] K. Franich and H. Keupdjio, "The Influence of Tone on the Alignment of Speech and Co-Speech Gesture," in *Proc. Speech Prosody 2022*, 2022, pp. 307–311. DOI: 10.21437/SpeechProsody.2022-63.
- [4] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, 2012. DOI: 10.1515/lp-2012-0006. [Online]. Available: <https://doi.org/10.1515/lp-2012-0006>.
- [5] A. Watanabe and Y. Hirose, "Correlates between prosodic features and manual gesture in japanese spontaneous speech," *Language and Speech*, vol. 58, no. 2, pp. 225–243, 2015.
- [6] C. Lugaresi, J. Tang, H. Nash, *et al.*, "Mediapipe: A framework for building perception pipelines," 2019. DOI: 10.48550/ARXIV.1906.08172.
- [7] T. Haslwanter, "Data filtering," en, in *Hands-on Signal Analysis with Python*. Cham: Springer International Publishing, 2021, pp. 71–104, ISBN: 9783030579029. DOI: 10.1007/978-3-030-57903-6_5.
- [8] Max Planck Institute for Psycholinguistics, *Elan*, version 6.3, 2022. [Online]. Available: <https://archive.mpi.nl/tla/elan/download>.
- [9] W. Dych, K. Garvin, and K. Franich, "Comparing manual vs. semi-automated methods for the coding of co-speech gestures," in *Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences*, Guarant International, 2023.

Introducing *DiCE*: A novel approach to elicit and capture multimodal accommodation via 3D electromagnetic articulography, audio, and video

Lena Pagel¹, Simon Roessig², Doris Mücke¹

¹University of Cologne, Germany, ²University of York, United Kingdom

lena.pagel@uni-koeln.de

When engaging in conversation, interlocutors frequently accommodate to each other in their speech patterns and co-speech movements. This phenomenon has been observed in various domains, including facial expression [1], manual gestures [2], intonation [3], and supra-laryngeal articulation [4, 5, 6]. However, only a few studies have, so far, investigated both the visual and auditory modalities at the same time (but cf. [1, 7, 8, 9]). Additionally, a challenge for experimental research is to account for the effect of information structure (among others, focus), which not only influences the production of speech and co-speech motion within a speaker but can also affect the patterns of accommodation between speakers [4]. Due to the increased complexity, information structure is frequently not thoroughly considered in studies on interpersonal accommodation, and the question of how speakers (multimodally) accommodate to each other *in their patterns of focus structure marking* remains an open area for investigation.

We present a novel methodological approach to elicit and record multimodal accommodation using a setup with audio, video, and dual electromagnetic articulography (dual EMA, directly tracking 3D movements with a high temporal and spatial resolution). We introduce the cooperative game *DiCE* (*Dialogic Collecting Expedition*) to elicit data with controlled focus structure, and provide information on the recording setup, procedure, and technical details. We have successfully applied the methodological approach in recordings of 15 German-speaking dyads, which, to our knowledge, forms the largest existing corpus of dual EMA recordings.

Each recording session involves two speakers, who are initially recorded in a solo condition individually and then in a dialogue condition as a dyad. The card game *DiCE* (available at <https://osf.io/9fmqh/>) is designed to elicit the production of lexically and prosodically controlled utterances and co-speech movements in an engaging setting. Speakers collaborate to collect cards and interact with each other in question-answer sets. The question posed by one participant prompts the focus structure of the answer given by the other participant, in a way that pre-defined target words are produced either in corrective focus or in the background. Additionally, speakers produce pointing gestures to indicate the location of the intended card.

Speakers are recorded with dual 3D EMA (one articulograph per speaker, each with 16 sensors attached to capture speech and co-speech kinematics), head-mounted microphones (one per speaker), and three video cameras (one per speaker from the front plus one from the side). We will present practical information on the technical recording setup and synchronisation of the various signal streams. The combination of dual EMA, audio, and video enables future analyses within dynamical approaches regarding (i) acoustic speech cues (audio signal: F0, intensity, spectral properties of consonant and vowel productions), (ii) vocal tract kinematics (EMA signal: lip aperture and spreading, jaw, tongue tip, and tongue body movements), (iii) kinematics of co-speech body movements (EMA signal: head motion, eyebrow raising and furrowing, torso and shoulder movement; video signal: facial expression, pointing gestures), and (iv) kinematics of smiles and breathing (EMA signal: lip spreading, torso expansion; video signal: smiles). The recorded multimodal data from our corpus are illustrated in Figure 1 for one question-answer set of one dyad in the dialogue condition. Four parameters of speech and co-speech kinematics (lip aperture, vertical tongue, head, and eyebrow motion) are selected from the wide range of possible parameters to showcase the nature of the recorded data.

With this contribution and the presentation of *DiCE*, we aim to provide valuable insights and materials for future recordings with the goal of capturing multimodal dyadic accommodation with controlled (co-)speech material.

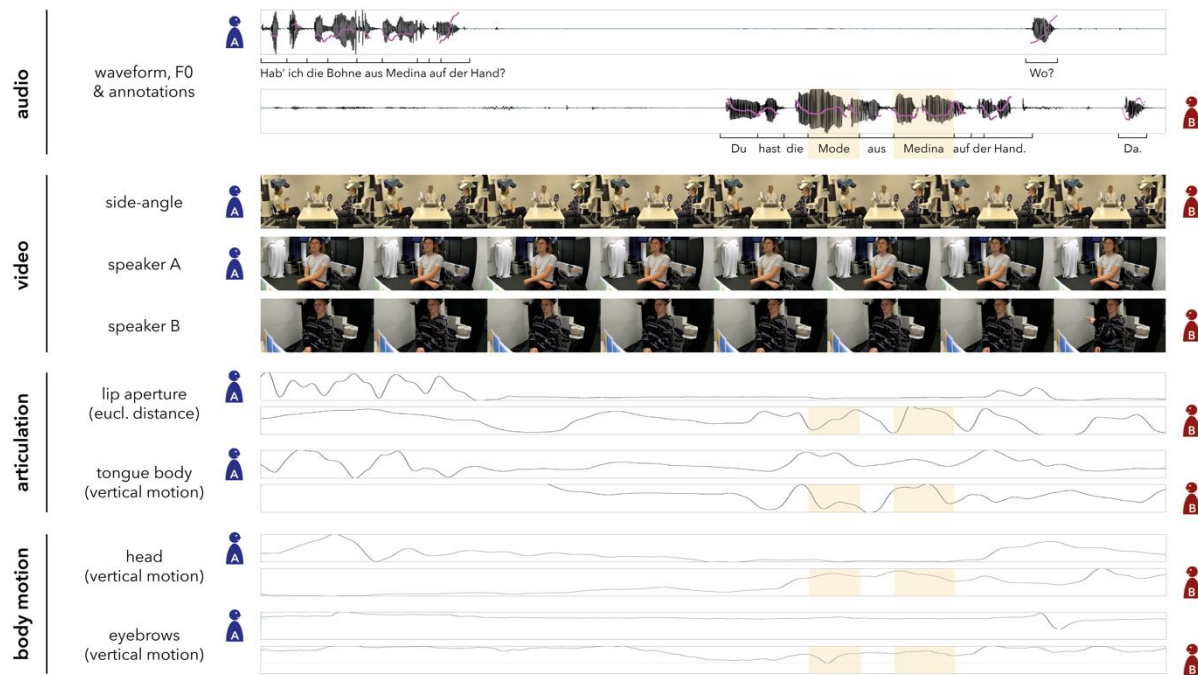


Figure 1: *Example of recorded multimodal data for selected parameters during one question-answer set by two speakers (A and B). Target words are indicated by yellow rectangles. Based on the question of speaker A, the answer of speaker B has a focus structure with the object in corrective focus and the city in the background.*

References

- [1] M. M. Louwerse, R. Dale, E. G. Bard and P. Jeuniaux, "Behavior matching in multimodal communication is synchronized," *Cognitive Science*, vol. 36, pp. 1404–1426, 2012. doi: 10.1111/j.1551-6709.2012.01269.x.
- [2] L. Mol, E. Krahmer, A. Maes and M. Swerts, "Adaptation in gesture: Converging hands or converging minds?," *Journal of Memory and Language*, vol. 66, pp. 249–264, 2012. doi: 10.1016/j.jml.2011.07.004.
- [3] M. Babel and D. Bulatov, "The Role of Fundamental Frequency in Phonetic Accommodation," *Language and Speech*, vol. 55, no. 2, pp. 231–248, 2012. doi: 10.1016/j.jml.2011.07.004.
- [4] Y. Lee, S. G. Danner, B. Parrell, S. Lee, L. Goldstein and D. Byrd, "Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task," *PLoS ONE*, vol. 13, no. 8, e0201444, pp. 1–26, 2018. doi: 10.1371/journal.pone.0201444.
- [5] S. Mukherjee, T. Legou, L. Lancia, P. Hilt, A. Tomassini, L. Fadiga, A. D'Ausilio, L. Badino and N. Nguyen, "Analyzing vocal tract movements during speech accommodation," *Proceedings of INTERSPEECH*, 2-6 September, Hyderabad, India, pp. 561–565, 2018. doi: 10.21437/Interspeech.2018-2084.
- [6] M. Tiede and C. Mooshammer, "Evidence for an articulatory component of phonetic convergence from dual electromagnetic articulometer observation of interacting talkers," *Proceedings of Meetings on Acoustics*, 2-7 June, Montreal, Canada, 3aSCa3, 2013. doi: 10.1121/1.4799497.
- [7] N. D. Duran and R. Fusaroli, "Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement," *PLoS ONE*, vol. 12, no. 6, e0178140, pp. 1–25, 2017. doi: 10.1371/journal.pone.0178140.
- [8] B. Oben, and G. Brône, "Explaining interactive alignment: A multimodal and multifactorial account," *Journal of Pragmatics*, vol. 104, pp. 32–51, 2016. doi: 10.1016/j.pragma.2016.07.002.
- [9] M. Tiede, R. Bundgaard-Nielsen, C. Kroos, G. Gibert, V. Attina, B. Kasisopa, E. Vatikiotis-Bateson and C. Best, "Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously," *Proceedings of Meetings on Acoustics*, 15-19 November, Cancun, Mexico, 4pSC10, 2010. doi: 10.1121/1.4772388.

Krajjat: A Python Toolbox for Analysing Body Movement and Investigating its Relationship with Speech

Romain Pastureau^{1,2}, Nicola Molinaro^{1,3}

1. Basque Center on Cognition, Brain and Language (BCBL), San Sebastián, Spain · 2.

Universidad del País Vasco/Euskal Herriko Unibertsitatea, San Sebastián, Spain · 3.

Ikerbasque, Basque Foundation for Science

r.pastureau@bcbl.eu · n.molinaro@bcbl.eu

Historically, the study of the production of co-speech gestures has been heavily dependent on the advances of technology. Most of the influential studies on the matter from the 20th century took advantage of video recordings to be able to dissect, categorize and quantify the movements of gesturing speakers [1, 2, 3]. The motion tracking technologies enable an even deeper exploration of this research area. While full motion capture sets are not readily available for everyone, in recent years, more portable devices became more accessible, and more affordable, via dedicated devices such as Microsoft Kinect, Intel RealSense or Orbbec Astra, or even via toolboxes such as OpenPose that can process regular video recordings to extract the joint positions from each frame. These systems offer a novel way to study the movements produced in parallel to speech, as they allow to extract measurements of the position and velocity of various parts of the body. However, despite the increased accessibility to the hardware, there is a lack of tools to help in the processing of the output motion tracking data.

Moreover, the use of automatic body movement tracking is particularly relevant in consideration with the widespread idea in the field of gesture studies stating that speech and gestures rely on common processes [4]: for example, both speech and gesture are subject to the Lombard effect in noise [5]; gestures are reliably produced during fluent speech in comparison to disfluent speech [6]. We can even push the research even farther by studying the gestures kinematics in the frequency domain, or in other words, the relation between the rhythmicity of body movements and speech – a connection that starts to raise interest in the literature [7, 8].

With Krajjat, we introduce a toolbox that offers the streamlining of parts of the processing of 3D motion capture data. The toolbox, taking the form of a Python module, makes it easily installable and usable on most machines. We designed it to be compatible with a wide range of inputs: while the toolbox natively accepts data collected with Kinect or Qualisys systems, it can also accept tables containing the timestamps and three-dimensional values or labelled joints or markers. The purposes of the toolbox are designed around three core functions. The pre-processing functions allow to smooth out the jitter and artifacts appearing with automatic joint position detection, to resample the data, and to re-reference the positions of the joints. The visualisation functions allow to observe the recorded data, and to compare raw and pre-processed data visually (see Figure 1). Finally, the analysis functions contain several signal processing methods (e.g. correlation, coherence, and principal component analysis) allowing to study the relationships between the kinematics of the joints and the acoustics of the speech. While still being a work in progress, we hope that the release of this toolbox will prove useful in the gesture research community.

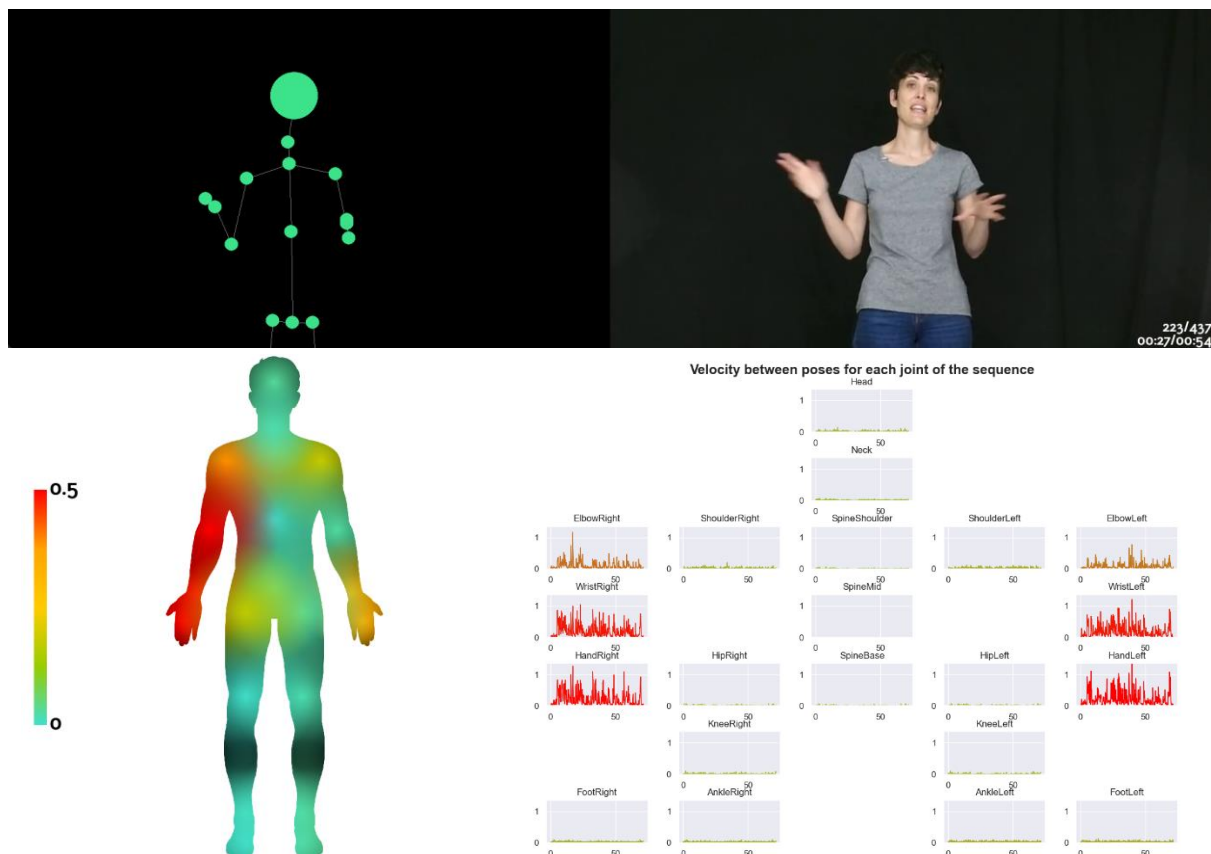


Figure 1: Three visualisation outputs from the Krajjat toolbox display functions. Top: side-by-side comparison of the motion tracking (left) and video (right) recorded by a Kinect camera. Bottom left: example of a “silhouette” visualisation with customizable values and colours for each of the joints. Bottom right: plotting of the velocity across time of each of the joints of a motion capture sequence.

References

- [1] Efron, D. (1941/1972). *Gestures, Race And Culture*. [First edition 1941 as *Gestures and environment*. New York: King’s Crown Press.ed.]. The Hague: Mouton.
- [2] McNeill, D. (1985). So you think gestures are nonverbal? *Psychol. Rev.* 92, 271–295.
- [3] Kendon, A. (1980). “Gesture and speech: two aspects of the process of utterance,” in *Nonverbal Communication and Language*, ed. M. R. Key (The Hague: Mouton), 207–227.
- [4] Krauss, R.M., Chen, Y., Gottesman, R.F. (2010). Lexical gestures and lexical access: a process model. *Lang Gesture*. Published online 2010:261-283. doi:10.1017/cbo9780511620850.017
- [5] Trujillo, J., Özyürek, A., Holler, J., Drijvers, L. (2021) Speakers exhibit a multimodal Lombard effect in noise. *Sci Rep.* 2021;11(1):1-12. doi:10.1038/s41598-021-95791-0
- [6] Graziano, M., Gullberg, M. (2018) When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Front Psychol.* 2018;9(JUN):1-17. doi:10.3389/fpsyg.2018.00879
- [7] Pouw, W., Trujillo, J.P., Dixon, J.A. (2020) The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behav Res Methods.* 2020;52(2):723-740. doi:10.3758/s13428-019-01271-9
- [8] Alviar, C., Dale, R., & Galati, A. (2019). Complex Communication Dynamics: Exploring the Structure of an Academic Talk. *Cognitive science*, 43(3), e12718. <https://doi.org/10.1111/cogs.12718>

MOBILE MULTIMODAL LAB

An Open-Source, Low-Cost and Portable Laboratory for the study of Multimodal Human Behavior

MMSYM 2024

Davide Ahmar¹, Šárka Kadavá^{1,2}, Wim Pouw¹

¹*Donders Institute for Brain, Cognition and Behavior*

²*Leibniz Centre for General Linguistics*

davide.ahmar@ru.nl

Interpersonal communication is inherently multimodal. This multimodality is obvious during our face-to-face encounters, where the complex exchanges of auditory, visual and (sometimes) somatosensory information allow us to successfully navigate intricate social situations [1-2]. Yet, due to the methodological complexity of multimodal research, most empirical investigations into human behaviour have relied on unimodal approaches until not so long ago [3]. Today, the main challenges preventing new developments in multimodal research entail setting-up, collecting and analyzing time varying signals with their own modality-characteristic differences that require careful integration and synchronization [4].

To address these challenges, we are creating a practical manual accompanied by a comprehensive coding library that will enable researchers to build a Mobile Multimodal Lab (MML). The guiding principles behind the MML project are threefold: using (i) open-source resources, researchers will be able to independently assemble a fully functional laboratory that is both (ii) low in monetary cost (i.e., less than 10K) and (iii) easily transportable across testing location to capture multimodal behaviors in a vast range of experimental settings.

Specifically, the manual will contain a step-by-step tutorial in setting up the MML using multiple frame-synced 2D videos (e.g., for 3D motion-tracking), audios (e.g., for prosodic or semantic analyses), and physiological signals (e.g., electrocardiogram, electromyography, and respiration) to record either individual participants or multiple interactants. The setup of the MML is modular, meaning that other measuring systems (such as electroencephalograms or eye-tracking devices) can also be incorporated with the above-mentioned recordings. To integrate all these recordings, the MML uses the Lab Streaming Layer (LSL, <https://github.com/scen/labstreaminglayer>), an open-sourced, networked middleware that allows to synchronize the different data streams with sub-millisecond precision, thus easing the process of centralized data collection (see Figure 1).

To demonstrate the potentialities of MML, the manual will also include a proof-of-concept experiment in which two participants engage in an interpersonal singing task whilst multiple videos, audios and physiological signals are recorded (Figure 1). The accompanying coding library will contain all the necessary programming steps employed in this experiment, from the experimental setup and synchronization of multiple recordings to the preprocessing and analysis pipelines of multidimensional timeseries data.

Thanks to its open-source nature, low-cost of materials, ease of transportation and modularity of recordings, we are confident that the MML project can provide valuable support for any researcher interested in capturing the full bandwidth of human behaviors in an ecologically valid, multimodal framework. We believe that the Multimodal Communication Symposium represents a perfect opportunity for us to introduce this MML and engage with researchers who can contribute to the next steps of this project.

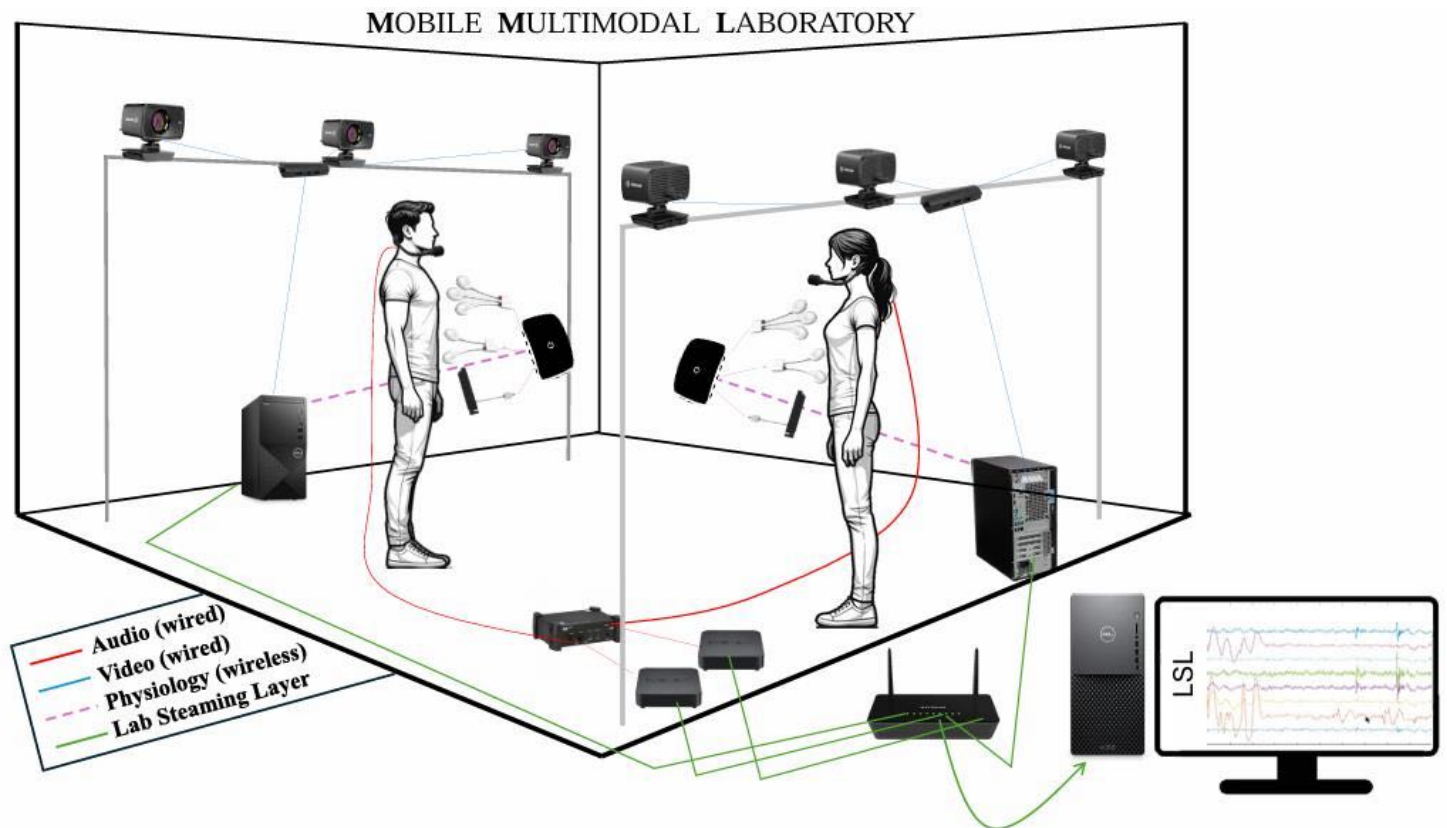


Figure 1. *The Mobile Multimodal Lab (MML)*. The figure shows the original setup of the MML employed in our proof-of-concept experiment. Two interactants are facing each other. Synchronous multimodal recordings are made using the Lab Streaming Layer (LSL, green). Audio (red): each interactant is wearing a cheek microphone, which feeds to an amplifier and Linux device before streaming to the LSL. Video (blue): each interactant is recorded by three arch-mounted cameras, feeding their frame-synced videos to a Windows PC, which then streams the three videos to the LSL. Physiology (purple): each interactant is wearing electrocardiogram (ECG), electromyography (EMG) and respiration (RSP) sensors, which send their data wirelessly (Bluetooth) to the PCs, finally streaming to the LSL.

References

- [1] I. Poggi. "Mind, hands, face and body". A goal and belief view of multimodal communication. Weidler, Berlin. 2007.
- [2] S.C. Levinson, and J., Holler. "The origin of human multi-modal communication." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), pp. 20130302. 2014.
- [3] C. Jewitt, J. Bezemer and K. O'Halloran. "Introducing multimodality." Routledge. 2016.
- [4] J. Bateman, J. Wildfeuer and T. Hiippala. "Multimodality: Foundations, research and analysis—A problem-oriented introduction". Walter de Gruyter GmbH & Co KG. 2017.

Postersession 1:

—

25.09.2024
15:00-16:20



Automatic Reconstruction of Dialogue Participants' Coordinating Gaze Behavior from Multiple Camera Perspectives

Alina Naomi Riechmann and Hendrik Buschmeier

Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University
hbuschme@uni-bielefeld.de

Gaze is an important modality in human face-to-face interaction with coordinative (e.g., turn-taking, feedback) and referential (e.g., communication of attention, deixis) functions. Access to gaze direction and gaze targets of interaction partners is important for research on human multimodal communication as well as for computational modeling of gaze behavior, e.g., in human-agent interaction. Reliable recording of gaze behavior in interaction is usually done with dedicated eye-tracking hardware. However, this can interfere with the natural movements and behaviors of interlocutors and may not be available for existing video-recorded interaction data.

This abstract presents an investigation into the feasibility of using vision-based eye tracking – based on the OpenFace software [1] – on ordinary video recordings, as a low-impact method for measuring gaze in human interaction data [2]. We developed the approach alongside the creation of a corpus of video recordings [MUNDEX; 3] of conversations in which one participant (the ‘explainer’, ER) explains a board game to another (the ‘examinee’, EE). The interactions were filmed with several cameras from different angles. In the experimental setup, the two dialogue partners sit across from each other at a table, with a camera behind each of them (C1 and C2 in Figure 1–A/B), filming from over-the-shoulder perspectives.

The basic idea of the proposed method is to automatically map the OpenFace-estimated gaze direction of the interlocutor shown in one camera perspective (e.g., gaze of EE in C2) to the target pixel location in the over-the-shoulder perspective of the other camera (e.g., C1; see Figure 1–A). For this to be possible, (i) each interlocutor needs to be recorded so that their face is visible, (ii) the camera recording the other interlocutor must be visible as a reference point in this video, and (iii) the gaze direction must be calibrated to a known target (ideally via a procedure).

A first task to test the approach is to automatically identify the intervals in which the EE looks at the ER (more precisely, at their face) and vice versa. As the interlocutors move during the interaction, the location of their face is also tracked, allowing the dynamic computation of gaze-at-interlocutor intervals based on gaze target and face position at each point in time. In addition, mutual gaze intervals are trivially derived by intersection. The videos and the generated annotations are automatically combined into an ELAN file [4] for easier visualization (see ELAN-tiers in Figure 1A) and further (manual) processing, and/or analysis.

While the method seems promising in general, the results are lacking in the current setup (Figure 1–C/D). The method could become viable if some aspects, especially gaze calibration (to get reference points for analysis) and camera positioning (to optimize the visibility of facial movements for OpenFace), were improved. Currently, to get a second reference point for calibration, participants in the interaction study were simply asked to look at each other at the beginning of the interaction (marked red in Figure 1–C). As an improvement, more specific points for calibration could be added, such as a fixed point visible in the environment or on the participants themselves (e.g., shoulders or hands).

Acknowledgments This research was supported by the German Research Foundation (DFG) in the Collaborative Research Center TRR 318/1 2021 ‘Constructing Explainability’ (438445824).

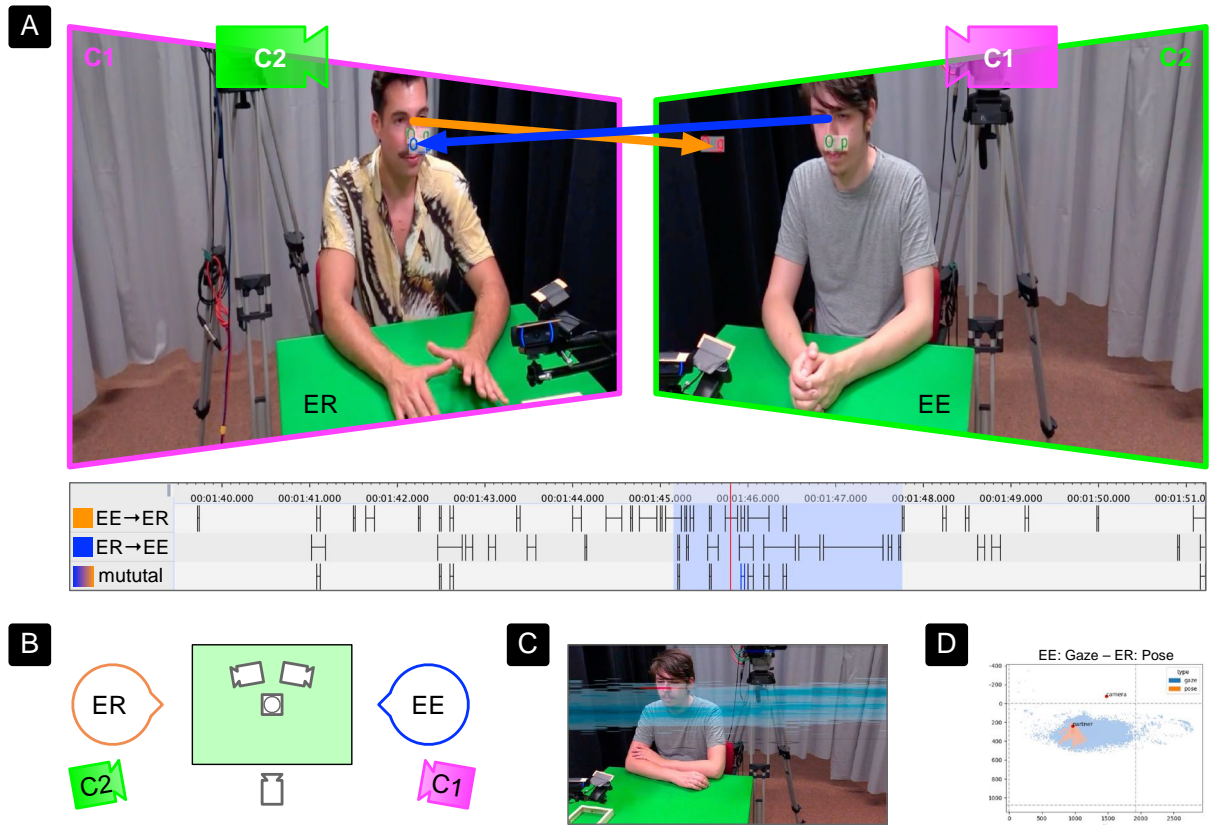


Figure 1: Gaze estimation of the interaction between an explainer (ER, left) and an explaine (EE, right). **A** EE's and ER's estimated gaze directions (blue and orange arrows) mapped onto the video frame showing their respective interlocutor, as well as corresponding ELAN tiers (where intervals indicate gaze on the tracked partner's face and mutual gaze). **B** Schematic of the experimental setup. The two interlocutors sit at a table facing each other and are recorded from six camera perspectives. Camera C1 (recording ER from over EE's shoulder) and camera C2 (recording EE from over ER's shoulder) are used for gaze estimation (the perspectives of the other cameras are not relevant here). **C** Heatmap of ER's gaze targets over time (blue) and gaze during calibration (red) mapped to camera C2's perspective. **D** Mapping of EE's tracked gaze and ER's tracked pose (i.e. head position) over time.

References

- [1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, Xi'an, China, 2018, pp. 59–66. doi: 10.1109/FG.2018.00019.
- [2] A. N. Riechmann, "Vision based reconstruction of dialogue participants' coordinating gaze behaviour from multiple camera perspectives," Master's Thesis, Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany, 2023.
- [3] O. Türk, P. Wagner, H. Buschmeier, A. Grimminger, Y. Wang, and S. Lazarov, "MUNDEX: A multimodal corpus for the study of the understanding of explanations," in *1st International Multimodal Communication Symposium*, Barcelona, Spain, 2023, pp. 63–64.
- [4] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research," in *Proceedings the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006, pp. 1556–1559.

The role of synchronization in face-to-face communication: A dual eye-tracking and motion capture study

Luca Béres^{1, 2}, Ádám Boncz¹, Péter Nagy¹, István Winkler¹

¹*HUN-REN Research Centre for Natural Sciences, Budapest, Hungary*

²*Budapest University of Technology and Economics, Department of Cognitive Science, Budapest, Hungary*

Corresponding author: Luca Béres, email: beres.luca@ttk.hu

Face-to-face communicative actions (e.g., speech, facial displays, gestures, etc.) by themselves are almost always ambiguous [1], [2], yet, interlocutors readily resolve the challenges of coordinating meaning and building a shared understanding during everyday interactions. How is this coordination achieved? There is ample evidence showing that people tend to mimic one another in interactions, specifically, they synchronize their movements, gaze direction and prosody (e.g., [3], [4]). Approaches focusing on shared conceptual space [5], [6] have assumed that (mostly verbal) behavioral synchrony indicates the extent of conceptual alignment and, thus, predicts communication success. However, there is conflicting evidence for a positive link between synchronized (verbal or nonverbal) behavior and successful communication (see e.g., [7]), questioning the explanatory power of the approach. Additionally, the notion of shared conceptual space does not address the richness of cues and modalities in face-to-face communication as it does not differentiate between communication channels, excluding a potentially important aspect of everyday interactions. Therefore, the goal of the current study was to 1) explore interpersonal synchronization [IS] across multiple aspects of behavior during a naturalistic face-to-face communicative setting, and 2) assess the putative (positive) relationship between IS and communication success.

In a series of experiments, we asked pairs of participants to solve a computer-mediated communicative task (“Bargaining Game” [BG]) involving negotiations, while head motion, gaze direction, pupil size, audio and video were collected from both participants. The BG was designed to elicit naturalistic face-to-face conversations (participants communicated via audio and video), while enabling the measure of communication success through task performance (see figure 1). IS of head motion (squared velocity) and pupil size was calculated using cross wavelet coherence, while IS of prosodic features (e.g., speech rate, vocal intensity, pitch) was computed using sliding window correlations. IS of gaze direction was defined as the amount of time participants spent fixating on the same areas of the computer screen. Significance of IS was estimated by comparisons against pseudo pairs (random pairings of participants from different pairs).

Our results indicate that pairs of participants (N=119) synchronized their behavior on multiple levels while communicating: IS of head movement, gaze direction, pupil size and several prosodic features were greater for real pairs compared to pseudo pairs. However, linear-mixed models revealed that only synchrony in terms of gaze direction was predictive of communication success in the BG, which can be partly explained by the task constraints.

61.611062

S H U V R Q ¶

we
that

Tokens marked with red squares are so-called „must-have„ items
tokens offered appear on the bottom, along with a button marked „Exchange„ and „End game„:
tokens and agree on prices. Own tokens are displayed on the right and left sides of the screen
streams (using full HD video cameras). Their task is to negotiate the exchange of available
communicating via low-latency audio (using headset microphones and loudspeakers) and video
Figure 1: Experimental setup of the BC. Participants are seated in sound-proof rooms and are

conversational dynamic could facilitate cooperation.

Spatial Narratives from Remote and Recent Memory in Individuals with Alzheimer's Disease and Healthy Older Adults: A Multimodal and Kinematic Perspective

Sharice Clough^{1,2} Beyza Sümer^{1,3} Kristel de Laat^{1,3} Annick Tanguay² Sarah Brown-Schmidt²
Melissa C. Duff² Aslı Özyürek¹

MPI for Psycholinguistics¹ Vanderbilt University Medical Center² University of Amsterdam³
sharice.clough@mpi.nl

Iconic gestures reflect properties of the visual world and our experience navigating space. Gesture conveys information holistically and simultaneously, offering unique affordances for communicating visuospatial relations. These gestures are posited to arise from mental simulations of motor and perceptual imagery that are directly related to the richness and integrity of a speaker's memory representations. To test this hypothesis, we examined the speech and gesture of healthy older adults (HA) and individuals with Alzheimer's disease (AD) in spatial narratives. Participants described the layout of their childhood bedroom (remote memory condition) and current bedroom (recent memory condition). HA and AD individuals exhibit opposite patterns of memory degradation: Whereas AD is characterized by a temporally graded memory loss with remote memories more intact than recent memories, in healthy aging, recent memory is better preserved than remote memory. Thus, we expect speech and gesture production to mimic these patterns of memory degradation for recent/remote memories.

Participants' spatial narratives were transcribed and imported into ELAN annotation tool, and co-speech gesture strokes were segmented. Data coding for this dataset is ongoing. Spatial content and visuospatial details of the narratives are coded in each modality separately. For example, in Figure 1, the participant produces an iconic gesture depicting two lamps with a raised index finger on each hand while simultaneously saying, "On top of the dressers, I have two tall lamps." Whereas some details are conveyed in both the speech and gesture modalities (e.g., *quantity* is expressed by "two" in speech and the depiction of lamp with both index fingers in gesture), other details are produced in speech only (e.g., *size* is expressed through the word "tall") or gesture only (e.g., *shape* is depicted by the upright pointed fingers). We also extracted motion tracking data to examine the kinematic properties of the gestures participants produced using key point estimation via OpenPose (Fig 1).

We report preliminary results from five participants with AD and five HA participants. As we predicted, AD participants were significantly less likely to produce visuospatial details in speech (Fig 2) than HA participants ($\hat{\beta}=-0.87$, $z=-3.45$, $p<.001$). HA participants were significantly more likely to produce visuospatial details when describing spatial layouts from recent compared to remote memory ($\hat{\beta}=0.69$, $z=3.75$, $p<.001$). A marginal group*condition interaction indicated that this effect was absent in the AD group ($\hat{\beta}=-0.50$, $z=-1.83$, $p=.07$). AD participants were significantly less likely to produce representative gestures (Fig 3) than HA participants ($\hat{\beta}=-1.29$, $z=-2.20$, $p=.03$). HA participants were significantly more likely to produce representative gestures when describing spatial layouts from recent compared to remote memory ($\hat{\beta}=0.89$, $z=2.09$, $p=.04$). The group*condition interaction was not significant ($\hat{\beta}=-1.01$, $z=-1.55$, $p=.12$).

Our findings suggest the memory degradation is linked to impoverished spatial narratives across modalities. This was evidenced by fewer visuospatial details and gestures in the AD compared to HA group and in the remote compared to recent memory condition within the HA group. Ongoing analyses examine differences in how the two groups utilize physical space in their spatial narratives using motion tracking (Fig 4), providing novel insights into the influence of memory on spatial communication in healthy and disordered aging.

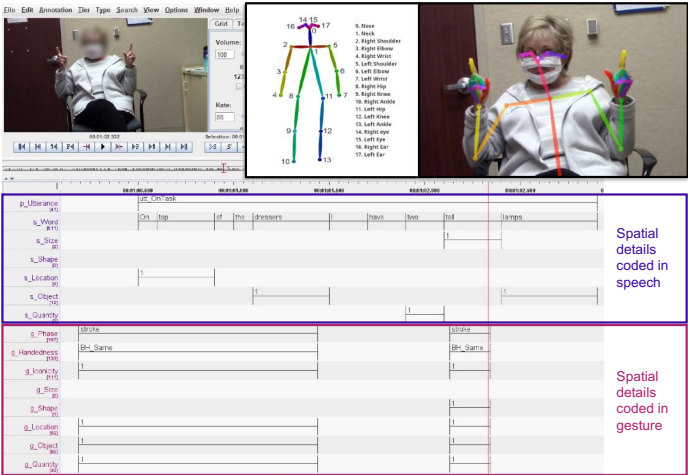


Figure 1. Example of qualitative coding of speech and gesture using ELAN annotation tool and quantitative motion tracking of gesture kinematics using OpenPose.

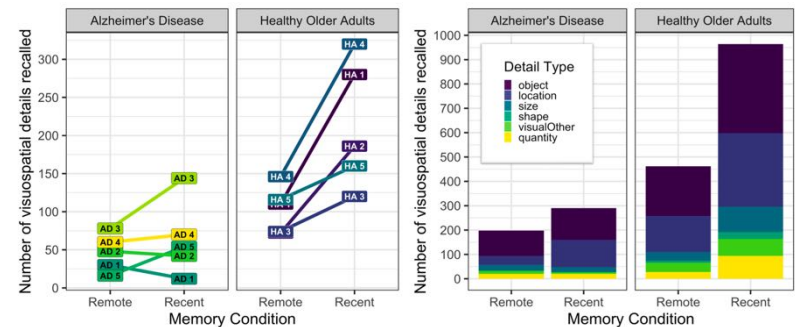


Figure 2. Frequency and type of visuospatial details produced in speech by group and condition

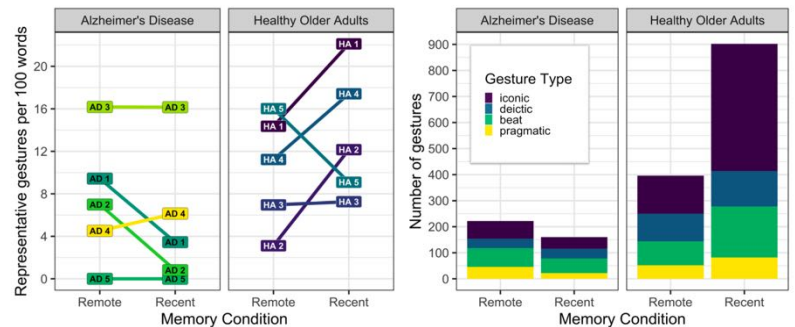


Figure 3. Frequency and type of gestures by group and condition

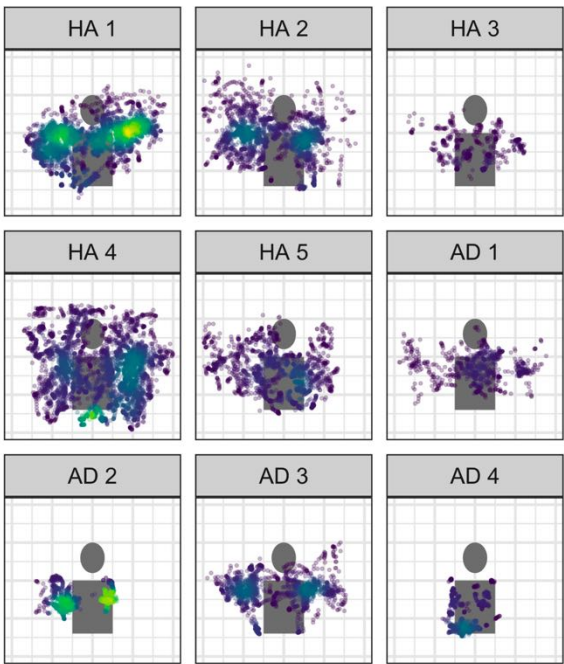


Figure 4. Gesture space used by participants based on coordinate locations of left and right key points during production of all gesture strokes. Participant AD 5 did not produce any utterances with gestures and is excluded from the plot.

Kinematic gestural evidence for higher-level prosodic constituents in speech

Stefanie Shattuck-Hufnagel, *MIT RLE Speech Communications Group*, sshuf@mit.edu

Ada Ren-Mitchell, *MIT Media Lab*, adarm@mit.edu

It has been proposed that sequences of co-speech gestures can be organized into higher-level constituents, whose boundaries are marked by kinematic changes, such as the choice of articulator, or the shape of the movement path [1, 2, 3]. In addition, these sequences, which Kendon called Gesture Units, are proposed to be marked by a gesture recovery phase at the end of the unit, in which the manual articulator returns to the onset position of the preparation phase. In this study we begin to address the question of whether such Gesture Units also align with other potential kinematic markers of higher-level constituents in the speech stream that they accompany. The kinematic marker of interest here is the presence of a rest phase following the recovery phase, i.e. a time interval characterized by the absence of intentional movement, in which the manual articulator returns not just to the onset position of the Gesture Unit, but to the speaker's preferred 'rest' position, and remains there for some time. (This rest phase is sharply distinguished from a post-stroke hold, in which there is no preceding recovery phase and the manual articulator sustains the intentional hand shape attained at the end of the stroke.)

We focus on the rest phase as a potential kinematic marker of a higher order gesture constituent boundary, in a 15-minute sample of academic-lecture-style speech from an American English single speaker. This speaking style was selected because it provides some evidence of higher order structure: longer pause durations evidently contribute to the perception of higher-level constituent boundaries in the speech signal (Figure 1). This allows us to ask whether the presence of a rest phase aligns with these longer pause durations, as a boundary cue to higher-level constituents in a multimodal speech-gesture stream. In the 15-minute sample, we analysed 698 individual Stroke-Defined Gestures (SDGs) (similar to Kendon's Gesture Phrases).

We investigated whether the locations with longer silences in the speech stream are also marked by the kinematic cue of a rest position in the gesture stream. The speech sample was annotated for Full Intonational Phrase Boundaries (4-breaks in the ToBI system), without access to the video, and the video sample was annotated for gesture phases of the SDGs (including Full and Partial Rest Phases), without access to the sound. Combining the spoken prosodic and gestural annotations, of the 698 SDGs, 231 were the final SDGs in ToBI Full Intonational Phrases. For this subset of 231 Intonational-Phrase-Final SDGs, when a Full Intonational Phrase is followed by a longer silence duration (greater than 500 milliseconds), it is more likely to be accompanied by a gestural rest than when it is followed by shorter silence duration (under 500 milliseconds; Figure 2). That is, the presence of a rest phase in the gesture stream aligns with the presence of a pause duration cue to a higher constituent boundary in the speech stream.

This preliminary finding for one speaker is consistent with the view that both the gesture stream and the speech stream are prosodically structured [3, 10], and as many researchers have noted, that the prosodic structures of the two streams are coordinated in time [4, 5, 6, 7, 8]. It also raises the question of whether other kinematic markers in the gesture stream, such as changes in the shape of the movement path, hand shape, and articulator choice (e.g. handedness) might also align with these boundaries. Finally, it suggests the value of exploring the possibility that gestural sequences are organized into a prosodic hierarchy, just as spoken utterances are. We are currently investigating these possibilities.

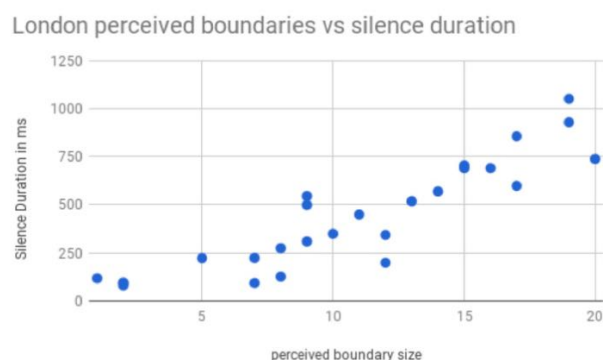


Figure 1: Increasing size of perceived spoken prosodic boundary in a sample “London” with increasing silence duration (from Shattuck-Hufnagel & Ren 2019)

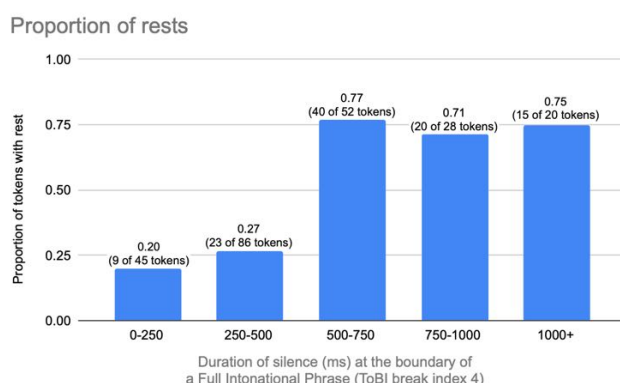


Figure 2: Proportion of rests at the boundaries of Full Intonational Phrases followed by silences of increasing durations

References

- [1] A. Kendon. “Some relationships between body motion and speech: An analysis of an example,” Aron W. Siegman & Benjamin Pope (eds.), *Studies in dyadic communication*, pp. 177–210. New York: Pergamon Press, 1972.
- [2] A. Kendon, “Gesticulation and speech: Two aspects of the process of utterance,” Mary Ritchie Key (ed.), *The relationship of verbal and nonverbal communication*, pp. 207–227. The Hague: Mouton, 1980.
- [3] A. Kendon, *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press, 2004.
- [4] D. Loehr, “Aspects of rhythm in gesture and speech,” *Gesture*, vol. 7, pp. 179-214, 2007.
- [5] D. Loehr, “Temporal, structural and pragmatic synchrony between intonation and gesture,” *Journal of Laboratory Phonology*, vol. 3, pp. 71-89, 2012.
- [6] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.
- [7] M. Renwick, S. Shattuck-Hufnagel, and Y. Yasinnik, “The timing of speech- accompanying gestures with respect to prosody,” *Journal of the Acoustic Society of America*, vol. 115, no. 5, p. 2397, 2004.
- [8] S. Shattuck-Hufnagel, Y. Yasinnik, N. Veilleux, and M. Renwick, “A method for studying the time-alignment of gestures and prosody in American English: ‘Hits’ and pitch accents in academic-lecture-style speech,” Anna Esposito, Maja Bratanić, Eric Keller & Maria Marinaro (eds.), *Fundamentals of verbal and nonverbal communication and the biometric issue*, pp. 34-44. Amsterdam: IOS Press, 2007.
- [9] S. Shattuck-Hufnagel and A. Ren, “The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech,” *Frontiers in Psychology*, vol. 9, no. 1514, 2019, doi:10.3389/fpsyg.2018.01514.
- [10] S. Shattuck-Hufnagel and P. Prieto, Dimensionalizing co-speech gestures. *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 1490-1494, 2019.

The Communicative Consequences of Multimodal Coordination

Aleksandra Ćwiek¹, Šárka Kadavá^{1,2,3}, Wim Pouw², Susanne Fuchs¹

Leibniz-Centre General Linguistics¹

Donders Institute for Brain, Cognition, and Behaviour²

Universität Göttingen³

cwiek@leibniz-zas.de

In communication, gestures and speech are time-coupled [1]. It is still unknown what regulates body-voice coordination, but it must serve a purpose: for instance, studies show that tighter multimodal coupling enhances prominence perception [2,3]. This coordination is nevertheless astonishing because mouth and limbs differ greatly. Anatomy shows that the jaw mandible is 21times lighter than the underarm and hand [4,5], allowing faster movement due to lighter mass. Furthermore, the voice system is entangled with the moving body: muscles responsible for posture and arm movement also affect expiratory flow crucial for vocal production [6,7]. Therefore, arm movements can leave ‘imprint’ on the voice [8].

Mastering multimodal coordination involves mastering these constraints. In the current study, we ask: Does coordination play a meaningful role? We recorded 60 dyads playing a charade game without the use of language, using only voice, only body gestures, or both. While one of the participants was performing the concepts, the other participant was guessing the meaning. Each participant communicated 7 concepts per condition.

Following a biomechanical framework, we have two hypotheses regarding the multimodal coordination. Hypothesis 1: Gestures precede vocalizations due to slower motor properties. This is due to the difference in weight between body parts and vocal articulators and it has been previously shown that the gesture onset tends to precede the onset in vocal modality [9]. Hypothesis 2: Modality-specific properties differ in uni- vs. multimodal utterances. We will compare a set of measures, such as duration, amplitude envelope slope/range, f_0 slope/range, and gesture kinematics (amplitude, submovements, acceleration) between the uni- and multimodal utterances. To increase the confirmatory evidential value of our research, we will pre-register our analyses plan based on exploratory analyses on a pilot dataset. These analyses will give us a characterization of what multimodal coordination requires of the gesture and vocal system.

Finally, we explore whether the temporal synchrony correlates with guessing success to understand the communicative advantages of combined multimodal utterances. Specifically, we will assess whether time-coupling in multimodal utterances is related to guessing accuracy. This study investigates whether motor control perspectives in multimodality provides a further understanding of communication and may provide insights into the evolutionary advantage of combining gesture and speech.

References

- [1] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014, doi: 10.1016/j.specom.2013.09.008.
- [2] P. Treffner, M. Peter, and M. Kleidon, "Gestures and phases: The dynamics of speech-hand communication", *Ecological Psychology*, vol. 20, no. 1, 2008, pp. 32-64.
- [3] H. R. Bosker, and D. Peeters, "Beat gestures influence which speech sounds you hear", *Proceedings of the Royal Society B*, vol. 288, no. 1943, 2021, 20202419.
- [4] M. Damavandi, N. Farahpour, and P. Allard. "Determination of body segment masses and centers of mass using a force plate method in individuals of different morphology". *Med Eng Phys* 31.9, 2009, pp. 1187–1194.
- [5] F. Zhang, C. C. Peck, and A. G. Hannam. "Mass properties of the human mandible". *J. Biomech.* 35.7, 2002, pp. 975–978.
- [6] W. Pouw, S. J. Harrison, N. Esteve-Gibert, and J. A. Dixon, "Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1231–1247, Sep. 2020, doi: 10.1121/10.0001730.
- [7] W. Pouw, R. Werner, L. Burchardt, and L. Selen, "The human voice aligns with whole-body kinetics." *bioRxiv*, p. 2023.11.28.568991, Nov. 28, 2023. doi: 10.1101/2023.11.28.568991.
- [8] Š. Kadavá, A. Ćwiek, K. Stoltmann, S. Fuchs, and W. Pouw, "Is gesture-speech physics at work in rhythmic pointing? Evidence from Polish counting-out rhymes," in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, Apr. 2023. doi: 10.31219/osf.io/67fzc.
- [9] V. Macuch Silva, J. Holler, A. Ozyurek, and S. G. Roberts, "Multimodality and the origin of a novel communication system in face-to-face interaction," *Royal Society Open Science*, vol. 7, no. 1, p. 182056, 2020, doi: 10.1098/rsos.182056.

Signaling discourse relations in multimodal communication

Schuyler Laparle¹, Merel Scholman^{2,3}

¹Tilburg University, ²Utrecht University, ³Saarland University
s.m.laparle@tilburguniversity.edu

There is substantial evidence that speakers ‘design’ their gestures for their addressee, depending on factors like visibility [1] and common ground [2]. However, we know significantly less about the uptake of gestural meaning by comprehenders, with existing work focusing largely on semantically-oriented gesture [3]. In the present work, we contribute to resolving this gap by focusing on the uptake of *pragmatic* gestural meaning, investigating *to what extent comprehenders can infer discourse relational meaning from a gesture*.

Discourse relations are the semantic-pragmatic links, such as CAUSE-CONSEQUENCE and CONTRAST, which hold between arguments [4], and are central to discourse processing. Discourse relations can be expressed by lexical connectives such as “also” and “on the other hand”, though approximately 50% of relations are not [5]. Where explicit connectives are not used, comprehenders must infer the relation through context or rely on “alternative” signals. This work considers gestural discourse markers as one such alternative signal available to comprehenders in face-to-face interaction. We focus on three types of relations for which recurrent hand gestures have been identified (see Figures 1-3): contrast relations, expressed by gesturing on opposing sides of the body [6], list relations, expressed by counting with fingers [7], and exception relations, expressed using a single raised finger (indicating singularity) [8]. We tested whether comprehenders can exploit the gestural discourse marker to inform their discourse relation predictions using an a multi-modal continuation study (an innovation on the mono-modal continuation paradigm frequently used in discourse research).

Methodology The experimental materials consisted of 18 prompts similar to the examples in Table 1 (six items per relation type). Two videos were created for every prompt: one video in which the relation was gestured, and one in which it was not – note that the same audio was used for both videos, and participants were only presented with one version of each video. In the videos, the speaker could be seen throughout the video, but the sound cut out at the second argument. Participants (n=48, recruited via Prolific) were asked to write a continuation of what the speaker might have said. The continuations were coded (blind to condition) to determine whether the participants constructed a relation that matches the gestural discourse marker.

Key findings Figure 4 displays the proportion of target responses per relation type and condition. Exception target items showed particularly few target continuations. The results were modeled using generalized mixed-effect regression models, with target continuation as binary response variable and fixed effects for condition, relation type and their interaction. The model showed that videos with a gestural discourse marker received a higher proportion of target continuations than videos without such markers ($\beta=0.56$, $SE=.19$, $z=2.95$, $p<.01$). Items in the contrast ($\beta=2.04$, $SE=.91$, $z=2.23$, $p<.05$) and list ($\beta=2.85$, $SE=.91$, $z=3.11$, $p<.01$) target conditions received more target continuations than exception items. Crucially, the effect of condition was significant for contrast ($\beta=-1.12$, $SE=.44$, $z=-2.53$, $p<.05$) and list items ($\beta=-1.56$, $SE=.45$, $z=-3.51$, $p<.001$), but not for exception items ($\beta=0.70$, $SE=.73$, $z=-.96$, $p=.34$).

In sum, the results indicate that comprehenders do take into account the provided gesture when interpreting discourse and inferring what a speaker might have said: when a contrastive or list gestural marker was present, participants were more likely to provide continuations with the corresponding relation.



Figure 1: *Gesture for contrast.*



Figure 2: *Gesture for lists.*

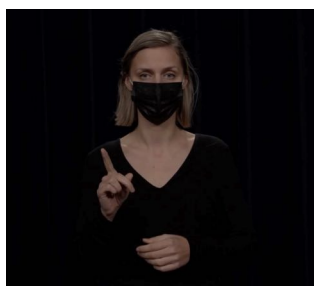


Figure 3: *Gesture for exceptions.*

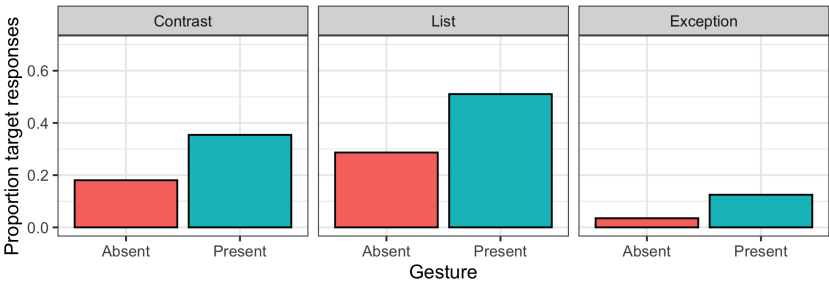


Figure 4: *Proportion of target responses per condition and relation type.*

Contrast	<p><i>Prompt:</i> Heather loves to travel solo. She is considering what activities to do while on vacation in Hawaii next month. She really loves the idea of learning how to surf...</p> <p><i>Example continuation:</i> On the other hand, relaxing at the beach sounds good too.</p>
----------	--

Table 1: *Example prompt and possible continuations.*

References

[1] J. Bavelas and S. Healing, “Reconciling the effects of mutual visibility on gesturing: A review,” *Gesture*, vol. 13, no. 1, pp. 63–92, 2013.

[2] A. Galati and S. E. Brennan, “Speakers adapt gestures to addressees’ knowledge: Implications for models of co-speech gesture,” *Language, Cognition and Neuroscience*, vol. 29, no. 4, pp. 435–451, 2014.

[3] C. Ebert, “Semantics of gesture,” *Annual Review of Linguistics*, vol. 10, pp. 169–189, 2024.

[4] T. J. M. Sanders, W. P. M. S. Spooren, and L. G. M. Noordman, “Toward a taxonomy of coherence relations,” *Discourse Processes*, vol. 15, no. 1, pp. 1–35, 1992.

[5] B. Webber, R. Prasad, A. Lee, and A. Joshi, *The Penn Discourse Treebank 3.0 annotation manual*. Philadelphia, University of Pennsylvania, 2019.

[6] J. Hinnell, “The verbal-kinesic enactment of contrast in North American English,” *The American Journal of Semiotics*, vol. 35, no. 1-2, pp. 55–92, 2019. DOI: 10.5840/ajs20198754.

[7] I. G. Rodrigues, “A tool at hand: Gestures and rhythm in listing events case studies of European and African Portuguese speakers,” *Oslo Studies in Language*, vol. 7, no. 1, pp. 253–281, 2015.

[8] A. Inbar, “The raised index finger gesture in hebrew multimodal interaction,” *Gesture*, vol. 21, no. 2-3, pp. 264–295, 2022.

Social meaning and multimodality: The performance of scientific authority

Marion Schulte

Universität Rostock

marion.schulte@uni-rostock.de

Prosody and gesture are assumed to be closely linked and may even be used to convey the same pragmatic and discursive meanings [1], to the extent that epistemic intonation and epistemic gesture can be considered “mutually co-expressive” [2]. Studies that investigate the interrelatedness of gesture and speech often focus on gestures with a referential function [3]. Gestures may, however, also have pragmatic functions that express speakers’ stances [4] - a key topic in contemporary sociolinguistics [5]. The present study puts social meaning at the centre of the investigation and analyses how an authority stance is performed by speakers of Irish English. Many prosodic signals have been found to express various social meanings and stances, e.g. non-phonemic clicks that function similarly to traditional pragmatic markers [6], silence that is essential to politeness [7], or non-modal voice qualities that can index various identities and positionings [8]. Some gestures have also been connected with stances taken by speakers, e.g. the precision-grip gesture employed by Barack Obama in his speeches [9]. It is, however, not yet well understood how these prosodic resources interact with each other and with gestural cues to enable speakers to take stances and create social meaning. The present study thus addresses the question how lexical grammatical, segmental phonetic, prosodic, and gestural cues interact to create social meaning.

The data for this investigation come from an Irish popular science podcast. The social meaning at the centre of this study is the co-creation of an authority stance by the scientists invited to the podcast and the hosts who interview them on their work. The data are authentic rather than elicited for a linguistic experiment, but the podcast recordings still offer a good-enough audio and video quality for instrumental phonetic measurements and a basic analysis of gestures. The speakers selected are Irish scientists working in Ireland. The podcast hosts are also Irish scientists and the audience can be assumed to be largely local as well. This ensures no cross-varietal differences in the performance of authority. Speakers of Irish English are known for their avoidance of directness and the explicit expression of an authority status vis-a-vis their interlocutor [10, 11], so this is a particularly interesting context to study authority positioning. All speakers are assumed to take a stance of authority and position themselves as experts especially in their introductory monologue. The first 5 minutes of each speaker talking about their research are transcribed. The audio features are annotated and analysed in Praat [12], and referential and pragmatic hand gestures are matched with the Praat annotations. The following features are taken into account: non-phonemic clicks, laughter, filled and unfilled periods of silence (> 150 ms), pragmatic markers, realisation of word-final /t/ as fricative, all visible hand gestures, and voice quality (especially creaky voice).

The results show that not all features are as aligned as the gesture-prosody link found in other studies would suggest. Instead, speakers layer signals from different modes to create an expert stance that does not challenge society-wide communication norms, which in this case discourage overt authority. A speaker who uses, for instance, a high amount of fricative /t/ - which signals authority in an Irish context [13] - may mitigate this by producing many pragmatic markers that function as hedges with a creaky voice quality. This behaviour seems to be gendered, as male and female speakers employ the resources across the modalities differently. Female speakers, for example, produce longer /t/ fricatives. The podcast hosts also play an important role as they choose between producing response tokens or remaining silent, and how they introduce a speaker, for example. Authority is thus clearly co-constructed in interaction, as well as multimodally layered.

References

- [1] L. Brown and P. Prieto. "Gesture and prosody in multimodal communication," In M. Haugh, D. Kádár and M. Terkourafi (eds.), *The Cambridge handbook of sociopragmatics*, Cambridge: Cambridge University Press, pp. 430-453, 2021.
- [2] J. Borràs-Comes, E. Kiagia and P. Prieto. "Epistemic intonation and epistemic gesture as mutually co-expressive: Empirical results from two intonation-gesture matching tasks," *Journal of Pragmatics* no. 150. pp. 39-52, 2019.
- [3] A. Kendon. "Gesture. Visible action as utterance," Cambridge: Cambridge University Press, 2004.
- [4] R. Lopez-Ozieblo. "Proposing a revised functional classification of pragmatic gestures," *Lingua*, vol. 247, 2020.
- [5] A. Jaffe (ed.). "Stance: Sociolinguistic perspectives," Oxford: Oxford University Press, 2009.
- [6] M. Schulte. "Dublin English and third-wave sociolinguistics," In R. Hickey (ed.), *The Oxford handbook of Irish English*, Oxford: Oxford University Press, pp. 339-360, 2023.
- [7] M. Lempert. "Barack Obama, being sharp: Indexical order in the pragmatics of precision-grip gesture," *Gesture*, vol. 11, no. 3, pp. 241-270.
- [8] J. Kallen. "Politeness in Ireland. 'In Ireland, it's done without being said'," In L. Hickey & M. Stewart (eds.), *Politeness in Europe*, Clevedon: Multilingual Matters, pp. 130-144, 2005.
- [9] R. J. Podesva and P. Callier. "Voice quality and identity," *Annual Review of Applied Linguistics*, vol. 35, pp. 173-194, 2015.
- [10] A. Barron and I. Pandarova, "The sociolinguistics of language use in Ireland," In R. Hickey (ed.), *Sociolinguistics in Ireland*, Basingstoke: Palgrave Macmillan, pp. 107-130, 2016.
- [11] E. Vaughan and B. Clancy. "The Pragmatics of Irish English," *English Today*, vol. 27, no. 2, pp. 47-52, 2011.
- [12] P. Boersma and D. Weenink. Praat: Doing phonetics by computer. Version 6.4.06, 2024.
- [13] F. O'Dwyer. "Linguistic variation and social practices of normative masculinity," London: Routledge, 2020.

The influence of different input types on the multimodal language processing of primary school children

Vera Wolfrum¹, Carina Lücke¹ & Simone Schaeffner¹
Julius-Maximilians University Würzburg, Germany¹
vera.wolfrum@uni-wuerzburg.de

In everyday communication, it is often necessary to switch between modalities, specifically between sensory-motor modality combinations, i.e. that something is perceived auditorily or visually and responded to vocally or manually. Preliminary evidence suggests that modality switching in speech processing is influenced by modality compatibility. Studies in adults show that switching between relatively incompatible modality combinations, such as auditory-manual and visual-vocal, is associated with longer reaction times and higher error rates than switching between more compatible combinations such as auditory-vocal and visual-manual (e.g., [1]). However, the role of modality compatibility in children's language processing remains unknown. Further, it is unclear whether potential effects are influenced by the type of input (i.e., less linguistic input in terms of pictures and sounds versus more linguistic input in terms of spoken language and gestures). In the present study we conducted two modality-switching experiments in order to investigate whether children show modality-compatibility effects and if these effects are influenced by the type of input. The findings will provide important insights into children's multimodal language processing and may be the starting point for new multimodal linguistic theories.

So far, a total of 59 typically developed primary school children from first to fourth grade took part in the study. Half of the participants performed Experiment 1 ($n=32$, $M_{Age} = 8;4$ years, $SD = 1,1$ years, 56% boys) and the other half performed Experiment 2 ($n=27$, $M_{Age} = 8;2$; $SD = 1,2$ years, 79% boys; final sample $n=32$). In both experiments, children switched between compatible and incompatible modality combinations (e.g., responding vocally to an auditory stimulus in the compatible condition and manually in the incompatible condition; see **Figure 1**) while they had to answer the question "Do you see or hear an animal?" by either pressing the left or right response key or saying the words "yes" or "no". The only difference between the two experiments was in the type of input. In Experiment 1, children had to categorize pictures and sounds. In Experiment 2 input consisted of gesture videos and spoken words. To investigate whether modality switching is influenced by modality compatibility we calculated proportional switch costs (i.e., performance differences between switch and repetition trials) for both conditions (i.e., for the compatible and for the incompatible switching condition).

A 2 x 2 MANOVA with the within-subject variable condition and the between-subject variable experiment revealed significantly higher proportional switch costs ($F(1, 57) = 3,88$, $p = .05$) for switching between incompatible combinations (11,7 %) compared to compatible combinations (7,5 %; see **Figure 2**) and no significant influence of experiment ($F(1, 57) = 0.06$, $p = .81$).

The results show effects of modality compatibility in primary school children, which seem to be independent of input type (sounds, pictures, spoken words and gestures).

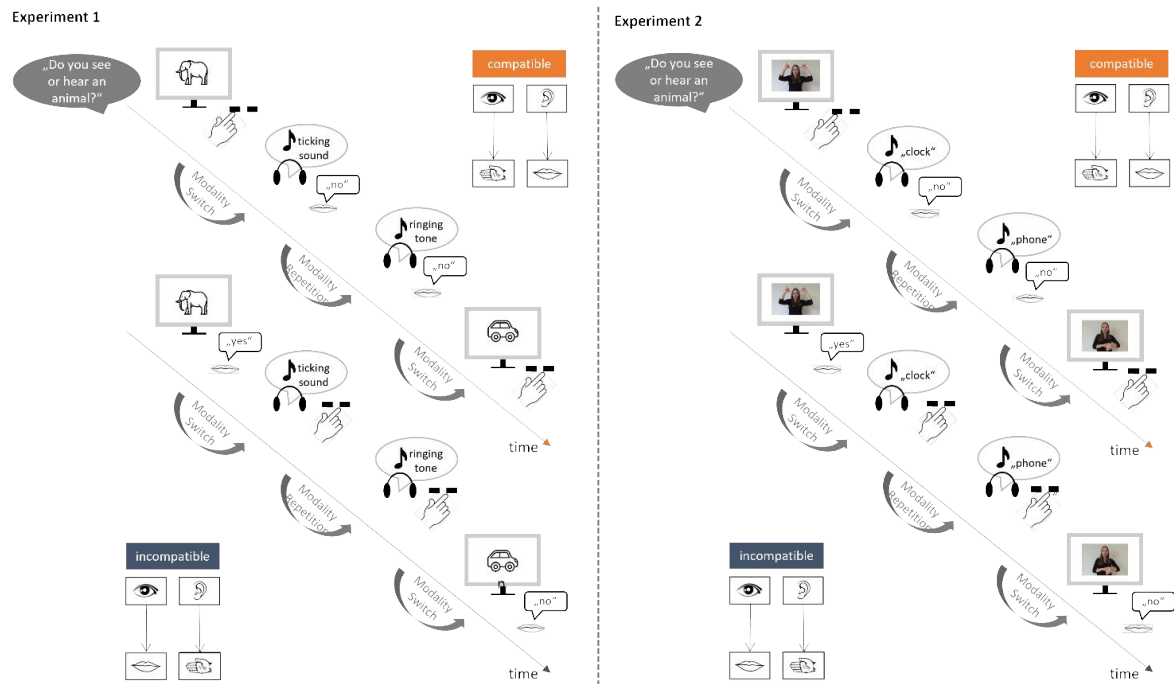


Figure 1: Presentation of the test procedure and item differences in the switching task of Experiment 1 and Experiment 2. Modality repetition: modality combination remains the same compared to the previous trial. Modality switch: modality combination changes compared to the previous trial.

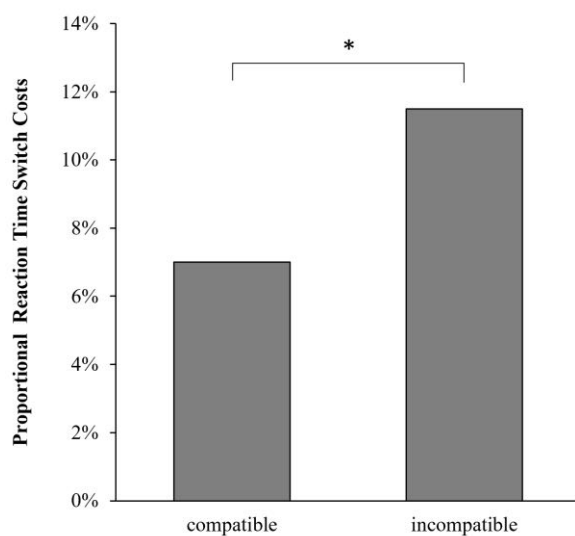


Figure 2: Main effect of modality compatibility.

References

- [1] Schaeffner, S., Koch, I., & Philipp, A. M. (2016). The role of sensory-motor modality compatibility in language processing. *Psychological Research*, 80(2), 212–223. <https://doi.org/10.1007/s00426-015-0661-1>

Verbal signals of understanding do not predict a decrease of gesture deixis

Stefan Lazarov¹, Angela Grimminger¹

Faculty of Arts and Humanities, TRR-318 “Constructing Explainability”

¹*Paderborn University*

stefan.lazarov@uni-paderborn.de

In explanatory dialogues, a more experienced interlocutor (explainer, henceforth EX) aims at increasing the understanding of a less experienced interlocutor (explainee, henceforth EE) about an entity or a process (i.e., explanandum) via co-constructions and scaffolding [1]. There are situations in which the explanandum is absent from the shared referential space between the EX and the EE, and EXs need to provide EEs with additional spatial orientation by using co-speech gestures indicating certain locations or the shape of invisible objects [2, 3].

In the present study, we investigated the relation between the EXs’ gesture deixis and EEs’ verbal signals of understanding in dyadic explanations of a board game, in a sample of 5 German-speaking adult EXs, each explaining a board game to 3 different adult EEs individually, resulting in 15 explanatory dialogues. The analyzed explanations are constituted by three phases (game absent, game present and game play) [4]. Initial observations of the video data indicated an increased gestural behavior by the different EXs during the game absent phase (i.e., the explanandum is not visible) compared to the other phases; therefore, only this phase was analyzed here. Also based on initial observations of the video data, we followed McNeill’s multidimensional view on gestures, including the dimension of deixis [5].

Our research question and hypothesis are motivated by previous research: In general, it was shown that co-speech gestures enhance addressees’ understanding [6]. More specifically, speakers’ deictic and iconic gesture rates were found to decrease significantly after addressees’ feedback of understanding [7]. Further, it was reported that teachers’ deictic and iconic gestures increase after detecting spots of students non-understanding [8]. Based on this, we hypothesized that EXs’ gesture deixis would decrease after EEs’ verbal signals of understanding and increase after EEs’ verbal signals of partial and non-understanding.

EEs’ verbal utterances were coded in relation to EEs’ understanding (e.g., backchannels *ok*, *yes*, *alright*, and also repetitions of EXs’ utterances), partial understanding (e.g., polar and tag questions), and non-understanding (e.g., open questions, corrections), based on a discourse annotation scheme ($k = 0.89$). EXs’ gesture phrases were coded based on the occurrence of gesture strokes and with respect to the dimension of gesture deixis, being observed in deictic, deictic-iconic, or deictic-beat gestures ($k = 0.94$). To incorporate the study design of 1 EX interacting with 3 different EEs and considering a non-normal data distribution, we conducted a Generalized Linear Mixed Effects Regression analyzing EXs’ raw frequencies of gesture deixis during the game absent phase.

The results (Tab.1) indicate a significant effect of the three levels of understanding signaled by EEs on the frequencies of EXs’ gesture deixis following these signals. Post-hoc comparisons (Tab. 2) reveal that the frequency of EXs’ gesture deixis after EEs’ signals of understanding is significantly higher than after EEs’ signals of partial and non-understanding (Fig. 1). Contrary to our hypothesis, our findings are not in line with previous research. One possible reason for the high frequencies of EXs’ gesture deixis after EEs’ signals of understanding could be related to the existing knowledge gap between the more experienced EXs and the novice EEs, who were not familiar with the physical appearance of the game components and their placement on the shared referential space. Another related reason could be that the EXs may have noticed a continuous high demand for spatial orientation on the invisible shared referential space in EEs’ (non-)verbal behavior during the game absent phase.

EX	gesture deixis after:	<i>M</i>	<i>SD</i>	β	<i>SE</i>	<i>z</i>	<i>p</i>
EE	understanding (Int.)	87.47	41.44	4.37	0.15	28.36	< 0.001
	partial understanding	21.13	23.10	-1.42	0.16	-22.74	< 0.001
	non-understanding	5.00	13.37	-2.86	0.12	-24.15	< 0.001

Table 1: A summary of descriptive statistics and fixed effects.

pairwise comparison:	β	<i>SE</i>	<i>z</i>	<i>p</i>
understanding – partial understanding	1.42	0.06	22.74	< 0.001
understanding – non-understanding	2.86	0.12	24.15	< 0.001
partial-understanding – non-understanding	1.44	0.13	11.25	< 0.001

Table 2: Post-hoc pairwise comparisons

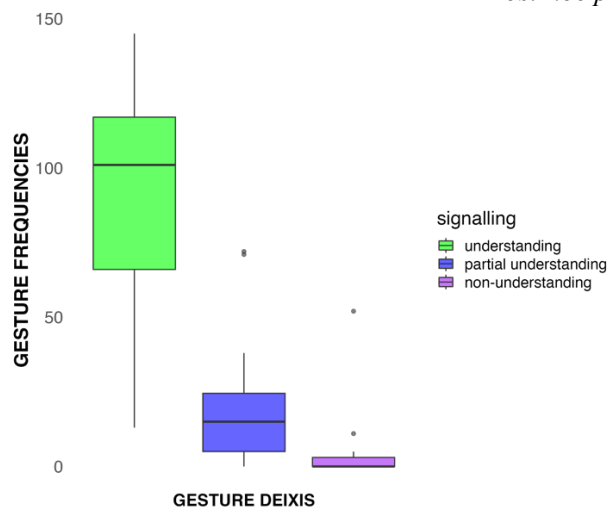


Figure 1: EXs' gesture deixis related to levels of EEs' understanding.

References

- [1] K. J. Rohlfing et al., "Explanation as a social practice: toward a conceptual framework for the social design of AI systems," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 717–728, Sep. 2021, doi: 10.1109/tcds.2020.3044366.
- [2] H. H. Clark, "Pointing and placing," in *Pointing: Where language, culture, and cognition meet*, S. Kita, Ed. Lawrence Erlbaum, 2003, pp. 251–276. doi: 10.4324/9781410607744-14.
- [3] N. McKern, N. Dargue, N. Sweller, K. Sekine, and E. Austin, "Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability," *Quarterly Journal of Experimental Psychology*, vol. 74, no. 10, pp. 1791–1805, Jun. 2021, doi:10.1177/17470218211024913.
- [4] O. Türk., P. Wagner, H. Buschmeier, A. Grimminger, Y. Wang, and S. Lazarov, "MUNDEX: A multimodal corpus for the study of the understanding of explanations" in *Book of Abstracts of the 1st International Multimodal Communication Symposium*, P. Paggio and P. Prieto, Eds., 2023, pp. 63–64.
- [5] D. McNeill, "Gesture and communication," in *Encyclopedia of Language & Linguistics*, K. Brown, Ed. Elsevier, 2006, pp. 58–66. doi: 10.1016/b0-08-044854-2/00798-7.
- [6] S. D. Kelly, A. Özyürek, and E. Maris, "Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension," *Psychological Science*, vol. 21, no. 2, pp. 260–267, Dec. 2010, doi: 10.1177/0956797609357327.
- [7] J. Holler and K. Wilkin, "An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses," *Journal of Pragmatics*, vol. 43, no. 14, pp. 3522–3536, Nov. 2011, doi: 10.1016/j.pragma.2011.08.002.
- [8] M. W. Alibali, M. J. Nathan, R. B. Church, M. S. Wolfgram, S. Kim and E. J. Knuth, "Teachers' gestures and speech in mathematics lessons: forging common ground by resolving trouble spots", *ZDM Mathematics Education*, vol. 45, pp. 425–440, Jan 2013, doi: 10.1007/s11858-012-0476-0.

Farsi-English bilinguals' gesture production while telling a story

Elena Nicoladis¹, Anahita Shokrkon², & Shiva Zarezadehkheibari²

University of British Columbia¹ University of Alberta²

elena.nicoladis@ubc.ca

Some researchers have argued that one function of gestures is to help speakers access words (particularly difficult ones) for production [1]. This argument has been referred to as the lexical retrieval hypothesis (LRH). Gestures refer to communicative hand/arm movements, often produced while speaking [2]. According to the LRH, gesturing helps speakers they activate the concept(s) they want to talk about. By activating the target concept(s), speakers increase activation of the targeted word(s) so they are more likely to retrieve them. The LRH focuses on representational gestures, or gestures that represent the referent with movement and/or handshape. For example, a speaker might pump their arms up and down at their sides to represent 'running'. In contrast, other gestures are non-representational, like beats (repetitive movements of hands/arms that often serve to highlight information).

Bilinguals often have greater difficulty with lexical retrieval than monolinguals [3] and greater difficulty retrieving words in their weaker language than their stronger language [4]. A prediction that follows from the LRH is that bilinguals might produce more representational gestures than monolinguals and more representational gestures in their weaker language than their stronger language. Some studies have found support for that prediction while others have not [review in 5]. The primary purpose of this study was to test whether Farsi-English bilinguals produce more representational gestures than English monolinguals and more representational gestures in their weaker language (English) than in their stronger language (Farsi).

There were two other research questions guiding this study: 1) do beats show the same pattern as representational gestures and 2) do we observe any evidence for culture impacting gesture frequency? Some studies have reported that beats are related to fluency in speech production in second language acquisition [6]. As for the effects of culture, some studies have found that gesture frequency is higher in some cultures than others [7]. To address the second research question, we compared the gesture frequency of Farsi-English bilinguals with that of French-English bilinguals.

Participants were 28 Farsi-English bilinguals, 46 English monolinguals, and 25 French-English bilinguals. Participants watched a cartoon and told the story back to a native speaker of the target language. Bilinguals did the story retell in both languages (on different days, with a different interlocutor), with order of the language sessions counterbalanced. Participants' stories were videotaped for later transcription and coding (representational and beat gestures). Disfluency was operationalized as the number of false starts and self-corrections.

The results showed that, consistent with the LRH, in English, all the bilinguals produced more gestures than the monolinguals and the Farsi-English bilinguals produced more beats in English than in Farsi. However, the Farsi-English bilinguals produced more beats and the French-English bilinguals produced more representational gestures than the English monolinguals. The Farsi-English produced more beats than the French-English bilinguals in English, but the rate of beats did not differ in Farsi/French. For the Farsi-English bilinguals, the number of beats in English was significantly positively correlated to disfluency.

In sum, these results did not support the LRH in a straightforward way. While the Farsi-English bilinguals did produce more gestures than English monolinguals, the difference was due to the production of more beats, not representational gestures. Moreover, the number of beats was correlated with fluency (that is, the less fluent speech, the more beats) for the Farsi-English bilinguals. We discuss these results in terms of how culture impacts the type of gesture produced when speech fluency becomes difficult.

	Referential gestures	Beats
English		
Monolinguals	3.2 (2.6)	1.6 (1.5)
French-English	6.2 (3.7)	2.7 (1.9)
Farsi-English	2.6 (3.7)	6.1 (5.7)
French/Farsi		
French-English	5.4 (3.2)	2.9 (2.0)
Farsi-English	3.5 (2.4)	2.5 (1.9)

Table 1: *Average (SD) Gesture Rate by Group.*

Table Note: Gesture rate was calculated as the number of gestures per 100 word tokens in order to control for individual differences in story length

References

- [1] K. J. Pine, H. Bird, and E. Kirk. "The effects of prohibiting gestures on children's lexical retrieval ability", *Developmental Science*, vol. 10, no. 6, pp. 747-754, 2007. doi: 10.1111/j.1467-7687.2007.00610.x
- [2] A. Kendon. "Do gestures communicate? A review", *Research on language and social interaction*, vol. 27, no. 3, pp. 175-200, 1994. doi: 10.1207/s15327973rlsi2703_2
- [3] M. D. Sullivan, G. J. Poarch, and E. Bialystok. "Why is lexical retrieval slower for bilinguals? Evidence from picture naming", *Bilingualism: Language and Cognition*, vol. 21, no. 3, pp. 479-488, 2018. doi: 10.1017/S1366728917000694
- [4] P. Ecke. "Words on the tip of the tongue: A study of lexical retrieval failures in Spanish-English bilinguals", *Southwest Journal of Linguistics*, vol. 23, no. 2, pp. 33-63, 2004.
- [5] E. Nicoladis and L. Smithson. "Gesture in bilingual language acquisition." (2022). In A. Morgenstern & S. Goldin-Meadow (Eds.), *Gesture in language: Development across the lifespan* (pp. 297–315). De Gruyter Mouton; American Psychological Association. 10.1037/0000269-012
- [6] S. G. McCafferty, "Gesture and the materialization of second language prosody", *International Review of Applied Linguistics in Language Teaching*, vol. 44, no. 2, pp. 197-209, 2006. doi: 10.1515/IRAL.2006.008
- [7] W. C. So. "Cross-cultural transfer in gesture frequency in Chinese–English bilinguals", *Language and Cognitive Processes* vol. 25, no. 10, pp. 1335-1353, 2010. doi: 10.1080/01690961003694268

Effects of Bowing During Japanese Telephone Conversation on Acoustic Properties

Kazuki Sekine^{1*}, Ikuko Nonaka¹

Waseda University¹

*ksekine@waseda.jp

Background: This study explores how bowing affects acoustic properties in business telephone conversations. Bowing is an important behavior in Japanese social interaction, often expressing respect, politeness, apologies, or greetings to the interaction partner. Previous studies [1][2] have shown that extended bowing durations correlate with increased perceptions of politeness and, consequently, influence assessments of the bowing individual's facial attractiveness—more pronouncedly in Japan than in other countries such as the U.S., Brazil, and India. This may be due to cognitive schemas and habitual associations between bowing and politeness. Interestingly, bowing is also practiced during telephone conversations where the interlocutor is not visible. Customer service training frequently advises trainees to bow while speaking, yet its effect on acoustic properties like speech pitch and intensity remains understudied. Previous studies have had mixed findings on the impact of co-speech gestures on these properties. Hoetjes et al. [3] showed that the presence or absence of gestures had no significant effect on acoustic properties. However, Cravotta et al. [4] reported that encouraging speakers to gesture while speaking heightened the pitch, but not intensity. If encouraging a nonverbal behavior enhances acoustic properties, we may observe the same phenomenon with bowing. Thus, our research aimed to determine how bowing might alter the acoustic characteristics of speech within the context of Japanese business telephone communication.

Method: 42 native Japanese speakers, who work at a printer manufacturing company, participated (16 females). The age range was from 31 to 64 years ($M = 49.6$, $SD = 9.22$).

The task for participants was to respond to a pre-recorded talk voice while holding a phone by reading out a script that had been set up for them (Figure 1). There were three types of scripts: Inquiry, Person in charge is not available, Response to a complaint. The average number of lines spoken by participants was 8 lines. Each participant read the three scripts under two conditions: with and without bowing. For the bowing condition, participants were instructed to read highlighted lines, such as four out of the eight provided, while bowing. We counterbalanced the order of bowing conditions and scripts across participants, although the three scripts were presented in a fixed order to each participant. In total, six were conducted. We recorded participants' movement and speech during the task, and analyzed the pitch, intensity, and duration of speech produced for each line using Praat [5].

Results: A two-way ANOVA was performed to analyze the effect of bowing (with vs. without) and script type (three scripts) on the pitch, intensity, and duration of speech, respectively (Table 1). The results revealed that there was a main effect of bowing on the maximum intensity, $F(1, 41) = 26.6$, $p < .001$, $partial \eta^2 = .39$, and on the duration of speech, $F(1, 41) = 38.2$, $p < .001$, $partial \eta^2 = .48$. Post hoc tests (Bonferroni, $p < .05$) indicated that the maximum intensity and duration of speech in the bowing condition were significantly greater than those in the no-bowing condition. However, no significant interaction or main effects for pitch were found.

Discussion: The present study demonstrates that the bowing during phone conversations influences acoustic properties by increasing the intensity and duration of speech. The increased intensity may result from a psychological head movement, speaking up because the mouth moves away from the telephone while bowing, or reflecting heightened politeness conveyed by the bow. Future studies will examine how these factors affect perception and whether listeners can detect the speaker's bows. The implication of the current study is that bowing enhances the perceived politeness, respect, and attractiveness of the speaker by altering voice dynamics.

Figure 1: *Experimental setting.*Table 1: *Mean and standard deviation of acoustic properties for each script and bowing conditions*

Acoustic property: Script type	With bowing		Without bowing	
	Mean	SD	Mean	SD
Intensity mean (dB): Inquiry	67.3	2.9	67.5	2.6
Intensity mean (dB): Absent	67.9	2.7	68.4	2.4
Intensity mean (dB): Complaint	67.5	3.0	68.5	2.3
Intensity max (dB): Inquiry	81.4	1.2	81.3	1.1
Intensity max (dB): Absent	81.4	1.2	81.2	1.1
Intensity max (dB): Complaint	81.4	1.1	81.2	1.1
Speech Duration (sec): Inquiry	4.5	0.4	4.4	0.4
Speech Duration (sec): Absent	4.5	0.4	4.3	0.3
Speech Duration (sec): Complaint	6.3	0.5	6.0	0.5
F0 mean (Hz): Inquiry	269.2	53.9	268.9	51.0
F0 mean (Hz): Absent	268.4	53.1	269.5	51.7
F0 mean (Hz): Complaint	283.5	48.7	277.6	53.9

References

- [1] Oosugi, T. & Kawahara, J. (2020). Effects of bowing and physical characteristics on perception of attractiveness. *The Japanese Journal of Cognitive Psychology*, 17(2), 69-77.
- [2] Oosugi, T. & Kawahara, J. (2022). Cultural difference in the effect of Japanese bowing on perception of attractiveness. In the Proceedings of the Japanese Society for Cognitive Psychology, 57-57.
- [3] Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication*, 57, 257–267. <https://doi.org/10.1016/j.specom.2013.06.007>
- [4] Cravotta, A., Busà, M. G., & Prieto, P. (2019). Effects of encouraging the use of gestures on speech. *Journal of Speech, Language, and Hearing Research*, 62(9), 3204-3219. https://doi.org/10.1044/2019_JSLHR-S-18-0493
- [5] Boersma, P. & Weenink, D. (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.05, retrieved 27 January 2024 from <http://www.praat.org/>

Session 3: Multimodality and Development

25.09.2024

16:50-18:10



Preschoolers' use of prosody and gesture in marking focus types

Sara Coego¹, Núria Estve-Gibert², Pilar Prieto^{3,1}

Universitat Pompeu Fabra¹, Universitat Oberta de Catalunya², Institució Catalana de Recerca i Estudis Avançats (ICREA)³
sara.coego@upf.edu

Previous research on adult speech has shown that prosodic prominence and co-speech gestures, often combined into multimodal ensembles, are consistent cues to focus crosslinguistically, especially to contrastive focus [1]. The developmental path leading to such an adult-like multimodal marking of focus, where both prosody and gestures are closely used, is still unclear. Developmental research has mainly looked at children's use of prosody independently from their use of co-speech gestures. This literature has shown that phonological cues to focus are acquired quite late in development, around 7-8 years of age, but are preceded by phonetic uses of prosody since age 2 [2]. However, considering a multimodal perspective, [3] showed that French-speaking 4- to 5-year-olds still made no significant use of the expected phonetic cues to focus, but used head gestures instead. To sum up, previous studies have shown that phonetic and phonological cues to focus are acquired at quite different stages in development, and have also pointed towards a precursor role of gestures in focus marking. In the present study, we aim at (1) exploring the interaction between prosody and gesture in distinguishing types of focus (information, contrastive, and corrective) in the developmental period ranging from 3 to 6 y.o.; and (2) testing whether the seemingly precursor role of gestures in focus marking can be replicated in stages of acquisition previous to 4-5.

A total of 116 Catalan-Spanish bilingual children (54 girls) belonging to three age groups (year 3, 4, and 5) were video recorded during a semi-controlled interactive task (adapted from [4]) in which they were encouraged to help a puppet select an object (target object) and place it inside a toy train by providing a verbal instruction containing a focus target word. By varying the color and number of objects displayed before the child, we were able to elicit words in three focus conditions: information, contrastive and corrective (see Figure 1 for the example stimuli and explanation of the experimental conditions). Data is currently being coded but will be ready for the conference. For prosody, we are coding intonation patterns following the Cat_ToBI guidelines [4] and perceived prosodic prominence following an adapted version of DIMA [5]. For gesture, we are coding in ELAN the presence of gestures, the articulator(s) used, and the perceived prominence of the gestures following M3D [6]. We will perform a prosodic analysis of target focus words in terms of prosodic structure (nuclear configuration) and perceived prosodic prominence. Regarding gesture production, we will analyze gestures overlapping with the target focus word in terms of gesture presence/absence, number of gesture articulators, and perceived gestural prominence. In all the analyses, we will compare the three age groups (year 3, 4, and 5) across the three different focus types (information, contrastive, corrective).

Preliminary results with 21 children (7 per each of the three groups) showed that words in corrective focus were significantly more prominent and aligned more often with a prominent gesture than words in the contrastive and information focus conditions. No significant differences were observed across age groups. Further analyses, which will be ready by the time of the conference, will confirm whether gestures have a precursor role in focus type marking in the developmental period studied. Final results will also show how prosodic and gestural prominence interact at each age group. Overall, this study will allow us to explore the developmental trajectory of the multimodal marking of focus types and assess the relevance of prosody and gesture for pragmatic development in early childhood. Moreover, it will help us evaluate the validity of innovative and newly implemented coding systems for the assessment of perceived prosodic and gestural prominence like [5] and [6].




Experimental conditions		
Information Focus	Contrastive Focus	Corrective Focus
		
<ul style="list-style-type: none"> • One object shown (target object) • Target object does not contrast with any object in the context • Expected production with underlined focus word: <i>Agafa les <u>ulleres liles</u></i> "Pick the <u>purple glasses</u>" 	<ul style="list-style-type: none"> • Two objects shown • Target object contrasts in color with competitor for which puppet has a preference • Expected production with underlined focus word: <i>Agafa l'estrella <u>lila</u></i> "Pick the <u>purple star</u>" 	<ul style="list-style-type: none"> • Two objects shown • Target object contrasts in color with competitor, which has been mistakenly collected by the puppet • Expected production with underlined focus word: <i>No, agafa l'estrella <u>lila</u></i> "No, pick the <u>purple star</u>"

Figure 1: Description of the experimental conditions and representation of stimuli with expected productions.

References

- [1] G. Ambrazaitis and D. House, "Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings," *Speech Communication*, vol. 95, pp. 100–113, 2017. doi: <https://doi.org/10.1016/j.specom.2017.08.008>.
- [2] A. Chen, "Get the focus right across languages," in *The Development of Prosody in First Language Acquisition*, P. Prieto and N. Esteve-Gibert, Eds. Amsterdam: John Benjamins, 2018, pp. 295–317.
- [3] N. Esteve-Gibert, H. Lævenbruck, M. Dohen, and M. D'Imperio, "Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech," *Developmental Science*, vol. 25, no. 1, pp. 1–12, 2021. doi: <https://doi.org/10.1111/desc.13154>
- [4] P. Prieto, J. Borràs-Comes, T. Cabré, V. Crespo-Sendra, I. Mascaró, P. Roseano, R. Sichel-Bazin, and M. M. Vanrell, "Intonational phonology of Catalan and its dialectal varieties," in *Intonation in Romance*, S. Frota and P. Prieto, Eds. Oxford: Oxford University Press, 2015, pp. 9–62.
- [5] F. Kügler, B. Smolibocki, D. Arnold, B. Braun, S. Baumann, M. Grice, and P. Wagner, "DIMA—Annotation guidelines for German intonation," in *Proc. 18th Int. Congr. Phonetic Sci.*, Glasgow, UK, 2015, pp. 1–5.
- [6] P. L. Rohrer, I. Vilà-Giménez, J. Florit-Pons, N. Esteve-Gibert, A. Ren, S. Shattuck-Hufnagel, and P. Prieto, "The MultiModal MultiDimensional (M3D) labeling scheme for the annotation of audiovisual corpora," 2021. Available: 10.17605/OSF.IO/ANKDX

Simultaneity in iconic two-handed gestures: a communicative strategy for children

Anita Slonimska¹, Alessia Giulimondi^{1,2}, Emanuela Campisi³, & Asli Ozyurek^{1,4}

¹Max Planck Institute for Psycholinguistics, ²Utrecht University, ³Catania University,

⁴Donders Institute for Brain, Cognition and Behavior

Corresponding author: anita.slonimska@mpi.nl

In face-to-face communication, humans adapt their multimodal utterances (i.e., speech+gesture) to meet the informational needs of their addressees. Indeed, research has shown that Italian speakers increase the rate of iconic gestures overall and two-handed iconic gestures with children, suggesting that it serves as a communicative strategy to increase the informativeness of their utterances (Campisi & Özyürek, 2013; Campisi et al., 2023). However, no systematic analysis has been conducted on whether the use of two-handed gestures actually leads to an increase in informativeness. Sign language studies show that signers exploit multiple body articulators (e.g., two hands) as a strategy to increase communicative efficiency by encoding multiple related semantic features of an event simultaneously, increasing the overall iconicity of the representation (Slonimska et al., 2020, 2021). As speakers might be recruiting similar strategies, we hypothesize that if two-handed gestures are used to increase informativeness for children, they should be used more with children than adults to represent more semantic elements simultaneously.

We analyzed iconic two-handed gestures produced by 16 native Italian adults explaining a board game (Fig.1) to a child (9-10 y.o.) and to another adult. We coded whether gestures represent two elements (e.g., two disks, Fig.2b and Fig.2c) as opposed to only one element (e.g., one disk, Fig.2a). Then, we annotated the type of information represented. If gestures represented a physical feature of an object (e.g., Fig.2a & b), it was coded as containing an imagistic component. If gestures represented relative position between two objects it was coded as containing spatial relationship, e.g., Fig.2c represents only spatial relationship & Fig.2b represents both objects and spatial relationship.

Results revealed that speakers depict two elements more often when talking to children (Fig.3). The use of the imagistic component was comparable in descriptions for both addressee age groups and in both types of gestures (1 element: ~78% vs. 2 elements: ~80%). Crucially, for both addressee age groups, the spatial relationship was depicted in ~80% of gestures representing 2 elements, compared to ~20% in gestures with one element.

In this study, we provide first insights into how adults modulate the number and type of information represented in their gestures with children. Our results show that speakers use two-handed gestures to represent more semantic units of information for children compared to adults. Furthermore, we show that the increase in informativeness in gestures for children is not only driven by the mere depiction of more elements but also the depiction of their spatial relationship. This research expands our understanding of the use of simultaneity in co-speech gestures as a communicative strategy, supporting the hypothesis that iconicity benefits from simultaneity to increase the informativeness of the representation.



Figure 1. Board game *Tower of Hanoi*

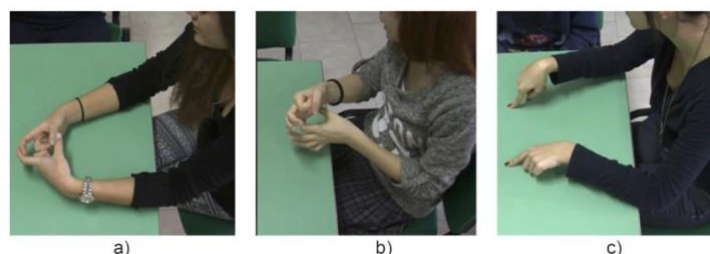


Figure 2. Two-handed gestures representing a) one disk, b) two disks on top of each other, c) only spatial relation between two objects.

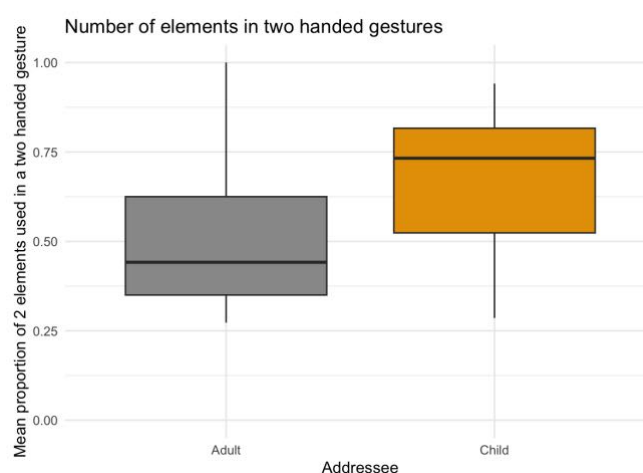


Figure 3. Proportion of two elements represented in two-handed gestures for adults and children.

References

- Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, 47(1), 14-27.
- Campisi, E., Slonimska, A., & Özyürek, A. (2023). *Cross-linguistic differences in the use of iconicity as a communicative strategy* [Conference presentation abstract] The 8th edition of Gesture and Speech in Interaction, GESPIN conference, Nijmegen, The Netherlands.
- Slonimska, A., Özyürek, A., & Capirci, O. (2020). The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS). *Cognition*, 200, 104246.
- Slonimska, A., Özyürek, A., & Capirci, O. (2021). Using depiction for efficient communication in LIS (Italian Sign Language). *Language and Cognition*, 13(3), 367-396.

The impact of a multimodal oral narrative intervention on boosting the frequency of use and the quality of children's non-dominant language

Joel Espejo-Álvarez¹, Júlia Florit-Pons¹, Claire Lien Luong¹, Mireia Gómez i Martínez²,
Alfonso Igualada³, Pilar Prieto^{4,1}

*Universitat Pompeu Fabra¹, University of Cork², Universitat Oberta de Catalunya³, Institució
Catalana de Recerca i Estudis Avançats⁴*
joel.espejo@upf.edu

Narrative-based interventions have been shown to improve children's oral narrative abilities [1][2][3], while also triggering gains in written narration and various academic outcomes [4][5]. Despite this, most interventions have been implemented in English-speaking countries and with monolingual children [2]. Other sociolinguistic conditions and languages have not been examined, such as officially bilingual communities with more than one societal language and with a variety of language dominance situations. Moreover, to our knowledge, multimodality understood as the communicative use of body gestures and voice has not been systematically incorporated into the design of narrative-based interventions, despite its known benefits in language development [6].

The goal of the present study is to investigate the impact of a novel 9-session narrative-based intervention, the MultiModal Narrative (MMN) [7][8], on improving the quality and frequency of use of Catalan in the narratives produced by preschoolers residing in the Spanish-dominant area of L'Hospitalet de Llobregat, Catalonia, Spain. Multimodality is incorporated into the intervention through three main components: 1) a video of a storyteller who retells a story using very expressive body gestures; 2) clear instructions to teachers, who are asked to enact the main actions and emotions of the story naturally; and 3) the participating children, who are also encouraged to enact the main actions and emotions of the story.

The MMN intervention was implemented with two groups of 5-to-6-year-old preschool children ($n = 115$; $M = 64.8$ months, $SD = 4.2$), one receiving the MMN intervention (the experimental group, $n = 77$), and the other being the control group following their usual school activities ($n = 38$). Before and after the intervention, children recounted three wordless animated cartoons. The language of testing was Catalan, but many children retold stories in Spanish or bilingual stories in Spanish and Catalan, especially before the intervention.

We analyzed all narratives using language productivity and complexity measures (i.e., frequency of use), including the total number of words (TNW), number of different words (NDW), and speech fluency (i.e., quality), in both Catalan and Spanish. Preliminary results with one story and 71 children show that, while the control group showed no improvements, the experimental group significantly increased narratives' TNW (see Figure 1) and NDW (see Figure 2) in Catalan after the intervention, regardless of language dominance. In contrast, neither TNW nor NDW in Spanish increased after the intervention. A complete statistical analysis of the language productivity and complexity of the three stories and an analysis of utterance fluency (i.e., quality) from the whole sample will be presented at the conference.

This study is pioneering in examining the impact of a multimodal narrative intervention in the use of Catalan in a sociolinguistic setting where the target language of the intervention is not the dominant language of the surrounding community. Preliminary results demonstrate the adequacy of a 9-session multimodal narrative intervention for improving the children's non-dominant language. The results have theoretical and pedagogical implications, as they inform on the field of narrative interventions and best practices for bi- and multilingual education. Although it is challenging to isolate the contribution of multimodality in a natural setting, these results demonstrate that multimodality can be beneficial for improving language outcomes in a real educational context.

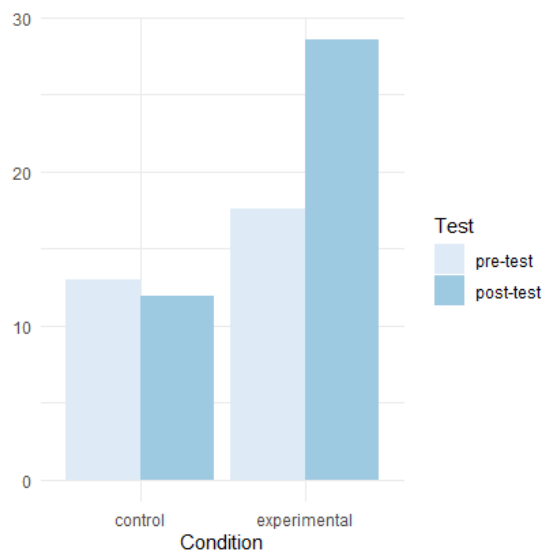


Figure 1. Mean TNW in Catalan by time (pre- and post-test) and group (control and experimental).

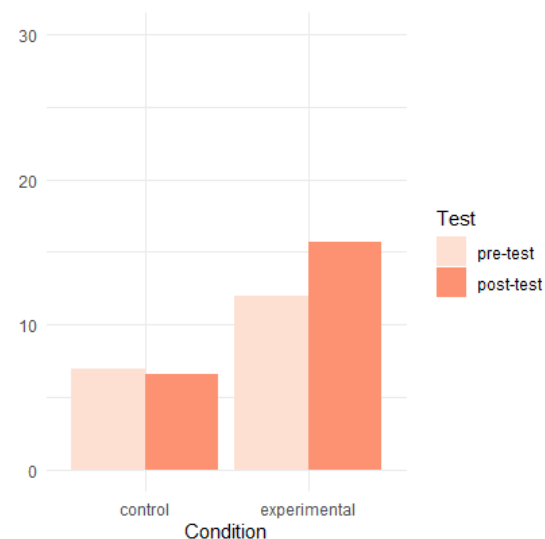


Figure 2. Mean NDW in Catalan by time (pre- and post-test) and group (control and experimental).

References

- [1] K. Favot, M. Carter and J. Stephenson, "The Effects of Oral Narrative Intervention on the Narratives of Children with Language Disorder: a Systematic Literature Review," *Journal of Developmental and Physical Disabilities*, vol. 33, no. 4, pp. 489–536, 2021.
- [2] D. L. Pico, A. Hessling Pahl, C. H. Biel, A. K. Peterson, E. J. Biel, C. Woods and V. A. Contesse, "Interventions Designed to Improve Narrative Language in School-Age Children: A Systematic Review With Meta-Analyses," *Language, Speech, and Hearing Services in Schools*, vol. 52, no. 4, pp. 1109–1126, 2021.
- [3] D. B. Petersen, "A systematic review of narrative-based language intervention with children who have language impairment," *Communication Disorders Quarterly*, vol. 32, no. 4, pp. 207–220, 2011.
- [4] S. Babayiğit, S. Roulstone and Y. Wren, (2021), "Linguistic comprehension and narrative skills predict reading ability: A 9-year longitudinal study," *British Journal of Educational Psychology*, vol. 91, pp. 148–168, 2021
- [5] D. K. Dickinson and A. McCabe, "Bringing it all together: The multiple origins, skills, and environmental supports of early literacy," *Learning Disabilities Research & Practice*, vol. 16, no. 4, pp. 186–202, 2001.
- [6] S. Goldin-Meadow, "How Gesture Promotes Learning Throughout Childhood," *Child Development Perspectives*, vol. 3, pp. 106–111, 2009.
- [7] J. Florit-Pons, A. Igualada and P. Prieto, "Evaluating the feasibility and preliminary effectiveness of a multi-tiered multimodal narrative intervention program for preschool children," under review.
- [8] J. Florit-Pons, P. Prieto and A. Igualada, "Co-creation of a narrative intervention program for speech-language pathology and educational settings," under review.

Different developmental paths of multimodal imitation in typically and non-typically developing preschool and primary school children

Mariia Pronina¹, Júlia Florit-Pons², Sara Coego², Pilar Prieto^{3,2}

Department of Catalan Philology and General Linguistics, The University of the Balearic Islands¹, Department of Translation and Language Sciences, Universitat Pompeu Fabra², Institució Catalana de Recerca i Estudis Avançats (ICREA)³

`mariia.pronina@uib.cat`

Imitation has been shown to act as a core mechanism for early social and language development [1-3], and imitation deficits have been linked to difficulties in communication skills [4]. While the crucial role of gesture and prosody imitation is well-documented and sentence imitation tasks are often used as diagnostic tools, few studies have compared imitation developmental paths and fewer still have assessed imitation considering both multimodal (gesture, prosody) and verbal aspects. This study compares the development of imitation skills in typically developing children (TD) and children with neurodevelopmental disorders (NDD) of preschool and early school age by focusing on multimodal imitation (sentences, prosody, gesture) and analyzes its link with oral language skills.

Participants were 290 Catalan-Spanish bilingual children (129 girls; 55 NDD) between ages 3 and 7 ($M=5$). Following a transdiagnostic approach, which suggests to soften the adherence to a diagnostic category [5], the NDD group included children with both Autism and DLD. All children undertook the Multimodal Imitation Task [6] and were asked to repeat contextualized sentences interacting with a toy while reproducing prosodic contours (i.e., statements, questions, exclamations) and imitating co-speech gestures (e.g., conventional, iconic). The accuracy of imitation of each component was assessed on a scale from 0-2. The children also undertook two language tasks: an expressive pragmatic test [7] and a narrative telling task.

LME models showed that all three imitation scores improved with age across all children and that the NDD group had overall significantly lower scores than the TD group and ($p<.001$, Fig.1). A significant interaction of age and group was found for gesture ($p<.001$), since only the TD group improved significantly in gesture imitation as they got older. We further analyzed separately the subset of responses when the children performed only gestural imitation but did not imitate the speech (gesture-only, NDD group: 11% of tokens, TD group: 3%) and the subset of responses when the children performed both gestural and speech imitation (gesture-speech integration). Gestural imitation of the NDD group was significantly better than that of the TD group in the gesture-only responses in the early ages (3-4) but not in gesture-speech integration responses (Fig.2). Moreover, in the TD group, all imitation scores were positively correlated ($p<.001$), while in the NDD group the gesture imitation did not correlate with sentence imitation. Similarly, for the TD group all imitation scores were correlated with pragmatic and narrative scores ($p<.001$), while in the NDD group, gesture did not.

These results show different imitation patterns for TD and NDD children. In TD children, all imitation skills are systematically improved with age, correlated with one another and with complex oral language skills (pragmatics and narrative). NDD children improve their imitation skills slower, with gesture showing no significant improvement over years. Gesture imitation in the NDD children behaves clearly differently from their other imitation abilities, since it was not correlated with speech imitation nor with other language abilities. Furthermore, higher percentage of gesture-only responses of the NDD children and higher gesture imitation scores in the early ages might indicate a compensatory gesture-based technique they apply to make up for the inability to imitate speech well. These results are of particular interest for the understanding of developmental paths of multimodal imitation and for their potential theoretical contribution to the understanding of gesture-language relationship in both typical and atypical populations.

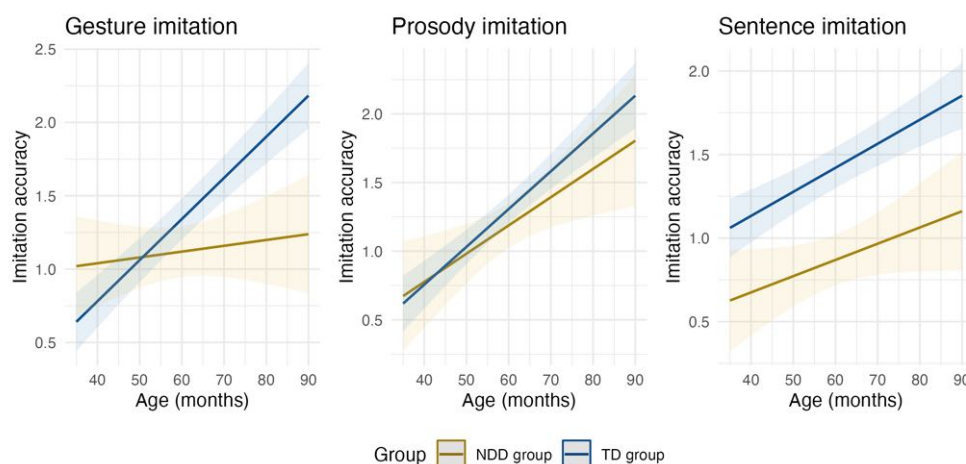


Figure 1: Predicted multimodal imitation scores over age in the typical and clinical groups.

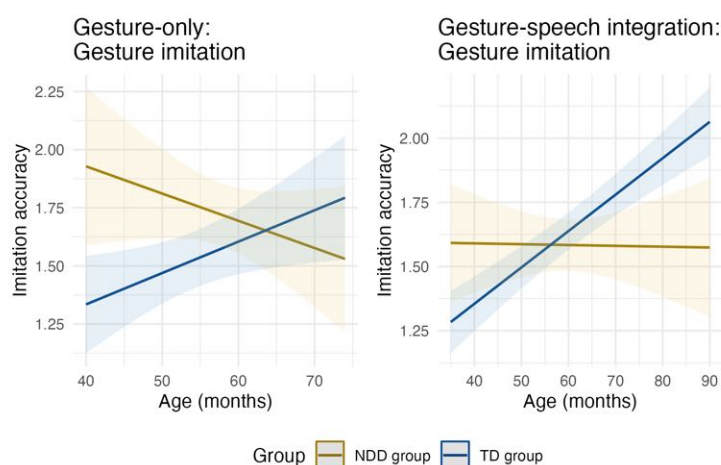


Figure 2: Predicted gesture imitation scores over age in the typical and clinical groups for gesture only performance and gesture-speech integration performance.

References

- [1] M. Carpenter, K. Nagell, and M. Tomasello, "Social cognition, joint attention, and communicative competence from 9 to 15 months of age," *Monogr. Soc. Res. Child Dev.*, vol. 63, no. 4, pp. i–vi, 1–143, 1998.
- [2] M. Carpenter, M. Tomasello, and T. Striano, "Role reversal imitation and language in typically developing infants and children with autism," *Infancy*, vol. 8, no. 3, pp. 253–278, Nov. 2005.
- [3] L. Hanika and W. Boyer, "Imitation and social communication in infants," *Early Child. Educ. J.*, vol. 47, no. 5, pp. 615–626, Sep. 2019.
- [4] J. Nadel, "How imitation boosts development: In infancy and autism spectrum disorder.," *How imitation boosts development: In infancy and autism spectrum disorder.* p. 237, 2014.
- [5] D. E. Astle, J. Holmes, R. Kievit, and S. E. Gathercole, "Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 63, no. 4, pp. 397–417, Apr. 2022.
- [6] E. Castillo, M. Pronina, I. Hübscher, and P. Prieto, "Narrative performance and sociopragmatic abilities in preschool children are linked to multimodal imitation skills," *J. Child Lang.*, vol. 50, no. 2, pp. 52–77, Dec. 2023.
- [7] M. Pronina, I. Hübscher, I. Vilà-Giménez, and P. Prieto, "A new tool to assess pragmatic prosody in children: Evidence from 3- to 4-year-olds," in *Proceedings of the 19th International Congress of Phonetic Sciences. 5-9 August 2019, Melbourne, Australia, 2019*, pp. 3145–3149.

Keynote 2: Judith Holler

26.09.2024

9:00-10:00



Producing and comprehending multimodal utterances in face-to-face conversation

Judith Holler

Donders Institute for Brain, Cognition & Behaviour, Radboud University

Max Planck Institute for Psycholinguistics, Nijmegen

judith.holler@donders.ru.nl

Face-to-face conversational interaction is at the very heart of human sociality and the natural ecological niche in which language has evolved and is acquired. Yet, we still know rather little about how utterances are produced and comprehended in this environment. This concerns especially the plethora of visual bodily signals that form part of utterances in face-to-face settings. In this talk, I will focus on how hand gestures, facial and head movements are organised to convey semantic and pragmatic meaning in conversation, as well as on how the presence and timing of these signals impacts utterance comprehension and responding. The basis for the studies I will discuss is a theoretical framework that situates language production and comprehension in face-to-face interaction and conversational turn-taking [1]. Conversational turn-taking is incredibly fast, thus creating a psycholinguistic bottleneck for processing incoming utterance information as well as preparing a timely response [2]. Visual bodily signals conveying semantic and pragmatic meaning that occur early on during an utterance are thus likely to be beneficial for early comprehension and fast responding. I will present studies based on complementary approaches, which feed into and inform one another. This includes qualitative and quantitative multimodal corpus studies showing that visual signals indeed often occur early [3, 4, 5], and experimental comprehension studies, which are based on and inspired by the corpus results, implementing controlled manipulations to test for causal effects between visual bodily signals and comprehension processes and mechanisms. These experiments include behavioural studies (using shadowing, categorical classification and reaction time paradigms [e.g., 6, 7, 8, 9, 10]) as well as EEG studies [e.g., 9, 10], most of them using multimodally animated virtual agents. Together, the findings provide evidence for the hypothesis that visual bodily signals form an integral part of semantic and pragmatic meaning communication in conversational interaction, and that they facilitate language processing, especially due to their timing and the predictive potential they gain through their temporal orchestration [1].

References

- [1] Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639-652.
- [2] Levinson, S. C. (2016). Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1), 6-14.
- [3] Ter Bekke, M., Drijvers, L., & Holler, J. (2024). Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech. *Cognitive Science*, 48(1): e13407.
- [4] Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11(8): 1017.
- [5] Nota, N., Trujillo, J. P., & Holler, J. (2023). Specific facial signals associate with categories of social actions conveyed through questions. *PLoS One*, 18(7): e0288104.
- [6] Ter Bekke, M., Drijvers, L., & Holler, J. (2024). Gestures speed up responses to questions. *Language, Cognition and Neuroscience*, 39(4), 423-430.
- [7] Ter Bekke, M., Levinson, S. C., Van Otterdijk, L., Kühn, M., & Holler, J. (2024). Visual bodily signals and conversational context benefit the anticipation of turn ends. *Cognition*, 248: 105806.
- [8] Trujillo, J. P., & Holler, J. (2024). Conversational facial signals combine into compositional meanings that change the interpretation of speaker intentions. *Scientific Reports*, 14: 2286.

- [9] Drijvers, L., & Holler, J. (2023). The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, 30(2), 792-801.
- [10] Nota, N., Trujillo, J. P., Jacobs, V., & Holler, J. (2023). Facilitating question identification through natural intensity eyebrow movements in virtual avatars. *Scientific Reports*, 13: 21295. doi:10.1038/s41598-023-48586-4.
- [11] Ter Bekke, M., Drijvers, L., & Holler, J. (in revision). Co-speech hand gestures are used to predict upcoming meaning.
- [12] Emmendorfer, A., & Holler, J. (in prep). Speaker gaze as a response mobilizing cue.

Session 4:

Embodiment and Arts

26.09.2024
10:00-11:20



Motif-Gesture Contiguity in Karnatak Vocal Performance: A Multimodal Computational Analysis

Lara Pearson¹, Thomas Nuttall², Wim Pouw³

Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany¹

Universitat Pompeu Fabra, Barcelona, Spain²

Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands³

lara.pearson@ae.mpg.de

Across a wide range of musical styles worldwide, vocalists tend to gesture while they sing. In Indian art music contexts, connections have been noted between bodily gestures and musical motifs [1], but systematic analyses remain few. Here we investigate whether performer gestures in Karnatak (South Indian) vocal performance systematically relate to co-occurring motifs: short musical patterns that act as building blocks of the style [2]. This research builds on work in gesture studies showing that semantically related gestures move alike [3], which found that both silent and co-speech gestures have similar kinematic trajectories when they convey a similar concept. In the current case we go a step further by assessing this in the challenging context of continuous Karnatak vocal performances. Conceptually we break novel ground in the understanding of multimodal communication outside the classical linguistic context, looking instead at multimodal meaning-making in musical vocal performance.

In this study, we ask whether there is a systematic relationship between sonic similarity of motifs and kinematic similarity of the co-occurring gestures. Through this inquiry, we also seek to better characterize the multidimensional codependencies of body movement and vocalizations. We analyze a dataset of 3.79 hours of Karnatak vocal performances (audio, video, motion-capture). Using a machine learning methodology tailored for Karnatak music, we locate regions of repeated melodic patterns across the dataset [4]. Dynamic time warping (DTW) distances between audio features (f_0 , Δf_0 , loudness, spectral centroid) and gesture (3d position, acceleration, velocity of hand/head motion) event trajectories are computed for each pairwise combination. We use correlation and regression analysis on these DTW values to assess whether acoustic motifs covary with spatiotemporal patterns of gesture (see Figure 1 for an overview of the analysis pipeline). In addition, we create an interactive visualization to further explore motif-gesture relationships.

Across all performers and performances, we find a significant positive correlation between all kinematic distances, and f_0 , Δf_0 and loudness distances (up to 0.42 r_s , $ps < .0001$). For individual performers, these correlations are greater (up to 0.53 r_s , $ps < .0001$), with notable individual differences observed. Three gradient boosted regression models trained to predict each sonic feature using kinematic features of hand, head, and combined head-hand evaluate on an unseen test set with results that in each case show combined head-hand features to be more strongly predictive than either head or hand alone.

The results show that sound and body movement are systematically related at the motif level. This suggests the potential for multimodal meaning-making through contiguity, a relation with semiotic potential where meaning is formed through bordering (spatial and/or temporal) of one thing against another [5]. The regression results imply that sound-gesture relationships are better understood when hand and head motion is combined. Through our analyses we also see how individual performers differ in the way they co-structure sound and movement, using differing characteristic salient dimensions (e.g., position change over acceleration, or loudness over pitch change).

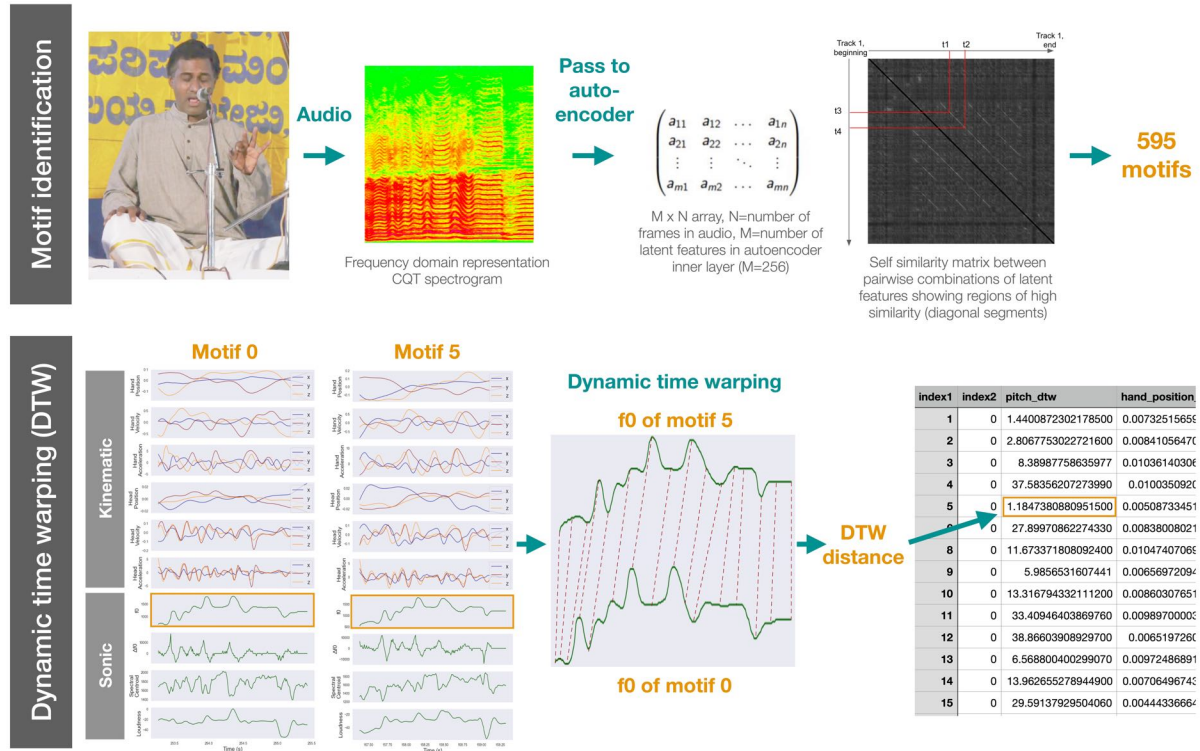


Figure 1: An overview of the first two stages in the analysis pipeline. The upper row shows the motif identification process, wherein pairwise regions of consistently high melodic similarity are identified as repeated motifs using features learnt by an autoencoder. The lower row visualizes the dynamic time warping process, in which DTW distances are calculated for pairs of all 10 sonic and kinematic features and placed in the DTW distance dataframe. The photograph shows the Karnatak vocalist, Hemmige S Prashanth, performing on stage in Mangalore in 2014.

References

- [1] M. Rahaim, *Musicking Bodies: Gesture and Voice in Hindustani Music*. Middletown, Conn.: Wesleyan University Press, 2012.
- [2] T. Viswanathan, "The Analysis of Rāga Ālāpana in South Indian Music," *Asian Music*, vol. 9, no. 1, pp. 13–71, 1977. doi: [10.2307/833817](https://doi.org/10.2307/833817)
- [3] W. Pouw, J. de Wit, S. Bögels, M. Rasenberg, B. Milivojevic, and A. Ozyurek, "Semantically Related Gestures Move Alike: Towards a Distributional Semantics of Gesture Kinematics," in *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior*, Springer, Cham, 2021, pp. 269–287. doi: [10.1007/978-3-030-77817-0_20](https://doi.org/10.1007/978-3-030-77817-0_20).
- [4] T. Nuttall, G. Plaja-Roglans, L. Pearson, and X. Serra, "In Search of Sancarar: Tradition-Informed Repeated Melodic Pattern Recognition in Carnatic Music," *In Proceedings of the 23rd International Conference on Music Information Retrieval (ISMIR)*, Bengaluru, India, pp. 337–344, 2022. <https://repositori.upf.edu/handle/10230/56440>
- [5] Mittelberg I, Hinnell J. *Gesture Studies and Semiotics*. In: Pelkey J, Cobley P (eds) *Bloomsbury Semiotics: Volume 4: Semiotic Movements*. London: Bloomsbury Academic, 2023. DOI: [10.5040/9781350139435](https://doi.org/10.5040/9781350139435).

Orofacial signals beyond sight: A study of expressive faces and whispered voices in GermanNasim Mahdinazhad Sardhaei¹, Marzena Zygis², Hamid Sharifzadeh³*Leibniz-Zentrum für Allgemeine Sprachwissenschaft¹, Humboldt Universität²,**Unitec Institute of Technology³*

na.mehdinejad@gmail.com

Human communication involves a multimodal system in which gestures play an integral role. In this domain, one of the intriguing questions is about the nature of the relationship between gesture and speech. Closely related to this, there are two influential hypotheses. One possible conjecture is that there is a trade-off relation between gesture and speech in terms of the communicative load [1], [2], [3]. Another alternative account is a hand-in-hand hypothesis viewing the relation between gestures and speech in parallel rather than compensatory [4], [5]. These two hypotheses largely depend on type of gesture as well as the communicative settings [6], [7]. In this study, we focus on measuring the orofacial expressions including eyebrow movements, eye opening, and lip aperture in polar questions with rising intonation vs. statements with falling intonation. The varying intonation can enable us to find out whether and to what extent speech with varying prosody interacts with the oro-facial expressions. Furthermore, taking “(semi-)whispered speech” and “invisibility” of speakers as two communicative difficulties into account, we aim to investigate what happens to speech and gesture when speakers (semi-)whisper and do not see each other.

We conducted an audio and video recorded experiment with 15 native German speakers producing 20 pairs of statements and questions, identical in content but differing in punctuation, i.e. a question mark in questions and a dot in statements. Each sentence was composed of 4 content words. The target word, which was the focus of our study, appeared at the sentence’s final position. All the target words were bisyllabic with the stress falling on the initial syllable. The stressed syllables had CVC structure containing one of the bilabial stops /p/, /b/, /m/ followed by an unrounded vowel of /e/, /a/, /i/. The experiment took part in the interaction between a confederate and a participant. The confederate, consistently the same speaker, either asked questions or made statements. Participants had to respond by turning questions into statements or statements into questions, adjusting their intonation accordingly. (see appendix). The data were double checked with respect to intonation perceptually by two native speakers of German. The experiment consisted of four stimulus blocks linking two conditions, i.e., speech mode [*normal, semi-whispered, and whispered speech*] and visibility [*visible vs invisible mode*]. Orofacial expressions were measured using Openface2 [8], which mapped 68 facial landmarks in each video.

Based on the results of linear mixed effect models, the three-way interaction between *Speech Mode*(In)visibility*Sentence Type* for both the right eyebrow ($t= 2.773$, $p<.01$, see Figure 1) and left eyebrow ($t= 2.248$, $p<.05$, see Figure 2) was significant indicating that speakers raise their eyebrows the highest when they produce statements in whispered speech and when they are visible. The results also revealed a significant effect between *Speech Mode*(In)visibility* for opening of both eyes ($t= 2.885$, $p < 0.01$ for the left eye, and $t= 2.758$, $p < .01$ for the right eye) with the larger opening in the (semi)whispered speech modes in visible condition as compared to the smaller opening in normal speech mode in visible condition. For the lip aperture, there was also a significant three-way interaction between *Speech Mode*(In)visibility*Sentence Type* ($t= 2.166$, $p<.05$). Pairwise comparisons showed lips are opened the largest in questions produced by whispered speech in invisible condition.

Overall, the results reveal more pronounced oro-facial expressions in a communicatively marked situation, i.e. when speakers whisper. Also, more pronounced orofacial gestures are produced when speakers see each other. We will discuss these findings in terms of trade-off and hand-in-hand hypothesis.

Appendix

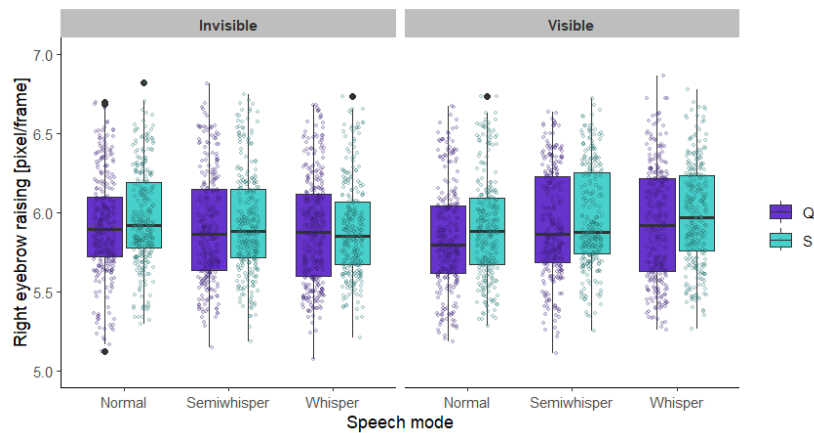


Figure 1: *Right eyebrow raising in the sentence-final word: interaction between **speech mode**, **(in)visibility**, and **sentence type***

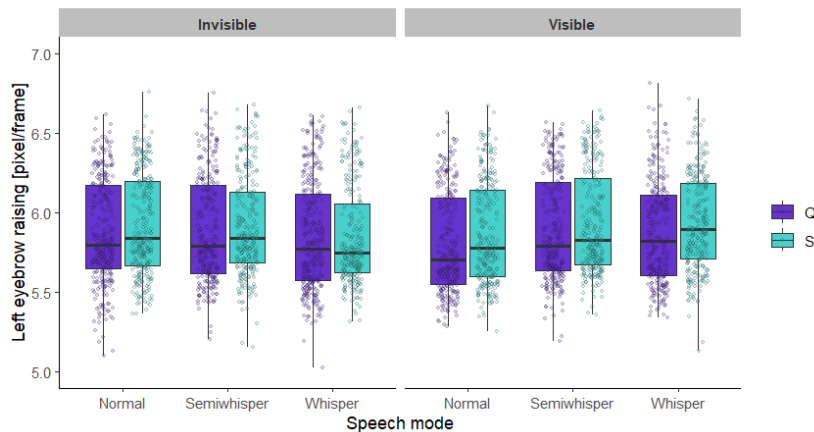


Figure 2: *Left eyebrow raising in the sentence-final word: interaction between **speech mode**, **(in)visibility**, and **sentence type***

References

- [1] A. Bangerter, "Using pointing and describing to achieve joint focus of attention in dialogue," *Psychological Science*, vol. 15, no. 6, pp. 415-419, 2004.
- [2] A. Melinger and W. J. Levelt, "Gesture and the communicative intention of the speaker," *Gesture*, vol. 4, no. 2, pp. 119-141, 2004.
- [3] J. P. De Ruiter, "Can gesticulation help aphasic people speak, or rather, communicate?," *Advances in Speech Language Pathology*, vol. 8, no. 2, pp. 124-127, 2006.
- [4] S. Goldin-Meadow, "How gesture promotes learning throughout childhood," *Child Development Perspectives*, vol. 3, no. 2, pp. 106-111, 2009.
- [5] W. C. So, S. Kita, and S. Goldin-Meadow, "Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand," *Cognitive Science*, vol. 33, no. 1, pp. 115-125, 2009.
- [6] J. Bavelas, "Gestures as part of speech: Methodological implications," *Research in Language and Social Interaction*, vol. 27, pp. 201-221, 1994.
- [7] J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost, "Gesturing on the telephone: Independent effects of dialogue and visibility," *Journal of Memory and Language*, vol. 58, pp. 495-520, 2008.
- [8] T. Baltrušaitis, A. Zadeh, Y. Ch. Lim, and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face and Gesture Recognition*, Xi'an, China, May 15-19, 2018.

The effects of familiarity on children's pantomimes

Elena Nicoladis¹
University of British Columbia¹
 elena.nicoladis@ubc.ca

When pantomiming actions performed with objects, young children (around three years of age) often produce a body-part-as-object (BPO), like an extended index finger for a toothbrush. In contrast, older children and adults often produce an imagined object (IO), such as pretending to hold a toothbrush. The most common explanation for this developmental change is in terms of children's symbolic understanding. That is, children do not initially grasp the abstract connection between arbitrary symbols and meaning. Iconic representations (such as onomatopoeia and iconic gestures) are therefore easier for children to relate to meaning than arbitrary representations (like words). Children therefore produce BPOs when pantomiming in order to provide concrete support for their understanding of the meaning [1]. As they develop better understanding of symbols, children start to produce IOs, as they no longer need the support of a concrete representation.

Researchers have pointed out a number of challenges in this explanation, including that even two-year-olds occasionally produce IOs and that the rate of BPOs/IOs varies considerably across items [2]. The latter result is surprising if the age-related shift toward IOs were related to global cognitive development (like symbolism). Weidinger et al. [3] proposed an alternative explanation for age-related changes in BPOs/IOs. They argued that IOs are produced when children have a rich conceptual understanding of the possible functions of a particular object. In other words, IOs emphasize which function one does with a particular object. In contrast, BPOs emphasize both the object and a function.

The purpose of the present study was to test a prediction that follows from Weidinger et al.'s [3] explanation. Namely, children should produce more IOs with familiar objects (because they have a rich understanding of their function) than with unfamiliar objects. Some research with adults has supported Weidinger et al.'s [3] explanation: England and Nicoladis [4] found that adults were more likely to produce IOs when pantomiming unfamiliar objects that they associated with multiple functions than for objects that they associated with a single function.

Preschool children between the ages of three and five years participated in this study. They were randomly assigned to pantomime either familiar objects (like a toothbrush) or unfamiliar objects (like a strawberry destemmer). The children in the two groups were matched on age, since age has been shown to be a strong predictor of BPO/IO production. The unfamiliar objects were taken from a previous study that had been with adults' pantomimes [4]. That study showed that these objects were unfamiliar even to adults. The children were asked to pantomime what to do with the objects. Their pantomiming was videotaped for later coding for use of either BPO or IO. The dependent variable was the percentage of IOs out of pantomimes with BPOs and IOs.

Consistent with predictions, the results showed that the children produced significantly more IOs with familiar objects ($M = 55\%$) than unfamiliar objects ($M = 17\%$).

These results are consistent with the argument that preschool children's pantomimes reflect the richness of their conceptual understanding of the particular target object [3]. Children's age-related increase in IOs could therefore be related to increased conceptual understanding first of particular objects and then, with even further experience, more generalized conceptual understanding of objects. Adults can likely infer possible functions of objects, even if they have had no experience with those particular objects [4]. As a result, they produce a lot of IOs when they pantomime.

References

- [1] A. S. Dick, W. F. Overton, and S. L. Kovacs, "The development of symbolic coordination: Representation of imagined objects, executive function, and theory of mind," *Journal of Cognition and Development*, vol. 6, no. 1, pp. 133-161, 2005. doi: 10.1207/s15327647jcd0601_8
- [2] P. Marentette, P. Pettenati, A. Bello, and V. Volterra, "Gesture and symbolic representation in Italian and English-speaking Canadian 2-year-olds." *Child Development*, vol. 87, no. 3, pp. 944-961, 2016. doi: 10.1111/cdev.12523
- [3] N. Weidinger, K. Lindner, K. Hogrefe, W. Ziegler, and G. Goldenberg, "Getting a grasp on children's representational capacities in pantomime of object use." *Journal of Cognition and Development*, vol. 18, no. 2, pp. 246-269, 2017. doi: 10.1080/15248372.2016.1255625
- [4] M. England and E. Nicoladis, "Functional fixedness and body-part-as-object production in pantomime." *Acta Psychologica*, vol. 190, pp. 174-187, 2018. doi: 10.1016/j.actpsy.2018.07.010

Session 5:

Sign languages

26.09.2024
11:30-12:50



Multimodal feedback in signed and spoken languages: Evidence for a shared infrastructure of conversation

Sonja Gipper*, Anastasia Bauer*, Jana Hosemann and Tobias-Alexander Herrmann
University of Cologne

**These authors contributed equally to this work.*

anastasia.bauer@uni-koeln.de

Any conversation among humans is rife with feedback, interactional moves that display some kind of stance towards another interlocutor's utterance [1]. Feedback signals are known to have different conversational functions: they may indicate a passive reciprocity, they may acknowledge and agree to what has been claimed, they may state a piece of information as new or evaluate a piece of information [2]. Pioneering research on feedback has provided valuable insights, focusing primarily on transcribed audio recordings of spoken language [3]–[5]. More recent research studying naturalistic conversational data unveiled feedback as a fundamentally multimodal phenomenon involving the coordination of different channels [6]–[10]. However, the use of multimodal cues as feedback in face-to-face interactions remains underresearched. Our understanding of how vocal, visual, manual and non-manual signals combine into complex feedback events in everyday conversation across different language modalities is limited. Therefore, we expand upon previous observations in the literature by investigating how feedback events vary in form and frequency in signed and spoken languages.

We examine feedback events from a multimodal and cross-linguistic perspective by utilizing corpora of casual conversations from four different languages: German Sign Language (DGS), Russian Sign Language (RSL), spoken German (GER) and spoken Russian (RUS) [11]–[15]. Our focus is on feedback signals may take the form of lexical cues (words like *ja* 'yes' or signs such as STIMMT 'right'), non-lexical cues (vocalizations like 'mm' or manual gestures such as palm-up), and non-manual cues such as head nods, eyebrow raise, or smile. Using parallel datasets and parallel annotation and analysis, we analyzed at least 45 minutes of face-to-face dyadic conversations in each of the four languages and identified ca. 1800 feedback events comprising roughly 3100 single feedback signals in total.

Our primary findings reveal: 1) Feedback is highly pervasive during interaction, occupying a significant portion of conversational time: roughly 3/4 of all feedback events occur within 5 seconds of the preceding feedback event in all four languages (1). Moreover, interlocutors generate feedback signals at a rate of up to 7–11 times per minute, on average. 2) there are remarkable similarities between signed and spoken languages in the form of feedback events, contrary to earlier assumptions of cross-modal variation [16]. The overwhelming majority of all feedback events are articulated with head or/and face: 80-99% were produced either only nonmanually or nonmanually in combination with spoken/signed elements across languages. This confirms and amplifies recent observations in the literature [17], [18].

We interpret these findings as contributing to the accumulating evidence supporting the existence of a shared interactional infrastructure of conversation among both signers and speakers [18]. Such cross-linguistic and cross-modal research is foundational to achieving a more comprehensive understanding of the use and interplay of multimodal cues in face-to-face interaction.

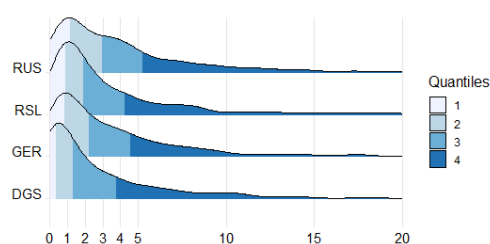


Figure 1: 3/4 of feedback events occur within 5 seconds after the end of the preceding one

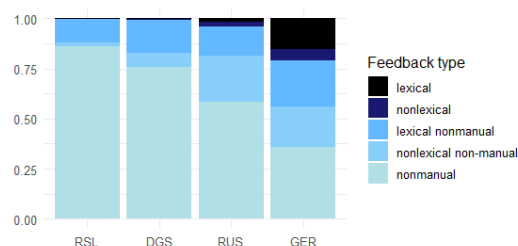


Figure 2: 80–99% of feedback events are constituted or accompanied by nonmanuals

References

- [1] J. Allwood, J. Nivre, and E. Ahlsén, “On the Semantics and Pragmatics of Linguistic Feedback,” *Journal of Semantics*, vol. 9, no. 1, pp. 1–26, 1992.
- [2] R. Gardner, *When Listeners Talk: Response tokens and listener stance* (Pragmatics & Beyond New Series). Amsterdam: John Benjamins Publishing Company, 2001, vol. 92.
- [3] V. H. Yngve, “On getting a word in edgewise,” in *Papers from the sixth regional meeting, Chicago Linguistic Society*, Chicago: Chicago Linguistic Society, 1970, pp. 567–577.
- [4] H. Sacks, E. A. Schegloff, and G. Jefferson, “A Simplest Systematics for the Organization of Turn-Taking for Conversation,” *Language*, vol. 50, no. 4, pp. 696–735, 1974.
- [5] E. A. Schegloff, “Discourse as an interactional achievement: Some uses of ‘uh huh’ and other things that come between sentences,” in *Analyzing discourse: Text and talk*, D. Tannen, Ed., Washington, D.C.: Georgetown University Press, 1982, pp. 71–93.
- [6] J. Allwood and L. Cerrato, “A Study of Gestural Feedback Expressions,” in *First Nordic Symposium on Multimodal Communication*, Copenhagen: Gothenburg university publications, 2003, pp. 7–22.
- [7] J. Allwood, S. Kopp, K. Grammer, E. Ahlsén, E. Oberzaucher, and M. Koppensteiner, “The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behavior simulation,” *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 255–272, Dec. 2007.
- [8] K. P. Truong, R. Poppe, I. D. Kok, and D. Heylen, “A multimodal analysis of vocal and visual backchannels in spontaneous dialogs,” *ISCA*, Aug. 2011, pp. 2973–2976.
- [9] C. Navarretta and P. Paggio, “Classification of Feedback Expressions in Multimodal Data,” Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 318–324.
- [10] Z. Malisz, M. Włodarczak, H. Buschmeier, J. Skubisz, S. Kopp, and P. Wagner, “The ALICO corpus: Analysing the active listener,” *Language Resources and Evaluation*, vol. 50, no. 2, pp. 411–442, Jun. 2016.
- [11] R. Konrad, T. Hanke, G. Langer, et al., *MY DGS – annotated. Public Corpus of German Sign Language*, 3rd release, 2020.
- [12] S. Burkova, *Russian Sign Language Corpus*, Novosibirsk University, 2015. [Online]. Available: <http://rsl.nstu.ru/>.
- [13] A. Bauer, *Russian multimodal conversational data*, 2023.
- [14] A. Bauer and R. Poryadin, *Russian Sign Language conversations*, 2023.
- [15] B. Hoffmann and N. Himmelmann, *Münster Videokorpus Alltagsgespräche. Unpublished corpus of spoken German*.
- [16] C. Backer, “Regulators and turn-taking in American Sign Language Discourse,” in *On the other hand: New perspectives on American Sign Language*. L. A. Friedman, Ed., vol. 81, 1977, pp. 138–139.
- [17] J. Mesch, “Manual backchannel responses in signers’ conversations in Swedish Sign Language,” *Language & Communication*, vol. 50, pp. 22–41, 2016.
- [18] H. Lutzenberger, L. D. Wael, R. Omardeen, and M. Dingemanse, “Interactional Infrastructure across Modalities: A Comparison of Repair Initiators and Continuers in British Sign Language and British English,” *Sign Language Studies*, vol. 24, no. 3, pp. 548–581, 2024.

Four seasons in one head:
The prosodic phrasing of enumerations in Portuguese Sign Language

Marisa Cruz & Sónia Frota
University of Lisbon
marisac@edu.ulisboa.pt

It is already known that the production of enumerations in signed languages makes use of three different strategies: spatial, linear, and digital [1]. Spatial enumeration consists in associating each element with a location in space; linear enumeration consists in producing a sequential list of elements, accompanied with nonmanual prosody; and digital enumeration involves the association of each element of the list with a finger of the nondominant hand. The latter strategy is considered the most used in signed languages to make lists [2]. However, as far as we know, the prosodic phrasing properties of enumerations remain unexplored.

Nonmanuals, in particular the head, have already been considered as markers of prosodic phrases in signed languages ([e.g., 3, 4]), and the amplitude of the head was shown to distinguish between prosodic domains (larger in Intonational Phrase boundaries than in Phonological Phrase boundaries) [5]. Our main goals are: (i) to determine whether the head plays a role in signaling prosodic boundaries in closed lists [6], and (ii) to establish whether this nonmanual behaves differently in linear and digital enumerations. Since in digital enumeration manuals are used to refer to each element of the list, we hypothesize that the head movement only plays a role in linear enumerations.

Using a corpus of role-play interviews in Portuguese Sign Language (LGP), obtained with an adapted version of the Discourse Completion Task [7, 8], we examined 10 utterances - closed lists corresponding to the four seasons of the year. The data was annotated in ELAN [9]. Four utterances were produced as linear enumerations; 6 were digital enumerations (Figure 1). Independently of the enumeration strategy used, all signers produced a falling head movement aligned with each element of the list as nonmanuals. A kinematic analysis of vertical head displacement (pixels - px) along the time series (ms) was conducted using *Kinovea* [10], to examine whether the amplitude of this nonmanual differed across enumeration strategies. The vertical displacement was automatically extracted from 100 datapoints time-normalized across utterances, thus resulting in a total of 1000 measurements for analysis.

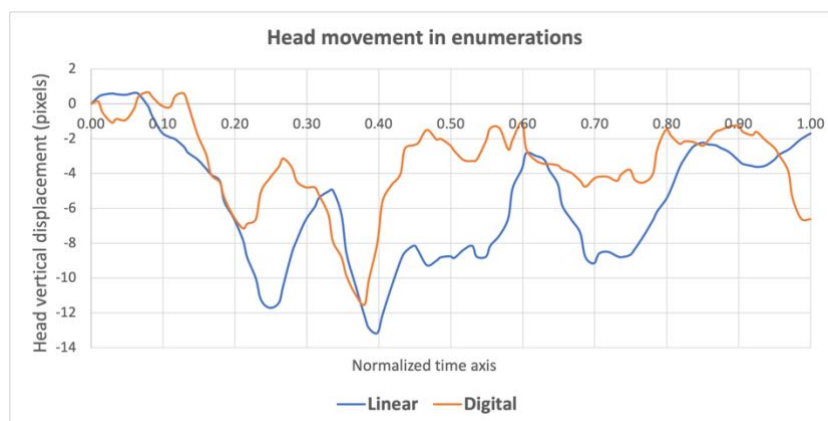
As shown in Figure 2, the pattern of the head movement is similar in linear and digital enumerations. However, the amplitude of the movement is larger in linear enumerations, thus suggesting that the head movement plays a relevant role when manuals are not used to mark prosodic boundaries. A Generalized Linear Mixed Model (GLMM) was run with participant as random factor and enumeration strategy as fixed factor. The dependent variable is the vertical displacement ($N=1000$). We found no significant effect of the random factor ($p>.05$), thus showing that the head vertical displacement did not vary across participants. Although the head amplitude is larger in linear enumerations ($M=-5.73\text{px}$, $SE=2.38\text{px}$) than in digital ones ($M=-3.41\text{px}$, $SE=1.94\text{px}$), it did not significantly differ between strategies [$F=.574(1, 8)$, $p=.470$].

We thus conclude that the head movement plays a role in signaling prosodic boundaries in enumerations, and that, although there was a trend for larger head displacement in linear enumerations, its role seems to be similar in linear and digital enumerations. However, further research is needed. We are now examining the exact prosodic domain(s) signaled by head movement in order to characterize the phrasing pattern of enumerations in LGP, as well as the role of the head (if any) as a prosodic boundary marker in enumerations produced in the spoken modality of Portuguese. This will add knowledge to the prosodic grammars of both signed and spoken modalities of Portuguese, and have implications to promote communication between the deaf and hearing communities.



Figure 1: *Frames of the digital enumeration of the four seasons of the year, produced by a native signer.*

Figure 2: *Head vertical displacement (pixels) along the normalized time axis for linear enumerations (blue line) and digital enumerations (orange line). This figure illustrates aggregated data.*



References (selected)

- [1] D. Pinsonneault and L. Lelièvre, "Enumeration in LSQ (Québec Sign Language): The use of fingertip loci", in I. Alhgren, B. Bergman and M. Brennan (Eds.), *Perspectives on sign language structure: Papers from the Fifth Symposium on sign Language Research* (Vol. 1, pp. 159–172), 1994. International Sign Linguistics Association.
- [2] S. Wilcox, A. Nogueira Xavier and S. Siltaloppi, "List constructions in two signed languages", *Language and Cognition*, vol. 16, no.1, pp. 57–92, 2024. doi:10.1017/langcog.2023.19
- [3] M. Nespore and W. Sandler, "Prosody in Israeli Sign Language", *Language and Speech*, vol. 42, no.2-3, pp. 143–176, 1999. <https://doi.org/10.1177/00238309990420020201>
- [4] S. Dachkovsky and W. Sandler, "Visual intonation in the prosody of a sign language", *Language and Speech*, vol. 52, no. 2/3, pp. 287-314, 2009. <https://doi.org/10.1177/0023830909103175>
- [5] M. Cruz and S. Frota, "Talking heads" in Portuguese Sign Language". Talk presented at the 35th Annual Conference on Human Sentence Processing, March 24-26, hosted by UC Santa Cruz, hybrid format, 2022.
- [6] M. Selting, "Lists as embedded structures and the prosody of list construction as an interactional resource", *Journal of Pragmatics*, vol. 39, pp. 483-526, 2007. doi:10.1016/j.pragma.2006.07.008
- [7] J. C. Félix-Brasdefer, "Data collection methods in speech act performance: DCTs, role plays, and verbal reports", in E. Usó Juan and A. Martínez-Flor (Eds.), *Speech act performance: Theoretical, Empirical, and Methodological Issues* (pp. 41-56), 2009. Amsterdam: John Benjamins Publishing.
- [8] K. Billmyer and M. Varghese, "Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests", *Applied Linguistics*, vol. 21, no. 4, pp. 517-552, 2000.

How visual cues make information units more prominent in spoken and signed languages: A case study on French and French Belgian Sign Language (LSFB)

Clara Lombart

University of Namur, NaLTT, LSFB-Lab

clara.lombart@unamur.be

With advancements in gesture studies, scholars (e.g., [1], [2]) began to advocate for comparisons between spoken and signed languages that consider the multimodal aspects of the former. However, such comparisons remain scarce (e.g., [3]). This study seeks to address this gap by investigating the resources used in Belgian French and LSFB (French Belgian Sign Language) to prosodically mark contrastive focus (CF), defined as the opposition between several explicit alternatives that form a limited set of possibilities [4]. At the pragmatic level, CF is deemed more prominent than information focus and background information (e.g., [5], [6]). Interestingly, the information unit of CF can be further divided into different subtypes (i.e., discourse opposition, selection, and correction; see Examples 1-3) and each correlates with an increase in pragmatic prominence (e.g., [4], [6]). Building on these considerations, the present study aims to answer the following questions: Is there a correlation between the increase in pragmatic prominence and the increase in prominence in marking for the encoding of CF and its subtypes? How are the prosodic cues of CF distributed in Belgian French and LSFB?

To investigate these questions, data from five Belgian French speakers (from the FRAPé Corpus [7]) and five LSFB signers (from the LSFB Corpus [8]) were examined. The participants were engaged in two spontaneous tasks: describing a face drawing and categorizing sets of similar objects. CF (and their subtypes) were identified from an informational perspective [4]. Inter-Pausal units containing, preceding, and following CF were annotated in both languages, to investigate how CF was marked compared to the surrounding non-contrastive context. In Belgian French, prosodic features, including syllabic duration, pitch mean and range, tone, and degree of prominence were annotated. As prosody in SpLs can also be multimodal (e.g., [9], [10]), hand, eyebrow, head and torso gestures were taken into account. In LSFB, the following prosodic features were considered: sign holding, sign repetition, dominance reversal, sign lengthening, and displacement, as well as non-manual cues such as eyebrow, head, and body movements, mouthings, and mouth gestures.

For this contribution, 380 instances of CF in each language (i.e., 3063 syllables and 375 manual gestures in Belgian French; 1424 manual signs in LSFB; 2559 non-manual cues for both languages) were examined. In Belgian French, CF shows greater prosodic prominence mainly through longer syllabic duration and wider pitch range, compared to non-contrastive elements. Similarly, gestures are more frequent in contrastive contexts (except for head movements), with a preference for synchronizing with more prominent CF instances. The same is true for LSFB, where CF is produced with greater prominence than non-contrastive elements through sign lengthening, sign holding, mouth articulations, body leans, and eyebrow movements. These findings can be explained by the fact that when a speaker or signer presents contrastive information, it creates challenges in updating the common ground because it requires the addressee to adjust their existing assumptions. Speakers and signers thus employ strategies, such as the ones outlined above, to facilitate this adjustment process [6].

Furthermore, for both languages, discourse opposition shows a higher degree of marking than correction, which was unexpected. A possible reason for this result can be found in the principle of least collaborative effort [11]: when communicating, individuals tend to resolve conversational problems in the most cost-efficient manner. This implies minimizing efforts as much as possible if context allows it [12]. By being less specific – because less pragmatic prominent – than correction, discourse opposition requires more marking to be understood in the context.

Examples

- (1) On one perspective, there is [A DUCK]_{DISCOURSE OPPOSITION} but from another perspective, there is [A RABBIT]_{DISCOURSE OPPOSITION}
- (2) Context – Participant A: For dessert, we eat that big round cake I was telling you about earlier.
Participant B: I prefer [CAKE]_{SELECTION} over ice cream.
- (3) You said that potatoes belong to the group of vegetables but they belong to [THE GROUP OF STARCHY FOOD]_{CORRECTION}

References

- [1] M. Vermeerbergen and E. Demey, ‘Sign + Gesture = Speech + Gesture?’, in *Simultaneity in Signed Languages: Form and function*, M. Vermeerbergen, L. Leeson, and O. Crasborn, Eds., Amsterdam: John Benjamins, 2007, pp. 257–282.
- [2] C. Müller, ‘Gesture and Sign: Cataclysmic Break or Dynamic Relations?’, *Front. Psychol.*, vol. 9, pp. 1–20, Sep. 2018, doi: 10.3389/fpsyg.2018.01651.
- [3] S. Gabarró-López and L. Meurant, ‘Contrasting signed and spoken languages: Towards a renewed perspective on language’, *LiC*, vol. 22, no. 2, pp. 169–194, 2022, doi: 10.1075/lic.00024.gab.
- [4] S. Repp, ‘Contrast: Dissecting an Elusive Information-structural Notion and its Role in Grammar’, in *The Oxford Handbook of Information Structure*, C. Féry and I. Shinichiro, Eds., Oxford: Oxford University Press, 2016, pp. 270–289.
- [5] V. Molnár, ‘Contrast – from a contrastive perspective’, in *Information Structure in a Cross-Linguistic Perspective*, H. Hasselgård, S. Johansson, B. Behrens, and Fabricius-Hansen, Eds., Leiden: Brill, 2002, pp. 147–161. doi: 10.1163/9789004334250_010.
- [6] M. Zimmermann and E. Onea, ‘Focus marking and focus interpretation’, *Lingua*, vol. 121, no. 11, pp. 1651–1670, 2011, doi: 10.1016/j.lingua.2011.06.002.
- [7] L. Meurant *et al.*, ‘Corpus de français parlé : vers la construction d’un corpus comparable LSFB - Français de Belgique.’ Université de Namur: Laboratoire de Langue des Signes de Belgique francophone (LSFB-Lab), Under Construction.
- [8] L. Meurant, ‘Corpus LSFB. Un corpus informatisé en libre accès de vidéos et d’annotations de la langue des signes de Belgique francophone (LSFB)’. FRS-F.N.R.S et Université de Namur., 2015. [Online]. Available: <http://www.corpus-lsfb.be>.
- [9] L. Brown and P. Prieto, ‘Gesture and Prosody in Multimodal Communication | Semantic Scholar’, in *The Cambridge Handbook of Sociopragmatics*, M. Haugh, D. Kádár, and M. Terkourafi, Eds., Cambridge: Cambridge University Press, 2021, pp. 430–453.
- [10] G. Ambrazaitis and D. House, ‘The multimodal nature of prominence: some directions for the study of the relation between gestures and pitch accents’, in *Proceedings of the 13th International Conference of Nordic Prosody*, Sciendo, 2023, pp. 262–273. doi: 10.2478/9788366675728-024.
- [11] M. Rasenberg, W. Pouw, A. Özyürek, and M. Dingemanse, ‘The multimodal nature of communicative efficiency in social interaction’, *Sci Rep*, vol. 12, no.19111, 2022, doi: 10.1038/s41598-022-22883-w.
- [12] H. H. Clark, *Using Language*. in ‘Using’ Linguistic Books. Cambridge: Cambridge University Press, 1996. doi: 10.1017/CBO9780511620539.

The phonetics of addressee's head nods in signed and spoken interaction using a computer vision solution

Anastasia Bauer¹, Anna Kuder¹, Marc Schulder², Job Schepens¹

1 Department of Linguistics, General Linguistics, University of Cologne, Cologne, Germany

2 Institute for German Sign Language and Communication of the Deaf, University of Hamburg, Hamburg, Germany
 anastasia.bauer@uni-koeln.de

Head nod is one of the most commonly produced bodily signals in interaction, an up-and-down movement of the head, often repeated. Both signers/speakers as well as addressees produce head nods during face-to-face interaction. Head nod is associated with a number of different communicative functions in interaction such as affirmation, emphasis, affiliation, and feedback among other [1]. In sign language linguistics, head nods are recognized as aspectual or prosodic non-manual markers to signal clause and constituent boundaries and to mark phonological and intonational phrases in narratives [2], [3]. However, most claims about the phonetic properties of head nods have been based on manual annotation without reference to naturalistic text types and the head nods produced by the addressee have been largely ignored (with notable exception of the work by Pupponen and Mesch [4], [5]). We thus lack detailed information about the phonetic properties of addressee's head nods and not much is known about whether linguistic functions of head nods influence their phonetic form.

This study presents findings about the phonetic properties of the addressee's head nods in natural dyadic signed and spoken interaction. The aim is to find out whether head nods serving different pragmatic functions in interaction vary in their phonetic/kinematic characteristics. We hypothesize that affirmation nods differ from feedback nods in both language modalities. We use the term 'affirmation' to describe somebody's positive reaction to a preceding question. We define a nod as fulfilling the function of feedback, when it functions as an interactional behavior that displays interlocutors' perception or understanding of the course of the conversation. We focus on the very common feedback mechanisms which can signal a non-uptake of a conversation turn, acknowledge a prior statement or demonstrate understanding of the information represented by another signer (aka continuers, backchannels, minimal responses) [5], [6].

To test the hypothesis, we combine manual annotation in ELAN with quantitative analysis of body pose information generated using the computer vision toolkit OpenPose [7] to extract head nod measurements from video recordings and examine head nods in terms of their duration, amplitude and velocity. The applicability of computer vision tools for phonetic analysis of non-manuals has been successfully tested for sign languages [8]. This study provides a cross-modal analysis of head nods. We use the publicly available data from the DGS (German Sign Language) Corpus [9] and newly collected multimodal data from spoken German conversations.

We investigate ca. 4 hours of naturalistic dyadic interaction per each language and identify more than 600 occurrences of nods in each dataset. While the quantitative data analysis for spoken German is still ongoing, our DGS results show that phonetic properties of affirmative nods differ from those of feedback nods in velocity and maximal amplitude. Feedback nods appear to be on average slower in production and smaller in amplitude than affirmation nods. We attribute the variation in phonetic properties of head nods to the distinct roles these cues fulfill in the turn-taking system (feedback nods are usually passive reciprocity signals and affirmation nods signal turn initiation). Our preliminary results reveal no cross-modal differences in phonetic characteristics of addressee's head nods.

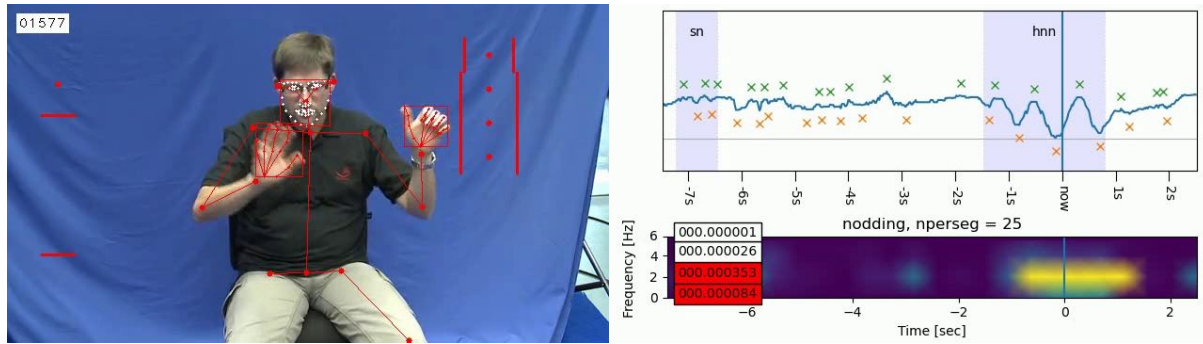


Figure 1: Visualization of head nods. The source video (left) is overlaid with the OpenPose body points used for the calculations. On the line graph (upper right) the upper (blue) line represents the vertical motion of the nose relative to body position, while the lower (red) line indicates the nose location prediction confidence of OpenPose (worse during blur or when occluded). Light blue boxes indicate durations manually labeled as head nods. The spectrogram (lower right) visualizes the spectrum of frequencies of vertical nose movement, with brighter areas indicating repeated up and down motion as during nodding.

References

- [1] L. Cerrato, “Linguistic functions of head nods,” in *Proceedings from The Second Nordic Conference on Multimodal Communication*, Gothenburg, Sweden: Göteborg University, 2005, pp. 137–152.
- [2] S. K. Liddell, *American Sign Language Syntax* (Approaches to Semiotics 52). Berlin: De Gruyter Mouton, 1980.
- [3] R. B. Wilbur, “Non-manual markers: Theoretical and experimental perspectives,” in *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, J. Quer, R. Pfau, and A. Herrmann, Eds., London: Routledge, 2021, pp. 530–565.
- [4] A. Puupponen, T. Wainio, B. Burger, and T. Jantunen, “Head movements in finnish sign language on the basis of motion capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls,” *Sign Language & Linguistics*, vol. 18, no. 1, pp. 41–89, 2015. DOI: 10.1075/s11.18.1.02puu.
- [5] J. Mesch, “Manual backchannel responses in signers’ conversations in swedish sign language,” *Language & Communication*, vol. 50, pp. 22–41, 2016. DOI: 10.1016/j.langcom.2016.08.011.
- [6] R. Gardner, *When Listeners Talk: Response tokens and listener stance* (Pragmatics & Beyond New Series). Amsterdam: John Benjamins Publishing Company, 2001, vol. 92.
- [7] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, “OpenPose: Realtime multi-person 2d pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021. DOI: 10.1109/TPAMI.2019.2929257.
- [8] A. Chizhikova and V. Kimmelman, “Phonetics of negative headshake in russian sign language: A small-scale corpus study,” in *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, E. Efthimiou, S.-E. Fotinea, T. Hanke, et al., Eds., European Language Resources Association (ELRA), 2022, pp. 29–36.
- [9] R. Konrad, T. Hanke, G. Langer, et al., *MY DGS – annotated. public corpus of german sign language, 3rd release*, version 3.0, 2020. DOI: 10.25592/dgs.corpus-3.0.

Session 6:

Methodical Perspectives

26.09.2024
14:20-15:20



Looking together.

An eye-tracking corpus of museum visitors' shared experience and joint attention

Geert Brône, Bert Oben, Julie Janssens

University of Leuven, Department of Linguistics, MIDI

geert.brone@kuleuven.be

Whereas experimental and corpus-based studies on multimodal communication traditionally resort to lab-based environments with task-based interactions, an increasing number of studies is moving towards authentic real-life settings [1]. Doing so potentially increases the ecological validity of empirical research on multimodal communication, but collecting and analyzing naturalistic data also comes with obvious challenges, ranging from privacy issues when recording in public space, over dealing with noisy data to (technically) managing the unpredictability of dynamic interactions with moving (“walking and talking”) participants. What is largely missing to date, is a description of the workflow for multimodal data collection in such real-life dynamic settings. In this contribution, we describe the pipeline for one such recently collected corpus that is unique through the use of mobile eye-tracking technology to capture participants' gaze behavior and other features of embodied behavior while they interact with artefacts, space and fellow visitors in an art museum.

Rationale – museum settings provide a particularly interesting context for multimodal analysis: (1) visitors typically visit museums and exhibitions in the company of others, as a consequence of which they need to oscillate their attention between companions, exhibited artefacts and other visitors; (2) museums are places of aesthetic experience and discovery, making them a prime locus for the study of the interface between (joint) attention and stance-taking, which are both known to be interactionally negotiated with various semiotic resources (pointing gestures, head movements, facial expressions, gaze, etc.) [2].

Participant info – in order to collect authentic visitor interactions during a real-life museum experience, we invited participants into the Royal Museum of Fine Arts in Antwerp, to visit a temporary museum dedicated to portrait painting (the *Turning Heads* exhibition). In collaboration with the museum, which used their social media network, we recruited visitor pairs (as well as individual participants) who participated in the study. They were briefly informed about the goal of the recordings and signed an informed consent but they were given no particular instructions and could visit the museum at their own pace.

Recording set-up – in order to get fine-grained access to the visitors' verbal and embodied interactions as well as their navigation through space, we combined multiple camera systems. Each of the participants wore a head-mounted eye-tracking system (Tobii 3 Glasses), providing information on the participants' gaze behavior (including moments of mutual gaze). In addition, we followed the participants from a distance (approx.. 5 meters) using a hand-held GoPro camera. Using this shadowing technique, we gain an external but at the same time dynamic perspective on the participants' interactions and movements through space (Fig. 1)

Data processing – the data collection, organized in Dec. 2023 and Jan. 2024, resulted in a corpus of 30 visitor pairs and 20 individual participants, with an average recording time of 35 minutes per visit. All recordings were synchronized to trivid videos (Fig. 1) and transcribed using ELAN. The continuous gaze data generated by the eye-tracking systems were annotated into discrete annotation categories to allow for the quantitative analysis of co-occurrence patterns of eye gaze between participants, and multimodal patterns across different modalities.

In this presentation, we raise issues and suggest solutions to the challenges involved in this type of multimodal data gathering. This discussion includes the usefulness of pre- and post-test questionnaires, overcoming the observer's paradox, managing privacy, synchronizing data, storing and sharing large volumes of data, and annotating data at multiple levels.



Figure 1: screenshot of the synchronized video files, with two ‘internalized’ perspectives generated by the mobile eye-tracking systems (top left and right) and one external camera perspective (bottom, shadowing technique using hand-held camera). The red dots in the eye-tracking data (the gaze cursor) show the participants’ gaze fixations.

References

- [1] S. Barnes and F. Possemato, “Multimodal analysis of interaction,” *The Handbook of Clinical Linguistics* (Eds. M. Ball, N. Müller and E. Spencer), 2024. doi: 10.1002/9781119875949.ch9
- [2] F. Andries, K. Meissl, C. de Vries, K. Feyaerts, B. Oben, P. Sambre, M. Vermeerbergen and G. Brône, “Multimodal stance-taking. A systematic literature review,” *Frontiers in Communication*, 8, doi: 10.3389/fcomm.2023.1187977.

Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction

Sam O'Connor Russell and Naomi Harte

Dept. of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

`russelsa@tcd.ie`

At key moments in conversation, the floor is either *held* by the speaker or *shifted* to the listener [1]. Thus, robots need to make fast and accurate hold/shift (H/S) decisions in order to converse naturally with humans. Recent turn-taking models enable H/S decisions 150 ms after a turn [2]. However, this involved telephone speech where interlocutors had no visual cues. As visual cues provide a wealth of turn-taking information when participants can see one another [1], our work investigates if they should be incorporated into turn-taking models.

We report on our ongoing experiments. We train a causal, transformer-based turn-taking model first introduced by Ekstedt and Skantze (see [2]). We use the Switchboard corpus (260 hours of telephone speech) [3] and the audio from the Candor corpus (850 hours of videoconferencing speech) [4]. We identify the Switchboard H/S times using the ground-truth alignment. We use the Speechmatics ASR toolkit for Candor, as the provided alignment is inaccurate. We also obtained an equivalent ASR transcription for Switchboard. We train three separate models with: 1) Switchboard and the ground-truth alignment, 2) Switchboard with the ASR alignment and 3) Candor audio with the ASR alignment. We withhold 20% of sessions for testing.

We identify all periods of silence greater than 200 ms with a speaker change (a shift) and where the speaker stays the same (a hold). We report performance in Table 1 ('full set'). The 'reduced set' in Table 1 represents the original method for identifying H/S times from [2]. They impose a constraint that only a single speaker should be active around a H/S time, reducing the number and complexity of H/S considered. Our F1 scores reproduce their results (0.94 S 0.53 H) on Switchboard. The performances drop on the 'full set' reflects the inclusion of more challenging turn shifts. The use of ASR labels also reduces performance on Switchboard. This is due to the error between the ASR and ground-truth alignments, which we compute as 70 ms per word on average. We find a notable performance drop when we deploy the Switchboard-trained model on the Candor corpus and vice-versa (Table 2). The F1 score drops by 11% for holds (Switchboard-trained on Candor) and by 40% for shifts (Candor-trained on Switchboard, Table 2). We are working on the following hypothesis to explain this finding: as interlocutors can see one another in Candor, they change the way in which they exploit the audio channel for signalling their intentions. Thus, a telephone-speech trained model cannot predict turn-taking accurately by using audio from a videoconferencing interaction and vice-versa.

We find evidence to support the visual signalling of turn-taking by extracting facial action units (FAUs) from the Candor videos using OpenFace [5]. We compute the percentage of frames in which each FAU is active in 500 ms windows. Windows are considered before and after shifts, before and after holds, and during random periods when speaking and not speaking. We use a Mann-Whitney U test to compare median percentages. The visual channel is clearly used to signal shifts and holds. We see in Table 3, that interlocutor lip, jaw and mouth movement are more likely to occur just before a shift (i.e. before a speaker begins to speak) than at any other period when not speaking. This indicates that the visual modality is used to signal in advance that an interlocutor wishes to 'grab the turn'. We find less pronounced differences in FAU activations for holds which could indicate that holds are less reliant on the visual channel, reflecting our smaller performance drop in (Table 2). We demonstrate the turn-taking model's prediction with an example in Figure 1. We are currently working on incorporating visual information into a turn-taking model. Our workshop presentation will report up-to-date results in exploiting relevant visual cues within a neural turn-taking model.

Model trained + deployed on		Balanced F1 score						
		Full set			Reduced set			
		H	S	# shifts	H	S	# shifts	
Switchboard (telephone)	Ground-truth	0.89	0.57	3939	0.94	0.53	1253	
Switchboard (telephone)	ASR	0.88	0.51	1391	0.91	0.49	1013	
Candor (videoconference)	ASR	0.82	0.55	7219	0.88	0.55	3595	

Table 1: *F1 values for shift (S) and hold (H) predictions of models trained on the audio from the Switchboard or Candor corpus*

Model trained on		Model deployed on		Balanced F1 Score	
				Full set	
				H	S
Switchboard (telephone)	Candor (videoconference)			0.73 (↓ 11%)	0.52 (↓ 5%)
Candor (videoconference)	Switchboard (telephone)			0.85 (↓ 3.4%)	0.31 (↓ 40%)

Table 2: *Comparing the performance of a model trained on one corpus, deployed on another. The drop in performance relative to Table 1 indicates audio cues are used differently in these corpora.*

Comparing facial action units in Candor (videoconference) when					
Facial action unit	Not speaking vs. when speaking	Not speaking vs. before a shift	Random speech vs. after a shift	Random speech vs. before a hold	Random speech vs. before a hold
Inner brow raiser		+ 6%	+1%	+4%	4%
Cheek raiser		+8%	+9%	+2%	+2%
Nose wrinkler			+4%	+2%	+3%
Upper lip puller		+11 %	+9%		
Lip corner puller		+11 %	+12%		
Jaw drop	+ 10%	+17%	+8%		
Mouth stretch	+ 10 %	+8%			
Blink	+ 4 %	+6%			

Table 3: *This table shows that certain FAUs are more likely during turn-taking events. For example, the jaw drop FAU is observed 17% more when the interlocutor is about to speak (i.e. just before a shift) than when the interlocutor is not about to speak. Only statistically significant results are displayed.*

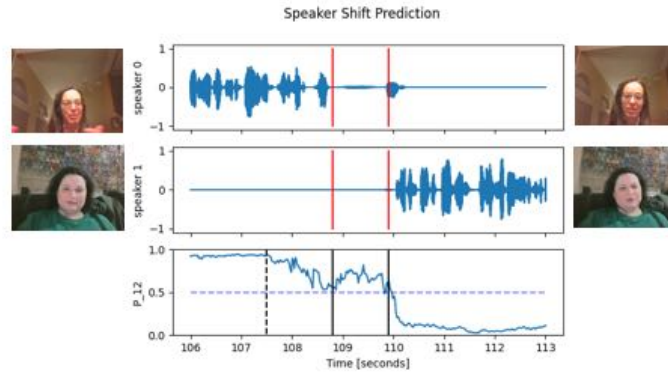


Figure 1: *Candor turn-taking model predicting a shift, with 1 sec silence in red. P_{12} is the probability that Speaker 0 will speak 2 sec from the present time. It begins to fall two secs before the silence (dashed line), illustrating the model’s predictive ability. During the silence, the model incorrectly favours speaker 0. We believe that visual information (note speaker 0 gaze aversion) could improve prediction.*

References

- [1] G. Skantze, “Turn-taking in Conversational Systems and Human-Robot Interaction: A Review,” *Computer Speech & Language*, vol. 67, p. 101 178, May 2021.
- [2] E. Ekstedt and G. Skantze, “Voice activity projection: Self-supervised learning of turn-taking events,” *INTERSPEECH 2022*, pp. 5190–5194, 2022.
- [3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “Switchboard: Telephone speech corpus for research and development,” in *Acoustics, speech, and signal processing, IEEE international conference on*, IEEE Computer Society, vol. 1, 1992, pp. 517–520.
- [4] A. Reece, G. Cooney, P. Bull, *et al.*, “The CANDOR corpus: Insights from a large multi-modal dataset of naturalistic conversation,” vol. 9, Mar. 2023, ISSN: 2375-2548.
- [5] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: An open source facial behavior analysis toolkit,” in *IEEE Winter Conference on Computer Vision*, 2016.

Navigating the topical landscape: Pointing at others as an embodied backlinking device in multi-party interaction

Mojenn Schubert¹

Leibniz-Institute for the German Language, Mannheim¹

schubert@ids-mannheim.de

In multi-party interaction, individual participants are involved with the topics discussed to varying degrees. Associating certain topics with those who are most committed to them therefore could be a natural way to structure conversations with such dynamic participation frameworks [1]. This paper presents an embodied strategy for making topical connections transparent by utilizing the co-presence of interlocutors: Pointing at co-participants. In detail, I analyze moments of self-selection in which a speaker initiates a new contribution whose relation to prior talk and activities is made clear through pointing. This can be done in sequential proximity to the conversational point of reference, but also across longer stretches of talk so that a more distant previous topic is made relevant again. An example of the focus phenomenon under study can be seen in the following extract coming from a dinner conversation among three friends (GS, NG and ZF). Just before the transcript, GS talked about his grandmother who keeps her household items for a very long time. In l. 01, he begins to formulate a closing assessment of this behavior.

Ex. 1 FOLK_E_00293_SE_01_T_02 c801

- | | | | |
|------|--|---|---------------|
| 1 GS | isch äh (.) manschma ich schieb_s immer auf_n krieg; ich sag die is | | |
| | <i>I uhm (.) sometimes- I always blame that on the war I say she</i> | | |
| 2 GS | im krieg gro[ß gewor#]den, (0.22) +[d#ie+ hat #+] | | |
| | <i>grew up in the war she has</i> | | |
| 3 NG | [hm: hm, #] | + [m#h+: ; | #+] (.) +.hh+ |
| | <i>uh huh</i> | | |
| | ng | +...+points at GS+-----+, , , + | |
| | Fig. #1 | #2 | #3 |
| 4 ZF | die sind (halt/ja) so; (ne,) | | |
| | <i>they are like this (aren't they)</i> | | |
| 5 | (0.22) | | |
| 6 GS | die (.) +die pflegt +die +sache noch;+ | | |
| | <i>she (.) she still takes care of her belongings</i> | | |
| | ng | +points at GS+-----+, , , , , , , , , , + | |
| 7 NG | zu dem <u>thema</u> , | | |
| | <i>on this topic</i> | | |
| 8 | (1.89) | | |
| 9 NG | <u>eins</u> zu eins es gleiche bei mir; mein opa is ja gestorben | | |
| | <i>one-to-one the same for me my grandpa passed away</i> | | |

NG, up to that point behaving as recipient to the unfolding story (see his continuer in l. 3), points twice at the current speaker GS (l. 3 and l. 6) to help claim speakership in a relatively fixed participation framework of the storytelling activity. Then, he formulates a connection to what has just been said (*on this topic*, l. 7) and further frames his upcoming contribution as being about a similar experience (*one-to-one the same for me*, l. 9). The pointing gesture not only facilitates claiming the floor [2], but it also establishes a deictic reference to a specific preceding speaker before the newly initiated turn is even formulated. Building on previous research on interactive pointing gestures that refer to common ground [3] [4] [5], I argue that the deictic nature of pointing can be used to refer back to parts of the conversation associated with the specific person being pointed to. Drawing on video recordings of multi-party interactions in German (16 hours, FOLK [6]), this study analyses moments (72 cases) in which participants use index finger pointing as an embodied backlinking device [7]. Using CA [8] [9] and multimodal interaction analysis [10] [11] [12], it is shown that pointing gestures are an effective instrument for navigating the topical landscape that builds up over the course of a conversation by using the people involved as thematic anchor points.



Figure 1

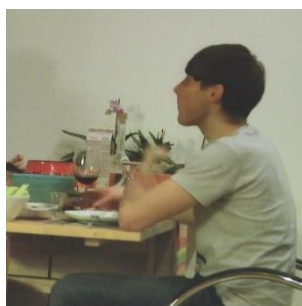


Figure 2



Figure 3

References

- [1] E. Goffman, "Forms of Talk", University of Pennsylvania Press, 1981.
- [2] L. Mondada, "Multimodal resources for turn-taking: pointing and the emergence of possible next speakers", *Discourse Studies*, vol. 9, no. 2, pp. 194–225, 2007.
- [3] J. B. Bavelas, N. Chovil, D. A. Lawrie, & A. Wade, "Interactive gestures", *Discourse Processes*, vol. 15, pp. 469–489, 1992.
- [4] J. B. Bavelas, N. Chovil, L. Coates, and L. Roe, "Gestures specialized for dialogue", *Personality and Social Psychology Bulletin*, vol. 21, pp. 394–405, 1995.
- [5] J. Holler, "Speakers' 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), Springer, 2010.
- [6] T. Schmidt, "Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German", In: J. M. Kirk & G. Andersen (Eds.), *Compilation, transcription, markup and annotation of spoken corpora. Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3]* (pp. 396–418), 2016.
- [7] E. A. Schegloff, "Turn organization: one intersection of grammar and interaction", In: E. Ochs, E. A. Schegloff, & S. Thompson (Eds.), *Interaction and Grammar* (pp. 52–133), Cambridge University Press, 1996.
- [8] H. Sacks, E. A. Schegloff, & G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation", *Language*, vol. 50 (4/1), pp. 696–735, 1974.
- [9] J. Sidnell, "Basic conversation analytic methods", In: J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 77–99), Wiley-Blackwell, 2013.
- [10] A. Deppermann & J. Streeck, "The body in interaction: Its multiple modalities and temporalities", In: A. Deppermann & J. Streeck (Eds.), *Time in Embodied Interaction: Synchronicity and Sequentiality of Multimodal Resources* (pp. 1–29), John Benjamins, 2018.
- [11] L. Mondada, "Conversation analysis: Talk and bodily resources for the organization of social interaction", In: C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill & S. Tessendorf (Eds.), *Body - Language - Communication: An international Handbook on Multimodality in Human Interaction* (Vol. 1) (pp. 218–226), De Gruyter, 2013.
- [12] L. Mondada, "Challenges of multimodality: Language and the body in social interaction", *Journal of Sociolinguistics*, vol. 20, no. 3, pp. 336–366, 2016.

Postersession 2:

–

26.09.2024

15:20-16:40



Virtually Restricting Modalities in Interactions: Va.Si.Li-Lab for Experimental Multimodal Research

Alexander Henlein¹, Alexander Mehler¹ and Andy Lücking¹,

¹*Goethe University Frankfurt, Text Technology Lab*

henlein@em.uni-frankfurt.de

Purpose: Research in multimodal communication usually focuses on how communication benefits from the inclusion of several modalities (e.g., [1], [2]). Here, by making use of “non-factual” simulation facilities of Virtual Reality (VR) settings, we investigate the contrary: how do modality-specific restrictions influence task-oriented, multi-party communication. To this end, we present a collaborative VR scenario where multiple users are asked to furnish a shared home. We compare the behavior of the users under different restrictions: vision restriction (blurry eyesight), hearing restriction (distorted hearing), and interaction restriction (not able to grab any objects).

Method: VR glasses have become quite sophisticated tracking devices in recent years. For example, the Meta Quest Pro¹ can now not only fully track hands, but also features upper/lower face tracking and eye tracking. By using Meta Avatars for player representation, these modalities can also be displayed² and taken into account in interactions – see Figure 1. In addition to the extensive tracking capabilities of these glasses, a VR world allows absolute control over the desired scenario, the objects that can be interacted with, and how, and thus, to a certain extent, the experience of users/interlocutors. To this end, in this abstract, we present our new system, a VR-based scenario tool for tracking and analyzing interpersonal communication. It is built on Va.Si.Li-Lab [3], [4] and Ubiq [5]. The scenarios support multiple users, each of which can be assigned different roles and thus different functions and restrictions. A total of 44 participants took part in our experiment (m=35, w=7, n/a=2, mainly aged between 21 and 26 years), divided into 15 groups. 5 groups participated in the scenario without any restriction and 10 groups were imposed with the above mentioned restrictions. In total, 14 participants had an unrestricted experience, 11 participants had a hearing restriction, 10 participants had an interaction restriction and 9 participants had a vision restriction. All tracking data (e.g. hand, face & eye tracking with the Meta Quest Pro) and audio data gets stored in Va.Si.Li-Lab’s MongoDB, aligned to a common timeline. These data are quantitatively analyzed with respect to motion, dialogue turns, and – by using NLP on speech recording – content. An example can be seen in Figure 2.

Results: The three-part evaluation reveals global and local results: On the global scale, we were – as expected – able to show that restricted participants behave significantly differently than those without restrictions. These differences, however, are not equally pronounced on a local scale. In our experiments, for example, the strongest effect was observed in people with interaction restrictions, who moved significantly less in the scenario. Restriction of interaction capacity seems to be more disruptive than other restrictions, indicating that the pivotal role of participation transfers into multiplayer VR settings.

Conclusion: Firstly, the results of our study confirm that the VR methodology provides a useful experimental setting for studying multimodal interaction. VR allows in particular to modify a user’s sense experiences (within the confines of immersion, at least). Secondly, the results point to the importance of *inclusion*: restricted communication facilities seem to be more tolerable than not taking part in interaction in the first place.

¹<https://www.meta.com/quest/quest-pro/tech-specs/#tech-specs>

²<https://www.meta.com/de/avatars/>



Figure 1: *Meta Avatars examples for representing the participants.*



Figure 2: *Interaction between three people in the scenario from two different perspectives [4].*

References

- [1] J. P. Trujillo and J. Holler, “Interactionally embedded gestalt principles of multimodal human communication,” *Perspectives on Psychological Science*, 2023, Online first. DOI: 10.1177/17456916221141422.
- [2] J. Holler and S. C. Levinson, “Multimodal language processing in human communication,” *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, 2019.
- [3] A. Mehler, M. Bagci, A. Henlein, *et al.*, “A multimodal data model for simulation-based learning with Va.Si.Li-Lab,” in *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. HCII 2023*, V. G. Duffy, Ed., Cham: Springer, 2023. DOI: 10.1007/978-3-031-35741-1_39.
- [4] A. Mehler, M. Bagci, P. Schrottenbacher, *et al.*, “Towards new data spaces for the study of multiple documents with Va.Si.Li-Lab,” in *Students’, Graduates’ and Young Professionals’ Critical Use of Online Information: Digital Performance Assessment and Training within and across Domains*, O. Troitschanskaia-Zlatkin, M.-T. Nagel, V. Klose, and A. Mehler, Eds., In press, Berlin: Springer, 2024.
- [5] S. J. Friston, B. J. Congdon, D. Swapp, *et al.*, “Ubiq: A system to build flexible social virtual reality experiences,” in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, 2021, pp. 1–11.

A Theoretical Model for Analyzing Metaphors in Multimodal Communication: Exemplified by Pictorial and Verbo-Pictorial Metaphors in Editorial Cartoons

Han Zhou

Heidelberg University, Germany

zhiujin@163.com

Multimodal metaphor research has received significant attention in recent years, particularly with numerous empirical studies focusing on metaphors within various genres like editorial cartoons, advertisements and films. However, a comprehensive theoretical framework for the qualitative analysis of metaphors in multimodal communication is still lacking. By comparing theoretical approaches from social semiotics and cognitive linguistics in the context of pictorial and verbo-pictorial metaphors research, this paper emphasizes the complementarity between these two theoretical perspectives and proposes the Social Semiotic Integration Model. This model will be demonstrated through analysis of several editorial cartoons.

The first theoretical framework, social semiotics, views metaphors as social practices. Based on Halliday's three metafunctions of language [1], Kress and van Leeuwen proposed Visual Grammar [2], indicating that images also possess ideational, interpersonal and textual metafunctions, and systematically presenting how to analyze an image's structure, meaning, and function in terms of these three aspects. As images serve as the medium for pictorial and verbo-pictorial metaphors, detailed image analysis can facilitate the identification and interpretation of metaphors in visual communication. However, this approach tends to look at images in isolation without considering pragmatic factors such as the author's intention and background knowledge, which may lead to a distorted or incomplete interpretation. Moreover, it cannot explain the cognitive mechanisms behind the metaphorical meaning.

The second theoretical framework is cognitive linguistics, which regards metaphors as cognitive phenomena. A widely used model is Fauconnier and Turner's Blending Theory [3] (Figure 1). It can not only visualize the process of generating metaphorical meanings, but also be effectively applied to analyze creative and short-lived metaphors. To analyze metaphorical process in multimodal contexts, Zhao further proposed a Multimodal Metaphor Integration Model [4] (Figure 2) based on Blending Theory and the work of Brandt and Brandt [5]. Its strength lies in explaining the cognitive mechanisms of metaphors while considering the influence of semiotic representations and pragmatic factors. However, it does not explain how these semiotic representations function and influence metaphorical meanings.

By comparing these theoretical approaches, it is clear that social semiotics and cognitive linguistics are complementary in analyzing metaphors in multimodal communication. Therefore, this paper proposes the Social Semiotic Integration Model (Figure 3), which integrates these two theoretical perspectives. Beginning with the semiotic base space, this model adds a metafunction space to Zhao's model, which guides the analyzer in interpreting the semiotic representations. Additionally, the author suggests considering the influence of pragmatic factors and other related elements when interpreting meanings and functions of the signs, rather than only revising the metaphorical meanings after they have been derived in isolation. Thus, the relevant space points to both the metafunction space and the blended space. This model can be employed for qualitative metaphor analysis in multimodal contexts, thereby facilitating a more precise understanding of the overall meaning in multimodal communication.

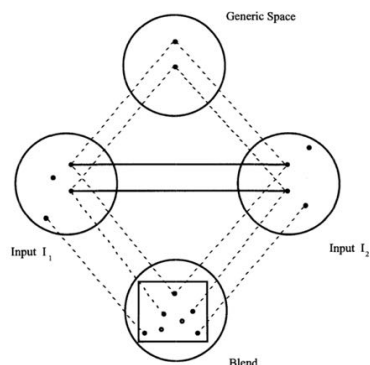


Figure 1: Blending Theory

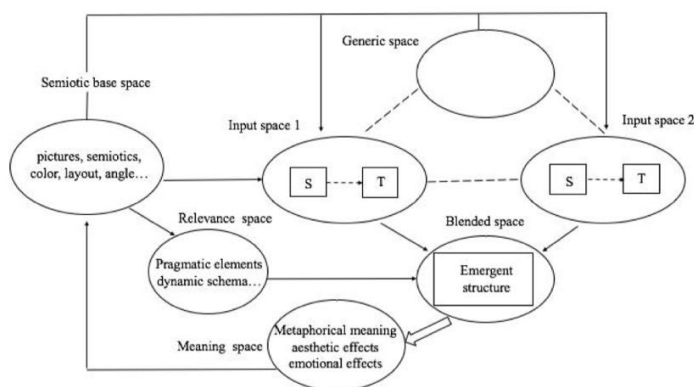


Figure 2: Multimodal Metaphor Integration Model

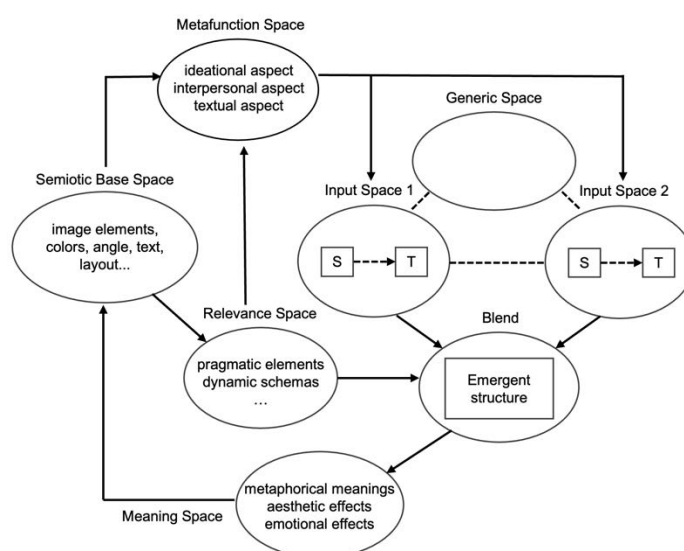


Figure 3: Social Semiotic Integration Model

References

- [1] M. A. K. Halliday, *Language as social semiotic: the social interpretation of language and meaning*. London: Arnold, 1978.
- [2] G. Kress and T. van Leeuwen, *Reading Images: The Grammar of Visual Design*. New York: Routledge, 1996.
- [3] G. Fauconnier and M. Turner, *The way we think: conceptual blending and the mind's hidden complexities*. New York: Basic Books, 2002.
- [4] X. Zhao, "The Conceptual Integration Model of Multimodal Metaphor Construction: A Case Study of a Political Cartoon," *Foreign Languages Research*, no. 05, pp. 1-8, 2013. (in Chinese: 赵秀凤, "多模态隐喻构建的整合模型——以政治漫画为例". 外语研究, 05, 1-8.)
- [5] L. Brandt and P. A. Brandt, "Cognitive poetics and imagery," *European journal of English studies*, vol. 9, pp. 117-130, 2005.

Movement entrainment in online meetings

Patrizia Paggio^{1,2}, Manex Agirrezabal¹ and Bart Jongejan¹

¹University of Copenhagen, ²University of Malta

(paggio|manex.agirrezabal|bartj)@hum.ku.dk

Two decades ago, Pickering and Garrod [1] proposed an *interactive alignment account* of dialogue according to which speakers align their linguistic representations at many levels as a consequence of a need to simplify language processing while interacting. Since then, several studies have investigated this phenomenon looking not only at speech [2], [3], but also at gestural and facial behaviour [4]–[6]. See [7] for a review of approaches from a multimodal view.

One interesting methodological issue is how to find evidence of the phenomenon using fully automatic methods. In essence, two main approaches are possible: either looking at the repetition of discrete elements (words, gestures, head movements, etc.) between or within speakers over temporal sequences [8], or modelling the phenomenon in terms of continuous variables, e.g. using prosodic measurements [2], [9]. The term *entrainment* has been used in the latter approaches to refer to the convergence of patterns of behaviour across speakers. The advantage of looking at continuous variables is that they may be easier to extract using automatic methods compared to having to annotate the data with linguistically meaningful discrete elements.

In this study, we investigate head movement entrainment between speakers in online Zoom meeting recordings by means of visual features extracted using OpenPose. Following the methodology used to measure prosodic entrainment in [2], *movement* feature value differences within a group of speakers are compared across time for entrainment convergence. We work with continuous variables for two reasons: i) we wanted to apply a method from prosodic analysis to visual coordinates (theoretically, both prosody and gesturing are suprasegmental phenomena); ii) there are no discrete labels (annotated head movements) in our dataset.

For each speaker, we extract x and y coordinate values for 6 different visual keypoints relating to head movements (Nose, Neck, Left and Right Eye, Left and Right Ear). We then compute the average difference across all speaker pairs in two 3-minute meeting intervals at meeting beginning and end (after having discarded initial and final greeting sessions). We do this for x and y coordinates of each keypoint separately. This set of measures reflects the degree to which speakers move in a similar way: the lower the values, the more similarly they move (horizontally and vertically). For each keypoint, we then compare the two averages from the initial and final meeting intervals. If the values decrease, we interpret this as an indication of the fact that speakers are showing entrainment convergence at the level of the feature in question.

We have so far applied the methodology to one meeting recording including 6 speakers. The results show different tendencies for the two types of feature and the different keypoints. On average, we see entrainment increase in y coordinate values for all keypoints with the exception of Neck (Figure 1) without, however, reaching significance for any of the keypoints (Table 1). The trend is less clear for x coordinate values, with differences becoming higher or lower depending on the keypoint. (These results are not visualised for lack of space.) The analysis has to be expanded to the entire dataset (12 meetings). It is tempting to hypothesise, however, that the speakers in this video seem to show a tendency to increased entrainment of head movement along the vertical line, which might be due to nodding behaviour. In addition to analysing the entire dataset, thereby also assessing the potential correlations between the five different measures, we also plan to follow the interesting suggestion given by one of the reviewers and look into Dynamic Time Warping as an alternative way to measure growing similarity of movement.

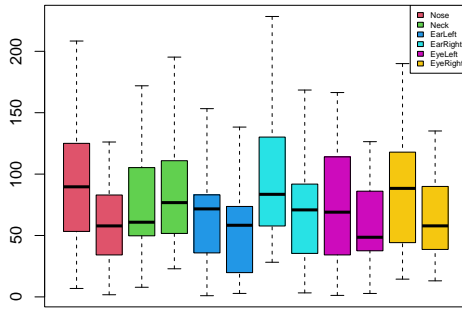


Figure 1: Average differences between speaker pairs in y coordinate values for six visual keypoints. For each keypoint, the boxplot on the left refers to an initial and the one on the right to a final meeting interval.

Keypoint	t-values	
	x coord	y coord
Nose	0.59	1.79
Neck	-0.82	-0.83
EarLeft	0.94	1.09
EarRight	-1.02	1.47
EyeLeft	1.03	1.26
EyeLeft	-0.17	1.34

Table 1: *T-values from Welsh Two Sample t-tests comparing average movement differences across speakers between two meeting intervals. Positive t-values show convergence, while negative t-values indicate divergence.*

References

- [1] M. J. Pickering and S. Garrod, “Toward a mechanistic psychology of dialogue,” *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.
- [2] D. Litman, S. Paletz, Z. Rahimi, S. Allegretti, and C. Rice, “The teams corpus and entrainment in multi-party spoken dialogues,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1421–1431.
- [3] C. Dideriksen, R. Fusaroli, K. Tylén, M. Dingemanse, and M. H. Christiansen, “Contextualizing conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations,” in *CogSci’19*, Cognitive Science Society, 2019, pp. 261–267.
- [4] J. Holler and K. Wilkin, “Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue,” *Journal of Nonverbal Behavior*, vol. 35, pp. 133–153, 2011.
- [5] K. Bergmann and S. Kopp, “Gestural alignment in natural dialogue,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, 2012.
- [6] C. Navarretta, “Mirroring facial expressions and emotions in dyadic conversations,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 2016, pp. 469–474.
- [7] M. Rasenberg, A. Özyürek, and M. Dingemanse, “Alignment in multimodal interaction: An integrative framework,” *Cognitive science*, vol. 44, no. 11, e12911, 2020.
- [8] M. M. Louwerse, R. Dale, E. G. Bard, and P. Jeuniaux, “Behavior matching in multimodal communication is synchronized,” *Cognitive science*, vol. 36, no. 8, pp. 1404–1426, 2012.
- [9] R. Levitan and J. Hirschberg, “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions,” in *Interspeech*, 2011. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14147636>.

Moving Meetings by Moving Prosody and Gesture

Alina Gregori¹ & Susanne Fuchs²

Goethe University Frankfurt¹, Leibniz-Zentrum für Allgemeine Sprachwissenschaft, Berlin²
gregori@lingua.uni-frankfurt.de

Prosody and gesture have been shown to coordinate with each other, forming an integrated system in communication [1]. Several studies have found evidence for the temporal coordination of prosodic and gestural cues, which focus on gestural strokes and prominent syllables in speech [2], [3], [4], [5], [6]. In this study, we want to assess how prosody-gesture (mis)alignment can influence the disambiguation of an ambiguous sentence using the “Next Wednesday Question” (NWQ) paradigm [7], [8]. In NWQ studies, participants answer the question “Next Wednesday’s meeting has been moved forward two days. On what day is it now?” (or similar) and based on their response, conclusions can be drawn. The NWQ paradigm has been used to test intuitions about time: e.g., whether a person perceives time from a “Moving Ego” or from a “Moving Time” perspective (cf. [9], [10]). The answer to the question thus depends on how the adverbial is interpreted. While in general, responses to the NWQ are divided evenly into “Monday” and “Friday” responses, it has been found that when an additional gesture is presented with the question, it can influence the response. The gesture amplifies a direction in time that is expressed in speech with the adverbial [11], [12], [13]. However, whether and how prosody (accentuation) and its alignment with gesture influences the perception of time in this paradigm has not been investigated. Thus, this study assesses the role of prosody, gesture and their synchronization on the disambiguation of adverbials.

Running a multimodal NWQ paradigm, we conduct a perception study in English in a between-subjects design (similar to [13]) using stimuli produced by a native speaker of British English and recruiting participants via Prolific. The study comprises two sub-studies. Study 1 investigates a gestural influence on the interpretation of the question, varying the spatial adverbial. A horizontal gesture is produced during the NWQ, with the speakers’ hand moving either towards or away from their body (see Fig. 1). The gesture is produced in one of three positions: On the adverbial when present, on the verb “moved” when there is not adverbial or produced as a pro-speech gesture in an added speech break. Sentence accentuation is placed on the adverbial when present, and on the verb “moved” otherwise. This study thus applies a 2 x 3 design with the factors *Gesture* (towards / away from speaker) and *Adverbial* (present, absent, replaced by pro-speech gesture). Participants answer the NWQ with “Monday” or “Friday”, which represents the investigated measure. If gesture-speech mismatches have an influence on responses to the NWQ paradigm in this study, gestures are interpreted to have an amplifying contribution to the perception of time expressed by an adverbial.

Study 2 addresses the question whether the placement of pitch accents and gestures influences the responses to the NWQ. To test this, the NWQ is posed similarly to study 1 but the adverbial is additionally varied in directionality (“forward” vs. “backward”) to create a mismatch between gesture and speech (e.g. saying “forward” with a gesture towards the speakers’ body). In addition, either the pitch accent or the gesture is moved away from the adverbial (where they are placed by default) to the verb “moved”, creating a temporal mismatch between prosody and gesture. This allows to investigate whether these cues then still act together or whether one of them is more important for disambiguation. The second study thus applies a 2 x 2 x 2 design with the factors *Gesture* (towards / away from speaker), *Language direction* (forward / backward) and *Moved cue* (prosody / gesture). The investigated measure is the response “Monday” or “Friday”. If the temporal misalignment of gesture and prosody has an influence on responses to the NWQ paradigm in this study, gestures are interpreted to have a contribution (independent to prosody) to the perception of time expressed by an adverbial. We plan on collecting data from 250 participants per sub-study (500 in total).

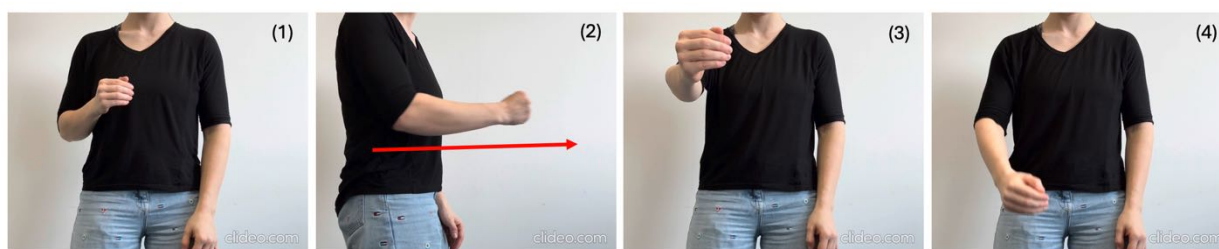


Figure 1: Illustration of the horizontally moving gesture used for the studies. Produced on the spatial adverbials “forward” or “backward” or on the verb “moved”. Varied in movement away from or towards the speaker. From left to right: (1) Starting position of the stroke; (2) Horizontal hand movement in front of the speakers’ body; (3) Gesture apex (farthest extended point of the gesture); (4) Retraction, return to rest.

References

- [1] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.
- [2] D. P. Loehr, “Temporal, structural, and pragmatic synchrony between intonation and gesture,” *Lab. Phonol.*, vol. 3, no. 1, Jan. 2012, doi: 10.1515/lp-2012-0006.
- [3] N. Esteve-Gibert and P. Prieto, “Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements,” *J. Speech Lang. Hear. Res.*, vol. 56, no. 3, pp. 850–864, Jun. 2013, doi: 10.1044/1092-4388(2012/12-0049).
- [4] S. Shattuck-Hufnagel and A. Ren, “The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech,” *Front. Psychol.*, vol. 9, p. 1514, 2018.
- [5] W. Pouw and J. A. Dixon, “Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback,” *Cogn. Sci.*, vol. 43, no. 3, p. e12721, Mar. 2019, doi: 10.1111/cogs.12721.
- [6] T. Leonard and F. Cummins, “The temporal relation between beat gestures and speech,” *Lang. Cogn. Process.*, vol. 26, no. 10, pp. 1457–1471, Dec. 2011, doi: 10.1080/01690965.2010.500218.
- [7] L. Boroditsky, “Metaphoric structuring: Understanding time through spatial metaphors,” *Cognition*, vol. 75, no. 1, pp. 1–28, 2000.
- [8] M. S. McGlone and J. L. Harding, “Back (or forward?) to the future: The role of perspective in temporal language comprehension,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 24, no. 5, p. 1211, 1998.
- [9] E. V. Clark, “Cognitive development and the acquisition of language,” *Whats Word Childs Acquis. Semant. His First Lang.*, 1973.
- [10] B. Tversky, “Spatial perspective in descriptions,” 1996.
- [11] A. Jamalian and B. Tversky, “Gestures alter thinking about time,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2012.
- [12] T. N. Lewis and E. Stickles, “Gestural modality and addressee perspective influence how we reason about time,” *Cogn. Linguist.*, vol. 28, no. 1, pp. 45–76, Feb. 2017, doi: 10.1515/cog-2015-0137.
- [13] B. Winter and S. E. Duffy, “Can co-speech gestures alone carry the mental time line?,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 46, no. 9, p. 1768, 2020.

Show me the choice

Marion Bonnet¹, Cornelia Ebert², Kurt Erbach², Markus Steinbach¹
 Georg-August Universität Göttingen¹, Goethe-Universität Frankfurt²
 marion.bonnet@uni-goettingen.de

In the present experimental study, we investigated the interaction of co-speech palm up gestures and prosody in sentences typically triggering free choice interpretation (FC). Results show that multimodal sentences boosted the exclusive interpretation compared to utterances displayed only in the spoken modality which only received the FC interpretation.

Prosody plays a crucial role in the interpretation of disjunctive sentences. Pruitt & Roelofsen [1] have shown that final contour and pitch accents allow to disambiguate between alternative and yes-no questions. At the semantic-pragmatic interface, sentences containing disjunction in the scope of a possibility modal give rise to the FC interpretation [2]. Therefore, a sentence like *Alex can have ice cream or cake* implies that Alex can freely choose to have one of the two options or both. It is still under debate which theory would best account for such data. Tieu et al. [3] investigated which of the two main approaches makes the best predictions. Their results favor the homogeneity approach or suggest revising the implicature theory. So far, however, all theories and empirical studies have neglected the influence of co-speech gestures on the interpretation of utterances. Recent studies on sign languages and/or gestures show that visual cues affect the interpretation and argue that theoretical accounts of disambiguation and enrichment should consider the semantic and pragmatic impact of the visual modality. More specifically, iconicity and speech and gestures alignment could provide interesting insight to such phenomena [4][5]. Finally, the resort to palm up gestures to express possibility is widely attested in sign languages and gestures accompanying speech [6]. One interesting question in our context is whether the standard FC interpretation could be modified when visual information and phrasing suggest a different reading?

We investigate the contribution of palm up co-speech gestures to German FC sentences using a picture selection task (*Figure 1*). Stimuli were displayed through short video clips in which a native speaker performed the gesture while saying the target utterance (except for the baseline experiment in which we presented audio only items). We tested three types of palm up gestures as well as a no gesture condition. Hand movements were systematically aligned with the disjuncts (following literature) but in different fashions (*Figure 2*). As a second factor, we introduced two phrasing patterns, adapted from Pruitt and Roelofsen [1]. The first pattern, “disjunctive phrasing”, displays the main pitch accent on the first disjunct and a short pause before the word *oder* (‘or’). The second one, “conjunctive phrasing”, does not show any pause between the disjuncts and bears the main pitch accent on the second disjunct. We tested 6 target sentences resulting in 48 target stimuli. Participants were either presented with the disjunctive or the conjunctive phrasing, and with the audio only or multimodal input. The overall study included 110 participants.

Results show that multimodal sentences boost the exclusive interpretation (close to 50%) whereas items displayed only in the spoken modality exclusively receive the FC interpretation (*Figure 3*). The type of palm up gesture does not create a significant difference in the interpretation. More surprisingly, the no gesture condition behaves like the other gesture conditions, that is boosting the exclusive reading (*Figure 4*). Finally, no significant difference was attested between the disjunctive and the conjunctive phrasing in both the audio only and the multimodal condition. We propose to analyze multimodal disjunctive sentences as complex disjunctions [7] in which the visual component nuances the FC interpretation. We finally suggest that traditional semantic-pragmatic theories should be refined to account for the contribution of multimodality.

Figures and tables

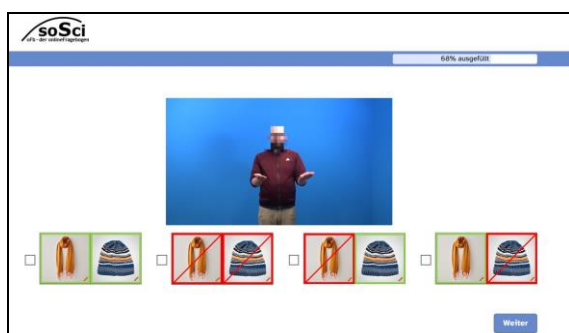


Figure 1. screenshot of the experimental platform illustrating the picture selection task



Figure 2. illustration of the palm up gestures and their alignment with the disjuncts

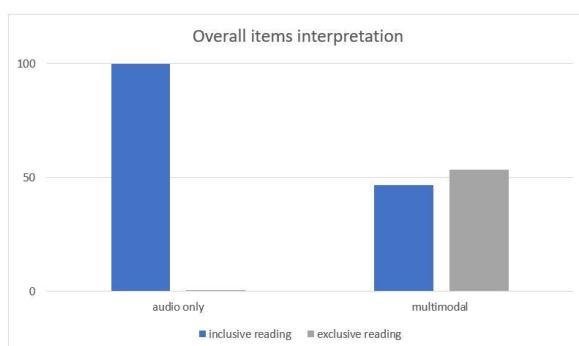


Figure 3. Items interpretation depending on the modality for both the disjunctive and conjunctive phrasing

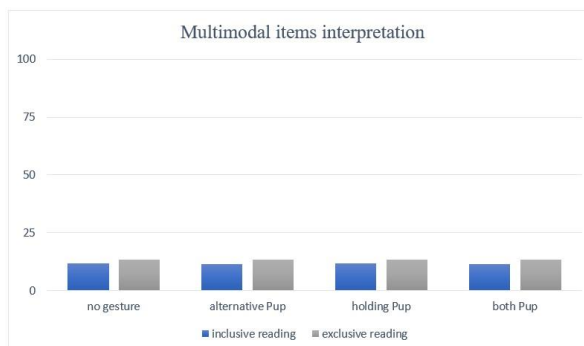


Figure 4. Multimodal items interpretation per type of gesture for both the disjunctive and conjunctive phrasing

References

- [1] K. Pruitt & F. Roelofsen, (2013). The interpretation of prosody in disjunctive questions. *Linguistic inquiry*, 44(4), 632-650.
- [2] H. Kamp, (2013). Free choice permission. In *Meaning and the Dynamics of Interpretation* (pp. 169-184). Brill.
- [3] L.Tieu, C. Bill & J.Romoli, (2019, December). Homogeneity or implicature: And experimental investigation of free choice. In *Semantics and linguistic theory* (Vol. 29, pp. 706-726).
- [4] P. Schlenker, (2023). On the typology of iconic contributions. *Theoretical Linguistics*, 49(3-4), 269-290.
- [5] C. Ebert & C. Ebert, "Gestures, demonstratives, and the attributive/referential distinction. " *Handout of a talk given at Semantics and Philosophy in Europe (SPE 7), Berlin 28* (2014).
- [6] K. Cooperrider, N. Abner & S. Goldin-Meadow (2018) The Palm-Up Puzzle: Meanings and Origins of a Widespread Form in Gesture and Sign. *Front. Commun.* 3:23.
- [7] B. Spector, "Global positive polarity items and obligatory exhaustivity." *Semantics and Pragmatics* 7 (2014): 11-1.

The effect of gesture expressivity on emotional resonance in storytelling interaction

Christoph Rühlemann & James Trujillo
University of Freiburg, University of Amsterdam
 chrisruehlemann@gmail.com

Storytelling is driven by emotion. Its key function is a meeting of hearts: a resonance in the recipient(s) of the storyteller's emotion towards the story events [1]. How emotions are expressed gesturally is still seriously underresearched. This paper focuses on the role of gestures in emotion expression and emotion resonance in storytelling. The data come from the Freiburg Multimodal Interaction Corpus (FreMIC), which features not only CA transcriptions of video-recorded talk-in-interaction but also Electrodermal Activity (EDA) data on storytellers and story recipients [2].

Specifically, the paper asks three questions: Does storytellers' gesture expressivity increase from story onset to climax offset (RQ #1)? Does gesture expressivity predict specific EDA responses in story participants (RQ #2)? How important is the contribution of gesture expressivity to emotional resonance compared to the contribution of other predictors of resonance (RQ #3)?

The analyses, based on 44 stories (collected in 9 recordings, total run time 7.55 hrs, with 949 gestures and 13 distinct participants), were annotated for variables that may potentially impact emotion arousal. These include (i) *Protagonist* (is the story's protagonist the storyteller vs. a non-present person), (ii) *Recency* (did the story events occur far in the past vs. are they occurring at/close to storytelling time), (iii) *Group_composition* (were participants all-female, all-male, or mixed), and (iv) *Group_size* (was the story told in a dyad or triad). Further, gestures were examined for whether they co-occurred with a quote (variable *G_quote*). The gestures were further coded for gesture phases [3] as well as for seven gesture-dynamic parameters: (i) Size (*SO*), (ii) Force (*FO*), (iii) Character view-point (*CV*) [4], (iv) Silence during gesture (*SL*), (v) Presence of hold phase (*HO*), (vi) Co-articulation with other bodily organs (*MA*) and (vii) Nucleus duration (*ND*). The binary annotations were aggregated in the Gesture Expressivity Index (GEI), which computes for each gesture an average value across all ratings; its values are stored in the variable *G_expressivity*, one of the key variables in the models. Interrater agreement for the coding of the GEI parameters (tested on c. 24% of all gestures) ranged between 79% for Force (*FO*) and 94% for Character viewpoint (*CV*).

To account for response latency, EDA responses were measured during the duration of the gesture as well as 1.5 sec post-gesture; further, they were classified as *specific* (i.e., as indexing a stimulus-related emotional response) if larger than 0.05 μ Siemens. Finally, *resonating gesture* were identified, i.e., gestures exhibiting *concurrent* specific EDA responses by two or more participants, resulting in a binary variable *EDA_G_resonance*, the dependent variable in the Random Forest model.

The first model, which addresses RQ #1, was a mixed-effects model with a relative positional measure *G_position_rel* for each gesture in each story (independent variable) and *G_expressivity* (dependent variable). The model suggested that storytellers' gestures become more expressive from story onset to climax offset.

To address RQ #2, a second linear mixed-effects regression model was constructed, with *EDA_specific_response_binary* as the dependent and *G_expressivity* as the independent

variable. This model suggested that increased gesture expressivity increases the probability of specific EDA responses.

To address RQ #3 a Random Forest ($n_{tree}=1,500$, $m_{try}=3$) for emotional resonance ($EDA_G_resonance$) as outcome variable and the seven GEI parameters as well as six more variables as predictors (G_quote , *Protagonist*, *Group_compose*, *Group_size*, *Role* (storyteller or story recipient), and *Recency*) exhibited a very good fit, significantly better than chance/baseline ($p < .001$), with a (traditional) R^2 of 0.86, and McFadden's R^2 of 0.37.

All but one predictor (*Role*) were found to impact $EDA_G_resonance$. Analysis of variable importance showed *Group_composition* to be the most impactful predictor, followed by *Recency*, *Group_size*, *ND* (nucleus duration), *Protagonist*, *FO* (gesture force), *SZ* (gesture size), *G_quote*, *HO* (hold phase), *CV* (character viewpoint), and *MA* (*multiple articulators*).

Inspection of ICE plots clearly indicated combined effects of individual GEI parameters and other factors, including *Group_size* and *Group_compose*. Fig. 1 depicts the effect on emotional resonance of gesture force (*FO*) interacting with group composition (*Group_compose*).

Methodologically, this study opens up new avenues of multimodal corpus linguistic research by examining the interplay of emotion-related metrics and gesture at micro-analytic levels and using advanced machine-learning methods to deal with the inherent collinearity of multimodal variables. More good is expected to come from this fruitful combination of qualitative and quantitative research.

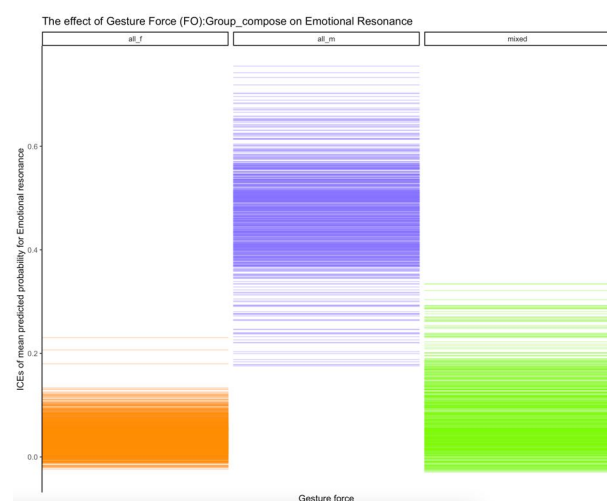


Fig. 1: ICE plot of effect of interaction of Gesture force (*FO*):Group composition (*Group_compose*) on Emotional resonance ($EDA_G_resonance$); y-axis represents jittered means of predicted probabilities for emotional resonance ($EDA_G_resonance$)

References

1. Stivers, T. (2008). Stance, Alignment, and Affiliation during Storytelling: When Nodding Is a Token of Affiliation. *Res. Lang. Soc. Interact.* 41,31–57.
2. Rühlemann, C. and A. Ptak. (2023). Reaching below the tip of the iceberg: A guide to the Freiburg Multimodal Interaction Corpus (FreMIC). Open Linguistics: <https://doi.org/10.1515/opli-2022-0245>
3. Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge, MA: Cambridge University Press.
4. Beattie, G. 2016. Rethinking body language: How hand movements reveal hidden thoughts. London/New York: Routledge

Does gesture play a similar role in the communication of second language learners in face-to-face and online interactions?

Himmet Sarıtaş¹ & Şeyda Özçalışkan¹
Georgia State University¹
hsaritas@gsu.edu - seyda@gsu.edu

Gesture and speech act together in communication in not only first language (L1) [1] but also second language (L2) [2] production contexts. Previous research showed that L2 learners frequently use gesture along with speech in extended speech contexts, such as telling narratives [3]. More importantly, gesture provides a particularly useful tool when L2 learners encounter word finding difficulties in their communications [4]. Learners use gesture more when their speech is disfluent [5], namely when they pause in their speech to find words or repair what they have expressed incorrectly [6]. Most of the earlier work on gesturing during disfluent speech in L2 learners focused on face-to-face interactions. However, we do not yet know whether gesture plays a similar or a different role when accompanying disfluent speech in online production contexts.

In this study, we aimed to fill in this gap by studying the role of gesture during disfluent speech in face-to-face vs. online communication, using a narrative production task. The participants included 20 (10 females) Turkish L2 learners ($M=21.25$; $SD=3.94$), residing in Turkey, who had intermediate (B2) level of proficiency in Turkish. Each participant was interviewed in Turkish with an L1 Turkish speaker, first face-to-face and two weeks later online, using a narrative task. The participants first watched two cartoons that are known to elicit gestures, one at a time; they were then asked to narrate each cartoon to the experimenter. The procedure for the online elicitation was the same, but the researcher and participant met online via the Zoom platform. Each elicitation included two cartoons; the first cartoon was the same in both elicitations, while the second one differed to make the task more engaging for the participants. Each participant also completed the Edinburg Handedness Inventory [7], a language experience questionnaire (LEAP-Q; [8]) that provides details on their experience in different languages, a demographic questionnaire that provides information about participant age, gender, and race, along with several other demographic information and short word generation task to provide a fuller understanding of language and other characteristics of each participant. All spoken responses were videotaped and transcribed using CHAT guidelines [9] and further coded for gesture (e.g., iconic, beat, deictic, emblems) following earlier works [5] [10].

Our preliminary findings ($n=10$ /per condition) showed that L2 learners frequently produced disfluent speech both in face-to-face and online communications, accounting for 72%-to-75% of speech production across contexts. Most of this disfluent speech was also accompanied by gestures (face-to-face: 78%; online: 75%), typically with iconic gestures that characterized entities or actions (e.g., moving empty hand forward as if throwing). Importantly, the context of communication (face-to-face vs. online) did not have an effect on the production of disfluent speech with or without gesture (see Figures 1, 2). In both contexts, gesture served multiple functions with disfluent speech. These included repairing speech, searching vocabulary, and mitigating grammatical errors. Overall, our results highlight the robust role gesture plays in the communications of L2 learners, evident in not only face-to-face but also online interactions.

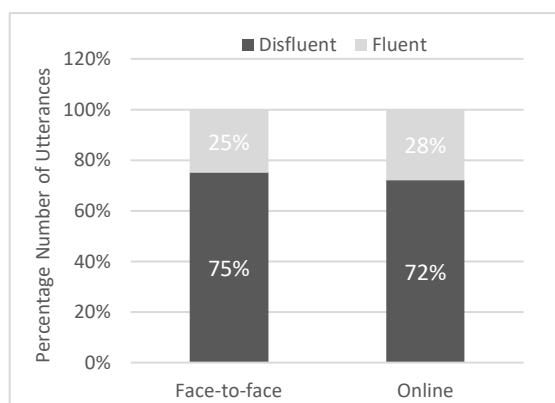


Figure 1: Percentage of fluent and disfluent utterances produced by Turkish L2 learners in face-to-face (left bar) and online interactions.

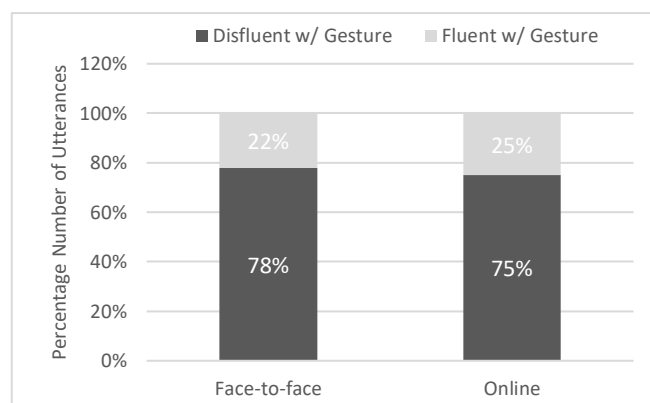


Figure 2: Percentage of fluent and disfluent utterances produced by Turkish L2 learners with or without gesture in face-to-face (left bar) and online interactions.

References

- [1] D. McNeill, *Hand and mind: What gestures reveal about language and thought*. Chicago, IL, USA: University of Chicago Press, 1992.
- [2] M. Gullberg, "Gestures and second language acquisition," in *Body – Language – Communication (HSK 38.2)*, C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill and J. Bressemer, Eds., Berlin, Germany: de Gruyter, 2014, pp. 1868-1875.
- [3] M. Gullberg, *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund, Sweden: Lund University Press, 1998.
- [4] M. Gullberg, "Multilingual multimodality: communicative difficulties and their solutions in second language use," in *Embodied Interaction: Language and Body in the Material World*, J. Streeck, C. Goodwin and C. LeBaron (eds.), Cambridge, UK: Cambridge University Press, 2011, pp. 137–151.
- [5] M. Graziano and M. Gullberg, "When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons," *Front. Psychol.*, 9:879, pp:1-17, 2018. doi: 10.3389/fpsyg.2018.00879
- [6] J. Wong and H. Z. Waring, *Conversation analysis and second language pedagogy*. New York, NY, USA: Routledge Taylor & Francis Group, 2010.
- [7] R. C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," *Neuropsychologia*, Vol. 9, pp. 97-113, 1971.
- [8] V. Marian, H.K. Blumenfeld and M. Kaushanskaya, "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals," *Journal of Speech Language and Hearing Research*, 50 (4), pp: 940-967, 2007.
- [9] B. MacWhinney, "The CHILDES project: Tools for analyzing talk. 3rd Edition, Mahwah, NJ, USA: Lawrence Erlbaum Associates, (2000).
- [10] Ş. Özçalışkan and S. Goldin-Meadow, "Do parents lead their children by the hand?" *Journal of Child Language*, 32(3), pp: 481-505, 2005.

Human language comprehenders appear to integrate rapidly gestural and verbal expressions of “yes” and “no”: Evidence from a two-choice response time task

Emanuel Schütt¹, Merle Weicker², and Carolin Dudschig^{1,3}

University of Tübingen¹, Goethe University Frankfurt², University of Cologne³

emanuel.schuettt@uni-tuebingen.de

In psycholinguistics, negation has traditionally been investigated in terms of a purely verbal operator (e.g., “not”) reversing the truth value of a proposition. A core finding from this research is that negation is often more difficult to process than affirmation, which is usually reflected in longer response times and higher error rates [1]. At the same time, natural communication settings (e.g., dialogues) are characterized by the simultaneous occurrence of verbal and nonverbal expressions of negation such as gestures like the head shake [2, 3]. A recent meta-analysis showed that speech-gesture combinations can indeed moderately improve comprehension [4]. In our preregistered study, we explored whether comprehenders integrate verbal and gestural information on the response particles “yes” and “no” instantly, thus investigating multimodal communication of rejection and affirmation. Participants (146 adult native speakers of German) performed a two-choice response time task adapted from Feiman et al. [5]. Each trial started with the presentation of a question (“Is the ball in the blue/green box?”). Concurrently, a blue and a green box appeared next to each other on the screen, with the left-right spatial arrangement of the boxes randomly changing from trial to trial. Participants then saw a short video clip of a male or female actor answering the question by uttering “yes” or “no” and performing a gesture referring to affirmation (head nod; thumbs up) or rejection (head shake; thumbs down). Critically, the verbal and gestural information included in the videos matched (e.g., “no” and head shake) or mismatched (e.g., “no” and head nod). Participants were asked to choose the correct box based on the verbal clue in one half of the experiment and based on the gestural clue in the other half of the experiment. We expected to see match-mismatch effects if verbal and gestural information are integrated instantly. The performance in the two-choice decision task was analyzed by repeated measures ANOVAs on response times (RTs) and error rates (see Figure 1), with the factors compatibility (match vs. mismatch), response type (affirmation vs. rejection), target modality (gesture vs. speech), and gesture type (head vs. hand). Crucially for the hypothesis under investigation, there was a significant main effect of the factor compatibility, $F_{RT}(1, 145) = 270.36, p < .001, F_{Error}(1, 145) = 99.65, p < .001$. Responses were faster and more accurate when verbal and gestural information matched. This effect interacted with response type, $F_{RT}(1, 145) = 129.24, p < .001, F_{Error}(1, 145) = 22.92, p < .001$, revealing that compatibility effects were larger in the affirmation than in the rejection condition. These interaction effects were further modulated by gesture type, $F_{RT}(1, 145) = 31.09, p < .001, F_{Error}(1, 145) = 4.41, p = .038$, with interaction effects being significantly more pronounced for head gestures than for hand gestures. The results suggest that comprehenders instantly integrate verbal and gestural expressions for “yes” and “no” answers. Interestingly, effects were smaller for rejection than for affirmation. This finding is in line with the observation that negation is harder to process than affirmation [1], indicating that multimodal contexts can only slightly mitigate processing costs. Likewise, non-linguistic inhibitory control mechanisms that are known to be involved in negation processing [6] might have been activated when encountering expressions of rejection in the task-relevant modality, thus hampering the process of integrating other information stemming from the task-irrelevant modality. Modulating effects of response type and gesture type also highlight the need to evaluate the potential role of functional differences of affirmative and negated verbal and gestural signals – also in more natural settings and across cultures.

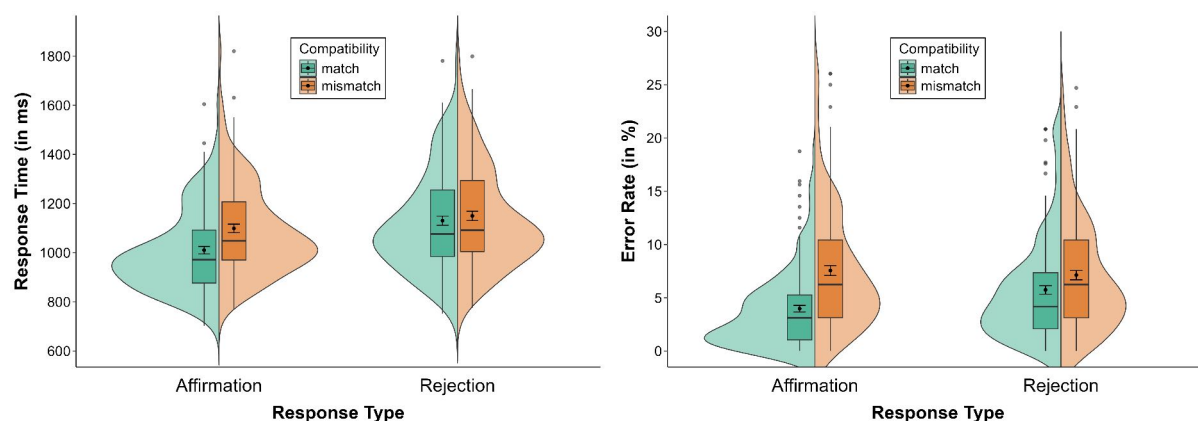


Figure 1: Split violin plots for response times (on the left) and error rates (on the right) as a function of response type and compatibility. Error bars represent the standard error of the mean.

References

- [1] B. Kaup and C. Dudschig, “Understanding negation: Issues in the processing of negation,” in *The Oxford handbook of negation*, V. Déprez and M. Espinal, Eds., Oxford: Oxford University Press, 2020, pp. 635–655.
- [2] A. Kendon, “Some uses of the head shake,” *Gesture*, vol. 2, no. 2, pp. 147–182, 2002. doi: 10.1075/gest.2.2.03ken.
- [3] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.
- [4] N. Dargue, N. Sweller, and M. P. Jones, “When our hands help us understand: A meta-analysis into the effects of gesture on comprehension,” *Psychological Bulletin*, vol. 145, no. 8, pp. 765–784, 2019. doi: 10.1037/bul0000202.
- [5] R. Feiman, S. Mody, S. Sanborn, and S. Carey, “What do you mean, no? Toddlers’ comprehension of logical ‘no’ and ‘not’,” *Language Learning and Development*, vol. 13, no. 4, pp. 430–450, 2017. doi: 10.1080/15475441.2017.1317253.
- [6] D. Beltrán, B. Liu, and M. de Vega, “Inhibitory mechanisms in the processing of negations: A neural reuse hypothesis,” *Journal of Psycholinguistic Research*, vol. 50, no. 6, pp. 1243–1260, 2021. doi: 10.1007/s10936-021-09796-x.

Exploring children's storytelling: The link between narrative abilities, receptive vocabulary and gesture rate in 7- to 9-year-olds

Ingrid Vilà-Giménez¹, Mariia Pronina², Pilar Prieto^{3,4}

*Universitat de Girona¹, Universitat de les Illes Balears², Institució Catalana de Recerca i
Estudis Avançats³, Universitat Pompeu Fabra⁴*
ingrid.vila@udg.edu

Oral narrative abilities and vocabulary knowledge are important precursors to literacy in childhood [1]. Previous research has highlighted the robust link between both receptive and expressive vocabulary skills and narrative performance [2]. Furthermore, longitudinal evidence indicates that referential iconic gestures produced in narrative retellings can serve as predictors of narrative performance [3]. However, the concurrent association between narrative abilities and gesture rate remains ambiguous, yielding mixed findings. For instance, [4] found that, when controlling for age, gesture use in a context-based gesture elicitation task emerged as a significant negative predictor of the narrative scores in 3- to 4-year-olds. Our study expands the developmental window by focusing on 7- to 9-year-old children's narrative retellings to examine the concurrent link between their narrative abilities (in terms of narrative macrostructure and speech fluency scores), and their vocabulary knowledge and production of fluent referential iconic and non-referential gestures. Importantly, we exclusively analyze gestures used in fluent speech, excluding those associated with a disfluency function. Considering the nature of narrative discourse, we hypothesize that (1) children's vocabulary will serve as a strong predictor of their narrative macrostructure scores, and (2) their production of referential iconic gestures in discourse will be associated with their narrative abilities.

Participants were 83 Catalan-Spanish bilingual children (43 girls) aged 7 to 9 who completed a narrative retelling task [5; available at OSF]. All narratives ($n = 166$) were coded for *duration*, *narrative macrostructure* and *speech fluency* using standard scales, and for *referential iconic* (referring to semantic content in speech) and *non-referential* (lacking semantic content) gesture rates. Receptive vocabulary was assessed using the Peabody Picture Vocabulary Test-III [6] adapted into Catalan, the children's dominant language (M Catalan exposure = 86.37%; $SD = 9.38$).

To address the aim of this study, six GLMM analyses were conducted, with narrative macrostructure and speech fluency as dependent variables, and vocabulary knowledge and gesture rate (referential iconic or non-referential gestures) as predictors. Two initial models showed that vocabulary emerged as a significant positive predictor of narrative macrostructure scores ($R^2 = 60\%$) and that narrative duration negatively predicted speech fluency scores ($R^2 = 74\%$). When referential iconic gesture rate ($n = 212$) was introduced in the first model, both vocabulary and referential iconics positively predicted narrative macrostructure ($R^2 = 48\%$). Referential iconics were also positively associated with speech fluency, while narrative duration showed a negative association with speech fluency ($R^2 = 69\%$). Conversely, the addition of non-referential gesture rate ($n = 236$) did not account for additional variance in either narrative macrostructure or speech fluency scores.

While our findings corroborate prior research in highlighting the importance of vocabulary knowledge as a reliable predictor of well-structured narratives in children (particularly between the ages of 7 and 9), they also provide novel evidence for the predictive role of referential iconic gesture rate in both narrative structure and speech fluency scores. Interestingly, while referential iconic gestures show a positive association with narrative performance during this developmental phase, non-referential gestures do not. These results could be influenced by both the nature of narrative discourse and the differing developmental trajectories of these two types of gestures within narrative contexts. The study pinpoints more complex relations between vocabulary and gesture use in children's narrative development at later stages of development.

References

- [1] D. Dickinson and P. Tabors, “Beginning literacy with language: Young children learning at home and school,” Brookes, 2001.
- [2] P. Uccelli and M. M. Páez, “Narrative and Vocabulary Development of Bilingual Children From Kindergarten to First Grade: Developmental Changes and Associations Among English and Spanish skills,” *Language, Speech and Hearing Services in Schools*, vol. 38, no. 3, pp. 225–236, 2007. doi: 10.1044/0161-1461(2007/024)
- [3] Ö. E. Demir, S. C. Levine and S. Goldin-Meadow, “A tale of two hands: children’s early gesture use in narrative production predicts later narrative structure in speech,” *Journal of Child Language*, vol. 42, no. 3, pp. 662–681, 2015. doi: 10.1017/S0305000914000415
- [4] M. Pronina, J. Grofulovic, E. Castillo, P. Prieto and A. Igualada, “Preschoolers narrative skills are related to gesture accuracy in an imitation task,” *Journal of Speech, Language and Hearing Research*, vol. 66, pp. 951–965, 2023. doi: 10.1044/2022_JSLHR-21-00414
- [5] I. Vilà-Giménez, J. Florit-Pons, P. L. Rohrer, S. Coego, G. Gurrado and P. Prieto, “Audiovisual corpus of Catalan children’s narrative discourse development”, 2021. <<https://doi.org/10.17605/OSF.IO/NPZ3W>>
- [6] Ll. M. Dunn, L. M. Dunn and D. Arribas, “PPVT-III Peabody: test de vocabulario en imágenes,” TEA Ediciones, 2010.

Multimodal insights into the lexical development of mono- and multilingual children with SLCN

Nathalie Frey & Carina Lücke
University of Würzburg
 nathalie.frey@uni-wuerzburg.de

Novel word learning is a crucial task for all children, but especially for children growing up multilingual with little to no exposure to the surrounding language, children with speech, language, and communication needs (SLCN, [1]) or children with a diagnosed developmental language disorder (DLD, [1]). On the background of the increasing number of children with difficulties in language acquisition [2] and the shortage of specialists, those children should be already supported in a group setting from an early age. Iconic gestures visualize characteristics of a word and can be integrated into effective group support training as a “*semantically enrichment cue*” [3]. There is evidence for the facilitative effect of iconic gestures in novel word learning in mono- and multilingual children as well as in children with and without difficulties in language acquisition in one-on-one settings [4].

We investigate whether (1) iconic gesture presentation leads to a growth in novel word learning in kindergarten children (2 - 6 years of age) and (2) whether there are children who particularly benefit from this multimodal approach.

An intervention study (waiting control group design) was conducted in two kindergartens with a total of $N = 80$ children (71 % multilingual, 56 % boys, 44 % girls, $M_{Age} = 55,93$ months, $SD = 10,914$). Children grew up with a total of 20 spoken languages and German as surrounding language. The implementation of the multimodal approach within the Intervention Group (IG) commenced with a six-week period wherein a member of the research team (Speech and Language Therapist, SLT) introduced iconic gestures in everyday life at the kindergarten. Subsequently, kindergarten teachers were guided by the SLT in utilizing gestures for five weeks embedded in daily kindergarten activities, followed by a three-month period wherein the intervention was solely administered by the kindergarten teachers (Fig. 1).

A target vocabulary of 50 words, determined based on predefined criteria including word frequency, iconicity, and applicability to daily kindergarten activities, was categorized into five kindergarten contexts (Fig. 2). Iconic gestures were derived from the surrounding sign language. Only kindergarten teachers of the IG were introduced to the iconic gesture presentation; the kindergarten teachers of the waiting control group (WCG) did not get any intervention during their status as control group. Receptive and expressive language proficiency of children in both groups was evaluated at four measurement points as well as their gesture productions in the target vocabulary.

The children acquired in mean 8.3 words ($SD = 5.429$, $t(79) = -13,673$, $p = <.001$) in the expressive and in mean 6.14 words ($SD = 4.665$, $t(79) = -11,767$, $p = <.001$) in the receptive target vocabulary test from pretest to follow up. A comparison of extreme groups with the largest and smallest vocabulary increase showed that children achieving the largest learning increase expressively ($M = 15$ words, $SD = 3.323$), were those with the lowest linguistic performance in the standardized procedures at pretest (lexical skills: $U = 25.00$, $Z = -4.037$, $p = <.001$; phonological working memory skills, $U = 63.000$, $Z = -2.553$, $p = .011$). Individual time of attendance in the kindergarten during the intervention had no influence on this effect ($U = 88.00$, $Z = -1.550$, $p = .121$). In the target vocabulary test, some children resorted to iconic gestures when they could not name a word. Currently, these properties are being analyzed.

Iconic gestures emerge as a viable and readily applicable language support strategy for promoting the acquisition of novel words within the everyday routines of kindergarten. Additionally, they function as a cueing strategy and naming opportunity, and enable the children to communicate despite their still lower linguistic abilities.

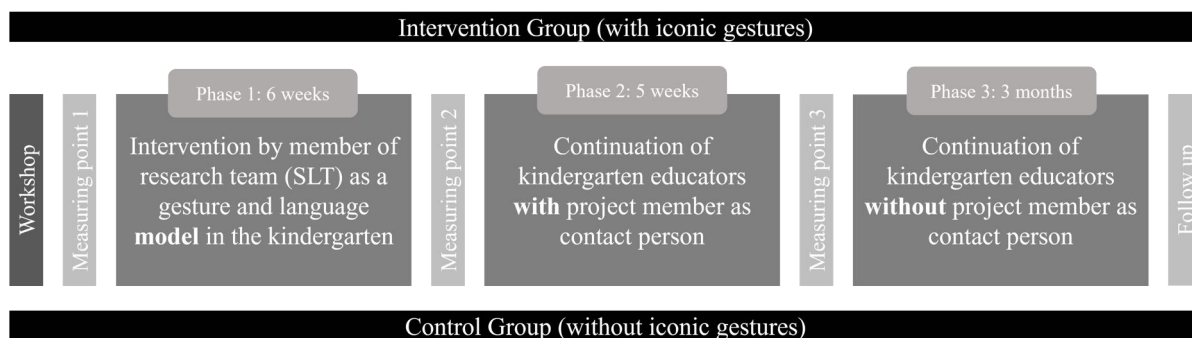


Figure 1: Design of the procedure of the intervention study.

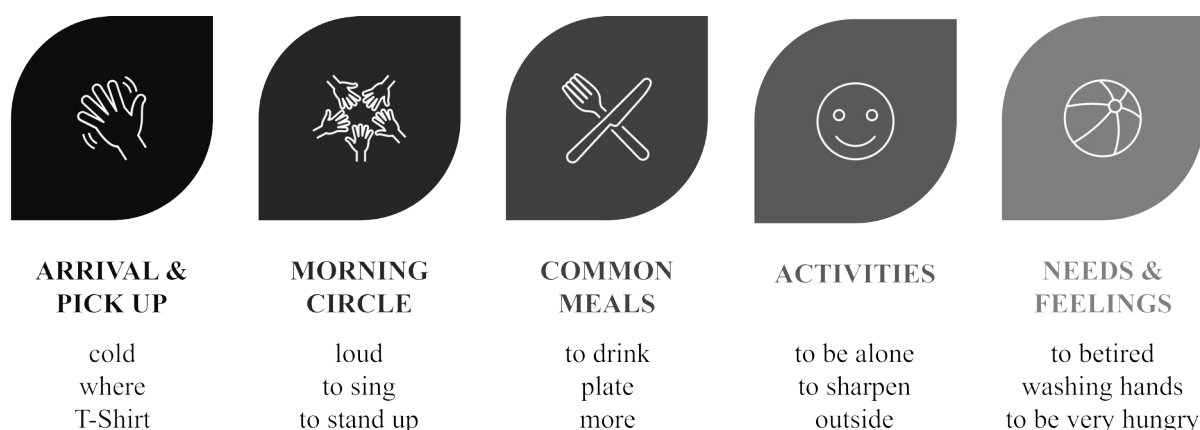


Figure 2: Examples of target vocabulary items that are tailored to the five kindergarten situations.

References

- [1] Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, 58(10), 1068–1080. DOI: 10.1111/jcpp.12721
- [2] Waltersbacher, A. (2020). Heilmittelbericht 2020. Ergotherapie, Sprachtherapie, Physiotherapie, Podologie. Wissenschaftliches Institut der AOK (WiDO), Berlin.
- [3] Capone Singleton, N. (2012). Can semantic enrichment lead to naming in a word extension task? *American Journal of Speech-Language Pathology*, 21(4), 279–292. DOI: 10.1044/1058-0360(2012/11-0019)
- [4] van Berkel-van Hoof, L., Hermans, D., Knoors, H., & Verhoeven, L. (2019). Effects of signs on word learning by children with Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, 62(6), 1798–1812. https://doi.org/10.1044/2019_JSLHR-L-18-0275

A multimodal narrative intervention for boosting NDD children's oral narrative skills

Júlia Florit-Pons¹, Pilar Prieto^{2,1}, Alfonso Igualada^{3,4}

Universitat Pompeu Fabra¹, Institució Catalana de Recerca i Estudis Avançats², Universitat

Oberta de Catalunya³, Institut Guttmann⁴

julia.florit@upf.edu

Children with neurodevelopmental disorders (NDD) like Autism or Developmental Language Disorder (DLD) have impairments in narrative and pragmatic skills [1-2]. Thus, many interventions have been developed to address these impairments (see [3] for a review). However, most intervention programs do not systematically integrate multimodal (gestural and prosodic) strategies, despite the great amount of evidence showing that multimodality can be beneficial for children's linguistic and cognitive abilities [4-5]. For this, the current study aims to assess whether a multimodal narrative intervention can boost the narrative abilities (i.e., macrostructure and perspective-taking) of NDD children and explore NDD children's narrative learning ability throughout the intervention.

The MultiModal Narrative (MMN) is a multi-tiered intervention program co-creatively designed with more than 90 preschool teachers and speech-language therapists to train both narrative macrostructure (i.e., the structure of the narrative in terms of story grammar elements) and perspective-taking (i.e., the ability to express story characters' perspectives and emotions). Narrative macrostructure and perspective-taking were trained using multimodality by providing models of multimodal narratives through 1) a video-recording of a storyteller, 2) by the interventionists (i.e., the speech-language therapists, who enacted the main actions and emotions) and 3) by children themselves (who were asked to enact actions and emotions).

To achieve our aim, 50 children participated in this study. Using a between-subjects research design, we established 3 groups: an experimental group with NDD children ($n = 16$; $M_{age} = 5.20$), an NDD control group ($n = 17$; $M_{age} = 4.75$) and a typically developing (TD) control group ($n = 17$; $M_{age} = 5.45$). Children in the experimental group received 1 weekly individualized MMN intervention session for 10 weeks, while those in the control groups continued with their usual intervention sessions (either at speech-therapy or school level). Children's oral narrative skills were evaluated before and around a week after the end of the intervention with a retelling task with trained and untrained stories with a scoring system of 0-6 (see Table 1 for the scoring criteria). Also, after each intervention session, two dynamic assessment (DA) measures were administered to evaluate children's learning ability by a) monitoring the ability to retell the trained story and b) using a hierarchy of prompts to correctly answer questions about the story elements.

Results revealed that children in the experimental group significantly improved in their ability to retell both trained and untrained stories from pre- to post-intervention ($ps < .01$) and that at post-intervention they outperformed their NDD-matched peers not receiving the intervention ($ps < .05$), and that their scores were similar to their TD peers (see Table 2 for all descriptives). No significant effects were observed for perspective-taking skills. Results from the two dynamic assessment measures revealed, first, that there was a significant improvement in macrostructure skills between sessions 2 and 3 ($p = .015$)—an improvement which was maintained throughout the rest of the sessions—, while the significant improvement for perspective-taking skills was reached at session 6 ($p = .015$). Second, children were able to answer questions about the story but needed support prompts, such as two-option questions or images ($p < .001$). Finally, we observed that the average of support prompts children needed to answer these questions throughout the sessions significantly predicted their narrative outcomes at post-test, suggesting that those who needed more support showed smaller improvements at post-test ($\beta = -.214, p = .002$).

Our findings indicate that the MMN intervention, which systematically integrates multimodal cues, can effectively boost children's oral narrative skills, particularly macrostructure skills. These results also highlight the importance of using DA measures assessing learning ability, such as monitoring and prompts, as these are sensitive to the narrative training gains.

Narrative macrostructure	Narrative perspective-taking
0) The retelling does not include any descriptive sequence. 1) It includes 1 descriptive sequence with no temporal sequence. 2) It includes an action sequence (e.g., main character + problem). 3) It is incomplete and lacks 2 or more of the macrostructure elements (character, problem, attempt, solution, final). 4) It is incomplete and lacks 1 of the elements. 5) It is a complete and includes all elements. 6) It is complete (includes all elements) and also adds details about the story.	0) The retelling does not include any emotion. 1) It includes 1 emotion. 2) It includes 2 or more emotions. 3) It includes 1 emotion + its cause. 4) It includes 2 or more emotions + the cause of at least 2 emotions. +1) It includes 1 mental term (such as <i>thinking, realizing, willing, wanting</i>). +2) It includes 2 or more mental terms.

Table 1: *Scoring criteria for narrative macrostructure and narrative perspective-taking*

		Control TD		Control NDD		Experimental NDD	
		PRE	POST	PRE	POST	PRE	POST
Narrative macrostructure	Trained story	4.35 (0.70)	3.71 (0.92)	1.82 (1.78)	2.76 (1.35)	2.63 (1.02)	4.00 (1.32)
	Untrained stories	4.29 (1.14)	4.04 (0.63)	2.22 (1.41)	2.29 (1.31)	2.81 (1.09)	3.77 (1.19)
Narrative perspective-taking	Trained story	0 (0)	0 (0)	0.24 (0.75)	0.06 (0.24)	0 (0)	0.31 (0.60)
	Untrained stories	0.39 (0.40)	0.06 (0.13)	0.29 (0.44)	0.02 (0.08)	0.26 (0.29)	0.23 (0.23)

Table 2: *Means (SD) narrative macrostructure scores and narrative perspective-taking for trained and untrained stories (upper panel) and trained story (lower panel) broken down by Test and Group. Bold indicates significant effects.*

References

- [1] C. Norbury and D. V. Bishop, "Narrative skills of children with communication impairments," *International Journal of Language & Communication Disorders*, vol. 38, no. 3, pp.287–313, 2003. doi: 10.1080/136820310000108133.
- [2] C. Norbury, T. Gemmell, and R. Paul, "Pragmatics abilities in narrative production: a cross-disorder comparison," *Journal of Child Language*, vol. 41, no. 3, pp. 485–510, 2014. doi: 10.1017/S030500091300007X.
- [3] K. Favot, M. Carter and J. Stephenson, "The Effects of Oral Narrative Intervention on the Narratives of Children with Language Disorder: a Systematic Literature Review," *Journal of Developmental and Physical Disabilities*, vol. 33, pp. 489–536, 2021. doi: 10.1007/s10882-020-09763-9
- [4] S. Goldin-Meadow, "Widening the lens: What the manual modality reveals about language, learning and cognition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, 2014. doi: 10.1098/rstb.2013.0295.
- [5] I. Vilà-Giménez and P. Prieto, "The Value of Non-Referential Gestures: A Systematic Review of Their Cognitive and Linguistic Effects in Children's Language Development," *Children*, vol. 8, no. 2, 2021. doi: 10.3390/children8020148.

Session 7:

Phonetic aspects of gestures

26.09.2024
17:10-18:30



Pitch accent realization as a function of accompanying manual or eyebrow gestures in spontaneous Swedish dialogue

Gilbert Ambrazaitis¹, Margaret Zellers², David House³

¹*Linnaeus University, Växjö, Sweden*, ²*Kiel University, Germany*,

³*KTH Royal Institute of Technology, Stockholm, Sweden*

`gilbert.ambrazaitis@lnu.se`

Prosodic research on speech-gesture integration has shown that gestures temporally align with prominent units in speech, e.g., [1][2][3], as formulated in terms of the phonological synchronization rule by McNeill [4]. Some studies have also provided evidence that speech and gesture may converge not only in the temporal, but also in the “spatial” domain, displaying correlations between the presence and strength of gestures (magnitude or complexity of gestures) with the strength of acoustic parameters in the production of prosodic prominence (as reflected, for instance, in the accentual *fundamental frequency* [f_0] range), e.g., [5][6][7][8]. This spatial convergence has been formulated in terms of the *Cumulative-Cue Hypothesis* [7][9] and has been argued to result from an underlying compulsion to express prominence in both speech and gesture, all else being equal. This compulsion could be understood as part of a *revised Effort Code* [10]: To signal prominence, we tend to produce vocal *and gestural* signals indicating an increased level of effort [9]. However, evidence in favor of the Cumulative-Cue Hypothesis is still rather sparse and diverse and stems mostly from studies involving instructed or elicited movements, rather than naturally occurring co-speech gestures. Also, most studies have strictly focused on arm or hand gestures, and hardly any studies have considered gestural clustering (e.g., combined hand and head gestures) as a possible dimension of gestural strength.

The present study extends this line of research, studying the realization of phrase-level pitch accents in Swedish (so-called ‘big accents’, see Fig. 1) as a function of accompanying manual gesture strokes and eyebrow movements. Our materials consist of Swedish spontaneous dyadic conversations taken from the Spontal Corpus [11]. So far, data from eight speakers (four female, four male; 20 minutes in total, or 4294 words) have been included in our preliminary analysis (Tab.1, Fig. 2), but more data are currently being processed. Big accents (BA) were manually labelled with access to the audio channel and an f_0 display, but without using the video channel. Manual gestures (MG) and eyebrow movements (EB) were manually labelled with access to the video only. All events (BA, MG, EB) were labelled, at least partially, by two annotators, revealing acceptable inter-rater reliabilities ($\kappa_{BA}=.78$; $\kappa_{EB}=.70$; $\kappa_{MG}=.82$). For BAs, f_0 landmarks were annotated manually (Fig. 1), following the criteria specified in [7]. Based on these landmarks, two dependent variables were calculated: the range (in semitones) of the accentual fall (if present) of the potentially two-peaked BA (see Fig. 1), and the range of the subsequent big-accent rise. Linear mixed modeling and likelihood ratio tests were used to assess how well the ranges of the fall and the rise are predictable by the presence of gesture (clusters), operationalized as a predictor MMP (*multimodal prominence*), comprising the four levels BA (= accent only, no gesture), BA+MG, BA+EB, BA+MG+EB.

In this preliminary data set, EBs and MGs seldom clustered (Tab. 1), that is, pitch accents most often occurred either with a manual or an eyebrow gesture, or without any gesture. The preliminary results reveal a significant trend for larger f_0 rises when an eyebrow movement accompanies the accented word, as indicated by a significant contribution of the predictor MMP ($\chi^2=19.93$, $df=3$, $p<.001$), and significant post-hoc comparisons for BA vs. BA+EB ($t=4.96$, $df=758$, $p<.001$) and BA+MG vs. BA+EB ($t=4.75$, $df=758$, $p<.001$). This suggests new, partial evidence for the Cumulative-Cue Hypothesis, although results for the BA+MG+EB cluster seem to suggest counterevidence. However, this preliminary data set contains very few data points for BA+MG+EB. At the conference, we will also discuss these results in relation to other hypotheses characterized by trading relationships.

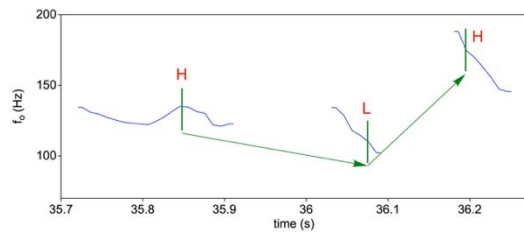


Figure 1: f_0 landmarks, illustrated for a word from the current materials, produced with a 'big accent' (BA). The arrows indicate the dependent variables (the fall and the rise) that were calculated from the labelled landmarks. The two-peaked nature of the BA depends on the lexical-prosodic features of the word; the fall can be absent.

MMP	Fall	Rise
BA	154	488
BA+MG	81	209
BA+EB	10	46
BA+MG+EB	2	9

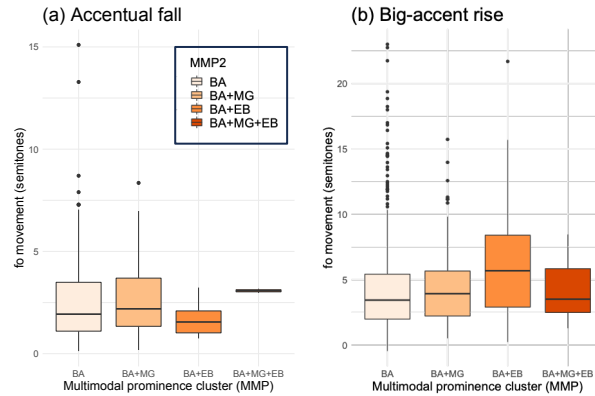


Figure 2: Boxplots for the accentual fall (a) and the following big-accent rise (b) measured in semitones as a function of the multimodal prominence cluster (MMP); for sample sizes, see Tab. 1.

Table 1: Sample sizes for the accentual fall and the subsequent rise of BAs per multimodal prominence cluster (MMP); see text for explanations.

References

- [1] M. L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English," *Speech Communication*, vol. 52, no. 6, pp. 542–554, 2010. doi: 10.1016/j.specom.2009.12.003.
- [2] D. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 3, no. 1, pp. 71–89, 2012. doi: 10.1515/lp-2012-0006.
- [3] N. Esteve-Gibert, J. Borràs-Comes, E. Asor, M. Swerts, and P. Prieto, "The timing of head movements: The role of prosodic heads and edges," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4727–4739, 2017. doi: 10.1121/1.4986649.
- [4] D. McNeill, "Hand and mind: What gestures reveal about thought," University of Chicago Press, 1992.
- [5] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Spatiotemporal coupling between speech and manual motor actions," *Journal of Phonetics*, vol. 42, pp. 1–11, 2014. doi: 10.1016/j.wocn.2013.11.002.
- [6] W. Pouw, L. de Jonge-Hoekstra, S. J. Harrison, A. Paxton, and J. A. Dixon, "Gesture–speech physics in fluent speech and rhythmic upper limb movements," *Annals of the New York Academy of Sciences*, vol. 1491, no. 1, pp. 89–105, 2021. doi: 10.1111/nyas.14532.
- [7] G. Ambrazaitis and D. House, "Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters," *Laboratory Phonology*, vol. 24, no. 1, 2022. doi: 10.16995/labphon.6430.
- [8] S. Berger and M. Zellers, "Multimodal prominence marking in semi-spontaneous YouTube monologs: The interaction of intonation and eyebrow movements," *Frontiers in Communication*, vol. 7, 2022. doi: 10.3389/fcomm.2022.903015.
- [9] G. Ambrazaitis and D. House, "The multimodal nature of prominence: some directions for the study of the relation between gestures and pitch accents", *Proceedings of the 13th International Conference of Nordic Prosody*, Sciendo, pp. 262–273, 2023. doi: 10.2478/9788366675728-024
- [10] C. Gussenhoven, "The phonology of tone and intonation," Cambridge University Press, 2004. doi: 10.1017/CBO9780511616983.
- [11] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture," *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 2992–2995, 2010. Valetta, Malta.

Gesture apex coordination with prosodic structure and tonal events in Maltese EnglishMartine Grice¹, Alexandra Vella², Maria Lialiou¹, Florence Bails³, Aviad Albert¹,Petra B. Schumacher¹, Nadia Pelageina¹ & Solveigh Janzen¹¹University of Cologne, ²University of Malta, ³Universitat de Lleida

martine.grice@uni-koeln.de

This study investigates the co-occurrence of manual gestures with prosodic structure and associated tonal events in Maltese English (MaltE). In this variety of English, H tones can be on a metrically stressed syllable, in which case they constitute pitch accents, but, similarly to Turkish [1], tones can also occur on an initial unstressed syllable of a prosodic-word-sized constituent, which can be the initial syllable of a content word or a (monosyllabic) function word preceding the content word. When there is an initial H tone (an 'early peak' [2]), tone and lexical stress compete as cues to prominence, contributing to a reduced sensitivity to stress in sequence recall tasks ("stress deafness") in bilingual speakers of Maltese and MaltE [3].

It has been proposed that the strokes and apices of referential and non-referential manual gestures co-occur with pitch accents or, more broadly, the stressed syllables associated with them [4,5,6,7,8]. To shed more light on the prominence cueing potential and prosodic status of early peaks in MaltE, we ask whether these early peaks can function as gestural anchors in a similar way to pitch accents. To address this question, we analysed a 14-minute long TEDx talk in MaltE [9]. We annotated H* pitch accents and early H tones [10] and, in a separate step, gestural strokes and apices [11]. First, to assess the alignment between gestures and tones, we calculated the distance between the annotated gesture apices and the F0 peaks corresponding to H* and early H tones within the same content word (as well as H on preceding function words; H and H* were never on the same word). In Fig. 1, the distribution that links the gesture apex to a H* pitch accent peak is narrow and centred around 0 (blue), indicating a robust link between gestural apices and H*. Since our prosodic annotation was restricted to H and H*, this robust link covers less than half of the gestural apices in our data (202 out of 445). The distributions that link the gesture apex to an early H tone within the content word (pink) are broader and reflect even fewer cases. Moreover, when the H tone occurs on a preceding function word (green), the timing appears to be normally distributed around 250 ms after the H peak rather than being skewed leftwards to reflect any link between the gesture apex and the early H peak.

The picture changes when we switch our timing perspective from the earlyH/H* peaks to the lexically stressed syllable, with substantially more cases appearing to be linked in this way: regardless of the type of tonal event linked with gesture-syllable pairs, they all seem to be normally distributed within the stressed syllable (376 out of 445 apices are distributed around the centre of the stressed syllable). Fig. 2 shows the 272 cases that we could link to an early H or H* peak. Crucially, the subset of cases related to the early H on a preceding function word (green) have a similar distribution to those related to H* (blue). Early H peaks, however, do seem to slightly attract the gesture apex when occurring on the initial weak syllable of a content word: these results show a leftward trend relative to the stressed syllable (pink), indicating a possible additional coupling with the early H in this case [8].

We show that the co-occurrence of gestural and prosodic units contributes to our analysis of the intonational phonology of MaltE: the fact that gesture apices align differently with H* and early H provides evidence for the two tones being phonologically distinct categories. Moreover, MaltE is different from Turkish [1], in which the lack of a discernible H peak on the content word leads to the gesture apex being aligned with the beginning of the prosodic word. Our results indicate that gesture apices may not be directly linked to H tones in terms of turning points in the F0 curve, but to potentially tone bearing syllables, i.e. to metrically strong syllables. We are currently planning follow-up studies that will explore gestural apex co-occurrence with syllabic landmarks in the acoustic signal using the ProPer toolbox [12].

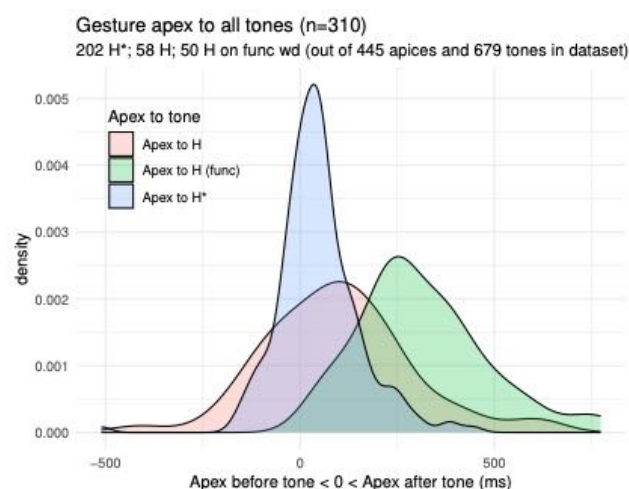


Figure 1: *Alignment of gesture apex with F0 peaks corresponding to H*, early H (on content word) and early H (on preceding function word).*

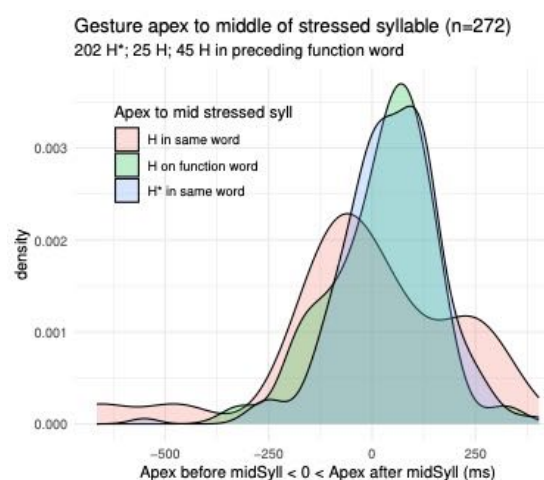


Figure 2: *Alignment of gesture apex with centre of stressed syllable when there is a H* on this syllable, and when there is an early peak on the same (content) word or on a preceding function word.*

References

- [1] O. Turk and S. Calhoun, “Multimodal cues to intonational categories: Gesture apex coordination with tonal events”, *Laboratory Phonology* 14(1), 2023.
- [2] A. Vella, “Alignment of the ‘early’ HL sequence in Maltese falling tune wh-questions”, *Proc. 15th International Congress of Phonetic Sciences* pp. 2062-2065, 2011.
- [3] M. Lialiou, A. Bruggeman, A. Vella, S. Grech, P. B. Schumacher and M. Grice, “Word-level prominence and ‘stress deafness’ in Maltese-English bilinguals”, *Proc. 20th International Congress of Phonetic Sciences*, pp. 132-136, 2023.
- [4] S. Shattuck-Hufnagel and A. Ren, “The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech”, *Frontiers in Psychology* 9, 2018.
- [5] P. Rohrer, E. Delais-Roussarie and P. Prieto, “Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses”, *Lingua* 293, 2023.
- [6] A. Gregori and F. Kügler, “The distribution of co-speech gestures, information structure and prosody: A corpus study on prominence peak alignment” Talk at *Phonetik und Phonologie im deutschsprachigen Raum*. <https://doi.org/10.11576/pundp2022-1029>, 2022.
- [7] S. Repp, L. Muhtz and J. Heim, “Alignment of beat gestures and prosodic prominence in German”, *Proc. Interspeech*, 107–111, 2023.
- [8] H. R. Bosker, M. Hoetjes, W. Pouw, W. and L. van Maastricht, “Gesture-speech coupling in L2 lexical stress production: A pre-registration of a speech acoustic and gesture kinematic study”. 2021, <https://doi.org/10.31219/osf.io/w2ezs>
- [9] TEDx University of Malta: “Till death do us part”, <https://www.youtube.com/watch?v=SVHda2x4B6Y>
- [10] A. Amalia, and J. Fletcher, “The Autosegmental-Metrical Theory of Intonational Phonology”, in C. Gussenhoven, and A. Chen (eds), *The Oxford Handbook of Language Prosody*, Oxford Handbooks, 2020.
- [11] P. Rohrer, U. Tütüncübasi, I. Vilà-Giménez, J. Florit-Pons, G. Gurrado, N. Esteve-Gibert, A. Ren-Mitchell, S. Shattuck-Hufnagel and P. Prieto, “The MultiModal MultiDimensional (M3D) labeling system”. doi: 10.17605/osf.io/ankdx/
- [12] A. Albert, F. Cangemi, T. M. Ellison, and M. Grice. ProPer: PROsodic analysis with PERiodic energy [Computer software], 2023. <https://osf.io/28ea5/>

Timing of Co-Speech Gesture in Igbo: Influence of Metrical Prominence and Tonal Melody

Kathryn Franich¹ and Vincent Nwosu²

Harvard University¹

University of Calgary²

kfranich@fas.harvard.edu

Introduction: Co-speech gestures are timed to occur with metrically-prominent syllables in several languages [1,2,3]. Little research has examined the temporal alignment of co-speech gestures in African tonal languages, where metrical prominence is often hard to identify [5]. Existing findings for African languages indicate that co-speech gestures gravitate to stem-initial position [4], which in some languages corresponds with metrical prominence [6]. Here, we look at the timing of co-speech gestures in Igbo, a Niger-Congo language with High and Low tones. Findings indicate that metrical structure, phrase position and tone melody—specifically, melodies involving sequences of two H tones—all play an important role in gesture alignment.

Background/Hypotheses: Evidence for typical acoustic correlates of linguistic stress (e.g. increased duration and intensity) is lacking in Igbo, as is the case for many other African tonal languages [6]. However, metrical foot structure is still posited to occur in the language: initial syllables in trisyllabic words—but not disyllabic words—are protected from undergoing downstep in certain grammatical contexts. Clark [7] attributes this fact to metrical prominence on odd numbered syllables in Igbo words, counting from the word-final syllable. We therefore predict that word-final syllables (and odd numbered syllables) will be more likely to attract a gesture than non-final syllables in Igbo. We also explore the possibility that high tones are more likely to attract a gesture than low tones, in line with findings from English [8].

Methodology: Data consist of conversational speech produced by 4 Igbo speakers recorded in pairs in Northern Nigeria. 1,073 gestures of the hands were coded manually in video data by a team of coders, with inter-coder reliability established [9]. Apexes of co-speech gestures—defined as peak velocity of manual movement—were extracted along with the phones with which they overlapped. We focus here on polysyllabic words of up to 3 syllables (total of ~750 tokens). Data were coded for word-, intonational phrase-, and stem-position, and tone.

Results: There was a statistically-credible effect of word position on gesture occurrence, such that word-final syllables were more likely to be targeted for gestures than non-final syllables (Figure 1; $\beta = 1.08$; 95% CIs [0.27, 1.85]). However, an interaction between word position and phrase position for trisyllabic words suggested that word-medial syllables were actually more likely to be targeted for a gesture phrase-medially (Figure 2; $\beta = 1.48$, 95% CIs [0.03, 2.95]). We suggest that gesture position may shift here due to the fact that vowel coalescence leads to ambiguity between word-final and word-initial syllables phrase-medially. While there was no effect of aligned tone on gesture occurrence, tone melody appears to play a role: gestures on disyllabic words were proportionally more likely to occur in word-initial position if that position was also the first in a HH sequence (Fig. 3). For trisyllables, preferential alignment was to the left member of a HH sequence in the word, unless broken by a downstep (Fig. 4).

Discussion: While metrical prominence is an important factor in determining gesture alignment in Igbo, other factors also play a role in gesture attraction. Most striking was the effect of tone melody, which reveals a preference for gestures on syllables which initiate a HH sequence not broken up by a downstep. These results are consistent with arguments for the existence of *tonal feet* in African tone languages, which have been used to explain the tendency in many languages for high tones to come in pairs [10] and for high tone spreading to be constrained to bisyllabic domains [11]. In conclusion, gestures in Igbo, as in other languages, highlight metrical structure at various levels of phonological representation.

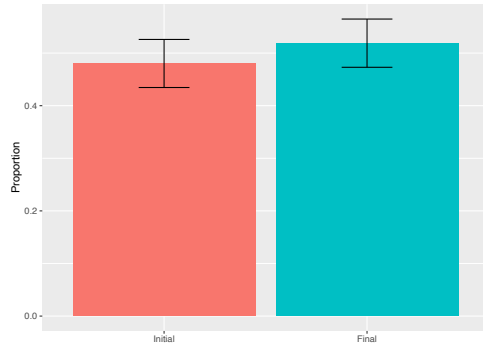


Fig. 1: Gesture occurrence by word position, disyllabic words

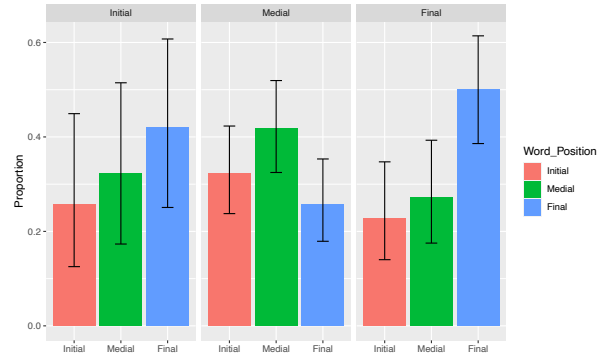


Fig. 2: Gesture occurrence by word position (indicated in colors) and phrase position (indicated in vertical panels), trisyllabic words

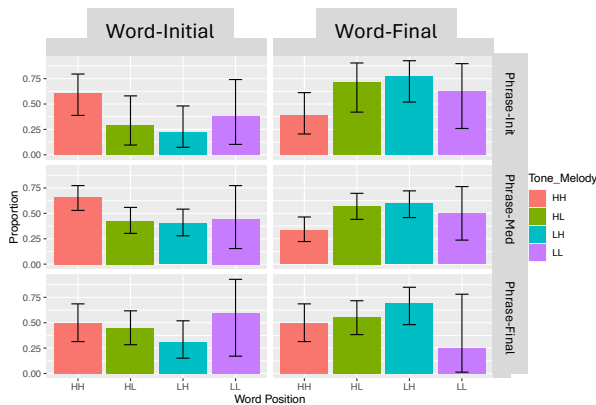


Fig. 3: Gesture by word position, phrase position, and tone melody, disyllabic words

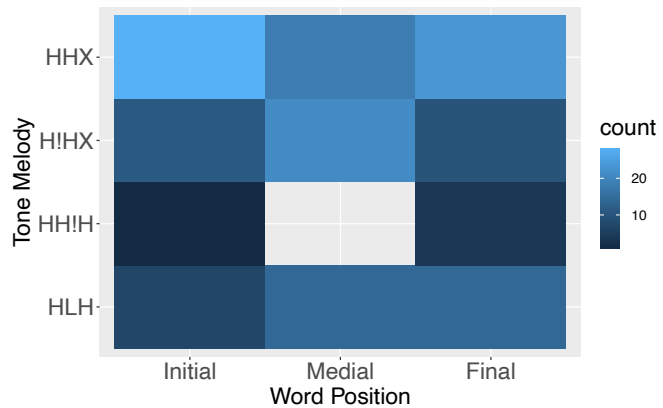


Fig. 4: Gesture by word position and tone melody, trisyllabic words

- [1] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech." *Language and Cognitive Processes*, 26, 10, 2011, pp. 1457-1471.
- [2] D. Loehr, *Gesture and Intonation*. Washington, D.C.: Georgetown University dissertation, 2012.
- [3] N. Esteve-Gibert and P. Prieto, "Prosodic structure shapes the temporal realization of intonation and manual gesture movements." *JSLR*, 56, 3, 2013, pp. 850-64.
- [4] Author, 2022
- [5] Hyman, L. (2014). Do all languages have word accent? In H. Van der Hulst (Ed.), *Word Stress: Theoretical and Typological Issues* (pp. 56-82). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139600408.004.
- [6] Downing, Laura J. 2010. Accent in African languages. A survey of word accentual patterns in the languages of the world, ed. by Harry van der Hulst, Rob Goedemans, and Ellen van Zanten, 381-427. Berlin: De Gruyter Mouton.
- [7] Clark, M. 1990. *The Tonal System of Igbo*. Dordrecht: Foris Publications.
- [8] Im, Suyeon & Baumann, Stefan. (2020). Probabilistic relation between co-speech gestures, pitch accents and information status. *Proceedings of the Linguistic Society of America*. 5. 685. 10.3765/plsa.v5i1.4755.
- [9] Kita, S., van Gijn, I., and van der Hulst, H. 2006. Movement phases in signs and co-speech gestures, and their transcription by human coders. *Gesture and Sign Language in Human-Computer Interaction*, 23-35.
- [10] Leben, W.R. 1996. Tonal feet and the adaptation of English borrowings into Hausa. *Studies in African Linguistics*, 25, 139-154.
- [11] Bickmore, L. 2003. The use of feet to account for binary tone spreading. In *Frankfurter Afrikanistische Blätter* vol. 15, R. J. Anyanwu (ed.), Rudiger Koepe Verlag, Koln

Quantifying the visual salience of Swedish vowels: A computer vision approach MMSYM 2024

Helene Springer¹, Henrik Garde², Frida Splendido¹, Marianne Gullberg^{1,2}
Lund University¹, Lund University Humanities Lab²

Corresponding author Email: helene.springer@ling.lu.se

Speech is a multimodal phenomenon that incorporates a variety of acoustic and visual cues that are part of our daily interactions. A wide range of studies has shown that visual information from articulatory gestures and lip movements facilitate the processing of speech sounds under both adverse and clear conditions, and in both first (L1) and second language (L2) listeners [1], [2], [3]. These studies suggest that this is not a constant effect but is modulated by several factors, one of them being the visual salience of speech sounds. However, the notion of visual salience is generally poorly operationalized. Therefore, the current study sets out to operationalize it using a computer vision approach.

In multimodal perception research, descriptions of visual salience are often made based on binary, phonological features from the articulatory perspective [3], [4] or based on speechreading performances [5], [6]. However, building feature matrices based on the assumed visually salient features ‘openness’ and ‘roundedness’ risks making incorrect predictions about the visual salience of sound contrasts (Table 1). In contrast, fine-grained statements on visual salience based on continuous measures of visual lip parameters would allow for vowel contrasts to be grouped into visually high- and low-salient from the addressee’s perspective. Such statements can be made by extracting the two-dimensional visual parameters area, height, and width of the mouth opening from video material. To explore these issues, we customized a script detecting Dlib’s facial landmarks on the mouth of a Swedish L1 speaker producing long vowels in continuous speech (Figure 1), thereby extending previous research that has focused on sustained vowels and monosyllabic words [7], [8]. The measurements were based on picture frames of the manually annotated static vowel midpoints. Area was defined as the number of pixels within the polygon that connects all landmarks along the inner mouth outline, the distance between the landmarks 63 and 67 determined the height, and the distance between the landmarks 61 and 65 the width in pixels (Figure 1, left panel).

The results reflect the gradual variation encountered by listeners in natural speech and listening settings. Specifically, the measurements reveal significant visual differences between vowels with similar phonological features (e.g. /u:/ and /ø:/), as well as a lack of significant visual differences between vowels with different features (e.g., /y:/ and /i:/). The results thus challenge approaches to visual salience based on binary, phonological features. The computer vision approach proved to be a good starting point for visual speech measurements, although landmark inaccuracies make post-corrections of the data inevitable.

Potential implications include educational contexts (e.g., visual instructions), the support of speech perception for vulnerable listener populations (including hard of hearing and L2 populations), and the development of systems recognizing and synthesizing speech. Finally, this work can provide insights into the salience and interaction of the acoustic and visual cues in multimodal speech perception. Sound contrasts that are predicted to be visually salient based on phonological features may look similar in the continuous speech signal, which may diminish the facilitatory effect of visual cues and therefore, input-driven perspectives on L2 acquisition can benefit from quantified descriptions of visual salience in multimodal speech perception.

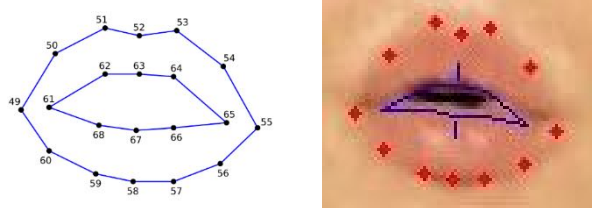


Figure 1: Dlib's facial landmarks (left) and an example of the detected landmarks on one of the video frames (right).

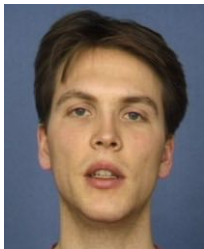
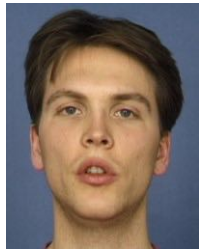

	/i:/	/y:/	/u:/
Phonological feature			
±open	–	–	–
± rounded	–	+	+
± front	+	+	+ ¹
Visual similarity			

Table 1: Phonological feature descriptions of the Swedish long vowel phonemes /i:/, /y:/, and /u:/, and their visual appearance.

References

- [1] P. Arnold and F. Hill, “Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact,” *British Journal of Psychology*, vol. 92, no. 2, pp. 339–355, May 2001, doi: 10.1348/000712601162220.
- [2] L. Drijvers and A. Özyürek, “Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension,” *Lang Speech*, vol. 63, no. 2, pp. 209–220, Jun. 2020, doi: 10.1177/0023830919831311.
- [3] H. Traunmüller and N. Öhrström, “Audiovisual perception of openness and lip rounding in front vowels,” *Journal of Phonetics*, vol. 35, no. 2, pp. 244–258, Apr. 2007, doi: 10.1016/j.wocn.2006.03.002.
- [4] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier, “Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise,” *The Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3677–3689, Jun. 1998, doi: 10.1121/1.423069.
- [5] S. Amcoff, “Visuell perception av taljud och avläsestöd för hörselskadade,” Lärarhögskolan i Uppsala, Pedagogiska Institutionen, Uppsala, 7, 1970.
- [6] J. Mártony, “On speechreading of Swedish consonants and vowels,” *STL-QPSR*, vol. 15, no. 2–3, p. 37, 1974.
- [7] W. Linker, “Articulatory and acoustic correlates of labial activity in vowels: A crosslinguistic study,” Los Angeles, Working Papers in Phonetics 56, 1982.
- [8] C. Ericsson, “Articulatory-acoustic relationships in Swedish vowel sounds,” Doctoral Dissertation, Stockholm University, Stockholm, 2005.

¹ Phonological feature accounts differ in terms of this classification, some state it is a central vowel. However, in strictly binary descriptions, it is classified as +front.

Session 8: Semantics / Pragmatics of gestures

27.09.2024

9:00-10:20



The Gesture–Prosody Link in Multimodal Grammar

Andy Lücking¹, Alexander Mehler¹ and Alexander Henlein¹


¹*Goethe University Frankfurt, Text Technology Lab*

luecking@em.uni-frankfurt.de

Speech and manual gesture are means of communication on different channels, but can still interact semantically. Furthermore, the (non-)interaction of speech and gesture shows characteristics of temporal and semantic well-formedness. Accordingly, theoretical linguistics developed multimodal grammars that regiment some of the interaction of speech and gesture. However, since gestures and their affiliated expressions in speech (usually words) can be temporally offset signals, temporal alignment cannot be the only means of linking them. Therefore, two additional constraints have been identified: (i) the gesture–prosody link plays a vital role as a kinematic–phonetic affiliation interface, and (ii) lexicalized visual models – *conceptual vector meanings* (CVM) in *dual coding* – act as “semantic filters” on multimodal integration. Within a temporal window, a gesture attaches to a prosodically marked expression, *if* the imagistic representation of that expression matches the visual percept gained from the gesture.

As an example, consider (1), taken from SaGA dialogue 6, around 1m51s; square brackets indicate temporal overlap of speech and gesture, uppercase indicates primary stress [1]:



- (1) beziehungsweise auf der [ rechten Seite wird so’ne SÄUle sein]
Or rather, on the [right-hand side there will be a pillar like this]

The speaker moves his hands up and down while saying *rechten Seite* ‘right-hand side’ and repeats the gesture while continuing to say *so’ne Säule* ‘a pillar like this’. Hence, “right-hand side” is the *prima facie* affiliate. But the gesture obviously shapes the outline of the pillar. How to capture this in multimodal grammars? Two features prevent the gesture from being integrated with “right-hand side”: the main stress is carried by “pillar”, not by “right-hand side”, and the gesture is not performed on the right-hand side (i.e., does not fit to the visual image of *right*, but to that of the axis vector of “pillar”).

This formal contribution introduces a constraint-based multimodal grammar extending on Head-driven Phrase Structure Grammar [2]. The technical notion of channel crossing chart parsers is introduced [3], and how it backs up multimodal integration schemes in grammar [4], [5] – see Figure 1 for an example. Lexicalized visual models are represented as vector sequences, following work in psychophysics and spatial semantics [6], [7]. Unification of vector representations is employed as a formal implementation of the *extended exemplification* relation from iconic gesture semantics [8]. In sum, multimodal grammars develop a precise spell-out of the rationale that speech and gesture integration follows the triplet of “temporal + prosody + CVM” to project temporally offset, semantically congruent, channel-crossing constituents.

Building on this compositional multimodal semantics, the talk discusses three additional topics: (i) incongruities due to gesture–CVM mismatches, (ii) frame-based enrichments triggered by non-overt affiliates, and (iii) cases challenging the gesture–prosody link such as gestural holds and repetitions.

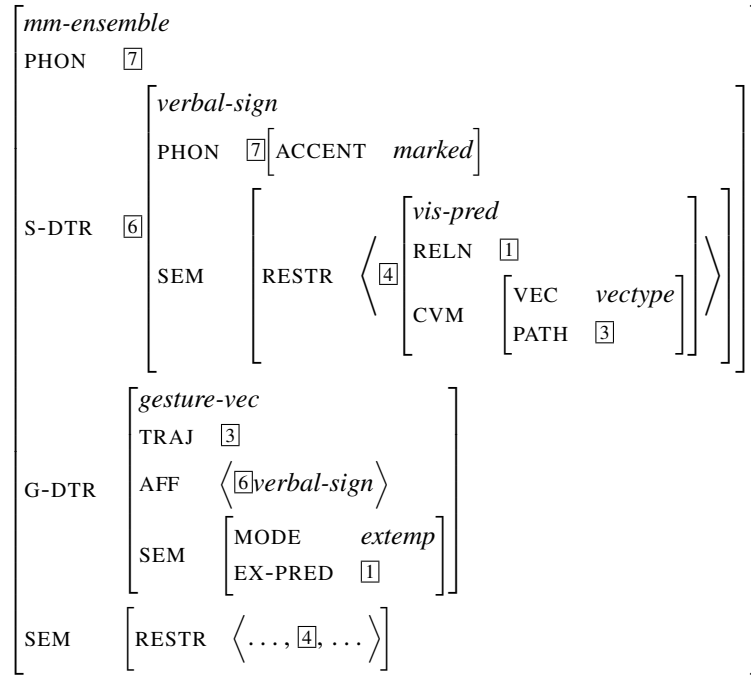


Figure 1: *Basic speech–gesture integration scheme. Within a certain temporal window, a gesture daughter (G-DTR) attaches (feature AFF and tag [6]) to a verbal expression (S-DTR) if the speech daughter is phonetically marked (tag [7]) and carries a conceptual vector meaning (CVM) that is compatible to the trajectory performed by the gesture daughter (tag [3]). The mother construction – a multimodal ensemble (mm-ensemble) – inherits the syntactic-semantic properties of the speech daughter (outmost PHON and SEM, SYN not shown due to space restrictions).*

References

- [1] A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser, “The Bielefeld speech and gesture alignment corpus (SaGA),” in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, ser. LREC 2010, 7th International Conference for Language Resources and Evaluation, Malta, 2010, pp. 92–98. DOI: 10.13140/2.1.4216.1922.
- [2] C. Pollard and I. A. Sag, *Head-Driven Phrase Structure Grammar*. Stanford, CA: CSLI Publications, 1994.
- [3] M. Johnston, “Unification-based multimodal parsing,” in *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics – Volume I*, Montreal, Quebec, Canada, 1998, pp. 624–630.
- [4] K. Alahverdzhieva, A. Lascarides, and D. Flickinger, “Aligning speech and co-speech gesture in a constraint-based grammar,” *Journal of Language Modelling*, vol. 5, no. 3, pp. 421–464, 2017.
- [5] A. Lücking, *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. Berlin and Boston: De Gruyter, 2013.
- [6] G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973. DOI: doi.org/10.3758/BF03212378.
- [7] J. Zwarts, “Vectors across spatial domains: From place to size, orientation, shape, and parts,” in *Representing Direction in Language and Space*, ser. Explorations in Language and Space 1, E. van der Zee and J. Slack, Eds., Oxford, NY: Oxford University Press, 2003, ch. 3, pp. 39–68.
- [8] A. Lücking, A. Henlein, and A. Mehler, “Iconic gesture semantics,” *L&P*, 2024, lingbuzz/007916, pre-published.

From Hand to Discourse: The Stabilization of the Slicing Gesture and its Metapragmatic Function in Mediated Discourse

Silva H. Ladewig
University of Goettingen
silva.ladewig@uni-goettingen.de

This paper explores the significant role of the Slicing gesture within the context of mediated political discourse and its implications for understanding gesture stabilization. This “recurrent gesture” [1-3] has not been previously described for German speakers but has been documented for French as the “Cutting gesture” [4] and English [5, 6], particularly in political discourse.

Drawing upon a rich corpus of political talk shows spanning five hours, this study documents the prominence of the Slicing gesture, characterized by a flat hand with fingers either held together or slightly spread apart. This gesture stands out in its frequency and application among 27 speakers. In total, 4,084 gestures were documented in the data, of which 3,038 are classified as recurrent gestures. The Slicing gesture was the most common, with 640 instances. This gesture exhibits three form variants: the palm facing towards the speaker's body (Figure 1a), diagonally towards the speaker's central gesture space (Figure 1b), or aligned with the sagittal plane (Figure 1c). While the variant with the palm facing towards the speaker's body is most frequently aligned with the speaker's own view, all variants perform the following functions: defining discourse objects, metapragmatic functions, performative functions, meta-comments, and modal functions. Among these, defining discourse objects and metapragmatic functions were the most common (Figure 2).

The study to be presented will focus on the intricate interplay between this recurrent gesture and speech, as well as its behavior in sequences of recurrent gestures. An in-depth examination using a linguistic approach to gestures [7, 8] reveals that beyond punctuating singular arguments with single uses of the Slicing gestures or embedded in short sequences of different recurrent gestures, the Slicing gesture gains momentum in extended sequences where variations of this gesture are used. In these cases, the gesture acquires a meta-pragmatic meaning, conveying a speaker's clarity in positioning and embodying the rhetorical quality of making an argument. This indicates that the speaker uses the gesture to present themselves as someone who clearly names things and positions themselves in the discussion by differentiating from their fellow discussants (interpersonal stance). The prevalence of this gesture in political talk shows, as opposed to its rarity in private settings (data: 3 hours of everyday conversations), supports Kiesling's observation that “how stances are taken, and which stances are taken, are often habitually repeated by people with similar identities [9].

In conclusion, this paper emphasizes the Slicing gesture as a distinctive semiotic resource in political communication, one that shapes the dynamics of public speaking and reflects practices linked to identity. Its utilization in political settings reveals the forms and functions of this recurrent gesture in German, offering insights into the complex interplay of gesture, language, and social interaction.



Figure 1: *Three form variants of the Slicing gesture*

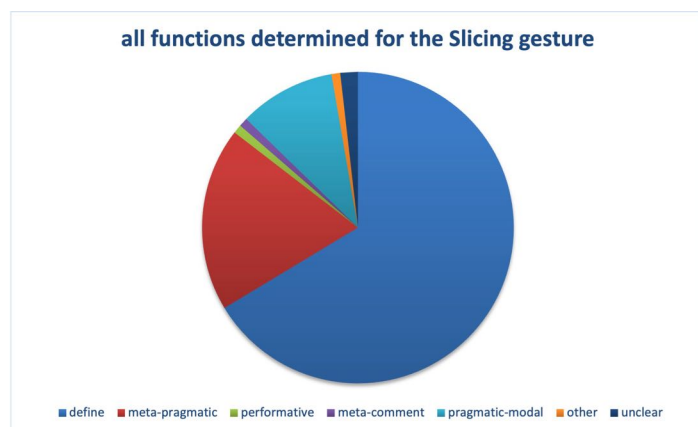


Figure 2: *Three form variants of the Slicing gesture*

References

- [1] Ladewig, S.H., Recurrent gestures, in *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction*, C. Müller, et al., Editors. 2014, De Gruyter Mouton.: Berlin, Boston. p. 1558–1575.
- [2] Müller, C., How recurrent gestures mean: Conventionalized contexts-of-use and embodied motivation. *Gesture*, 2017. 16(2): p. 278–306.
- [3] Ladewig, S.H., Recurrent Gestures: Cultural, Individual, and Linguistic Dimensions of Meaning-Making, in *The Cambridge Handbook of Gesture Studies*, A. Cienki, Editor. 2024, Cambridge University Press: Cambridge. p. 32-55.
- [4] Calbris, G., From cutting an object to a clear cut analysis. *Gesture as the representation of a preconceptual schema linking concrete actions to abstract notions*. *Gesture*, 2003. 3(1): p. 19-46.
- [5] Streeck, J., *Gesturecraft. The manu-facture of meaning*. 2009, Amsterdam, Philadelphia: John Benjamins.
- [6] Lempert, M., Uncommon resemblance: Pragmatic affinity in political gesture. *Gesture*, 2017. 16(1): p. 35-67.
- [7] Bressemer, J., S.H. Ladewig, and C. Müller, A linguistic annotation system for gestures (LASG), in *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*, C. Müller, et al., Editors. 2013, De Gruyter Mouton: Berlin, Boston. p. 1098-1125.
- [8] Müller, C., S.H. Ladewig, and J. Bressemer, Gestures and speech from a linguistic perspective: a new field and its history, in *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*, C. Müller, et al., Editors. 2013, De Gruyter Mouton: Berlin, Boston. p. 55-81.
- [9] Kiesling, S.F., Stance and stancetaking. *Annual Review of Linguistics*, 2022. 8: p. 409-426.

On the Role of Co-speech Gesture with ʔayʔajuθəm D Elements

Daniel K. E. Reisinger
University of British Columbia
daniel.reisinger@ubc.ca

Marianne Huijsmans
University of Alberta
huijsman@ualberta.ca

In this paper, we present a small experimental study that examines the role of co-speech gesture for demonstratives and determiners in ʔayʔajuθəm (ISO 639-3: coo), an endangered Salish language spoken by approximately 45 fluent speakers in British Columbia, Canada [1]. Just like other languages in the family [2, 3], ʔayʔajuθəm is known for its remarkably rich system of D elements, containing 17 distinct demonstratives and five distinct determiners [4, 5]. For the demonstratives, Reisinger and Huijsmans [4] distinguish between “gesture demonstratives” (GDEMs), which they claim create joint attention via the use of obligatory co-speech gesture, and “salience demonstratives” (SDEMs), which they claim refer to entities already salient in the discourse context and, consequently, do not to require gesture [4]. Using an experiment designed after similar work by Ebert et al. [7] on German, Reisinger and Huijsmans [6] recently argued that gesture is at-issue for the GDEMs but not for the determiners. This supports Ebert et al.’s claim that the contribution of gesture accompanying demonstratives is shifted into the at-issue dimension [7].

While Reisinger and Huijsmans [6]’s experiment provided initial insight on the use of co-speech gesture in ʔayʔajuθəm, it also raised new questions. First, their experiment did not include any SDEMs, so it is not yet known whether SDEMs also shift the contribution of co-speech gesture to the at-issue dimension. Secondly, their results cast doubt on whether gesture is in fact obligatory for GDEMs, since they found that participants often did not respond to gestureless uses as if they were infelicitous.

We designed a follow-up experiment comparing the role of co-speech gesture with GDEMs, SDEMs, and determiners. Four fluent participants were shown 72 test items, each of which consisted of a video in which an interviewer asks a yes/no question in ʔayʔajuθəm about one of five objects on a table in front of them (e.g., *təlosa Felipe {təʔta/taŋ/tə} pukʷ?* ‘Is Felipe reading {GDEM/SDEM/DET} book?’ + pointing gesture to a blue book) as well as a picture in which someone interacts with either the target item or another item (e.g., Felipe is holding a {blue / red} book). The experiment included three conditions: (i) a match condition in which the gesture referent in the video matches the item shown in the picture, (ii) a mismatch condition in which the gesture referent in the video does not match the item shown in the picture, and (iii) a no-gesture condition in which the video did not include any co-speech gesture. Participants were instructed to answer *ʔe* ‘yes’, *xʷaʔ* ‘no’, or *xʷač toχʷnexʷən* ‘I don’t know’. We hypothesized that they would answer ‘no’ more often in the mismatch condition when gesture is at-issue (accompanying demonstratives), and that they would be more likely to flag a missing gesture by responding ‘I don’t know’ when the question used a GDEM. The test items were interspersed by 34 filler items. Table 1 summarizes the results.

We find a high rate of ‘no’ responses in the mismatch condition with both GDEMs (96.7%) and SDEMs (93.1%), and a slightly lower rate for the determiners (82.8%). This result suggests a contrast in at-issueness for gesture accompanying demonstratives vs. determiners. On the other hand, the results challenge our hypothesis that absence of gesture should be difficult to interpret with GDEMs, resulting in ‘I don’t know’ responses. Participants still generally answered ‘yes’ (mostly) or ‘no’ (occasionally) (83.3%), though occasionally accompanied by comments flagging the missing gesture as problematic. While SDEMs reach a similar acceptance rate in this condition (81.0%), the number of ‘yes’/‘no’ responses for determiners is considerably higher (96.1%). Overall, our results support the hypothesis that demonstratives serve as “dimension shifters” [7, 8], extending empirical coverage across types of demonstratives, but challenge Reisinger and Huijsmans’ [6] claim that GDEMs require gesture.

	Match effect (= ‘yes’ answers in matching condition)	Mismatch effect (‘no’ answers in mismatch condition)	Acceptance effect (‘yes’/‘no’ answers in no-gesture condition)
GDEM (<i>təyʔta</i>)	100.0%	96.7%	83.3%
SDEM (<i>taŋ</i>)	100.0%	93.1%	81.0%
DET (<i>tə</i>)	100.0%	82.8%	96.1%

Table 1: Results of the experiment for the GDEM *təyʔta*, the SDEM *taŋ*, and the DET *tə*.

References

- [1] FPCC. 2022. Report on the status of B.C. First Nations languages. URL: <https://fpcc.ca/wp-content/uploads/2023/02/FPCC-LanguageReport-23.02.14-FINAL.pdf>
- [2] Montler, T. 2007. Klallam demonstratives. *Papers for ICSNL* 42:409–425.
- [3] Gillon, C. 2006 [2013]. *The Semantics of Determiners: Domain Restriction in Skwxwú7mesh*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.
- [4] Huijsmans, M. & D. K. E. Reisinger. 2022. Demonstratives in ʔayʔajuθəm Managing joint attention through gesture and salience. *Proceedings of Sinn und Bedeutung* 26:432–450.
- [5] Reisinger, D. K. E., M. Huijsmans, and L. Matthewson. 2021. Evidentials in the nominal domain a Speasian analysis of ʔayʔajuθəm determiners: *Proceedings of Sinn und Bedeutung* 25:751–768.
- [6] Reisinger, D. K. E. & M. Huijsmans. 2023. The Role of Gesture in ʔayʔajuθəm Determiners and Demonstratives. Paper presented at *Sinn und Bedeutung* 28, Ruhr-Universität Bochum, Germany.
- [7] Ebert, C., C. Ebert, & R. Hörnig. 2020. Demonstratives as dimension shifters. *Proceedings of Sinn und Bedeutung* 24(1):161–178.
- [8] Ebert, C. & C. Ebert. 2014. Gestures, demonstratives, and the attributive / referential distinction. Talk at Semantics and Philosophy in Europe 7, ZAS, Berlin, Germany.

The many ways to mark agreement & rejection: Multimodal polar responses in German

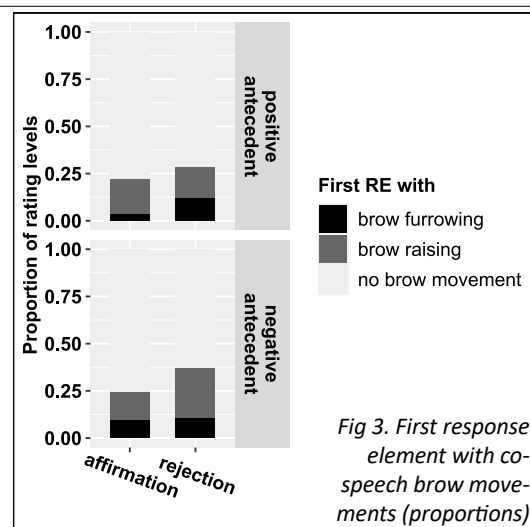
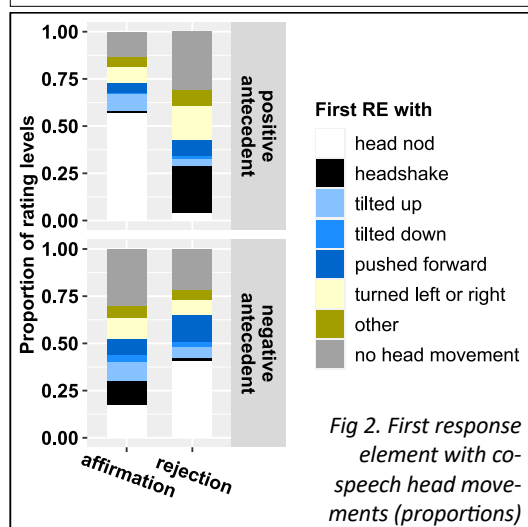
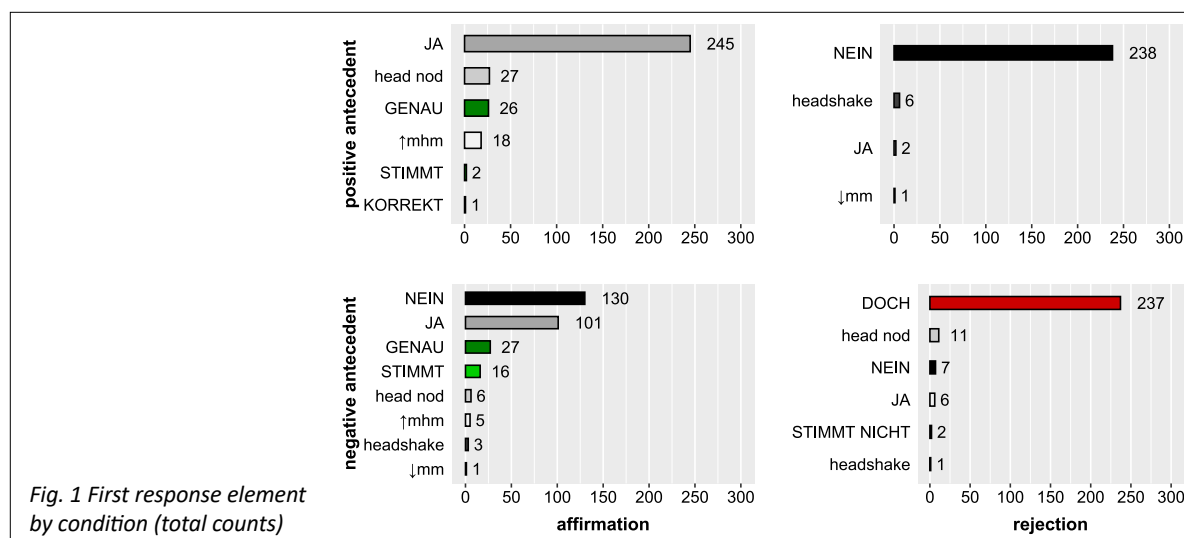
Cornelia Loos¹ & Sophie Repp²

Universität Hamburg¹, Universität zu Köln²

`cornelia.loos@uni-hamburg.de`

In German, affirmation and rejection of an assertion are typically expressed by response particles like *ja* ‘yes’ or *nein* ‘no’, followed by a response clause. Aside from marking these speech acts, response particles may also indicate the positive or negative polarity of the response. In responses to negative antecedents, the two functions come apart and response particles become ambiguous, see (1). An interesting question is if co-speech gestures are used for disambiguation, for instance, by marking different speech acts. Head movements have been claimed to encode affirmation and rejection [5][6], but also polarity ([3] for Russian). Data from Mandarin Chinese [7], Catalan, and Russian [4] suggest that head nods can encode positive polarity in rejections of negative antecedents. However, in visual gestural languages like German Sign Language (DGS), head movements are not used for disambiguation: they typically encode the same function as the manual response element they accompany, showing a preference for speech act marking (e.g., headshake for rejections) [8]. Brow movements also are sensitive to response type across languages, brow raising being frequent in rejections of negative antecedents [4][7][8], brow furrowing in rejections in general [8]. We present a multimodal production study of affirming and rejecting responses to positive and negative assertions in German, examining the contribution of co-speech gestures to encoding affirmation/rejection vs. response polarity. In a discourse completion task, 28 speakers (14m, 14f) watched videos (48 lexicalizations) to which they reacted. The experiment had a 2×2 design (speech act × antecedent polarity). Each video started with a narrator introducing a situation involving two characters. Then the first character appeared on screen and made an assertion. Participants assumed the role of the second character and affirmed or rejected the assertion while being video-recorded. We annotated vocal and non-vocal response elements (REs), response clauses, and co-speech movements of nine (non-)manual articulators: head, torso, shoulder, foot, gaze, eye lid, brows, mouth, hands. 1277 (of 1344) recordings were used for analysis. Participants produced 11 different REs including vocalizations (*hm*, *mm*) and head nods/shakes, see Fig.1. Most vocal REs were accompanied by three to four gestural articulators (independent of condition). Fig. 2 & 3 illustrate head and brows movements. The statistical analysis (linear mixed & multinomial models) revealed that some gestures aligned more with the speech acts while others aligned with response polarity. *Speech act alignment*: Up or down head tilts and smiling occurred more often in affirmations than rejections. Head protrusion, brow furrowing, and manual gestures (esp. smaller ones articulated at wrist or knuckle joints) were more common in rejections than in affirmations. *Polarity alignment*: Stand-alone and accompanying head nods and brow raises were more frequent in affirmations of positive antecedents and in rejections of negative antecedents, thus aligning with positive response polarity. Headshakes aligned with negative polarity. The polarity alignment of nods and shakes was mostly isofunctional with the REs they occurred on (*ja*–nod, *nein*–shakes) but crucially, there were also exceptions. The specialized particle *doch* occurred with a nod. Head turns also associated with polarity: Participants turned away from the addressee less often after a negative antecedent. Overall, co-speech gesture on German REs reflects patterns observed for other spoken languages, and importantly, headshakes/nods seem to differ between German and DGS, indicating different conventionalizations of shakes/nods. The distribution of brow furrowing points to associated negative attitudinal meanings in rejections, which is plausible for this face-threatening act (cp. incredulity, uncertainty [1][2][9]). Raising seems to be associated with positive polarity. In general, gesture is used to disambiguate REs but not in a highly conventionalized extent (no (near-)categorical distribution, no stark differences).

- (1) *Antecedent*: Peter hat die Wette **nicht** gewonnen. ‘Pete has **not** won the bet.’
Response: a. **Ja/nein**, hat er nicht. *ja* = affirmation, *nein* = negative polarity
 b. **Ja/nein**, hat er. *nein* = rejection, *ja* = positive polarity



References

- [1] Brown, L., and P. Prieto (2021). “Gesture and prosody in multimodal communication,” In *The Cambridge Handbook of Sociopragmatics*. 430–453. Cambridge University Press.
- [2] Crespo-Sendra, V., K. Kaland, M. Swerts and P. Prieto (2013). “Perceiving incredulity: The role of intonation and facial gestures,” *Journal of Pragmatics*, 47, 1–13.
- [3] Esipova, M. (2021). “Polar Responses in Russian across Modalities and across Interfaces. *Journal of Slavic Linguistics*, 29, FASL extra issue. <http://ojs.ung.si/index.php/JSL/article/view/151>.
- [4] González-Fuente, S., S. Tubau, M. Espinal and P. Prieto (2015). Is there a universal answering strategy for rejecting negative propositions? Typological evidence on the use of prosody and gesture. *Frontiers in Psychology* 6(899), 1-16.
- [5] Jakobson, R. (1972). “Motor signs for ‘yes’ and ‘no’.” *Language in Society*, 1, 91–96.
- [6] Kendon, A. (2002). “Some uses of the head shake,” *Gesture*, 2, 147–182.
- [7] Li, F., S. González-Fuente, P. Prieto and M.T. Espinal (2016). “Is Mandarin Chinese a truth-based language? Rejecting responses to negative assertions and questions,” *Front. Psych.*, 7, 1967, 1-10.
- [8] Loos, C., Steinbach, M., S. Repp. (accepted). “Polar response strategies across modalities: Evidence from German Sign Language (DGS),” *Language*.
- [9] Żygis, M., N. Sarhaei and M. Krifka (2023) “Oro-facial expressions and acoustic cues in German questions,” *Proc. ICPhS*. 4135-4139.

Postersession 3:

—

27.09.2024
10:20-11:40



Cross-situational learning of word-pseudosign pairs in children and adults: a behavioral and event-related potential study

Arianna Colombani^{1,2,3}, Varghese Peter⁴, Quian Yin Mai⁵, Outi Tuomainen³,
Natalie Boll-Avetisyan³, Amanda Saksida⁶, Mridula Sharma⁵

¹ International Doctorate for Experimental Approaches to Language and Brain (IDEALAB)

² School of Psychological Sciences, Macquarie University, Sydney, Australia

³ Department of Linguistics, University of Potsdam, Potsdam, Germany

⁴ Discipline of Psychology, School of Health, University of the Sunshine Coast, Brisbane, Australia

⁵ Department of Linguistics, Macquarie University, Sydney, Australia

⁶ Institute for Maternal and Child Health-IRCCS "Burlo Garofolo", Trieste, Italy

Corresponding author: arianna.colombani@mq.edu.au

Human communication is innately multimodal, comprising both auditory and visual input. In this integrated system, language develops on a continuum from gestures to speech. In word learning, evidence shows that children can rely on their inherent ability to detect regularities across varying and ambiguous environmental inputs to acquire new words (cross-situational learning [1]). However, this mechanism is rarely studied in the context of language learning in the visual modality, and its neurophysiological underpinnings remain largely unknown. Using behavioral and electroencephalography (EEG) measures, we investigated cross-situational learning of pseudosigns that stood for familiar spoken words to understand whether children and adults could (1) form word-pseudosigns associations and (2) associate pseudosigns with word meaning. We hypothesized that, due to the multimodality of language and despite the ambiguity of the learning context, pseudosigns could be associated with words and understood as gestural labels for their referents. This study included both children and adults to explore developmental differences in this learning process.

In a familiarization phase, 25 children (8–11 y.o) and 19 adults (18–35 y.o) (all English speakers, naïve to the learning nature of the phase), were exposed to 8 word-pseudosigns pairs across 48 trials. Target words were nouns of familiar objects (*bed, dog, car, cold, cup, pink, shirt, toe*) from 8 different semantic categories (*furniture, animals, vehicles, weather, kitchenware, colors, clothes, body parts*). Visual stimuli were non-iconic pseudosigns, created based on sign language phonotactics to ensure visual salience and linguistic relevance while preventing the risk of prior exposure and cultural meaning. Each target word was matched with a pseudosigns to create a pair that remained consistent throughout the experiment. In each trial, two pseudosigns were presented simultaneously on the screen and played side by side with their matched spoken words, played one after the other. To maintain ambiguity, in half of the trials, the first spoken word referred to the video on the left and, the second, to the video on the right. Following this phase, participants were assessed in (1) the learning of word-pseudosigns association (*recognition task*), and (2) their ability to correctly categorize the pseudosigns in appropriate categories (*categorization task*, e.g.: “Is [pseudosigns for *dog*] in the same category as [spoken word *cat*]?”). During the entire experiment, EEG activity was recorded. To detect learning, we investigated the amplitude and latency of the N400 component of the event-related potential (ERP), a neural indicator of lexical/semantic processing [2]. Behavioral responses were analyzed in terms of accuracy (% of correct trials) and d-prime scores, calculated to account for detection sensitivity and a potential response bias through signal detection analysis. For both groups, both measures were above chance in both the *recognition* and *categorization* tasks confirming the learning of pseudosigns forms and their meaning. A linear mixed model analysis revealed an effect of group, with adults performing better than children, but no effect of task nor a task*group interaction (Table 1). Cluster-based permutation test [3] on ERPs showed a significant N400 response followed by a late positive P600-like response in both groups during the *recognition task*. In the *categorization task*, an N400 was found in the adult

group only. However, additional analysis of children's ERPs from correctly identified trials only also revealed an N400 (Table 2, Fig 1).

Overall, our findings suggest that cross-situational learning of pseudosigns is possible, with adults outperforming children in recognition and categorization, likely due to differences in higher cognitive abilities like memory and attention. Importantly, all participants successfully associated pseudosigns with meanings, treating pseudosigns as labels for referents, as evidenced by accurate *recognition* and *categorization*, and supported by the N400 results. These results highlight the multimodality of language, suggesting that sign-like gestures are highly salient linguistic inputs perceived as meaningful communication, likely to be learned implicitly through statistical computations. Future research should test low-level or nonbiological visual stimuli to test whether any movement can be mapped into spoken words or whether this is a phenomenon unique to gestural inputs.

Group	Task	Accuracy (%)			D-prime (d')			Response Bias
		M	SD	Wilcoxon W	M	SD	t-statistics	M
Adults (N = 19)	Recognition	87.5	13.6	190, p < .001	2.76	1.27	9.47, p < .001	0.398
	Categorisation	83.8	16.6	170, p < .001	2.48	1.34	8.04, p < .001	
Children (N = 25)	Recognition	72.0	17.5	314, p < .001	1.47	1.33	5.52, p < .001	0.292
	Categorisation	69.0	18.2	295, p < .001	1.29	1.34	4.79, p < .001	

Table 1. Descriptive statistics of behavioral performance of the two groups. M = mean, SD = standard deviation, and correlation coefficient, respectively. D-prime scores were analyzed using one-sample t-tests to determine if they were significantly different from zero. Accuracy scores were analyzed using the Wilcoxon signed-rank test to determine if they were significantly above chance levels. Bonferroni corrections were applied to p-values to control for multiple comparisons.

Group	Task	Cluster type	Latency	p	Cohen's d
Adults (N=19)	Recognition	Negative	285-497	.001	-1.77
		Positive	625-997	.001	1.62
	Categorisation	Negative	513-673	.005	-1.44
Children (N=25)	Recognition	Negative	156-464	.001	-1.39
		Positive	572-896	.002	1.23
	Categorisation *	Negative	760-876	.043	-1.44

Table 2. Cluster statistics results with effect sizes of the ERPs of the two groups in recognition) and categorization task. * Results of the additional analysis on ERPs from correctly identified trials only.

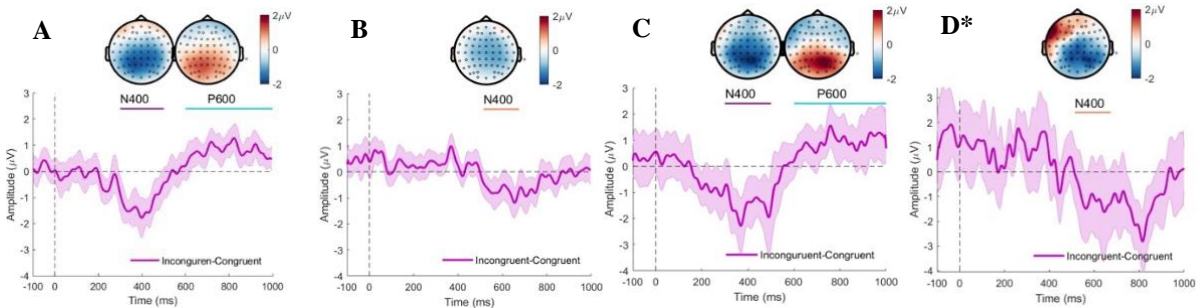


Figure 1. N400 and P600 effects from the difference waveform of congruent minus incongruent trials. The dark lines show grand averaged waveform, and the shading encompasses 95% confidence intervals. A and B: results of the adult group in recognition and categorization task, respectively. C and D: same tasks in children. * Results of the additional analysis on ERPs from correctly identified trials only.

References

[1] Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics: Research article. *Psychological Science*, 18(5), 414–420. <https://doi.org/10.1111/j.1467-9280.2007.01915.x>.

[2] Friederici, A. D. (2004). Event-related brain potential studies in language. In *Current Neurology and Neuroscience Reports* (Vol. 4, Issue 6, pp. 466–470). Curr Neurol Neurosci Rep. <https://doi.org/10.1007/s11910-004-0070-0>

[3] Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of neuroscience methods*, 164(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>

Participation of deaf children with a cochlear implant in family dinner interactions: the role of gesture

Stéphanie CAËT¹², Loulou KOSMALA³⁴, Carla FERRAN¹, Marine LAVAL¹

Université de Lille¹

UMR 8163 Savoirs, Textes, Langage²

Université Paris Nanterre³

EA 370 CREA⁴

stephanie.caet@univ-lille.fr

Children acquire language thanks to the language their caregivers address them [1], from taking an active part in dialogues as speakers [2] and from listening and observing others in conversation [3]. Deaf children who were born in hearing families, even when they received a cochlear implant, may encounter difficulties in participating in interactions [4] and may thus lack language learning [5] as well as socialisation opportunities [6], which may further impact their well-being [7].

In this study, we investigate the extent to which deaf children who received a cochlear implant participate in the first and ideal locus of language socialisation [8]: family dinners. More specifically, we question the role of multimodality on deaf children's occupation of different statuses in those interactions: speakers, addressed and non-addressed participants.

Five French-speaking families composed of 2 hearing parents, 1 deaf implanted child and 1 older sibling were video-recorded while having dinner. We used the ELAN program [9] to systematically annotate the language productions of each family member as being vocal, gestural (manual as well as non manual gestures including facial expressions) or multimodal, and who their addressee was. Focusing on the deaf child, we coded whether the child was visually attending the speaker (in the case of directed as well as non-directed language) or the speaker's addressee (in the case of non-directed language).

Our preliminary results suggest that, in our data, implanted deaf children occupy most of the discursive space as speakers. They use gestures more than their hearing sibling, whether they acquired signs from a sign language or not. Gestures or signs enable them to initiate an exchange or semantically complement their vocal productions. Our data also shows that these children tend to be the preferred addressees. Parents use more gestures when addressing their deaf child than his/her hearing sibling. As addressees, deaf children orient their attention to gestures more when joint attention has already been established. As non-addressed participants, younger implanted deaf children tend not to look at the main speaker nor at the addressee; however, as they become older and more proficient in managing their gaze in this complex multiparty and multi-activity setting, implanted deaf children look more at the speaker or addressee, especially when speakers produce gestural or multimodal utterances.

These results are compared with data coming from signing and speaking families recorded in the same situation, following the same template for analysis. Qualitative analyses of selected sequences underline the potential impact for implanted deaf children not to visually access non-addressed language, as well as the way deaf parents finely scaffold their deaf children's visual attention. Implications for hearing parents and professionals on the use of gestures in interaction with (implanted) deaf children are discussed.

References

- [1] Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions : Interactive, linguistic, and conceptual. *Journal of Child Language*, 47(1), 5-21.
<https://doi.org/10.1017/S0305000919000655>

- [2] Salazar Orvig, A. (2019). Approches théoriques actuelles de l'acquisition du langage. In S. Kern (Éd.), *Le développement du langage chez le jeune enfant* (p. 13-51). De Boeck Supérieur. <https://hal.archives-ouvertes.fr/hal-03562359>
- [3] Fitch, A., Lieberman, A. M., Luyster, R. J., & Arunachalam, S. (2020). Toddlers' word learning through overhearing : Others' attention matters. *Journal of Experimental Child Psychology*, 193, 104793. <https://doi.org/10.1016/j.jecp.2019.104793>
- [4] Crowe, K., & Dammeyer, J. (2021). A Review of the Conversational Pragmatic Skills of Children With Cochlear Implants. *Journal of Deaf Studies and Deaf Education*, 26(2), 171-186. <https://doi.org/10.1093/deafed/enab001>
- [5] Cheng, Q., Roth, A., Halgren, E., & Mayberry, R. I. (2019). Effects of Early Language Deprivation on Brain Connectivity : Language Pathways in Deaf Native and Late First-Language Learners of American Sign Language. *Frontiers in Human Neuroscience*, 13, 320. <https://doi.org/10.3389/fnhum.2019.00320>
- [6] Meek, D. R. (2020). Dinner table syndrome : A phenomenological study of deaf individuals' experiences with inaccessible communication. *The Qualitative Report*, 6(25), 1676-1694.
- [7] Hall, W. C., Smith, S. R., Sutter, E. J., DeWindt, L. A., & Dye, T. D. V. (2018). Considering parental hearing status as a social determinant of deaf population health : Insights from experiences of the « dinner table syndrome ». *PLOS ONE*, 13(9), e0202169. <https://doi.org/10.1371/journal.pone.0202169>
- [8] Blum-Kulka, S. (1997). *Dinner talk : Cultural patterns of sociability and socialization in family discourse*. Lawrence Erlbaum.
- [9] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

Using social robots for cross-cultural gesture elicitation in children: Psycholinguistic considerations on dialogue design

Katharina J. Rohlfing¹, Nils Tolksdorf¹, Angela Grimminger¹, Koki Honda², Kazuki Sekine²
Paderborn University, Germany,¹ Waseda University, Japan²
 katharina.rohlfing@upb.de

Can social robots contribute to the investigation of children's gestural behavior? And if so, do they have a similar effect across cultures? In this work, we propose social robots as a useful methodological tool for engaging children in collaborative social interactions and eliciting interactional behavior from them. Using robots as a tool has the advantage of controlling their behavior and keeping the interactional properties systematically constant across participants, thus benefiting the generalizability of the comparison. Prior research has shown that children not only socially conform to social robots [1], but also view them as trustworthy interaction partners [2]. Importantly, due to their embodied nature, social robots are able to communicate multimodally and enrich an interaction by using different communicative signals (e.g., gaze or gesture). Additionally, children seem to attend to a robot's social cues in the same way as they do with a human partner when engaged in a collaborative task [3]. Thus, we argue that there is considerable potential for social robots to be used as effective tools for investigating human behavior across cultural contexts. Despite a substantial body of work on crosscultural variation in gestures indicating distinct differences in emblematic gestures or that representational gestures expressing spatial concepts may be culturally specific, clear findings on children's culturally specific use of gestures in interaction remain scarce (e.g., [4]). Below, we specify our methodological approach to developing a dialogue design for a child-robot interaction with preschoolers to elicit their gestures across different cultures: Germany and Japan.

We based our design on the intersection of the task that is requested as well as the partner's involvement in the task and tested whether it elicited gestures from children in both cultural environments. Adopting an interactionist perspective on task design [5], our rationale for using social robots hinges on the assumption that they allow for the establishment of a sequential structure of a task-oriented interaction. This sequential structure facilitates consistent replication across participants, advantageous in cross-cultural studies of children's gestures, where numerous contextual elements additionally influence human behavior. We designed the overall task so that each trial elicited a repetition of an interactional sequence in which both partners clearly contributed to the overall goal (see Fig. 1). The goal of the overall task was to learn how to perform everyday actions, building on findings that actions elicit gestures in children [6]. Each trial consisted of an action, such as constructing a paper airplane. With respect to the partner's involvement, employing robots as social agents in interpersonal encounters makes it possible to precisely tailor robots' behaviors to specific social roles [7]. The robot's role in the interaction was that of a learner that was instructed by the child. In our pilot study with Japanese ($N = 4$) and German ($N = 5$) children (mean age 5.14, $SD = 0.5$), we investigated whether (a) children recognized the task and (b) they engaged in the task with gestural behavior that was expected within a particular "slot" in the interactional sequence. Results showed that children successfully engaged in the task. Crucially, by presenting the same interactional context, the developed dialogue design effectively elicited gestural behavior in children from both cultural environments. In this light, this approach achieved a high degree of experimental control, a notable benefit in research settings traditionally challenged by the need for controlled but ecologically valid methods. While these results are preliminary and will be further enriched by a detailed comparative analysis of children's gestures, this pilot study has demonstrated a promising methodological approach in exploring cultural variations experimentally. Building on these initial findings, we discuss the benefits and limitations associated with the use of robots in the cross-cultural study of children's gestural development.

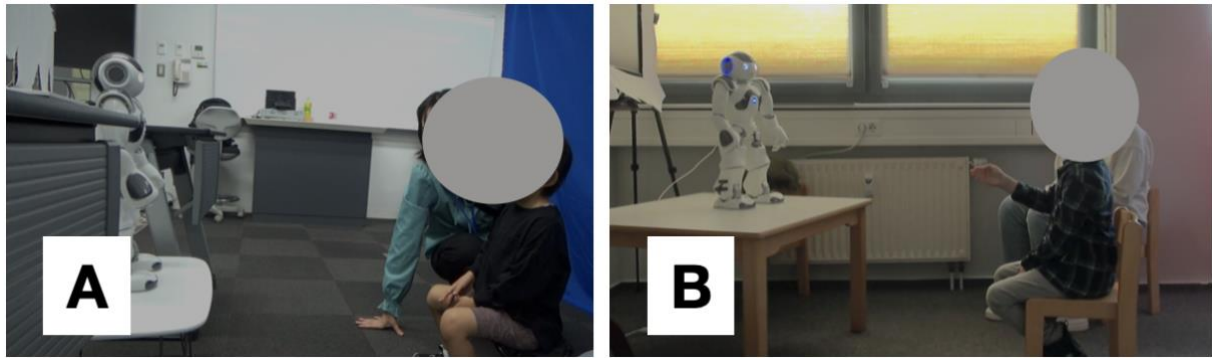


Figure 1: *The design developed to elicit the child's gestural communicative behavior across cultures. Example of the experimental setting in (A) Japan and (B) Germany.*

References

- [1] A.-L. Vollmer, R. Read, D. Trippas, and T. Belpaeme, "Children conform, adults resist: A robot group induced peer pressure on normative social conformity," *Science Robotics*, vol. 3, no. 21, p. eaat7111, 2018. doi: 10.1126/scirobotics.aat7111.
- [2] C. Oranç and A. C. Küntay, "Children's perception of social robots as a source of information across different domains of knowledge," *Cognitive Development*, vol. 54, p. 100875, 2020. doi: 10.1016/j.cogdev.2020.100875.
- [3] J. M. Kory Westlund, L. Dickens, S. Jeong, P. L. Harris, D. DeSteno, and C. L. Breazeal, "Children use non-verbal cues to learn new words from robots as well as people," *International Journal of Child-Computer Interaction*, vol. 13, pp. 1–9, 2017. doi: 10.1016/j.ijcci.2017.04.001.
- [4] A. Cattani, C. Floccia, E. Kidd, P. Pettenati, D. Onofrio, and V. Volterra, "Gestures and words in naming: Evidence from crosslinguistic and crosscultural comparison," *Language Learning*, vol. 69, no. 3, pp. 709–746, 2019. doi: 10.1111/lang.12346.
- [5] K. J. Rohlfing, B. Wrede, A.-L. Vollmer, and P.-Y. Oudeyer, "An alternative to mapping a word onto a concept in language acquisition: Pragmatic frames," *Front. Psychol.*, vol. 7, no. 470, pp. 1–18, 2016. doi: 10.3389/fpsyg.2016.00470.
- [6] E. L. Congdon, M. A. Novack, and E. M. Wakefield, "Exploring individual differences: A case for measuring children's spontaneous gesture production as a predictor of learning from gesture instruction," *Topics in Cognitive Science*, p. tops.12722, 2024. doi: 10.1111/tops.12722.
- [7] K. J. Rohlfing et al., "Social/dialogical roles of social robots in supporting children's learning of language and literacy—A review and analysis of innovative roles," *Front. Robot. AI*, vol. 9, no. 971749, pp. 1–15, 2022. doi: 10.3389/frobt.2022.971749.

Exploring gesture distribution over disfluency markers in competent speakers and language learners

Maria Graziano¹, Joost van de Weijer¹, Marianne Gullberg^{1,2}

¹ *Lund University Humanities Lab, Centre for Languages and Literature, Lund University*

Corresponding author: maria.graziano@humlab.lu.se

When speakers experience expressive difficulties while speaking, they produce a variety of disfluency markers (e.g., silent pauses, filled pauses, repetitions, lengthenings). It has been suggested that different disfluency markers have different functions and signal different production problems (e.g., lexical access, planning or formulating troubles) (e.g., [1]). Previous research has shown that speakers only rarely gesture during disfluencies and more interestingly that these few gestures are not only referential (i.e., gestures convey information about referents' size, shape, movement or location) but also pragmatic (i.e., gestures emphasising parts of the speaker's discourse, expressing speech acts, indicating speaker's stance towards his discourse) [2, 3, 4], indicating that speakers do not only use gestures to alleviate lexical difficulties by representing the sought word gesturally, but also use them to comment on the breakdown itself (cf. [5]). It remains unknown how the function of gestures (i.e., referential or pragmatic, [6]) produced during disfluencies is distributed, whether the distribution depends on the type of disfluency marker, and whether such functions and distribution can vary depending on the language and the language competence (competent speakers vs. learners). If gestures and speech form an integrated system, we hypothesise that this is the case. The present study aimed to explore this issue.

Extending previous work [2], analyses were conducted on narrative retellings produced in dyadic, interactive settings by adult Italian ($n=11$) and adult Dutch speakers ($n=11$); Italian children aged 4-5 ($n=11$), 6-7 ($n=11$), and 8-10 years ($n=11$), and by adult Dutch learners of French as a second language ($n=11$) at low to intermediate levels of proficiency. All disfluencies in the narratives were identified and classified as filled or unfilled pauses, interruptions, or lengthenings. Further, the function of the few gestures that occurred during disfluencies was coded as either referential or pragmatic.

Our observations indicate that, descriptively, the adult Italian speakers produced many lengthenings, that were accompanied mainly by pragmatic gestures. In contrast, the adult Dutch speakers produced more filled pauses that were accompanied by referential gestures, and unfilled pauses accompanied by pragmatic gestures. The Italian children produced many lengthenings (a trend especially clear in 9-year-olds who are similar to Italian adults) that were accompanied by referential gestures (unlike the Italian adults). The adult second language speakers of French, instead, mainly produced filled and unfilled pauses accompanied by both referential and pragmatic gestures.

Overall, the data point to differences in the use of disfluency markers used by adult speakers of different languages (Italians preferring lengthenings, and the Dutch preferring filled and unfilled pauses), but also between children and adult second language learners (with Italian children preferring lengthenings, and adult Dutch learners of French preferring filled and unfilled pauses). As for the gesture distribution relative to the disfluency markers, the data do not show a clear pattern: while the Italian adults accompanied lengthenings mainly by pragmatic gestures, the Italian children predominantly used referential gestures. The Dutch speakers preferred filled and unfilled pauses both in their first and second language, but their gesture distribution only showed a clear pattern in their second language, where they accompanied unfilled pauses mainly by pragmatic gestures. We discuss the implications of these preliminary findings which suggest crosslinguistic differences in the preference for vocal disfluency markers, and developmental patterns in the distribution of gestural functions during disfluent speech.

References

- [1] Clark, H. H., and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition* 84, 73–111.
- [2] Graziano, M. & Gullberg, M. (2018). When speech stops, gesture stops: evidence from crosslinguistic and developmental comparisons. *Frontiers in Psychology*, 9: 879.
- [3] Gullberg, M. (2006). Handling discourse: gestures, reference tracking, and communication strategies in early L2. *Lang. Learn.* 56, 155–196.
- [4] Gullberg, M. (1998). *Gesture as a Communication Strategy in Second Language Discourse: A Study of Learners of French and Swedish*. Lund: Lund University Press.
- [5] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: Chicago University Press.
- [6] Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Multimodal Behavior in Native Hearing PJM-Polish Bilinguals Using Spoken Polish

Joanna Wójcicka¹, Anna Kuder², Justyna Kotowicz³

¹*Department of General Linguistics, Sign Language Linguistics and Baltic Studies, University of Warsaw, Poland;* ²*Department of Linguistics, General Linguistics, University of Cologne, Cologne, Germany;* ³*Institute of Pedagogy, University of Silesia Katowice, Poland*
akuder@uni-koeln.de

This paper presents an analysis of the multimodal behavior among bimodal native hearing users of spoken Polish and Polish Sign Language (PJM, *polski język migowy*) during the production of language tasks performed in spoken Polish. The aims of the study were: (1) to investigate the use of manual gestures, code-switching and code-blending occurring in the participant's productions and (2) to examine the sign language suppression present in the participants' productions.

Data sample for (1) included two types of elicitation tasks. The first task was to re-tell a video clip from the cartoon 'Tweety and Sylvester' (*Canary row*, 1950), the second one was to provide a route description based on the shown map [2]. The informants were video-recorded, while a moderator fluent in PJM and Polish was present in the room. We examined linguistic material from 7 informants (6 F, 1 M; age M=33,5, SD=11,3), who use both languages in their everyday communication. In the dataset we identified all instances of meaningful hand movements. Each instance was assigned to one of the following categories: non-referential gestures, referential gestures, and signs. By 'non-referential gestures' we mean those gestures which do not show a clear link with the semantic content of concurrently produced speech. They include beat gestures, speech-act gestures, and markers of discourse organization and interaction. In contrast, 'referential gestures', which can be iconic, metaphoric, or deictic, do exhibit a direct link to the content of speech [3]. By 'signs' we mean lexicalized PJM signs. Data sample for (2) included answers from a) self-assessment questionnaires on language mixing practices and b) language background questionnaires from all participants.

As a result of the annotation process, we have identified 138 manual activities: 92 were categorized as non-referential gestures, 46 – as referential gestures. The distribution of manual activities performed by each participant is depicted in Fig. 1. None of the manual activities in the data were interpreted as PJM signs, which means that no instances of code-mixing, code-switching or code-blending were found.

The obtained results suggest that the participants do not use PJM during spoken language monolingual tasks. This suppression of sign language was observed even though the tasks were managed in a signed-spoken bilingual environment. This absence does not align with the results reported in the previous literature, e.g. by Emmorey et al. [4], who show that American hearing bimodal bilinguals frequently code-blend and code-switch during their communication. While such multimodal behaviors are known to appear in Polish hearing native bilinguals' face-to-face communication, they were not present in the analyzed setup.

As there is evidence that mixing of signed and spoken language is generally seen as inappropriate in the Polish Deaf community [5], we speculate that participants' language suppression might stem from their internalized language ideologies. This is supported by the observation that their behavior is in line with their self-assessment overtly expressed in the questionnaire answers. However, more research is needed to explore this hypothesis in depth.

We then tentatively explain the participants' lack of language mixing as good language control based on a more general enhanced cognitive control mechanism [6]. We also address some of the limitations of the study, i.e. the influence of the elicitation task, the participant's professional background etc.

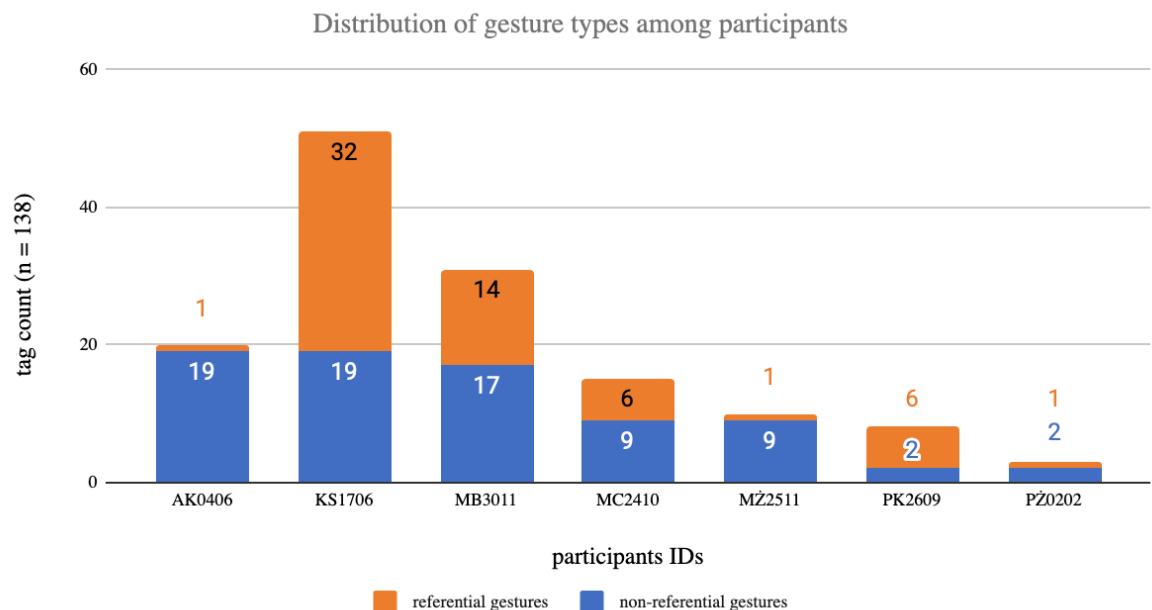


Figure 1: *Distribution of gesture types among participants*

References

- [1] F. M. Branzi, P. A. Della Rosa, M. Canini, A. Costa and J. Abutalebi, “Language control in bilinguals: monitoring and response selection,” *Cereb Cortex*, vol. 26, no. 6, pp. 2367–80, 2016. doi: 10.1093/cercor/bhv052.
- [2] S. Matthes, T. Hanke, A. Regen, J. Storz, S. Wörseck, E. Efthimiou, A. L. Dimou, A. Braffort, J. Glauert and E. Safar, “Dicta-Sign – building a multilingual sign language corpus”, 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, pp. 117–122, Istanbul, 2012.
- [3] P. Rohrer, I. Vilà-Giménez, J. Florit-Pons, N. Esteve-Gibert, A. Ren, S. Shattuck-Hufnagel and P. Prieto, “The MultiModal MultiDimensional (M3D) labeling scheme for the annotation of audiovisual corpora,” *Proceedings of the 7th Gesture and Speech in Interaction (GESPIN)*, Stockholm, 2020.
- [4] K. Emmorey, H. Borinstein and R. Thompson, “Bimodal Bilingualism: Code-blending between Spoken English and American Sign Language,” *Proceedings of the 4th International Symposium on Bilingualism*, Somerville MA, 2005.
- [5] P. Tomaszewski, T. Gałkowski and P. Rosik, „Nauczanie polskiego języka migowego jako obcego języka: Czy osoby słyszące mogą przyswoić język wizualny?,” *Studia nad kompetencją językową i komunikacją niesłyszących*, Warszawa, 2003.
- [6] M. R. Giezen, H. K. Blumenfeld, A. Shook, V. Marian and K. Emmorey, „Parallel language activation and inhibitory control in bimodal bilinguals,” *Cognition*, vol. 141, pp. 9–25, 2015. doi: 10.1016/j.cognition.2015.04.009.

Mocking enactments: a case-study of multimodal stance-stacking

Fien Andries, Katharina Meissl, Clarissa de Vries

KU Leuven, Belgium

fien.andries@kuleuven.be

In interaction, people frequently take a stance: they express an evaluation of a stance object, and employ a wide variety of semiotic resources in doing so. Stance-taking as a multimodal phenomenon has gained interest over the past years [1]. One frequently used resource for the expression of stance is enactment, i.e. the use of “bodily movements, postures and eye gaze to ‘construct’ actions and dialogue in order to ‘show’ characters, events and points of view” [2, p. 373]. Enactments allow interactants to simultaneously express a “representation of linguistic actions” and, on the other hand, “commentaries about these actions” [4, p. 161]. The result is a construction of stacked stance acts [3], comprising the reported stance of a character as well as the stance from the interactant as a narrator.

Given this inherent layering, enactments constitute a particularly convenient resource for the expression of mockery. During mocking, participants express a stance on a serious layer that can be heightened, diminished, or inverted on a non-serious layer. Consider the following example from one of our data sets (Figure 1): a participant is telling a story about a time she spent a weekend with a group of scouts in a cabin in the forest, that could not be locked. She enacts the cabin landlady, who seemed to be indifferent about this issue. While using multiple shoulder shrugs, palm up open hand gestures, and head shakes, she says ‘yeah gosh, throughout the years, all those keys got lost’, thus stacking her own mocking stance on top of the landlady’s reported stance.

In the current study, also reported in [4], we investigate mocking enactments such as the one above, as a case study of stance-stacking in four different languages and three interactional settings, using the following datasets: Music instructions in Dutch, German and English [5]; Spontaneous face-to-face interactions among friends in Dutch [6], [7]; and narrations on past events in Flemish Sign Language (VGT) [8]. The aim of the study is twofold: 1) Exploring the use of enactments for mocking, and 2) Mapping out the multimodal construction and unfolding of mocking enactments. We take a holistic approach, taking into account all semiotic resources that become relevant for this construction within sequences of mocking enactments.

Regarding the multimodal construction of stacked stances, we found that mocking enactments are sequentially embedded in highly evaluative contexts, marked by the use of a variety of bodily-visual resources. Within mocking enactments, these resources can serve multiple functions, both constructing the enactment as well as contributing to the mocking character of the enactment sequence. We will present various examples, illustrating that mocking enactments do not only comprise exaggerations and stylized caricatures, but may as well create contrast with an expectation by evoking an absurd scenario. Furthermore, we found that mocking enactments go beyond enactments of the target of the mockery. Interactants include other characters and viewpoints in their depicted scenarios, so that the target of the mockery, the stance object and the enacted character do not necessarily overlap. As such, this study highlights the variety and complexity of the multimodal design of mocking enactments. More generally, it puts up for discussion the relation between exaggeration and the combined use of different semiotic resources.

Figures



Figure 1: Multimodal transcription and stills of example from our dataset

References

- [1] F. Andries *et al.*, “Multimodal stance-taking in interaction—A systematic literature review,” *Front. Commun.*, vol. 8, 2023, doi: <https://doi.org/10.3389/fcomm.2023.1187977>.
- [2] G. Hodge and L. Ferrara, “Showing the story : Enactment as performance in Auslan narratives,” in *Selected Papers From the 44th Conference of the Australian Linguistic Society (2014)*, 2014, pp. 372–397. Accessed: Mar. 31, 2022. [Online]. Available: <https://www.semanticscholar.org/paper/Showing-the-story-%3A-Enactment-as-performance-in-Gawne-Vaughan/cc5fae968168cca1d33f146364cddf40dbbcbddddd>
- [3] B. Dancygier, “Negation, stance verbs, and intersubjectivity,” in *Viewpoint in Language: A Multimodal Perspective*, B. Dancygier and E. Sweetser, Eds., Cambridge: Cambridge University Press, 2012, pp. 69–93.
- [4] C. de Vries, F. Andries, and K. Meissl, “Mocking enactments: a case-study of multimodal stance-stacking,” *Front. Psychol.*, vol. 15, 2024, doi: 10.3389/fpsyg.2024.1379593.
- [5] S. Schrooten and K. Feyaerts, “Conducting Fanfare Orchestras. A multimodal corpus.” KU Leuven MIDI, 2020.
- [6] C. de Vries, B. Oben, and G. Brône, “The coffee bar corpus: spontaneous triadic interactions between friends.” Leuven, Belgium.
- [7] G. Brône and B. Oben, “InSight Interaction: a multimodal and multifocal dialogue corpus,” *Lang. Resour. Eval.*, vol. 49, no. 1, pp. 195–214, Mar. 2015, doi: 10.1007/s10579-014-9283-2.
- [8] M. Van Herreweghe, M. Vermeerbergen, E. Demey, H. De Durpel, H. Nyffels, and S. Verstraete, “Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent ism KU Leuven. <www.corpusvgt.be>,” 2015, Accessed: Dec. 02, 2020. [Online]. Available: <http://hdl.handle.net/1854/LU-6973686>

Referential gestures and the management of turn-taking in conversation

Margaret Zellers¹, Jan Gorisch² and David House³

¹Kiel University, ²Leibniz-Institut für Deutsche Sprache, ³Kungliga Tekniska Högskolan
mzellers@isfas.uni-kiel.de

McNeill [1] introduced four categories to describe gesture function: *iconic*, *metaphoric*, *deictic* and *beat* gestures, though a growing body of behavioral and cognitive evidence demonstrates that there is not a clear divide between, e.g., beat gestures and metaphoric gestures [1], [2]. One aspect of this functional classification is to attempt to evaluate whether a gesture stroke refers to something (including metaphoric meaning) or if it does not; thus we can speak of “referential” and “non-referential” gestures.

The issue of gesture function remains unclear, and thus there is a need to observe real data, i.e. conversational interaction, and try to annotate gestures and to identify their verbal referents exhaustively. It is also unknown whether and to what degree referential and non-referential gestures are employed in the management of conversation. Since turn-taking principles such as minimizing gaps and overlaps are known to be universal [cf. 3], it is relevant to evaluate whether patterns of gestural marking of turn-taking also hold for more than one language. Thus, the current study attempts to identify which role the implementation of referential versus non-referential gesture may play in the management of turn-taking in conversation.

The data used in the current study are drawn from two multimodal corpora of conversational speech. The Swedish data come from the Spontal corpus [4], while the German data are taken from FOLK (Research and Teaching Corpus of Spoken German) [5]. Gesture and turn annotations were carried out using ELAN [6]. Spoken features were annotated in Praat [7]. We segmented the gesture phases *preparation*, *stroke*, *hold*, *retraction*, following [8]. For each gesture stroke, the annotator listened to the audio in the vicinity of the gesture to identify a lexical referent. We used a broad definition of referentiality which meant that our annotations were biased towards referentiality; however, there were still a substantial number of tokens for which no lexical referent could be identified.

Referential strokes were found to be very rare in the vicinity of turn ends compared to other locations in conversation; thus, we investigated whether there was a difference of distribution of referential gestures in the vicinity of turn ends versus at other locations in conversation. Assuming that referentiality features are “carried over” from the stroke to other phases of the gesture, we searched backwards starting from any gesture unit ongoing at the offset of speech through all gesture phases until a stroke was reached (to a maximum distance of 500ms, see Figure 1) to classify the whole gesture unit as referential or non-referential. The resulting distribution of referential and non-referential gestures and their location is shown in Table 1. For both languages, the distribution of referential and non-referential gestures was different at turn ends versus not at turn ends; the cross-linguistic comparison did not have a significant result; see full model results in Table 2. In German, non-referential strokes occur at a rate of 1.49/minute in the overall data, but a rate of 4.74/minute preceding turn ends. In Swedish, non-referential strokes occur at a rate of 2.42/minute in the overall data, but a rate of 3.26/minute preceding turn ends.

Thus we see that non-referential gestures become more frequent near turn ends. Since these gestures can still have pragmatic functions, it is unsurprising to find them arising where speaker transition becomes relevant. The differences in distribution support the argument that some kind of distinction between referential and non-referential gestures has validity, though more research is needed to clarify the functional differences that arise.

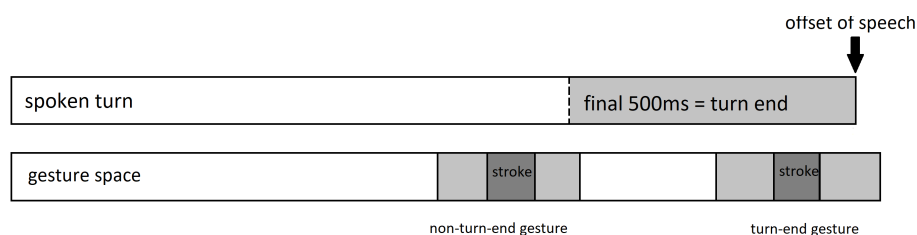


Figure 1: Schematic diagram of gesture location classification.

Table 1: Referentiality distribution of gestures in turn-end or non-turn-end locations; German: DE, Swedish: SW.

	Non-referential	Referential
Turn end	36 (DE) 31 (SW)	209 (DE) 74 (SW)
Not turn end	46 (DE) 44 (SW)	912 (DE) 308 (SW)

Table 2: Results of GLMM. The estimate reflects the log-likelihood increase in non-referential gestures compared to referential gestures when moving from non-turn-end to turn-end position. Since no significant interaction was found, a model with only main effects is reported here. Conditional R^2 (delta method) = 0.119.

	Est.	SE	z value	p value
(Intercept)	0.43	0.55	0.79	0.432
locationEnd	1.13	0.19	6.10	0.000
languageSW	-0.47	0.54	-0.88	0.378

glmer(referentiality~location+language+(1|speaker), family=binomial)

References

- [1] D. McNeill, “Gesture and communication,” in *The Encyclopedia of Language and Linguistics*, ser. Psycholinguistics, K. Brown and A. Anderson, Eds., 2nd ed., Amsterdam and Boston: Elsevier, 2006, pp. 58–66.
- [2] D. Casasanto, “When is a linguistic metaphor a conceptual metaphor,” *New Directions in Cognitive Linguistics*, vol. 24, pp. 127–146, 2009.
- [3] T. Stivers, N. J. Enfield, P. Brown, *et al.*, “Universals and cultural variation in turn-taking in conversation,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 26, pp. 10 587–10 592, 2009.
- [4] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, “Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture,” in *Proceedings of LREC 2010*, ELRA, 2010.
- [5] T. Schmidt, “The Research and Teaching Corpus of Spoken German – FOLK,” in *Proceedings of LREC 2014*, ELRA, 2014.
- [6] Max Planck Institute for Psycholinguistics, *ELAN (version 5.2)*, <https://tla.mpi.nl/tools/tla-tools/elan/>, Nijmegen, Netherlands, 2018.
- [7] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, praat.org, 2021.
- [8] A. Kendon, *Gesture: Visible action as utterance*. Cambridge, UK: CUP, 2004.

Multimodal profiles of different (negative) question types

Johannes Heim¹, Rebecca Woods², Franziska Busche³ & Sophie Repp³
University of Aberdeen¹, Newcastle University², University of Cologne³
 johannes.heim@abdn.ac.uk

There are many ways to ask a question. Some have been argued to be more marked than others because they are formally more complex (e.g., they contain negation) or because they are more prone for misinterpretation (e.g., due to a syntax-illocution mismatch). Marked forms come with special conventionalized discourse effects [1]. One such effect is the expression of a bias, i.e. a meaning component regarding previous assumptions or future expectations. In this talk we focus on the multimodal forms of different negative questions, which typically serve to double-check a proposition based on contextual evidence or a speaker's original belief resulting in a bias [2]. The examples in (1) illustrate variation in illocutionary meaning of a positive question (PQ, 1a) and three types of negative questions in the same context. The negative questions differ in the extent to which they can be understood as an information-seeking question, a suggestive question or a straightforward suggestion. While the first reading in principle but not preferred is available for negative polar questions (NegQ, 1c) and *Why-don't-you* questions (WhyQ, 1d), the second reading is easily available for NegQs and negative tag questions (NegQ, 1b), while the third reading is preferred for WhyQs (as a frozen expression).

An interesting issue arising here is whether the different question types, or the range of illocutionary meanings they may have, are characterized by different patterns of co-speech gesture and prosody that guide the addressee in arriving at the intended illocutionary interpretation. We know for PQs in both spoken and signed languages, that they frequently occur with raised manual and facial gestures [3,4] mirroring the often-upward trajectory of question intonation in spoken languages [5]. For negative questions, research is still in its infancy [6,7].

We present a corpus study using data from four actors from the American soap opera *Bold and Beautiful* (years 2012-2020). Actors aim to reflect recognisable conventions for expressing different meanings. This semi-natural format also yields sufficient instances of different question types for different speakers without posing data protection and privacy problems. We selected 20 questions per type (1a-d) and per actor. Of the 320 selected questions, only 84 had visible hands but all full head visibility. TagQs and NegQs had the largest numbers of multiple gesture events per utterance. We annotated all questions using the M3D guidelines [8] with additions for beatlikeness and brow movements. Our findings show that the different negative questions exhibit different gestural profiles, differing also from those of PQs. Regarding hand gestures, all negative questions are more beatlike than positive questions (Fig. 1). WhyQs and NegQs have more open hand shapes than TagQs and WhyQs (Fig. 2), but these open hand shapes only sometimes occur with palms up for WhyQs (9%). Instead, hand trajectories (not pictured) confirm that WhyQ are often accompanied by beat gestures. TagQs show the least number of manual gestures, and NegQs have a similar profile to PosQs (with said exception of the open hand shape). None of the questions have a distinct brow movement (raised or furrowed), but TagQs stand out by how often brows are relaxed (55%). Furthermore, TagQs typically have a similar gestural profile across anchor and tag, lending little support to conceptualizing them as hybrid speech acts consisting of assertion and question [9]. For head movements, NegQs and TagQs are the most distinct question types (Fig. 4): NegQs frequently have a sideward head turn (56%) while TagQs often come with functional nodding (50%). These findings suggest that the negative questions considered here do not form a natural class with a uniform gestural encoding. Some observations, such as the beatlike, vertical, open-handed gesture for WhyQs and NegQs or the asymmetry of head movements between WhyQs and NegQs are promising leads for a more detailed analysis. Prosodic measures included at the time of presenting will enrich the multimodal profiles of the different questions types.

Examples and Figures

(1) *Mary attends a party at Peters house. After a fun evening, Mary begins to worry about getting home. The last bus just left; taxis are unreliable. Peter says to Mary:*

- a. Do you want to stay? (Positive Polarity Question; PosQ)
- b. You want to stay, don't you? (Negative Tag questions; TagQ)
- c. Don't you want to stay? (Negative Polarity Question; NegQ)
- d. Why don't you stay? (Why-Don't-You Question; WhyQ)

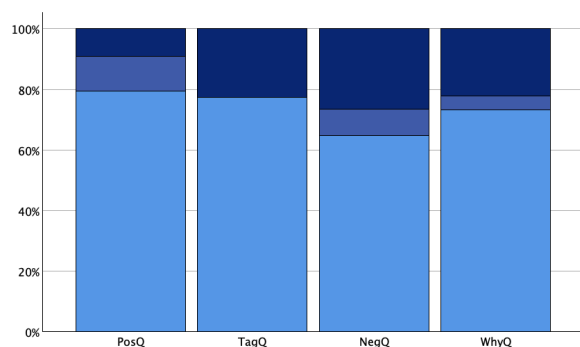


Figure 1: *Beatlikeness* (bottom to top: not, somewhat, very) by question type across all gestures.

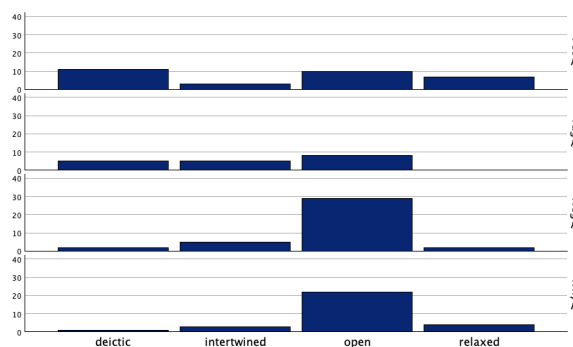


Figure 2: *Handshapes* ($n > 5$) by question type for dominant hand.

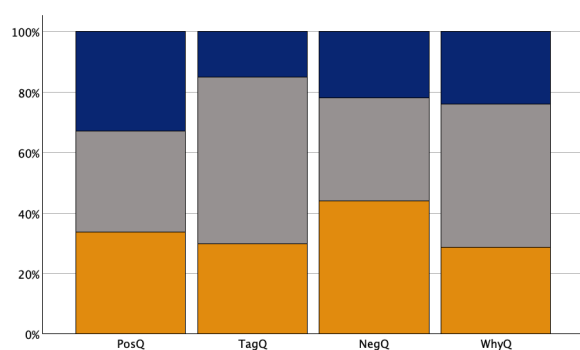


Figure 3: *Brow movement* (bottom to top: furrowed, relaxed, raised) by question type across all gestures.

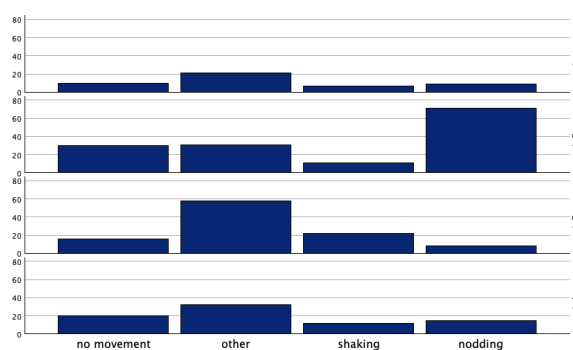


Figure 4: *Headshakes* by question type.

References

- [1] D. F. Farkas and F. Roelofsen. "Division of labor in the interpretation of declaratives and interrogatives," *Journal of Semantics* 34(2), 2017, pp. 237–289.
- [2] M. Romero. "Form and function of negative, tag, and rhetorical questions," In V. Dépraz & M. T. Espinal (Eds.), *The Oxford Handbook of Negation*, pp. 235–254. OUP.
- [3] U. Zeshan. "Interrogative constructions in signed languages: Crosslinguistic perspectives," *Language*, 80(1), 2004, pp. 7–39.
- [4] E. Krahmer and M. Swerts. "How children and adults produce and perceive uncertainty in audiovisual speech," *Language and Speech*, 48(Pt 1), 2005, pp. 29–53.
- [5] R. Ultan. "Some general characteristics of interrogative systems," in J. Greenberg (ed.), *Universals of Human Language*, 1987, pp. 83–124.
- [6] M. Ippolito. "The contribution of gestures to the semantics of non-canonical questions," *Journal of Semantics*, 38(3), 2021, pp. 363–392.
- [7] M. Oomen and F. Roelofsen. "Biased polar questions in Sign Language of the Netherlands-Methods description," 2002, <https://eprints.illc.uva.nl/id/eprint/2249/1/MoL-2023-08.text.pdf>.
- [8] P. Rohrer et al. "The MultiModal MultiDimensional (M3D) labelling system," 2023, doi: 10.17605/OSF.IO/ANKDX.
- [9] B. J. Reese. *Bias in questions*. PhD thesis. The University of Texas at Austin. 2007.

Embodied pronunciation training for the Swedish complementary length contrast

Federica Raschellà¹, Frida Splendido², Nadja Althaus³, Marieke Hoetjes⁴,
Gilbert Ambrazaitis¹

¹Linnaeus University, Växjö, Sweden, ²Lund University, Sweden, ³University of East Anglia, Norwich, UK, ⁴Radboud University, Nijmegen, The Netherlands
federica.raschella@lnu.se

Research has shown that gestures can have beneficial effects on second language (L2) pronunciation learning. However, different studies have investigated different gestures and usually either considered effects on learner's production [1][2][3][4] or perception [5][6] but hardly both [7], delivering mixed results. Moreover, previous studies have focused on a small number of L2s such as English [2], Spanish [1][3], French [4] Japanese [6] and Chinese [5][7].

This study aims to understand whether embodied pronunciation training has beneficial effects on learning the Swedish complementary length contrast (*vila* ≠ *villa*). In Swedish, a stressed syllable contains either a long vowel (V:(C)) or a long consonant following a short vowel (VC:). The study will assess adult Swedish learners' production and perception of this contrast through a pre-/ post-/ delayed post-test design, that is, before and after receiving pronunciation instruction with or without gestures. The instruction phase will consist of a video training learners on the given contrast in three different conditions (between subjects), plus a control group (no training): no gestures (audiovisual speech only), and two gesture conditions, testing two different sets of gestures, where participants will repeat the spoken words while imitating the gestures produced by the instructor in the training video. The gesture conditions are defined based on interviews conducted with nine teachers of Swedish as a second language. During these interviews, teachers were, among others, asked about their use of embodiment and gestures in teaching the Swedish length contrast. Most teachers indicated very similar gestures to illustrate the length contrast, usually involving a metaphorical illustration of temporal length through the depiction of a long horizontal distance in space using both hands (see Fig. 1). Notably, length was usually indicated this way for words with long vowels, rather than for words with long consonants. Words with long consonants (i.e., short vowels) were typically marked with a gesture illustrating brevity, although this was realized differently by different teachers, for instance using a hand clapping gesture, a simple beat gesture (see Fig. 2), or a gesture depicting a short distance (see Fig. 3). In one (of two) gesture conditions, we will thus contrast a *length gesture* with a *brevity gesture*, like those seen in Fig. 1 and 3. However, a current debate in the L2 Swedish pronunciation teaching context deals with the issue whether we should characterize the length contrast in terms of *long vs. short vowel*, or rather in terms of *long vowel vs. long consonant*. Therefore, in our second set of gestures two *length* gestures will be used, aligned with the vowel vs. the consonant.

Learners' production in the pre- and post-tests will be assessed through native speaker ratings and acoustic analysis. Perception will be measured through an identification task, but also using a visual-world eye-tracking experiment, where the time course of target looking will provide us with a continuous measure of learners' processing abilities – a novelty in this field of study. In addition, learners will be asked to fill out a language background questionnaire and perform auxiliary tests, such as a memory test or a speech imitation test. We aim to present the final study set-up and some pilot data at the conference.

This initial study is part of a five-year project, which studies effects of embodied pronunciation training in L2 Swedish learning, focusing on two known difficulties: the length contrast (*vila* ≠ *villa*) discussed above and the vowel contrast /i/≠/y/. We will thus compare a prosodic and a segmental feature, moreover, both in the lab and in an authentic classroom setting. We thus hope to be able to contribute to a broadened as well as deepened understanding of the role of gestures in pronunciation teaching.



Figure 1: Two teachers using similar gestures to illustrate a long vowel (a metaphorical illustration of temporal length through the depiction of a long horizontal distance in space using both hands).

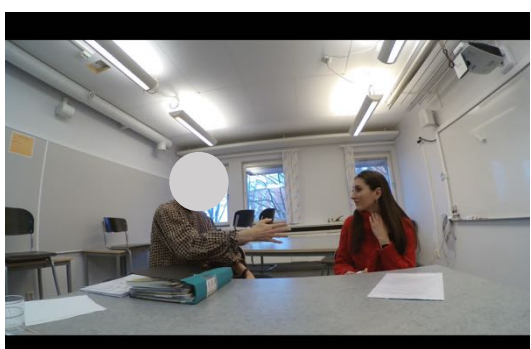


Figure 2: A teacher using a simple beat gesture to indicate a short vowel.

Figure 3: A teacher using a gesture depicting a short distance to indicate a short vowel.

References

- [1] M. Hoetjes and L. van Maastricht, "Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity," *Frontiers in Psychology*, vol. 11, 2020.
- [2] Y. Li and T. Somlak, "The effects of articulatory gestures on L2 pronunciation learning: A classroom-based study" *Language Teaching Research*, vol. 23, no. 3, pp. 352-371, 2017.
- [3] C. Yuan, S. González-Fuente, F. Baills, and P. Prieto, "Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers," *Studies in Second Language Acquisition*, vol. 41, no. 1, pp. 5-32, 2019.
- [4] Y. Zhang, F. Baills, and P. Prieto, "Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from training Chinese adolescents with French words," *Language Teaching Research*, vol. 24, no. 5, pp. 666-689, 2018.
- [5] F. Baills, N. Suárez-González, S. González-Fuente, and P. Prieto, "Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words," *Studies in Second Language Acquisition*, vol. 41, no. 1, pp. 33-58, 2019.
- [6] Y. Hirata, S.D. Kelly, J. Huang, and M. Manansala, (2014), "Effects of Hand Gestures on Auditory Learning of Second- Language Vowel Length Contrasts," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 6, pp. 2090-2101, 2014.
- [7] X. Xi, P. Li, F. Baills, and P. Prieto, "Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 11, pp. 3571-3585, 2020.

Pointing at the addressee in Hebrew face-to-face interaction

Anna Inbar, *The Academic College Levinsky-Wingate*

Yael Maschler, *University of Haifa*

While some studies of pointing gestures have addressed their deictic referential function [e.g., 1, 2, 3, 4, 5], other studies have revealed interactional practices that are accomplished by pointing [e.g., 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. The present study explores Hebrew data to focus on interactional practices which are accomplished by pointing that is directed at the prior (or current) speaker.

Our data are drawn from the *Haifa Multimodal Corpus of Spoken Hebrew*, which consists of video recordings of naturally occurring casual conversations collected during the years 2016–2023, with over 19 hours of recordings in total. We identified 89 occurrences of such addressee-directed pointing gestures, excluding from our analysis ambiguous cases in which the gesture was coordinated with an utterance including some reference to the second person. In general, pointing at the addressee for indexical purposes was found to be very infrequent throughout our data. Employing the methodologies of multimodal conversation analysis [e.g., 17, 18] and interactional linguistics [19], our study reveals two broad contexts in which interactional pointing gestures directed at the addressee occurred in our corpus, namely, affiliation and disaffiliation. In the context of affiliation, these gestures, with or without the verbal components of the utterances of which they formed a part, constituted social actions such as agreement, confirmation, or appreciation; in the context of disaffiliation, they constituted dispreferred actions, such as disagreement, correcting the interlocutor, or mocking. One of our main findings is that *index-finger pointing* (63 occurrences) is employed in disaffiliative contexts, whereas *whole-hand pointing* (26 occurrences) in affiliative contexts. Moreover, the deployment of such pointing gestures unaccompanied by any verbal components, manifests a high degree of conventionality.

In the present talk, we elaborate on social actions that are accomplished via index-finger pointing vs. whole-hand pointing gestures directed at the addressee with or without corresponding verbal components, considering additional morphological differences, such as differences in palm orientation, the degree of arm extension, and whether the gesture reaches into the addressee's gesture space physically or not. These morphological differences may carry an impact on the interaction between participants [cf. 4, 11] and yield further sub-classifications that reveal further form-function correlations.

References

- [1] H. H. Clark, "Pointing and placing," in: *Pointing: Where Language, Culture, and Cognition Meet*, ed S. Kita (Mahwah, NJ: Lawrence Erlbaum), pp. 243–268, 2003.
- [2] K. Cooperrider, "Reference in Action: Links between Pointing and Language," Ph.D. dissertation, University of California, 2011.
- [3] K. Cooperrider, "Body-directed gestures: Pointing to the self and beyond," *Journal of Pragmatics*, vol. 71, pp. 1–16, 2014.
- [4] A. Kendon, *Gesture: Visual Action as Utterance*, Cambridge University Press, 2004.
- [5] S. Kita, "Pointing: A foundational building block of human communication," in: *Pointing: Where Language, Culture, and Cognition Meet*, ed S. Kita (Mahwah, NJ: Lawrence Erlbaum), pp. 1–8, 2003.
- [6] A. Bangerter, "Using pointing and describing to achieve joint focus of attention in dialogue," *Psychological Science*, vol. 15, n. 6, pp. 415–419, 2004.

- [7] N. J. Enfield, S. Kita, and J. P. de Ruiter, "Primary and secondary pragmatic functions of pointing gestures," *Journal of Pragmatics*, vol. 39, n. 10, pp. 1722–1741, 2007.
- [8] C. Goodwin, "Pointing as situated practice," in: *Pointing: Where Language, Culture, and Cognition Meet*, ed S. Kita (Mahwah, NJ: Lawrence Erlbaum), pp. 217–241, 2003.
- [9] C. Healy, "Pointing to show agreement," *Semiotica*, vol. 192, pp. 175–195, 2012.
- [10] J. Hindmarsh and C. Heath, "Embodied reference: A study of deixis in workplace interaction," *Journal of Pragmatics*, vol. 32, n. 12, pp. 1855–1878, 2000.
- [11] J. Holler, "Speaker's use of interactional gestures as markers of common ground," in: *Gesture in embodied communication and human-computer interaction* (vol. 5934), eds S. Kopp and I. Wachsmuth (Springer Berlin Heidelberg), pp. 11–22, 2010.
- [12] L. Mondada, "Multimodal resources for turn-taking: pointing and the emergence of possible next speakers," *Discourse Studies*, vol. 9, n. 2, pp. 194–225, 2007.
- [13] L. Mondada, "Deixis: An integrated interactional multimodal analysis," in: *Prosody and Embodiment in Interactional Grammar*, eds P. Bergman and J. Brenning (Berlin: De Gruyter), pp. 173–206, 2012.
- [14] L. Mondada, "Pointing, talk, and the bodies: Reference and joint attention as embodied interactional achievements," in: *From Gesture in Conversation to Visible Action as Utterance: Essays in Honor of Adam Kendon*, eds M. Seyfeddinipur and M. Gullberg (Amsterdam/Philadelphia: John Benjamins Publishing Company), pp. 95–124, 2014.
- [15] J. Streeck, *Self-making Man: A Day of Action, Life, and Language*, Cambridge University Press, 2017.
- [16] E. Yasui, "Sequence-initial pointing: Spotlighting what just happened as a cause of a new sequence," *Discourse Studies*, vol. 25, n. 3, pp. 409–429, 2023.
- [17] C. Goodwin, *Co-operative Action*, Cambridge University Press, 2018.
- [18] L. Mondada, "Challenges of multimodality: Language and body in social interaction," *Journal of Sociolinguistics*, vol. 20, n. 3, pp. 336–366, 2016.
- [19] E. Couper-Kuhlen, and M. Selting, *Interactional Linguistics*. Cambridge University Press, 2018.

The role of interactive gestures in explanatory interactions

Vivien Lohmer, Prof. Dr. Friederike Kern

Faculty of Linguistics and Literature, Bielefeld University, Germany

(vivien.lohmer@uni-bielefeld.de)

While the discursive and interactive structure of explanatory interactions is relatively well researched [1], [2] less is known about the role of gestural behaviour and in particular, interactive gestures, i.e. abstract gestures with the function to structure the interaction globally [3, p. 472]. Explanatory interactions feature a series of jointly accomplished interactive tasks in a routine [2], [4]: *establishing topical relevance* for an explanation (1), *co-constructively constituting an explanandum* (2), *explicating the conceptual, causal, and/or procedural relations* concerning the explanandum (3), and finally, co-constructively agreeing on *closing* the activity (4) and pass over to the next topic (*5 transition*).

To investigate gestural behaviour in explanatory interactions and the function of interactive gestures and their link to semantic units in particular, we designed a study with the board-game Quarto! to elicit near-to-natural explanatory interactions. Two participants were seated face-to-face towards each other. One participant (Explainer, short: EX) explains the game to the other participant who is unfamiliar with the game (Explainee, short: EE). The interactions were audio- and videotaped from three camera perspectives. We recorded 26 dyadic explanations (52 participants in total) that were all transcribed following GAT 2 [5]. Four dyads were excluded for various reasons. The remaining 22 dyads were coded according to a scheme developed on the basis of the above-mentioned interactive tasks [6, p. 320]. The coding resulted in a Cohens Kappa of 0.7 (substantial) [7]. Deviations between both coders were smoothed afterwards.

The micro-analysis followed principles of Conversation Analysis (CA) [8], [9] and revealed that the core job (3) features smaller semantic units (explanation nodes, short: *nodes*), e.g. certain game materials and their rule-based use, with each node being systematically linked to the following one [4, p. 6]. From this observation, an additional coding-scheme was developed and applied to the data (ibid.). Starting from the observation that participants tend to use interactive gestures [3, p. 472] when a) explicating the connection between two nodes and b) marking the discursive structure within a node by highlighting or explicating relevant information, we analysed six data sets for gestural behaviour [3], [10], [11], looking closely at the nodes ‘board’, ‘figures’, and ‘goal’, and at the transition spaces between them.

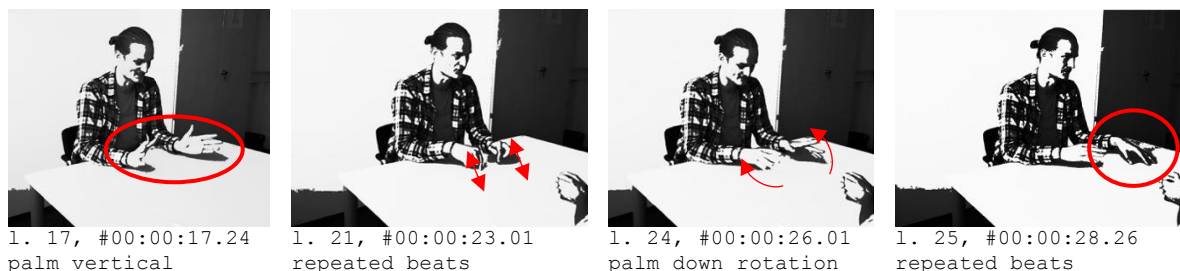
```

017 EX FOLgendermaßen;# (#00:00:17.24)
      like this
018 DA gibt's ein SPIELbrett,
      there is a game board
019 da gibt ES,
      there are
020 hm d? das VIER mal VIER,
      that four by four
021 so LÖCHER# drin; (#00:00:23.01)
      holes in it
022 EE hm_hm,
      Hm hm (affirmativ)
023 EX und?
      and
024 also sechszehn# LÖCHer, (#00:00:26.01)
      so sixteen holes

```

Node: board

025	und du hast auch sechszehn SPIELfiguren;# (#00:00:28.26)	
	and you also have sixteen game figures	
026	EE JA-ha,	Node: figure
	Yes	



First results show that a) participants use interactive gestures and mark verbally the connection between two nodes. When participants explicate the connection (cf. l. 24-25), they tend to use interactive gestures such as palm-vertical open hand gestures (cf. #00:00:26.01), or beat gestures (c.f. #00:00:28.26). Additionally, b) within a node, participants use interactive gestures when introducing something new (cf. l. 17, #00:00:17.24) or when elaborating on previously given information (cf. l. 21 #00:00:23.01).

The micro-analysis thus revealed that interactive gestures are systematically used to further structure semantic content according to semantic units, i.e. explanation nodes. They are doing so by highlighting the link between nodes, and to emphasize important information, thus providing more insights into gesture-speech-integration in particular settings, i.e. explanatory interactions.

References

- [1] V. Lohmer, L. Terfloeth, and F. Kern, 'Explaining the technical Artifact Quarto!:How Gestures are used in Everyday Explanations', presented at the 1st International Multimodal Communication Symposium, Barcelona, 28.04 2023, pp. 133–134. [Online]. Available: http://mmsym.org/wp-content/uploads/2023/04/Book_of_abstracts_MMSYM.pdf
- [2] U. Quasthoff, V. Heller, and M. Morek, 'On the sequential organization and genre-orientation of discourse units in interaction: An analytic framework', *Discourse Studies*, vol. 19, no. 1, pp. 84–110, Feb. 2017, doi: 10.1177/1461445616683596.
- [3] J. B. Bavelas, N. Chovil, D. A. Lawrie, and A. Wade, 'Interactive gestures', *Discourse Processes*, vol. 15, no. 4, pp. 469–489, Oct. 1992, doi: 10.1080/01638539209544823.
- [4] J. B. Fisher, A. Robrecht, K. Rohlfing, and S. Kopp, 'Exploring the Semantic Dialogue Patterns of Explanations - a Case Study of Game Explanations', 2023. Accessed: Aug. 09, 2023. [Online]. Available: <https://pub.uni-bielefeld.de/record/2980822#apa>
- [5] M. Selting *et al.*, 'Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)', *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*, 2009, Accessed: Apr. 05, 2023. [Online]. Available: <https://orbi.lu.uni.lu/handle/10993/4358>
- [6] J. B. Fisher, V. Lohmer, F. Kern, W. Barthlen, S. Gaus, and K. J. Rohlfing, 'Exploring Monological and Dialogical Phases in Naturally Occurring Explanations', *Künstl Intell*, vol. 36, no. 3–4, pp. 317–326, Dec. 2022, doi: 10.1007/s13218-022-00787-1.
- [7] J. R. Landis and G. G. Koch, 'The Measurement of Observer Agreement for Categorical Data', *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
- [8] C. Goodwin, *Co-Operative Action*, 1st ed. Cambridge University Press, 2017. doi: 10.1017/9781139016735.
- [9] F. Kern and M. Selting, 'Conversation Analysis and Interactional Linguistics', *C. A.*, 2020.
- [10] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.
- [11] D. McNeill, *Gesture and thought*. Chicago: University of Chicago press, 2005.

Keynote 3: Petra Wagner

27.09.2024

12:10-13:10



**The multimodal expression of (non-)understanding in dyadic explanations
- some lessons learned**

Petra Wagner

University of Bielefeld

petra.wagner@uni-bielefeld.de

During an explanation between an explainer (a person who explains) and an explainee (a person something is explained to), explainers crucially rely on the explainee's feedback about their current level of understanding as well as their level of cognitive load or attention. Based on the monitoring of a wide range of verbal and non-verbal feedback cues, an explainer can then dynamically adjust the explanation strategy, e.g., by changing the tempo of the ongoing explanation, repeat or skip parts of the explanation, or even shift the focus of the explanation.

In my talk, I will report first insights from the TRR318 „Constructing Explainability“ (<https://trr318.uni-paderborn.de/en/>) subproject A02 on „Monitoring the understanding of explanations“, in which we gather and investigate multimodal signals of (non-)understanding in explanations, see how they evolve in course of ongoing explanations, and how they are interpreted and reacted to. In particular, I will describe the recording and rich multimodal annotation of a corpus of 87 dyadic board game explanations, provide information about our annotation of different levels of (non-)understanding using a recall task, address the floor management dynamics across different phases of the explanations, present some insights on how explainers adapt their multimodal behavior to different explainees, and show how verbal and non-verbal information combine in a model of classifying (non-)understanding. Throughout, I will also address the various challenges we were faced with.



Index of Authors



A

Agirrezabal, Manex	26.09. 15:20 – 16:40	100
Ahmar, Davide	25.09. 13:40 – 15:00	35
Albert, Aviad	26.09. 17:10 – 18:30	121
Althaus, Nadja	27.09. 10:20 – 11:40	153
Ambrazaitis, Gilbert	26.09. 17:10 – 18:30	119
	27.09. 10:20 – 11:40	153
Andries, Fien	27.09. 10:20 – 11:40	147

B

Baills, Florence	25.09. 11:00 – 12:20	26
	26.09. 17:10 – 18:30	121
Bauer, Anastasia	26.09. 11:30 – 12:50	80
	26.09. 11:30 – 12:50	86
Baumann, Stefan	25.09. 11:00 – 12:20	26
Béres, Luca	25.09. 15:00 – 16:20	40
Boll-Avetisyan, Natalie	27.09. 10:20 – 11:40	137
Boncz, Ádám	25.09. 15:00 – 16:20	40
Bonnet, Marion	26.09. 15:20 – 16:40	104
Bosker, Hans Rutger	25.09. 11:00 – 12:20	20
Brône, Geert	26.09. 14:20 – 15:20	89
Brown-Schmidt, Sarah	25.09. 15:00 – 16:20	42
Bujok, Ronny	25.09. 11:00 – 12:20	20
Busche, Franziska	27.09. 10:20 – 11:40	151
Buschmeier, Hendrik	25.09. 15:00 – 16:20	38

C

Caet, Stéphanie	27.09. 10:20 – 11:40	139
Campisi, Emanuela	25.09. 16:50 – 18:10	63
Cantalini, Giorgia	25.09. 11:00 – 12:20	22
Clough, Sharice	25.09. 15:00 – 16:20	42
Colombani, Arianna	27.09. 10:20 – 11:40	137
Coego, Sara	25.09. 16:50 – 18:10	61
	25.09. 16:50 – 18:10	67
Cruz, Marisa	26.09. 11:30 – 12:50	82
Ćwiek, Aleksandra	25.09. 15:00 – 16:20	46

D

De Laat, Kristel	25.09. 15:00 – 16:20	42
De Vries, Clarissa	27.09. 10:20 – 11:40	147
Dudschig, Carolin	26.09. 15:20 – 16:40	110
Duff, Melissa C.	25.09. 15:00 – 16:20	42
Dych, Walter Philip	25.09. 13:40 – 15:00	29

E

Ebert, Cornelia	26.09. 15:20 – 16:40	104
-----------------	----------------------	-----

Erbach, Kurt	26.09. 15:20 – 16:40	104
Espejo-Álvarez, Joel	25.09. 16:50 – 18:10	65
Esteve-Gilbert, Núria	25.09. 16:50 – 18:10	61

F

Ferran, Carla	27.09. 10:20 – 11:40	139
Florit-Pons, Júlia	25.09. 16:50 – 18:10	65
	26.09. 15:20 – 16:40	67
	26.09. 15:20 – 16:40	116
Franich, Kathryn	25.09. 13:40 – 15:00	29
	26.09. 17:10 – 18:30	123
Frey, Nathalie	26.09. 15:20 – 16:40	114
Frota, Sónia	26.09. 11:30 – 12:50	82
Fuchs, Susanne	25.09. 15:00 – 16:20	46
	26.09. 15:20 – 16:40	102

G

Garde, Henrik	26.09. 17:10 – 18:30	125
Garvin, Karee	25.09. 13:40 – 15:00	29
Gipper, Sonja	26.09. 11:30 – 12:50	80
Giulimondi, Alessia	25.09. 16:50 – 18:10	63
Gómez i Martínez, Mireia	25.09. 16:50 – 18:10	65
Gorisch, Jan	27.09. 10:20 – 11:40	149
Graziano, Maria	27.09. 10:20 – 11:40	143
Gregori, Alina	26.09. 15:20 – 16:40	102
Grice, Martine	26.09. 17:10 – 18:30	121
Grimminger, Angela	25.09. 15:00 – 16:20	54
	27.09. 10:20 – 11:40	141
Gullberg, Marianne	26.09. 17:10 – 18:30	125
	27.09. 10:20 – 11:40	143

H

Harte, Naomi	26.09. 11:30 – 12:50	91
Henlein, Alexander	26.09. 15:20 – 16:40	96
	27.09. 09:00 – 10:20	128
Heim, Johannes	27.09. 10:20 – 11:40	151
Herrmann, Tobias-Alexander	26.09. 11:30 – 12:50	80
Hoetjes, Marieke	27.09. 10:20 – 11:40	153
Holler, Judith	26.09. 09:00 – 10:00	70
Honda, Koki	27.09. 10:20 – 11:40	141
Hosemann, Jana	26.09. 11:30 – 12:50	80
House, David	26.09. 17:10 – 18:30	119
	27.09. 10:20 – 11:40	149
Huijsmans, Marianne	27.09. 09:00 – 10:20	132
Hunter, Julie	25.09. 09:30 – 10:30	17

I

Igualada, Alfonso	25.09. 16:50 – 18:10	65
	26.09. 15:20 – 16:40	116
Inbar, Anna	27.09. 10:20 – 11:40	155

J

Janssens, Julie	26.09. 11:30 – 12:50	89
Janzen, Solbeigh	26.09. 17:10 – 18:30	121
Jongejan, Bart	26.09. 15:20 – 16:40	100

K

Kadavá, Šárka	25.09. 13:40 – 15:00	35
	25.09. 15:00 – 16:20	46
Kern, Friederike	27.09. 10:20 – 11:40	157
Kosmala, Loulou	27.09. 10:20 – 11:40	139
Kotowicz, Justyna	27.09. 10:20 – 11:40	145
Kuder, Anna	26.09. 11:30 – 12:50	86
	27.09. 10:20 – 11:40	145
Kügler, Frank	25.09. 11:00 – 12:20	24

L

Ladewig, Silva H.	27.09. 09:00 – 10:20	130
Laparle, Schuyler	25.09. 15:00 – 16:20	48
Laval, Marine	27.09. 10:20 – 11:40	139
Lazarov, Stefan	25.09. 15:00 – 16:20	54
Lialiou, Maria	26.09. 17:10 – 18:30	121
Lohmer, Vivien	27.09. 10:20 – 11:40	157
Lombart, Clara	26.09. 11:30 – 12:50	84
Loos, Cornelia	27.09. 09:00 – 10:20	134
Lücking, Andy	26.09. 15:20 – 16:40	96
	27.09. 09:00 – 10:20	128
Lüke, Carina	25.09. 15:00 – 16:20	52
	26.09. 15:20 – 16:40	114
Luong, Claire Lien	25.09. 16:15 – 18:10	65

M

Mai, Quian Yin	27.09. 10:20 – 11:40	137
Mahdinazhad Sardhaei, Nasim	26.09. 10:00 – 11:00	75
Maschler, Yael	27.09. 10:20 – 11:40	155
Mehler, Alexander	26.09. 15:20 – 16:40	96
	27.09. 09:00 – 10:20	128
Meissl, Katharina	27.09. 10:20 – 11:40	147
Molinaro, Nicola	25.09. 13:40 – 15:00	33
Moneglia, Massimo	25.09. 11:00 – 12:20	22
Mücke, Doris	25.09. 13:40 – 15:00	31

N

Nagy, Péter	25.09. 15:00 – 16:20	40
Nicoladis, Elena	25.09. 15:00 – 16:20	56
	26.09. 10:00 – 11:20	77
Nonaka, Ikuko	25.09. 15:00 – 16:20	58
Nuttall, Thomas	26.09. 10:00 – 11:00	73
Nwosu, Vincent	26.09. 17:10 – 18:30	123

O

O'Connor Russel, Sam	26.09. 14:20 – 15:20	91
Oben, Bert	26.09. 11:30 – 12:50	89
Özçalışkan, Şeyda	26.09. 15:20 – 16:40	108
Özyürek, Aslı	25.09. 15:00 – 16:20	42
	25.09. 16:50 – 18:10	63

P

Pagel, Lena	25.09. 13:40 – 15:00	31
Paggio, Patrizia	26.09. 15:20 – 16:40	100
Pastureau, Romain	25.09. 13:40 – 15:00	33
Pearson, Lara	26.09. 10:00 – 11:00	73
Pelageina, Nadia	26.09. 17:10 – 18:30	121
Peter, Varghese	27.09. 10:20 – 11:40	137
Pouw, Wim	25.09. 13:40 – 15:00	35
	25.09. 15:00 – 16:20	46
	26.09. 10:00 – 11:00	73
Prieto, Pilar	25.09. 11:00 – 12:20	24
	25.09. 16:50 – 18:10	61
	25.09. 16:50 – 18:10	65
	25.09. 16:50 – 18:10	67
	26.09. 15:20 – 16:40	112
	26.09. 15:20 – 16:40	116
Pronina, Mariia	25.09. 16:50 – 18:10	67
	26.09. 15:20 – 16:40	112

R

Raschellà, Federica	27.09. 10:20 – 11:40	153
Reisinger, Daniel K. E.	27.09. 09:00 – 10:20	132
Ren-Mitchell, Ada	25.09. 15:00 – 16:20	44
Repp, Sophie	27.09. 09:00 – 10:20	134
	27.09. 10:20 – 11:40	151
Riechmann, Alina Naomi	25.09. 15:00 – 16:20	38
Roessig, Simon	25.09. 13:40 – 15:00	31
Rohlfing, Katharina J.	27.09. 10:20 – 11:40	141
Rohrer, Patrick Louis	25.09. 11:00 – 12:20	20
Rühlemann, Christoph	26.09. 15:20 – 16:40	106

S

Saksida, Amanda	27.09. 10:20 – 11:40	137
Sánchez-Ramón, Paula G.	25.09. 11:00 – 12:20	24
Sarıtaş, Himmet	26.09. 15:20 – 16:40	108
Schaeffner, Simone	25.09. 15:00 – 16:20	52
Schepens, Job	26.09. 11:30 – 12:50	86
Scholman, Merel	25.09. 15:00 – 16:20	48
Schubert, Mojenn	26.09. 14:20 – 15:20	93
Schulder, Marc	26.09. 11:30 – 12:50	86
Schulte, Marion	25.09. 15:00 – 16:20	50
Schumacher, Petra B.	26.09. 17:10 – 18:30	121
Schütt, Emanuel	26.09. 15:20 – 16:40	110
Sekine, Kazuki	25.09. 15:00 – 16:20	58

	27.09. 10:20 – 11:40	141
Sharifzadeh, Hamid	26.09. 10:00 – 11:00	75
Sharma, Mridula	27.09. 10:20 – 11:40	137
Shattuck-Hufnagel, Stefanie	25.09. 15:00 – 16:20	44
Shokrkon, Anahita	25.09. 15:00 – 16:20	56
Slonimska, Anita	25.09. 16:50 – 18:10	63
Splendido, Frida	26.09. 17:10 – 18:30	125
	27.09. 10:20 – 11:40	153
Springer, Helene	26.09. 17:10 – 18:30	125
Steinbach, Markus	26.09. 15:20 – 16:40	104
Sümer, Beyza	25.09. 15:00 – 16:20	42
 <u>T</u>		
Tanguay, Annick	25.09. 15:00 – 16:20	42
Tolksdorf, Nils	27.09. 10:20 – 11:40	141
Tuomainen, Outi	27.09. 10:20 – 11:40	137
 <u>V</u>		
Van de Weijer, Joost	27.09. 10:20 – 11:40	143
Van Maastricht, Lieke	25.09. 11:00 – 12:20	20
Vella, Alexandra	26.09. 17:10 – 18:30	121
Vilà-Giménez, Ingrid	26.09. 15:20 – 16:40	112
 <u>W</u>		
Wagner, Petra	27.09. 12:10 – 13:10	160
Weicker, Merle	26.09. 15:20 – 16:40	110
Winkler, István	25.09. 15:00 – 16:20	40
Wójcicka, Joanna	27.09. 10:20 – 11:40	145
Wolfrum, Vera	25.09. 15:00 – 16:20	52
Woods, Rebecca	27.09. 10:20 – 11:40	151
 <u>Z</u>		
Zarezadehkheibari, Shiva	25.09. 15:00 – 16:20	56
Zellers, Margaret	26.09. 17:10 – 18:30	119
	27.09. 10:20 – 11:40	149
Zhou, Han	26.09. 15:20 – 16:40	98
Zygis, Marzena	26.09. 10:00 – 11:00	75