

Sociolinguistic and multimodal variation of turn-taking dynamics in conversation

Dissertation in fulfillment of the requirements for
the doctoral degree
at the Faculty of Arts and Humanities
at Kiel University
in Kiel

submitted by

Martina Rossi

Kiel

5 March 2024

First examiner: Prof. Dr. Margaret Zellers

Second examiner: Prof. Dr. Petra Wagner

Date of the oral examination: 12 June 2024

Approved for printing by the second Vice Dean,
Prof. Dr. Elmar Eggert: 19 June 2025

Open Access

The work is published under the Creative Commons license Attribution 4.0 International (CC-BY 4.0). The deed can be found here: <https://creativecommons.org/licenses/by/4.0/deed.en>.

The electronic version of this doctoral thesis is available on the open access publications server MACAU of the University Library Kiel (<https://macau.uni-kiel.de>)

URN: urn:nbn:de:gbv:8:3-2025-01165-3

©Martina Rossi, 2024

(ORCID: <https://orcid.org/0000-0001-5970-3366>)

Every second counts.

Christopher Storer, "The Bear"

Abstract

Turn-taking refers to the process by which interlocutors alternate between speakership and listenership in conversation. Through shared collaborative efforts, interlocutors manage who is to speak next. In informal interactions involving two participants the possible turn-taking actions are: a speaker change, i.e., the conversational floor passes from one interlocutor to the other; a turn keep, i.e., the current speaker continues talking; a backchannel, i.e., the non-speaking interlocutor, or the listener, does not take the floor but produces a brief and unobtrusive response to display attentiveness. While across languages a tendency has been observed to keep silent gaps and speech overlaps reduced to a minimum in the timing of the alternation between speakership and listenership, thus pointing towards the existence of a universal turn-taking system, gaps and overlaps are not at all absent from everyday spontaneous interaction. From previous investigations, in fact, they appear to be more or less frequent in specific languages and/or specific social groups or communicative contexts.

This thesis explores the hypothesis that these perceived deviations from a possible universal turn-taking system have a socio-cultural and contextual nature. Since informal conversation is the prime social environment for identity construction and language use, turn-taking dynamics, i.e., turn allocation techniques and the distribution and timing of turn-taking actions among interlocutors, can reveal information about the interaction's setting and the interlocutors' socio-cultural identity, attitudes and values. Specifically, this dissertation investigates the dynamics of turn-taking in spontaneous face-to-face conversation by focusing on the influence of the gender of the interlocutors, the availability of multiple communication modalities (visual and vocal), and the language in which the exchange takes place through a comparison of German and Swedish.

The work comprises analyses aimed at exploring these distinct aspects of turn-taking in relation to sociolinguistic variables and multimodality. The data used for the analyses come from speech corpora comprising two-party spontaneous interactions between native German or between native Swedish speakers. The influence of the gender of the speaker and the gender of the listener is tested in German through quantitative analyses on the turn-taking-related variation of acoustic cues regulating turn transition. Specifically, the acoustic turn-taking cues considered throughout the work are fundamental frequency (F0), intensity, speech rate and voice qual-

ity; the availability of different modalities of communication and its influence on the temporal organization of the interaction is explored through the use of backchannels, particularly considering the use of short verbal, non-verbal responses and head movements as part of listener feedback. The influence of gender is also tested on the timing and the distribution of backchannels across modalities in German. Finally, the use of turn-taking cues, the frequency and timing of speaker changes, turn-keeps and multimodal backchannels, as well as gender-specific distributional and timing patterns, are tested across languages by carrying out a comparison with Swedish dialogues.

The results indicate that gender influences the variation of acoustic cues signaling turn-taking intentions, with female and male speakers displaying gender-specific preferences in the use of F0, intensity and speech rate, and in their patterns of variation towards potential transition locations. Additionally, both male and female speakers adapt their acoustic cues differently based on the gender of their interlocutor, potentially reflecting gender-specific performance expectations. Backchannel placement also varies in relation to gender, with female listeners providing feedback at locations characterized by low F0 and slower speech rates, contrasting with male listeners who confirm attentiveness with feedback after high F0 and faster speech rates. The modality of backchannels further influences timing and distribution, with head movements occurring in overlap with the current speaker's utterance, given their unobtrusive nature, while vocal responses often follow speech offset. Furthermore, language-specific factors significantly shape turn-taking behaviors, evident in cross-linguistic comparisons between German and Swedish. Differences in the use of acoustic cues, backchannels, and multimodal alternation reflect distinct language-specific phonological constraints, conversational priorities and socio-cultural attitudes and values. Gender conversational dynamics also exhibit language-specific patterns, with German interlocutors displaying more traditional gendered behaviors compared to Swedish participants.

This dissertation contributes to the existing research on spontaneous conversation by incorporating sociolinguistic variables in a systematic and quantitative manner to understand how interactions are structured through turn-taking dynamics. The findings indicate strong links between how acoustic features are used to manage conversations and the gender of both the speaker and listener, suggesting that turn-taking-related variations of these cues may convey social significance and contribute to individuals' performance of their social identities in spontaneous interactions. Specifically, this work is the first to investigate in detail the possibility of sociophonic patterns of variation of turn-taking cues. Moreover, the study's cross-linguistic analysis sheds light on language-specific variations in listeners' multimodal behavior and gender-specific patterns across languages, emphasizing the importance of considering linguistic and cultural contexts in investigating turn-taking dynamics.

Finally, despite their typological and societal similarities, the comparison of German and Swedish reveals significant differences in how turn-taking is implemented by interlocutors in the two languages, indicating that turn-taking dynamics are highly sensitive to variations in the linguistic and sociolinguistic background of the speakers, and thus represent an ideal subject for the quantitative study of identity performance across contexts, social groups, languages and cultural backgrounds.

Kurzzusammenfassung

Turn-Taking bezeichnet den Vorgang, bei dem Gesprächspartner im Gespräch zwischen Sprecher- und Zuhörerschaft wechseln. Welche Person als nächstes spricht, wird dabei in Kooperation durch bestimmte Turn-Taking-Aktionen geregelt. Bei informellen Interaktionen mit zwei Teilnehmern sind folgende Aktionen des Turn-Taking möglich: ein Sprecherwechsel, d. h. die Sprecherrolle geht von einem Gesprächspartner auf den anderen über; ein Turn Keep, d. h. der aktuelle Sprecher spricht weiter; ein Backchannel (auch Rückmeldungssignal), d. h. der nicht-sprechende Gesprächspartner oder Zuhörer ergreift nicht das Wort, sondern gibt eine kurze, unaufdringliche Antwort, um Aufmerksamkeit zu signalisieren. Sprachübergreifend wurde eine Tendenz beobachtet, die Lücken und Überschneidungen im zeitlichen Ablauf des Wechsels zwischen Sprecher und Zuhörer im Gespräch auf ein Minimum zu reduzieren, was auf die Existenz eines universellen Turn-Taking-Systems hindeuten würde. Eine Abwesenheit von Lücken und Überschneidungen in der alltäglichen spontanen Interaktion ist jedoch selten. Aus früheren Untersuchungen geht tatsächlich hervor, dass ihr Vorkommen durch bestimmte Sprachen, bestimmte soziale Gruppen und/oder bestimmte Kommunikationskontexte bestimmt ist.

Diese Arbeit untersucht die Hypothese, dass diese beobachteten Abweichungen von einem möglichen universellen Turn-Taking-System soziokulturellen und kontextuellen Charakter haben. Da informelle Konversation das wichtigste soziale Umfeld für die Identitätskonstruktion und den Sprachgebrauch ist, kann die Turn-Taking-Dynamik, d. h. die Techniken der Sprecherzuweisung sowie die Verteilung und der Zeitpunkt von Turn-Taking-Aktionen unter den Gesprächspartnern, Informationen über den Kontext der Interaktion und die soziokulturelle Identität der Gesprächspartner liefern, sowie deren Einstellungen und Werte. Konkret untersucht diese Dissertation die Dynamik des Turn-Taking in spontanen persönlichen Gesprächen, indem sie den Einfluss des Geschlechts der Gesprächspartner und die Verfügbarkeit mehrerer Kommunikationsmodalitäten (verbale und non-verbale, d.h. durch Stimme und Gestik) in persönlichen Interaktionen in den Fokus nimmt, sowie das Turn-Taking-Verhalten in den Sprachen Deutsch und Schwedisch miteinander vergleicht.

Die Arbeit umfasst Analysen, die darauf abzielen, diese unterschiedlichen Aspekte des Turn-Taking in Bezug auf soziolinguistische Variablen und Multimodal-

ität zu untersuchen. Die für die Analysen verwendeten Daten stammen aus Sprachkorpora, die deutsche oder schwedische spontane Interaktionen von je zwei muttersprachlichen Personen umfassen. Der Einfluss des Geschlechts des Sprechers und des Geschlechts des Zuhörers wird im Deutschen durch quantitative Analysen der Variation akustischer Signale, die in Zusammenhang mit Turn-Taking auftreten und den Sprecherwechsel regulieren, getestet. Konkret handelt es sich bei den im gesamten Werk berücksichtigten akustischen Parametern um die Grundfrequenz (F0), die Intensität, die Segmentdauer, die Sprechgeschwindigkeit und die Stimmqualität. Die Verfügbarkeit verschiedener Kommunikationsmodalitäten und ihren Einfluss auf die zeitliche Organisation der Interaktion werden anhand von Backchannels untersucht, insbesondere unter Berücksichtigung der Verwendung kurzer verbaler (z. B. "genau") und nonverbaler Antworten (z. B. Kopfnicken) als Teil der Ressourcen auf Zuhörerseite zur Bereitstellung von Feedback. Der Einfluss des Geschlechts auf den Zeitpunkt und die Verteilung von Backchannels verschiedener Modalitäten im Deutschen wird ebenfalls getestet. Abschließend werden die Verwendung von akustischen Turn-Taking-Hinweisen, die Häufigkeit und der Zeitpunkt von Sprecherwechseln, Turn-Keeps und multimodalen Backchannel sowie geschlechtsspezifische Verteilungs- und Timing-Muster sprachübergreifend getestet, indem ein Vergleich mit schwedischen Dialogen durchgeführt wird.

Die Ergebnisse deuten darauf hin, dass das Geschlecht die Variation akustischer Signale, die Turn-Taking-Absichten signalisieren, beeinflusst, wobei weibliche und männliche Sprecher geschlechtsspezifische Präferenzen bei der Verwendung von F0, der Intensität und der Segmentdauer sowie bei ihren Variationsmustern in Richtung potenzieller Sprecherwechsel zeigen. Darüber hinaus passen sowohl männliche als auch weibliche Sprecher ihre akustischen Hinweise je nach Geschlecht ihres Gesprächspartners unterschiedlich an, was möglicherweise geschlechtsspezifische Leistungserwartungen widerspiegelt. Die Platzierung des Backchannels variiert auch in Abhängigkeit vom Geschlecht, wobei weibliche Zuhörer Feedback an Orten geben, die durch niedrige F0 und langsamere Sprechgeschwindigkeiten gekennzeichnet sind, im Gegensatz zu männlichen Zuhörern, die ihre Aufmerksamkeit durch Feedback bei hoher F0 und schnelleren Sprechgeschwindigkeiten bestätigen. Die Modalität der Backchannel hat weiteren Einfluss auf Timing und Verteilung, wobei Kopfbewegungen aufgrund ihrer unauffälligen Natur in Überschneidung mit der Äußerung des aktuellen Sprechers auftreten, während verbale Reaktionen oft dem Ende einer Äußerung, d.h. einer Stille folgen. Darüber hinaus beeinflussen sprachspezifische Faktoren das Turn-Taking-Verhalten maßgeblich, was in sprachübergreifenden Vergleichen zwischen Deutsch und Schwedisch deutlich wird. Unterschiede in der Verwendung von akustischen Turn-Taking-Hinweisen, Backchannels und Einsatz multimodaler Ressourcen spiegeln unterschiedliche sprachspezifische phonologische

Einschränkungen, Prioritäten in der Gesprächsführung sowie soziokulturelle Einstellungen und Werte wider. Die Gesprächsdynamik zwischen den Geschlechtern weist auch sprachspezifische Muster auf, wobei deutsche Gesprächspartner im Vergleich zu schwedischen Teilnehmern eher traditionelle geschlechtsspezifische Verhaltensweisen zeigen.

Diese Dissertation trägt zur bestehenden Forschung zu spontanen Gesprächen bei, indem sie soziolinguistische Variablen auf systematische und quantitative Weise einbezieht, um zu verstehen, wie Interaktionen durch Turn-Taking-Dynamiken strukturiert werden. Die Ergebnisse deuten auf einen starken Zusammenhang zwischen der Art und Weise hin, wie akustische Merkmale zur Steuerung von Gesprächen genutzt werden, und dem Geschlecht des Sprechers und Zuhörers, was darauf hindeutet, dass diese Variationen im Turn-Taking-Verhalten soziale Bedeutung vermitteln und dazu beitragen können, wie Einzelpersonen ihre soziale Identität in spontane Interaktionen konstruieren. Insbesondere ist diese Arbeit die erste, die die Möglichkeit soziophonetischer Variationsmuster von Turn-Taking-Hinweisen im Detail untersucht. Darüber hinaus beleuchtet die sprachübergreifende Analyse der Studie sprachspezifische Variationen im multimodalen Verhalten der Zuhörer und geschlechtsspezifische Muster in verschiedenen Sprachen und unterstreicht die Bedeutung der Berücksichtigung sprachlicher und kultureller Kontexte bei der Untersuchung des Turn-Taking-Verhaltens. Schließlich zeigt der Vergleich von Deutsch und Schwedisch trotz ihrer typologischen und gesellschaftlichen Ähnlichkeiten signifikante Unterschiede in der Art und Weise, wie das Turn-Taking umgesetzt wird, was darauf hindeutet, dass die Dynamik des Turn-Takings sehr empfindlich auf Unterschiede im sprachlichen und soziolinguistischen Hintergrund der Sprecher reagiert. Sie stellen somit ein ideales Thema für die quantitative Untersuchung der Identitätskonstruktion über Kontexte, soziale Gruppen, Sprachen und kulturelle Hintergründe hinweg dar.

Acknowledgments

Many people inspired and supported me over the past three years. First, I am deeply grateful to my supervisor, Meg Zellers. Meg, thank you for welcoming me into your project, for your guidance throughout all the work that went into this thesis, and for allowing me the freedom to pursue my own research interests and ideas. Thank you for all the invaluable opportunities and the responsibilities that you've given me these past three years. I learned so much from you during this time, and I look forward to continuing to do so in the future.

I would also like to express my gratitude to Kathrin Feindt. Thank you, Kathrin, for always taking the time to listen to me and discuss my ideas, and for all the help, the support, the advice, the guidance that you've given me, and continue to give me, in navigating academia (and a bunch of other areas). I will never not laugh when thinking about our adventures in Sweden, and all of our conversations and fun times together will always have a special place in my memories.

I gratefully acknowledge the financial support provided by the Deutsche Forschungsgemeinschaft (DFG) as part of the project "Sound patterns and linguistic structures at the transition space in conversation" (n. 444631148), which made this research possible.

I am incredibly thankful for my friends and fellow Ph.D. students at the ISFAS, Stephanie Berger, Marlene Böttcher and Ariba Khan. Thank you for always being there for me when I needed help or just a chat, or a walk, or a cup of tea, or all of it at the same time. Our meetings have definitely been one of the most fun parts of this journey. I feel so lucky and so grateful to have shared my Ph.D. years with you. I couldn't have asked for a better group.

A big thank you to the members and ex-members of the ISFAS of Kiel University, John Peterson, Benno Peters, Eugenie Stapert, Lee Pratchett, Zarina Molochieva, Erika Just, Csilla Kasz, Tobias Weber and Anke Baumann – thank you for welcoming me into the Institute and in Kiel, and for all the support and advice you've given me throughout my Ph.D. A special thank you to the student assistants with whom I had the chance to work over the past three years, in particular to Margareta Ehlers-Karmanova, whose dedication and hard work – and all of our conversations over some definitely-not-Italian coffee – contributed to the completion of this work.

I would like to express my gratitude to Maria Graziano. Maria, thank you for welcoming me into the Humanities Lab of Lund University, and for sharing with me

your deep knowledge of gesture research and your incredible passion and dedication for your work. Thank you for making my research stay in Lund so informative, fun and enjoyable. I would also like to thank Marianne Gullberg for opening the doors of her department and her research group to me, and all the other people I met in Lund, who made my stay so fun and pleasant. Moving further north, I would like to thank David House and Joakim Gustafsson for welcoming me into the Department of Speech, Music and Hearing of KTH in Stockholm, and to all the researchers who generously volunteered their time and efforts to help us setting up our experiments or to take part in them. Then, a big thank you goes to all the Ph.D. students that I met during my stays at KTH, who included me in their group and made my time in Stockholm fun and memorable.

My gratitude also goes to Chiara Meluzzi. Chiara, thank you for believing in me from day zero of this journey and for encouraging me to pursue a Ph.D. Your constant support, advice and suggestions played a significant role in the completion of this dissertation. Thank you for all the opportunities you gave me, for the fun times we spent together at conferences, and for introducing me to some great people who I am lucky to have as colleagues and friends: Lucia Sbacco, Nicholas Nese, Roberta Bianca Luzietti, Camilla Masullo, Massimo D'Aco, and the rest of the team. A special thank you also to Marina Castagneto and to Chiara Celata, whose input and comments on my presentations and papers greatly contributed to improving the quality of my work.

Then, I am greatly thankful for the amazing people who made Kiel feel like home over the past few years, as well as for all the old friends that always supported me from afar. I am especially grateful to Checo, Alessa, Marco, Ginevra, Duygu, Anja, and Sophia, who provided encouragement and much-needed distractions throughout the writing of this thesis.

A big thank you to Simona, Stefano and Guglielmo, who believed I would be able to complete a doctorate before I even started one. Thank you for always encouraging me and believing in my abilities.

I am deeply grateful to my family – mamma, papà, Ale – for allowing me to pursue my studies, for celebrating my big and small achievements with me and for encouraging me when I needed it. Thank you for putting so much work and effort into making things easy and pleasant for me without ever expecting a thank you. Even though we lived a bit far from each other over the past three years, you made me feel like I was there with you all the time.

And finally, thank you, Ollie. Thank you for believing in me every single day, for supporting me, for putting up with me, for cooking for me, for always being the most interesting, smart, caring, fun, sensitive person I could talk to. Thank you for helping me do what makes me happy. These years have been so special because you were by my side.

Contents

1	Introduction	1
I	Background	5
2	Turn-taking in conversation	7
3	Speaking and listening in interaction	13
3.1	Speaking and turn-taking cues	14
3.1.1	Turn-taking cues in German and Swedish	19
3.2	Listening and multimodal backchannels	22
3.2.1	Backchannel distribution in conversation	28
4	Sociolinguistic variation of turn-taking	33
4.1	Turn-taking dynamics across languages	34
4.2	Gender-specific variation in conversation	38
II	Study Overview	45
5	Studies and hypotheses	47
III	Analyses	53
6	Gender-specific turn-taking cues in German	55
6.1	Introduction	55
6.2	Gender variation in conversational behavior	57
6.3	Methods	58
6.3.1	Dataset	58
6.3.2	Annotation	59
6.3.3	Data extraction	62
6.4	Analysis and results	63
6.4.1	F0 values	64
6.4.2	Intensity	66

6.4.3	Segmental duration	67
6.4.4	Summary of the results	68
6.5	Discussion	70
6.6	Conclusions and outlook	72
7	Multimodal backchannels in German	75
7.1	Introduction	75
7.2	Methods	76
7.2.1	Dataset	76
7.2.2	Annotation	77
7.2.3	Phonetic features	78
7.3	Results	78
7.3.1	Head movements	79
7.3.2	Distribution	79
7.3.3	Phonetic variation	80
7.4	Discussion	81
7.5	Conclusions	82
8	Language-specific variation of turn-taking cues	85
8.1	Introduction	85
8.2	Methodology	86
8.2.1	Dataset	86
8.2.2	PTBs annotation	87
8.2.3	F0 points	88
8.3	Analysis and Results	88
8.4	Discussion	91
8.5	Conclusion	93
9	Gender-specific and multimodal variation in backchannel distribution	95
9.1	Introduction	95
9.2	Feedback in face-to-face conversation	96
9.3	Feedback and turn-taking cues	99
9.4	Backchannels and sociolinguistic variation	101
9.5	Objectives	104
9.6	Methods	106
9.6.1	Feedback annotation	107
9.6.2	Extraction of turn-taking cues	109
9.7	Data and analysis	111
9.8	Results: Backchannel timing	112
9.9	Results: Acoustic configurations preceding feedback	115
9.9.1	F0	116

9.9.2	Intensity	117
9.9.3	Speech rate	120
9.9.4	Voice quality	122
9.10	Discussion	122
9.11	Conclusions and outlook	129
10	Cross-linguistic and gender-specific variation of turn-taking dynamics	131
10.1	Introduction	131
10.2	Conversational dynamics across languages	132
10.3	Gender-related variation in speech	137
10.4	Previous comparisons of German and Swedish	139
10.5	Objectives	141
10.6	Methods	142
10.6.1	Datasets and comparability	143
10.6.2	Annotation	145
10.6.3	Labels preceding backchannels	146
10.7	Analysis and Results	148
10.7.1	Sequential structure distribution	148
10.7.2	Multimodal backchannel distribution	151
10.8	Discussion	154
10.8.1	Language-specific variation	156
10.8.2	Gender-specific variation	159
10.9	Conclusions and outlook	161
IV	Discussion	163
11	Summary of the results	165
12	General discussion	173
12.1	Gender and turn-taking dynamics	173
12.2	Listener behavior across modalities	177
12.3	Language-specific turn-taking dynamics	180
12.4	Potential limitations	186
12.4.1	Subjects	186
12.4.2	Data	187
12.4.3	Acoustic analysis	190
12.5	Outlook	191
13	Conclusions	193

V	Bibliography	195
VI	Appendix	219

List of Tables

6.1	Transition type (gaps, overlaps, no-gap-no-overlaps) distribution in same-gender (“MM” for male-to-male, “FF” for female-to-female) and mixed-gender (“FM” for female-to-male) dyads	63
6.2	Summary of the linear mixed model for F0 in speaker change cases. A significant three-way interaction is found between the value of the final F0 test location (End), the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(F0 \sim distance * gender * listener + (1 speaker))$	65
6.3	Summary of the linear mixed model for F0 in turn hold cases. A significant three-way interaction is found between the value of the penultimate F0 test location (-200), the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(F0 \sim distance * gender * listener + (1 speaker))$	65
6.4	Summary of the linear mixed model for intensity in speaker change cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(intensity \sim distance * gender * listener + (1 speaker))$	67
6.5	Summary of the linear mixed model for intensity in turn hold cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(intensity \sim distance * gender * listener + (1 speaker))$	67

6.6	Summary of the linear mixed model for segmental duration in speaker change cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $\text{lmer}(\text{duration} \sim \text{distance} * \text{gender} * \text{listener} + (1 \text{speaker}))$	69
7.1	Count of observations (and residuals) of the χ^2 test for the transition type relative to the feedback item (lexical, non-lexical, laughter, head movement).	79
9.1	Summary of the linear mixed model analyzing the variation of F0 in the speech preceding feedback: $f0\text{mean} \sim \text{listener} + \text{completeness} + \text{transitiontype} + (1 \text{listener}) + (1 \text{token}) + \text{listener} * \text{transitiontype}$	115
9.2	Summary of the linear mixed model analyzing the variation of intensity in the speech preceding feedback: $\text{intmean} \sim \text{speaker} + \text{modality} + \text{completeness} + \text{distance} + (1 \text{listener}) + (1 \text{token}) + \text{speaker} * \text{distance} + \text{transitiontype} * \text{distance}$	117
9.3	Summary of the linear mixed model analyzing the variation of speech rate in the speech preceding feedback: $\text{duration} \sim \text{listener} + \text{completeness} + \text{transitiontype} + \text{distance} + (1 \text{listener}) + (1 \text{token}) + \text{listener} * \text{transitiontype} + \text{completeness} * \text{distance} + \text{transitiontype} * \text{distance}$	120
9.4	Summary of the linear mixed model analyzing the variation of voice quality in the speech preceding feedback: $\text{cpp} \sim \text{completeness} + (1 \text{listener}) + (1 \text{token})$	122
10.1	Count of observations (and residuals) of the Chi-squared test for the transition type (<i>Gap</i> , <i>No-gap-no-overlap</i> , <i>Overlap</i>) relative to the sequential structure (<i>Backchannel</i> , <i>Change</i> , <i>Keep</i>) in German and in Swedish.	149
10.2	Count of observations (and residuals) of the Chi-squared test for the sequential structure (<i>Backchannel</i> , <i>Change</i> , <i>Keep</i>) relative to the dyad composition (<i>Female_Other</i> , <i>Female_Same</i> , <i>Male_Other</i> , <i>Male_Same</i>) in German and in Swedish.	150
10.3	Count of observations (and residuals) of the Chi-squared test for the transition type (<i>Gap</i> , <i>No-gap-no-overlap</i> , <i>Overlap</i>) relative to the dyad composition (<i>Female_Other</i> , <i>Female_Same</i> , <i>Male_Other</i> , <i>Male_Same</i>) in German and in Swedish.	151

List of Figures

6.1	Example of the annotation of PTLs in Praat, from one of the dialogues in the dataset. The first part of the label(s) refers to the completeness in context (“yes” or “no”), the following letter refers to the sentence type (“d” for declarative, “q” for questions, “t” for tag questions), the third part of the label refers to the sequential structure that followed (“c” for speaker change, “k” for keep and “b” for backchannel), and the final letter describes the transition type (“g” for gaps, “o” for overlaps, “n” for no-gap-no-overlaps).	61
6.2	F0 datapoints distribution (in st, above the speaker’s baseline) for speaker change and keep cases at the three test locations (“Distance”: -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other).	64
6.3	Intensity datapoints distribution (in dB, normalized to the speaker’s mean) for speaker change and keep cases at the three test locations (“Distance”: -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other).	66
6.4	Average segmental duration in ms for speaker change and keep cases over the three test locations (“Distance”: -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other).	68
7.1	Distribution of feedback items (lexical, non-lexical, laughter, head movements) with respect to the interlocutor’s current speaking turn (turn-external, turn-internal).	80
7.2	Maximum intensity (normalized with speaker’s mean) and CPP values for vocal feedback items (non-lexical, lexical, laughter).	80

8.1	Boxplots of normalized F0 for German speakers (SG01A, SG01B, SG02A, SG02B) and Swedish speakers (SS01A, SS01B, SS02A, SS02B) preceding a speaker change. Numbers “01” and “02” in the German and Swedish participants’ identification codes indicate the conversation in which they took part.	88
8.2	Boxplots of normalized F0 for German speakers (SG01A, SG01B, SG02A, SG02B) and Swedish speakers (SS01A, SS01B, SS02A, SS02B) preceding a keep. Numbers “01” and “02” in the German and Swedish participants’ identification codes indicate the conversation in which they took part.	89
9.1	Example of the FL (Feedback Location) annotation in Praat. In this example, speaker A has the floor. The 2nd tier contains the transcription of the utterance, and the 3rd tier the segmentation of the utterance into words. B is listening and producing backchannels in both the verbal and the visual modality, annotated on the 6th and 7th tier respectively. FL are labelled on the 4th tier, using a point label placed at the left boundary of the backchannel, in correspondence with the current utterance. The reported annotation does not include the syllable and segment tiers, nor the head movement annotation for speaker A.	108
9.2	Distribution of backchannels produced by either male or female listeners (<i>Listener</i>) in conversation with either male or female conversational partners (<i>Speaker</i>).	111
9.3	Frequency of listener responses in the different modalities (<i>Gestural, Verbal, Multimodal</i>) in our dataset produced either by a female listener (<i>Female</i> , Gestural = n. 48; Verbal = n. 78; Multimodal = n. 44) or a male listener (<i>Male</i> , Gestural = n. 50; Verbal = n. 74; Multimodal = n. 57).	112

9.4	Summary of the multinomial model including the transition types, <i>No-gap-no-overlap</i> , <i>Overlap</i> and <i>Gap</i> as the categorical outcomes and the completeness status of the FL (<i>Complete</i> or <i>Incomplete</i>), the modality of the backchannel, the gender of the listener (<i>Male</i> or <i>Female</i>), the gender of the interlocutor (<i>Same</i> or <i>Other</i>), the interactions between the gender of the listener and the gender of the speaker and between the gender of the listener and the modality of the feedback (<i>Gestural</i> , <i>Multimodal</i> or <i>Verbal</i>). The Odds Ratios on the x-axis represent the likelihood of transitioning to the category <i>No-gap-no-overlap</i> or <i>Overlap</i> compared to the reference category <i>Gap</i> for a one-unit increase in the corresponding predictor variables or variable interactions.	113
9.5	Effects plot of the multinomial model representing the likelihood of each transition type (<i>Gap</i> , <i>No-gap-no-overlap</i> , <i>Overlap</i>) occurring in relation to the modality of the backchannel (<i>Gestural</i> , <i>Multimodal</i> , <i>Verbal</i>).	113
9.6	Effects plot of the multinomial model representing the likelihood of each transition type (<i>Gap</i> , <i>No-gap-no-overlap</i> , <i>Overlap</i>) occurring in relation to the interaction between the gender of the listener (<i>Male</i> , <i>Female</i>) and that of the interlocutor (<i>Same</i> , <i>Other</i>).	114
9.7	Estimated marginal means plot for the variation of F0 in the post-hoc contrast for the interaction between the gender of the listener (<i>Female</i> , <i>Male</i>) and the transition type of the feedback (<i>No-gap-no-overlap</i> , <i>Gap</i> , <i>Overlap</i>).	116
9.8	Mean intensity values (in dB, normalized with the speaker's individual mean) in the speech preceding backchannels across modalities (<i>Verbal</i> , <i>Gestural</i> , <i>Multimodal</i>).	118
9.9	Intensity values' (in dB, normalized with the speaker's individual mean) variation over time approaching the FL in same-gender (<i>Same</i>) and mixed-gender (<i>Other</i>) dialogues.	119
9.10	Intensity values' (in dB, normalized with the speaker's individual mean) variation over time approaching FLs with different timings in relation to the current speaker's utterance (<i>No-gap-no-overlap</i> , <i>Gap</i> , <i>Overlap</i>).	119

9.11	The plot visualizes the relationship between the temporal distance from the FL and the variation of speech rate (operationalized as syllable duration in ms, normalized with the individual speaker's mean) before the different transition types (<i>No-gap-no-overlap</i> , <i>Gap</i> , <i>Overlap</i>) in the male and female listener groups. The lines are the result of a linear regression model (lm) fit to the data visualizing the overall trend in the data.	121
9.12	Speech rate's (operationalized as syllable duration in ms, normalized with the individual speaker's mean) variation over time approaching FLs with different syntactic/semantic completeness status (<i>Complete</i> , <i>Incomplete</i>).	122
10.1	Example of the turn-taking annotation of speaker "A" in one of the German dialogue participants in Praat. In this portion of the dialogue, speaker A is talking (annotated on the first tier) and speaker B (fifth tier) is listening. The 2nd tier contains the transcription of A's utterances, and the third tier the segmentation of the utterance into words (syllables and segments are not reported). B's contributions in this portions are two multimodal backchannels, from which the verbal and the visual component are annotated on the sixth and seventh tier respectively. The turn-taking labels are placed on the fourth tier, using a point label, placed either at the offset of speech for turn keeps, or where backchannels (and speaker changes) arose.	146
10.2	Distribution of the sequential structures (backchannels, speaker changes, turn keeps) in the German and Swedish speakers.	149
10.3	Frequency of the forms of verbal backchannels—combinations (<i>com</i>), complex lexical phrases (<i>cplx</i>), disyllabic lexical items (<i>dlx</i>), monosyllabic lexical items (<i>mlx</i>), non lexical expressions (<i>non</i>) and repetitions (<i>rep</i>)—in the German and in the Swedish conversations.	152
10.4	Distribution of gestural, multimodal and verbal backchannels based on their transition time (<i>Gap</i> , <i>No-gap-no-overlap</i> and <i>Overlap</i>) in the German and in the Swedish dialogues.	153

10.5	Summary of the multinomial model for German, including the transition types, <i>No-gap-no-overlap</i> , <i>Overlap</i> and <i>Gap</i> as the categorical outcomes and the completeness status of the label (<i>Complete</i> or <i>Incomplete</i>), the modality of the backchannel (<i>Gestural</i> , <i>Verbal</i> , <i>Multimodal</i>), and the dyad composition (<i>Female_Other</i> , <i>Female_Same</i> , <i>Male_Other</i> , <i>Male_Same</i>). The Odds Ratios on the x-axis indicate the likelihood of transitioning to the category <i>No-gap-no-overlap</i> or <i>Overlap</i> compared to the reference category <i>Gap</i> for a one-unit increase in the corresponding predictor variables.	153
10.6	Summary of the multinomial model for Swedish, including the transition types, <i>No-gap-no-overlap</i> , <i>Overlap</i> and <i>Gap</i> as the categorical outcomes and the completeness status of the label (<i>Complete</i> or <i>Incomplete</i>), the modality of the backchannel (<i>Gestural</i> , <i>Verbal</i> , <i>Multimodal</i>), and the dyad composition (<i>Female_Other</i> , <i>Female_Same</i> , <i>Male_Other</i> , <i>Male_Same</i>). The Odds Ratios on the x-axis indicate the likelihood of transitioning to the category <i>No-gap-no-overlap</i> or <i>Overlap</i> compared to the reference category <i>Gap</i> for a one-unit increase in the corresponding predictor variables.	154
10.7	Effects plots of the multinomial models for German and for Swedish representing the likelihood of the transition types (<i>Gap</i> , <i>No-gap-no-overlap</i> , <i>Overlap</i>) of backchannels occurring in relation to the dyad composition (<i>Female_Other</i> , <i>Female_Same</i> , <i>Male_Other</i> , <i>Male_Same</i>). .	155

Chapter 1

Introduction

Let's imagine an informal, face-to-face conversation between two people —a very casual, spontaneous exchange. In an imaginary prototypical two-party interaction, the first thing that can be observed is that the participants take turns at talk. Since this is an informal conversational exchange between peers, it is likely that the turns at talk will be more or less evenly distributed between the two. While one person is talking, the non-speaking interlocutor —the listener —does not simultaneously talk as well, but produces backchannels, i.e., short and unobtrusive attention signals to display attentiveness and understanding of what is being said. Then, when the speaker is finished with what they are saying, the floor passes to the other interlocutor without any apparent effort. Notably, the turn transition between interlocutors, in this prototypical exchange, occurs without any disruptive silent pauses or stretches of overlapped speech. However, in a real-life two-party interaction things might go slightly differently. In a real-life dialogue, in fact, while some of the main features from the previous imaginary situation still remain —speakers take turns and talk one at a time —it can very often happen that, for instance, one speaker holds the floor for much longer than the other, the listener produces very loud and salient attention signals and, before the current speaker has finished talking, the interlocutor starts with their own speech, so that the two audibly overlap with each other for a portion of their turns. However, in spite of these apparent deviations from the prototypical conversation that was earlier described, the real-life interaction continues flowing and does not break down, and the two participants continue building their exchange, each with their own conversational style. Additionally, not only do these deviations not cause major impairments in a cooperative and friendly exchange; they can also be particularly informative. In fact, interlocutors, both as

speakers and as listeners, rely on conversational behavior, together with other social acts, to perform specific aspects of their social and cultural identities, based on the situations in which they are involved, on their communicative goals, and on the communicative resources that they have available.

Overall, the main takeaway from this example should be that we are all well aware that, even in the most informal conversational exchanges, speakers do not *just speak*, and listeners do not *just listen*. In other words, speakers do not start talking at random, and listeners do not passively receive information, but rather they both do interactional work and implement specific turn-taking dynamics in order to 1) achieve conversational goals, and 2) convey their identities through interaction.

The argument at the base of this thesis is that the observation that interlocutors in conversation do not behave randomly has two co-existing explanations. The first one can be defined as being linguistic in nature, while the second one can be identified as social or cultural.

As for the first explanation, interlocutors involved in an interaction are able to signal and make predictions about what will happen next. More specifically, by relying on a complex set of turn-taking cues, current speakers are able to project their intentions to either cede the floor to the listener, or continue speaking; listeners, on their side, by closely monitoring the speech by the current speakers, are able to make early decisions about their next conversational actions, i.e., to either prepare their turn to speak, to remain silent and let the other continue and, in that case, possibly signaling, verbally or non-verbally, that they are listening and have no intentions of taking up the next turn. That is, interlocutors rely on turn-taking norms to carry out mostly smooth and cooperative exchanges. Which conversational action is most likely to arise after which combinations of cues, and with what timing, is one of the main focuses of the major studies on turn-taking.

The second explanation at the base of the observation that interlocutors in conversation do not behave randomly is that the turn-taking dynamics that they adopt in conversation are closely tied to their identities as speakers of a specific language, or members of a culture, community, age group, social group simultaneously. That is, through interaction with others and through their conversational behavior within a specific situation, interlocutors carefully perform, construct, deconstruct and reconstruct their social identities, both as speakers and as listeners. In particular, this

thesis will seek to demonstrate that speakers' and listeners' conversational actions are systematically influenced and shaped by the performance of the social identity of the speaker/listener, with a specific focus on the performance of gender, in interaction with the gender of the interlocutor and with the language and related culture. This sociolinguistic approach to the analysis of turn-taking is still relatively under-investigated, especially through the use of quantitative methods based on empirical data. Therefore, this dissertation addresses this research gap by presenting a series of studies that tackle turn-taking in spontaneous interactions from different perspectives, combining both linguistic and sociolinguistic factors in the analysis and in the interpretation of conversational exchanges.

The significance of the type of approach taken in this thesis stems from the basic observation that conversation "does not occur in a vacuum, it takes place in cultural contexts, that is, a system of norms and values" (Myers and Myers, 1976, p. 215); moreover, spontaneous verbal interaction is the main setting where language is used on a daily basis, as well as the core of social life (Hamilton and Holler, 2023; Levinson and Torreira, 2015; Stivers et al., 2009), and language and communication are intrinsically social (Goodwin, 1986; Knight, 2011). This means that turn-taking dynamics and conversational behaviors, that is, how we manage speakership and listenership in interaction, must have a fundamental social component (Beňuš, 2009) that should be observed, analyzed, measured, across different conversational contexts. The contribution of this thesis will be to define the extent to which observed turn-taking strategies, including the implementation of both verbal and non-verbal behavior (i.e., multimodal behavior, with a focus on the signals provided by the listener in conversation) can be connected to either linguistic factors or to the performance of the social identities of the speakers in relation to their genders. Additionally, this work will take a cross-linguistic and cross-cultural approach in the investigation of turn-taking by comparing German and Swedish. In spite of being typologically related languages and relatively close in terms of their societal structure and cultural norms, German and Swedish display language-specific features, as well as gender- and culture-specific tendencies that might influence the dynamics of turn-taking. Thus, the current dissertation also tests the ability of conversational turn-taking to function as a tool to detect language-specific and sociolinguistic vari-

ation, even when dealing with fine-grained and small-scale variability, and to study identity performance through speech adopting a quantitative approach.

The thesis is structured as follows: Chapter 2 is dedicated to the description of turn-taking and the fundamental concepts related to the structure of conversational turns and the processes regulating the alternation between speakership and listenership in interaction. Chapter 3 describes the turn-taking cues that are employed by speakers to manage turn transition and the use of backchannels by the listeners to confirm attentiveness without taking the floor. Chapter 4 reviews the literature on sociolinguistic variation in conversational behavior and sociophonetic variation in speech with a specific focus on culture-specific tendencies in turn-taking, and on the performance of gender through interaction and through patterns of acoustic variation. Chapter 5 reports the central thesis, the research questions and the hypotheses that will be investigated in the dissertation. Chapter 6 through Chapter 10 report the empirical analyses testing the influence of gender, language and the availability of multiple modalities of communication on different turn-taking aspects. Chapter 11 provides a detailed summary of the results obtained in the empirical section and confronts them with the hypotheses formulated in Chapter 5. Chapter 12 discusses the results, answers the research questions, and addresses the potential limitations of the analyses and the data used, also keeping in mind potential future developments in the investigation. Chapter 13 concludes the thesis by providing a brief overview of the findings obtained and the contributions of the work to the study of turn-taking.

Part I

Background

Chapter 2

Turn-taking in conversation

Everyday conversation has been defined as “the prime ecological niche for language” (Levinson and Torreira, 2015, p. 1), “the primary, fundamental embodiment of sociality” (Schegloff, 2020, p. 70) and the “core matrix for human social life” (Stivers et al., 2009, p. 10587). Informal interaction represents a staple of human life in society, with individuals being involved in face to face or mediated conversations for around 2 to 3 hours of speech on a daily basis, producing an average of 1200 talk spurts (Levinson and Torreira, 2015; Mehl et al., 2007).

Even in the most casual type of exchanges it is clear that the people involved do not speak at random, and that there should be a mechanism in action for which, generally, one speaker talks at a time, while the other(s) listens, produces backchannels (i.e., non-disruptive and non-turn-competitive listener responses), and waits for the current speaker to be finished talking to contribute themselves with more talk in a timely manner (e.g., Sacks et al., 1974). The alternation of speakership and listenership between interlocutors in conversation is regulated by the turn-taking mechanism. In other words, turn-taking defines who speaks and when, and its organization represents one of the foundations of all kinds of conversational exchanges.

Even though an alternation between speakership and listenership can be observed across all types of interactions that are part of social life, such as, for example, interviews in professional or media contexts, medical consultations, courtroom proceedings, and classroom discussions, it is precisely in informal, spontaneous interactions that it can be studied in its most intricate and complex form: in fact, even though such exchanges might appear as the most basic type, they are not constrained in terms of who speaks when and the actions specific participants are allowed to take, such as posing questions and providing answers (Gardner, 2004).

The course and the structure of informal interactions are determined by the speakers involved, by the context in which the exchange takes place, and by the conversational goal of the interaction. In fact, while turn-taking is defined as a system of rules, or norms —terms adopted in the current dissertation as well —that regulates the alternation between speakership and listenership (Sacks et al., 1974), it is not to be intended as providing absolute directions that are followed mechanically by interlocutors (Steensig, 2011) on who is to speak next, the length or the format of turns; in fact, turn-taking is interactionally managed by conversational participants, who determine floor transition by interpreting each other's local behavior (Sacks et al., 1974). In other words, turn-taking rules do not function like an external discussion leader selecting the next speaker (Baumann, 2013); rather, they identify systematic and recurrent patterns of behavior that are observed in conversation. With “turn-taking dynamics” the current dissertation refers to the structural, temporal and sequential aspects of those patterns of behavior in spontaneous dialogues, such as the duration of turns, the timing of turn transitions and the distribution of the turn-taking actions (turn-yields, turn keeps, backchannels) among conversational participants.

Across languages, turn-taking stands out as a universal mechanism, while simultaneously possessing the remarkable capacity to adapt in terms of both organization and dynamics, adjusting to the specific social and contextual features of each conversational situation. According to Sacks et al. (1974, p. 699), the system of turn-taking possesses fundamental “context-free” dynamics, while at the same time displaying exceptional “context-sensitive” flexibility. Thanks to these two fundamental aspects, turn-taking allows for the investigation of conversation both from a linguistic, as well as a socio-cultural perspective, not only to describe, e.g., the temporal features of the alternation between speakership and listenership, but also to investigate how and why different interlocutors behave in specific interactional contexts. In fact, beyond linguistics or sociolinguistics, turn-taking dynamics have been —and continue to be —a major source of interest for a number of scientific fields, such as anthropology, psychology or sociology (Steensig, 2011).

A foundational description of turn-taking in conversation was provided by Sacks et al. (1974)'s seminal paper “A Simplest Systematics for the Organization of Turn-Taking for Conversation”. In their work, Sacks and colleagues describe turn-taking

as being fundamentally defined by a turn-constructive component and by turn-allocation rules that are managed interactionally by the conversational participants. The turn-constructive component of the system defines the units that constitute turns, i.e., the turn-constructive units (TCUs). TCUs do not have a predefined size, and can be composed by single words, phrases, simple clauses or sentences that constitute syntactically, pragmatically or prosodically complete units, i.e., they are perceived as complete in context. In other words, as defined in Ford et al. (1996), a unit, or an utterance, can be considered as complete if it can be interpreted as such in context. Similarly, Meteor et al. (1995)'s unit definition identifies a syntactic sentence, or any construction below sentence-level, as complete if the hearer interprets it as such in the context where it occurs. The final boundary of a TCU arises when it reaches a completion point. Conversational turns are formed by one or multiple consecutive TCUs, at the end of which a Transition Relevance Place (TRP) arises (Sacks et al., 1974).

In the studies within this thesis, the units of turn-taking under consideration theoretically correspond to TCUs. Following Hough et al. (2016), who base their definition on Meteor et al. (1995), these units will also be referred to as "utterances," which may be part of longer conversational turns containing multiple utterances. Similarly, as described by Hough et al. (2016), "turns," comprising multiple utterances, are defined as a continuous stretch of speech by the same speaker until the floor is passed to the interlocutor.

Located at the end of TCUs, or at the end of an utterance, TRPs are defined as the locations in speech where the turn-allocation rules come into play, and thus speaker change becomes possible, though not mandatory (Sacks et al., 1974). Such locations within conversational turns have been addressed with different terminologies, such as Potential Transition Locations (PTLs; Schiffrin, 1987) or Potential Turn Boundaries (PTBs; Zellers, 2017).

At TRPs, according to Sacks et al. (1974), once one speaker has finished talking, they may directly select the next speaker or, if not, the next speaker may self-select. The alternative to the current speaker selecting the next at TRPs is not solely the interlocutor selecting themselves. If this were the case, speaker change would occur at the end of each single TCU. However, as Steensig (2011) points out, this is rarely the case in natural conversations, where turns are often longer and consist of

uninterrupted sequences of utterances. While speaker change might be considered a default option in conversation, both the current speaker and listener must engage in interactional work to maintain the floor at the TRP and properly signal that the channel remains open for continuation (e.g., Levinson and Torreira, 2015; Local and Walker, 2012; Sacks et al., 1974; Zellers, 2017).

In fact, generally, these turn-taking norms are collaboratively implemented by both the interlocutors involved in the interaction, with the common goal of achieving a smooth exchange of turns. To achieve the close timing that verbal exchanges normally exhibit, in fact, both the speaker and the listener, i.e., the momentarily non-speaking interlocutor, must pay close attention, respectively, to what they are saying, and to what is being said in the current turn, at all times, in order to take up the floor at the appropriate moment, or to remain silent and let the current speaker continue (Levinson and Torreira, 2015; Steensig, 2011). That is, smooth dialogues consist of the joint conversational efforts of both the speaker and the listener. On one hand, current speakers should signal their conversational intentions for the next utterance or the next turn, while the current listener should be able to respond in a timely manner, in order to keep silent gaps and speech overlaps reduced to a minimum (Sacks et al., 1974).

While Sacks et al. (1974) do not quantify the extent of such minimization, i.e., precise measurements of the length of naturally occurring overlaps or silent gaps in conversation are not provided, empirical explorations suggest that the majority of gaps between turns across different languages worldwide have a duration of around 200 milliseconds (e.g., Heldner and Edlund, 2010; Norwine and Murphy, 1938; Stivers et al., 2009; Strömbergsson et al., 2013). This data is particularly striking for its implications for the mechanisms regulating turn-taking. In fact, results from cognitive studies indicate that approximately 600 milliseconds (ms) are necessary, to the language production system, to encode a single word (Indefrey, 2011; Indefrey and Levelt, 2004; Schnur et al., 2006), and around 1500 milliseconds to encode a simple short clause (Griffin and Bock, 2000). The discrepancy between the language production system's latencies and the average silences observed between turns suggest that interlocutors, by closely monitoring the speech within the current turn, should be able to predict an upcoming turn yield and, in that case, start to prepare their speech well before the offset of the utterance by the current speaker.

Levinson and Torreira (2015) suggest that the cognitive processes for turn-taking involve early predictions of the speech act of the current utterance, enabling the interlocutor to initiate speech planning as soon as possible, and the use of cues, approximately 500 ms before the end of the turn—a time span that Zellers (2017) has proposed extends up to 1 second—during which the interlocutor monitors speech for turn-final cues anticipating the actual conclusion of the turn. Barthel et al. (2017) confirm this model by demonstrating, through the analysis of speech onset latencies and eye movements, that the content planning of the response begins as soon as the message of the current utterance can be understood, while response initiation is held until specific turn-final cues are recognized in speech. Therefore, smooth turn-taking is achieved based on the prediction ability of the interlocutor in comprehension by projecting the intentions of the current speaker in order to anticipate when to start speaking at the most appropriate time (Barthel et al., 2017; Garrod and Pickering, 2015; Levinson and Torreira, 2015; Sacks et al., 1974; Ward, 2019).

The presence, at turn ends, of syntactic, phonetic and gestural turn-taking cues, and the suggestion that the interlocutor would react to these cues by starting to speak when they indicate that the current speaker has finished or is about to complete their turn was first formulated by Duncan (1972, 1973, 1974) and Duncan and Fiske (1977). Specifically, they proposed that speakers exhibit various behavioral cues at turn endings, including changes in intonation, gestures, vocalizations, and completion of grammatical clauses, and that the likelihood of an interlocutor initiating a turn-taking attempt increases proportionally with the number of cues displayed. While this theory has faced criticism for its lack of formal, objective cue descriptions and the weak results from the statistical analysis (Gravano and Hirschberg, 2011; Levinson and Torreira, 2015), the identification of the possibility of a complex set of turn-taking signals was fundamental for the following research on conversation (the use of turn-taking cues for turn management will be addressed in detail in Section 3.1).

Studies supporting a “composite” view (Steensig, 2011, p. 511) for turn-taking organization, in line with the observations from Levinson and Torreira (2015) and Barthel et al. (2017), suggest that both mechanisms are active in regulating turn-taking in conversation, i.e., prediction mechanisms based on structural and contextual features enable the interlocutor to anticipate the possible upcoming conclusion

of the current speaker's turn, while the variation of turn-taking cues marking PTLs support the prediction about the possible upcoming completion of the turn (Barthel et al., 2017; Levinson and Torreira, 2015; Skantze, 2021; Steensig, 2011). In other words, this view of turn-taking can be summarized as follows: interlocutors in an interaction are able to make predictions about the approach of a TRP and about the conversational intentions of the current speaker for it by relying on different turn-taking cues, which also allow them to time their responses. The signals in the speech by the current speaker, as well as the active role of the listener, in either taking up the turn or allowing for continuation, are both fundamental components of the turn-taking mechanism in conversation.

Chapter 3

Speaking and listening in interaction

The present chapter focuses on describing the role of current speakers and that of current listeners within the turn-taking dynamics of spontaneous face-to-face dialogues. The smooth flow of the conversation is, in fact, a joint effort by both participants in the dialogical exchange, with one interlocutor being occupied in “speaking activities” and the other in “listening activities” (Yngve, 1970, p. 568), depending on whether they have the floor or not.

As the following Sections (3.1 and 3.2) will describe in detail, on one hand, the interlocutor that has the turn, i.e., the current speaker, manages the floor by signaling their conversational intentions to either cede the turn or to continue speaking, and marks such intentions through the use of turn-taking cues in the speech preceding a TRP. On the other hand, the non-speaking interlocutor, i.e., the current listener, is responsible for monitoring the speech by the current speaker and, optionally, signaling active attention and grounding through multimodal resources by producing backchannels. Specifically, Section 3.1 is organized as follows: first, turn-taking cues projecting turn-yielding and turn-holding intentions will be described, with a specific focus on acoustic cues. This is followed by a description of the language-specific use of acoustic cues for turn-taking purposes in German and Swedish (3.1.1). Section 3.2 describes the general features of backchannels in conversation, followed by a more detailed report on verbal, gestural and multimodal ones. Then, the final part of the section addresses their distribution in conversation and the acoustic forms of the speech preceding a backchannel within the current speaker’s turn (3.2.1).

3.1 Speaking and turn-taking cues

A number of communicative means have been found to be used by speakers to mark their conversational intentions at potential turn boundaries in conversation: the turn-taking-related variation of gestural, linguistic and phonetic resources in face-to-face conversation contribute to floor management by the current speaker. In fact, when the interlocutors are able to see each other, they rely significantly on both the visual and the verbal modalities of communication to coordinate their conversational actions.

Among the gestural means used in face-to-face interactions, manual gestures, gaze direction, head and eyebrow movements have been investigated in relation to turn-taking. For instance, Duncan (1972) finds that ending or relaxing hand gestures co-occur with turn-yields, while ongoing gestures are associated with turn-holds. Similarly, Kendrick et al. (2023) report that the beginning or continuation of a gesture, as well as averting gaze (also in Kendon, 1967 and Duncan, 1972), inhibit speaker change. Zellers et al. (2019) observe a higher frequency of hand gestures at TRPs followed by turn keeps, i.e., the current speaker produces a full turn and then continues speaking, than by speaker changes, i.e., the current speaker produces a full turn and then the other interlocutor takes the conversational floor. Kendon (1967) and Duncan (1972) find that gazing at the interlocutor is treated as a turn-yielding and response-requesting signal. Moreover, even though head movements are more frequent as part of listener behavior (Danner et al., 2021), they have also been observed to be used as a speakership request (Hadar et al., 1983) and in correspondence with head movements by the interlocutor (McClave, 2000), while eyebrow movements have been reported to occur more frequently in the vicinity of the beginning of a speaker's turn than in other locations of the conversation (Guaitella et al., 2009; Nota et al., 2021). Both head and eyebrow movements' density increase at speaker changes accompanied by overlapped speech (Danner et al., 2021). In the present work, the use of gestures will be taken into account from the perspective of the listener involved in the conversation (Section 3.2); the rest of the analyses comprising this thesis will place their focus on the linguistic and phonetic variation implemented by the speaker for turn-taking purposes.

Linguistic cues refer to the form and the syntactic/semantic completion of the ut-

terance in context. Syntactic completeness has been assigned a fundamental role to signal turn-yielding intentions (e.g., Bögels and Torreira, 2015; Duncan, 1972; Ford et al., 1996; Gravano and Hirschberg, 2011; Hjalmarsson, 2011; Sacks et al., 1974; Selting, 1996). However, the variation of acoustic cues appears to be just as decisive in influencing the turn-taking outcome following the boundary. In fact, the acoustic configuration of speech preceding a TRP contribute to inform the early predictions of the listener about the conversational intentions of the current speaker (Levinson and Torreira, 2015). Experimental studies have detected turn-taking related variation of acoustic cues as early as 500 milliseconds or even 1 second before turn ends (Gravano and Hirschberg, 2009b, 2011; Levinson and Torreira, 2015; Zellers, 2017).

The acoustic cues that have been observed to function as turn-taking cues towards TRPs are fundamental frequency (F0), intensity, speech rate, phonetic reduction and voice quality.

F0 is determined by the cycles per second of the periodic wave created by the vibration of the vocal folds, and is correlated with the perceived voice pitch (Johnson, 2003; Ladefoged and Johnson, 2006). When F0 is higher, perceived pitch is higher, and vice versa. The patterns of variation in the pitch of the voice, in combination with the lexical-syntactic context, determine intonation (Ladefoged and Johnson, 2006). While F0 is commonly measured in Hertz (Hz), i.e., a linear scale, the use of logarithmic scales, such as semitones (st) has been found to be more appropriate to account for the non-linear human perception of pitch differences (Hirst and de Looze, 2021; Simpson, 2009). This is particularly significant when the analysis of F0 involves speakers' gender. Due to the different sizes of men's and women's vocal folds, men's mean F0 tends to be lower than women's mean F0 (Fant, 1966; Simpson, 2009; Titze, 1989). Pitch movements that are perceived as equal, are actually physically different (Simpson, 2009), in that the higher the mean F0, the larger the F0 range that has to be used to achieve the same perceived pitch movement (Hirst and de Looze, 2021; Pépiot, 2014; Simpson, 2009). Thus, pitch movements measured in Hz will be larger for a female voice than for a male voice (Hirst and de Looze, 2021). However, the use of a logarithmic scale like semitones, in which a doubling Hz frequency (e.g., 150-300 Hz, 100-200 Hz) is represented by an equal st unit, "not only provides us with a better approximation to the way in which we perceive male and female pitch, it is important to provide an appropriate representation of male and

female pitch for a balanced discussion of the speech differences between the sexes” (Simpson, 2009, p. 624).

Acoustic intensity is correlated with the perceived voice loudness and is measured in decibels (dB), a logarithmic scale that, like the semitone scale for pitch, provides an approximation of humans’ non-linear perception of the loudness sensation (Johnson, 2003). Speech rate refers to the perceived speech tempo and can be measured by counting production units, such as phones or phonemes, syllables, morae or words, per unit time (Koreman, 2006), or also by relying on the average duration, e.g., in seconds or milliseconds, of phonetic segments or that of syllables, with higher duration correlating with the perception of slower speech, and vice versa (Plug et al., 2022a). Segmental reduction and deletion are, however, pervasive phenomena in natural speech (Ernestus, 2014; Plug et al., 2022b) and Koreman (2006) raised the question whether listeners rely on both surface forms (i.e., actually produced phones and syllables) or canonical forms (i.e., expected canonical forms connected to underlying representations in the mental lexicon) in speech rate perception. While his findings indicate that listeners rely on both forms, Plug et al. (2022a) present more nuanced results by indicating that, when syllables are taken as the unit of measurement, listeners used both surface and canonical forms to assess speech rate variations; when phones were taken into account, they find that listeners’ judgment was influenced only by the segments that were actually produced. Moreover, Plug et al. (2022a) report that mean F0 and mean intensity values also influenced speech rate perception, with higher F0 and louder speech being associated with the perception of faster speech tempo, confirming previous studies on the co-variation of these acoustic parameters (e.g., Kohler, 1986; Rietveld and Gussenhoven, 1987).

Finally, voice quality indicates the type of phonation determined by the degree of glottal aperture, i.e., creak, modal voice, breathy voice, harsh voice, falsetto (Eckert and Laver, 1994; Ladefoged and Johnson, 2006; Laver, 1980). Different acoustic features have been used to measure the perceived quality of the voice, such as the joint observation of jitter, shimmer, noise-to-harmonics ratio (NHR) and harmonic-to-noise ratio (HNR) (Bhuta et al., 2004; Susana Finger et al., 2009; Teixeira et al., 2013), and, more recently, the parameter of Cepstral Peak Prominence (CPP) (Fraile Muñoz and Godino Llorente, 2014; Hillenbrand and Houde, 1996). Jitter and shim-

mer are measure of F0 disturbance in terms of, respectively, frequency variation from cycle to cycle and amplitude variation of the waveform (Teixeira et al., 2013); NHR and HNR are inversely proportional values to measure the presence of noise components in a voice signal, with lower NHR and a higher HNR associated with a more modal phonation (Susana Finger et al., 2009; Teixeira et al., 2013). CPP is a measure extracted from the cepstrum of a signal, i.e., the log power spectrum of a log power spectrum. As defined by Hillenbrand and Houde (1996), in a periodic signal, the first power spectrum represents the frequency distribution of the energy of the signal, while the spectrum of the spectrum (the cepstrum) represents the regularity of the harmonic peaks; the cepstral peak is the dominant harmonic in the cepstrum, and its height, or prominence, in relation to a regression line across the overall cepstrum is the CPP (Hillenbrand and Houde, 1996; Murton et al., 2020). It is measured in dB and it has been found to be a more robust measurement for voice quality assessment and to be more reliable in continuous speech than the other parameters (Maryn et al., 2009; Murton et al., 2020).

Among turn-taking cues, the role of intonation has received particular attention in several investigations carried out in different languages. In general, high-rising or falling intonation approaching turn ends has been observed to be associated with speaker changes (Gravano and Hirschberg, 2009b, 2011; Koiso et al., 1998; Local et al., 1986; Niebuhr et al., 2013), while level F0, in the middle range of the individual speaker's values, towards the end of a turn have been reported to project turn-holding intentions (Duncan, 1972; Edlund and Heldner, 2005; Gravano and Hirschberg, 2009b, 2011; Koiso et al., 1998; Selting, 1996; Skantze, 2021).

Some studies have challenged the contribution of pitch variations to turn-taking, aiming to demonstrate the primacy of lexical-syntactic information over intonational features. For example, in their study comparing the relative roles of lexical syntactic structure and intonation in predicting turn endings, de Ruiter et al. (2006) conducted an online experiment in which subjects listened to segments of turns extracted from Dutch phone conversations and were requested to signal when the turn ended by pressing a button. The researchers manipulated either the lexical content or the intonation of the stimuli. Based on their results, they argue that the absence of an intonational contour did not hinder turn-end predictions and suggest that "lexico-syntactic structure is necessary (and possibly sufficient) for accurate end-of-turn

projection" (de Ruiter et al., 2006, p. 531). Furthermore, Keitel et al. (2013) also explored turn-end predictions using stimuli with natural or flattened intonation. They tested German adults and children by showing them videos of dialogues and monitoring their eye movements via eye-tracking. These movements served as indicators of predictions about when the turn was likely to end and to be passed to the interlocutor. The findings indicate that turns with natural intonations did not result in improved predictions about upcoming turn endings among adult participants.

However, the extreme claims of these studies in regard to the contribution of intonation, not only did not convince the successive literature to abandon the investigation of F0 as a turn-taking cues; their methodology emphasized the importance of the other acoustic cues that occur towards turn end and that, together with lexical information and F0 variations, should be considered in turn-taking studies. In fact, what de Ruiter et al. (2006) and Keitel et al. (2013) failed to take into account is that conversational turns are not structured around the variation of a single cue, but on a combination of cues relevant within the current interaction and the conversational participants involved (Local and Walker, 2005; Selting, 1996). Specifically, as Local and Walker (2012) and Bögels and Torreira (2015) point out, while de Ruiter et al. (2006) and Keitel et al. (2013)'s experimental materials excluded the information related to pitch, all the other acoustic cues commonly associated with turn-ends were left intact, such as phrase-final lengthening (e.g., Bogetic, 2009; Turk and Shattuck-Hufnagel, 2007; Zellers, 2017) and intensity decrease (e.g., Gravano and Hirschberg, 2009b, 2011; Niebuhr et al., 2013; Włodarczak and Heldner, 2022).

Investigating further into the non-pitch turn-taking cues, Local and Walker (2012) find that, at points of syntactic completion (SYNCOMPs) in British English conversation, the features that project more talk are: an avoidance of phrase-final lengthening, an anticipatory articulation of the following segmental content, continued voicing across the boundary, and vowel and consonant reduction. Observing American English, Gravano and Hirschberg (2009b, 2011) report that the cues that occur 500 milliseconds and 1 second before smooth speaker changes are: a falling or high-rising intonation at the end of the Inter Pausal Unit (i.e., a stretch of speech by the same speaker between two silences; IPU); a reduced lengthening of the final words in the IPU (i.e., a faster speech rate); lower intensity and lower pitch; a point of textual completion (i.e., completeness of the IPU in context, by taking into account

not only the syntactic form, but also its semantic meaning and intonation: Gravano and Hirschberg, 2011, p. 611); a higher value of the voice quality features of jitter, shimmer, and NHR; and an overall longer duration of the IPU. The likelihood of an interlocutor initiating a turn-taking attempt escalates linearly as multiple cues arise concurrently. On the other hand, to keep the floor, Gravano and Hirschberg, 2011's model indicates that the potential boundary should display the fewer floor-yielding cues as possible and include features such as a final flat intonation, high intensity and pitch levels and textual incompleteness.

3.1.1 Turn-taking cues in German and Swedish

Turn-taking-related variations of F0, intensity, duration, and voice quality have been reported in studies on different languages, including German and Swedish.

With the aim of investigating the interaction of syntax and prosody, Selting (1996) analyzed a corpus of North Western Standard German interactions and reports that, at the end of TCUs that are syntactically and semantically complete, a speaker's holding intonation characterized by a flat pitch, or a rush-through, i.e., a faster speech rate, are treated as turn-holding signals by the interlocutor. Furthermore, in line with previous findings by Schegloff (1987), who suggested that a syllable with a pitch peak and high intensity may precede a TRP, Selting (1996) observes that the final prominent syllable in a TCU appears to signal turn-yielding intentions. Specifically, it functions as what Barthel et al. (2017, p. 3) refer to as a "go-signal," prompting the interlocutor to either initiate their turn or provide a backchannel response. Niebuhr et al. (2013) explored the Kiel Corpus of Spontaneous Speech (Kohler et al., 2017), which contains dialogues in Northern Standard German. They conducted both production and perception experiments to examine the role of F0 movements, intensity, and segmental reduction in German turn-taking. In their production data, they found that segmental reduction is lower at turn-final boundaries compared to turn-internal ones, while intensity higher at turn-internal boundaries and steeply decreases at turn-final boundaries. In their perception study, native adult speakers were presented with yes/no questions with manipulated signals. Their results revealed that participants responded faster to the questions when the final rising F0 had a concave shape rather than a convex, the last final unstressed syllable showed a lower degree of reduction and the intensity was low. They observed that segmen-

tal reduction and intensity reduction play opposite roles in the turn-taking mechanism, with one serving as a turn-holding cue and the other as a turn-yielding cue. Köser (2014) specifically targets the contribution of voice quality changes, more in detail creaky voice, to turn-taking in German spontaneous two-party interactions. Building on Peters (2006)'s findings that creak in German appears to occur more frequently at turn-final boundaries than at turn-internal boundaries, together with a low-falling intonation, Köser (2014) reported that, in interaction, creaky voice appears to be correlated more frequently with turn changes, or backchannels. However, since she also observed it before turn keeps, she concluded that voice quality changes might be a weak turn-taking signal in German, and should be considered together with other turn-taking features, such as variations in F0.

In Swedish, similar results regarding the contribution of voice quality to turn-taking were reached by Włodarczak and Heldner (2022). They analyzed the variation in voice quality through the variation of the CPP values and, while they did find differences between turn-internal and turn-final boundaries, with the latter having lower CPP values, indicating a creaky or breathy phonation, the predictive power of voice quality in their model for turn-taking was significantly outperformed by F0 and intensity. Hjalmarsson (2011) investigated the additive effect of different turn-taking cues in Swedish dialogues involving both human participants and a synthetic voice by asking a group of raters to decide whether a certain token in the dialogue would be followed by a speaker change or not. The features that were most predictive for speaker change in her experiment were semantic completeness and a falling intonation, while a flat intonation and incompleteness were cues for turn-holds. Moreover, she reported finding no effect of phrase-final lengthening in the data and, in line with Edlund and Heldner (2005), also inconclusive results in regards to the contribution of rising intonation patterns in Swedish, found to be followed both by speaker changes and turn-holds. The turn-taking role of pre-boundary lengthening is observed by Hjalmarsson and Laskowski (2011). They manipulated Inter-Pausal Units' (IPUs) final lengthening by adding longer durations in turn-medial locations than in turn-final locations, which improved the performance of a dialogue system in speaker-independent next-speakership prediction.

Zellers (2017) aimed to test the contribution of different turn-taking cues in Swedish through perception and production experiments. She found that dura-

tional cues seem to play a more fundamental role in turn-taking in Swedish than F0 variation and voice quality. Specifically, in her perception experiment, she manipulated the length of the final unstressed syllable(s) of a dialogue portion, along with the F0 contour and final pitch height. Raters were asked to decide whether the stimulus was followed by a turn-hold or a speaker change. She found that the length of the final unstressed syllable(s) was the most important cue for predictions, with pre-boundary longer syllables being associated with continuation, a result confirmed by the analysis of production data in Swedish. Regarding speech rate, slower speech used to hold the turn, she discussed, does not necessarily contradict the practice of rush-through (e.g., Local and Walker, 2012; Selting, 1996) used to occupy the transition space between IPUs, as both methods serve to fill the transition space and signal turn hold intentions. Conversely, no significant patterns of variation in F0 were associated with specific turn-taking outcomes. Zellers (2017) connects this specific finding to the fact that, in Central Swedish, pitch movements are involved in the phonological system of the language, where content words can be distinguished by a two-way lexical pitch accent, and focused words or final words are marked with an additional high tone (Bruce, 1977; Féry, 2017; Gårding, 1989; Gussenhoven and Bruce, 1999). Thus, she hypothesizes, pitch “already bears a relatively heavy signaling load from the phonological system and taking advantage of other elements of the speech signal could help avoid saturation or confusion due to too many resources associated with pitch variation” (Zellers, 2017, p. 12).

This conclusion holds significance for cross-linguistic comparisons in turn-taking studies, especially when involving languages with different prosodic phonologies than Swedish, despite being typologically related. In a previous study, Zellers (2014) conducted a comparison between Swedish and Southern British English. She built a perception test containing stimuli manipulated for the duration of the final unstressed syllables, as well as the height of the final peak and the final pitch contour, and tested native speakers of both languages. Confirming her initial hypothesis, Zellers (2014) found that, due to the stronger role of intonation in English for marking conversational intentions, as well as the absence of tonal specifications for content words (unlike in Swedish), English speakers relied exclusively on pitch variations for turn-taking purposes. In contrast, Swedish speakers reacted more reliably to duration cues. Feindt et al. (2021) conducted a preliminary analysis

comparing turn-taking in German and Swedish, revealing similar cross-linguistic tendencies. German, like English, is an intonation language, where pitch is used to mark prominent syllables, and can be used to convey information about sentence and discourse structure. Additionally, boundary tones are observed at the end of intonation phrases. In languages such as German, speakers have the freedom to manipulate the intonation contours of pitch accents and boundary tones to convey prominence levels according to their informational status, as well as the pattern of individual tones for pragmatic purposes (Féry, 2017). On the other hand, while Swedish also uses boundary tones to mark the end of intonation phrases, their variation appears to be more restricted compared to German. In fact, low boundary tones (L%) are prevalent in Swedish, even marking the end of most questions (House, 2005). Feindt et al. (2021) report that the tonal specifications found in Swedish at the lexical level limits the freedom for pitch variation towards TRPs for turn-taking purposes. Their analysis of spontaneous dialogues in German and Swedish revealed a lower degree of F0 variations in Swedish compared to German, with F0 values in the Swedish dataset remaining close to the individual speakers' F0 baselines, while F0 values for German speakers tended to be higher and more variable. Building on these preliminary results, further investigation of the influence of cross-linguistic differences between German and Swedish in turn-taking-related variation of acoustic cues and the consequent influence on (multimodal) turn-taking dynamics will be part of the present work.

3.2 Listening and multimodal backchannels

For a long time, research on verbal interaction either considered listeners only marginally, or assumed them to be only passively waiting for their turn to speak (Bavelas et al., 2000). However, recently, the activity of listening in a conversation started to receive increased attention, being recognized as “a communicative act in its own right” (Erickson, 1986, p. 297). The listener is now often recognized as an active contributor to the construction and development of interaction, and the behavior of the non-speaking interlocutor(s) in conversation is being studied systematically in its different dimensions and in its patterns of within-language and cross-linguistic variation.

The term “backchannel” (Yngve, 1970) is used to refer to the signals produced by the listener while the current speaker is talking in order to display active listener-ship without interrupting and without claiming the turn (e.g., Ward and Tsukahara, 2000; Yngve, 1970). In face-to-face interactions, backchannels are multimodal, i.e., they arise both in the visual and in the vocal modality of communication. In fact, they can be short lexical (e.g., “yeah”, “okay”) and non-lexical (e.g., “mh-hm”, “uh-huh”, laughter) vocalizations, or gestures (e.g., head movements).

Besides being non-turn-competitive, further inherent features of backchannels are that they are optional, i.e., the listener may or may not decide to produce a backchannel response (thus short answers to questions are not considered as backchannels; Ozkan and Morency, 2013; Ward and Tsukahara, 2000), and that they tend to go unacknowledged in interaction, i.e., they are not actively acknowledged by the current speaker (in this sense, questions and turn-opening vocalizations are not considered as backchannels; Shelley and Gonzalez, 2013; Ward and Tsukahara, 2000). Besides “backchannels”, they have also been referred to with several other descriptive terms in the literature. For instance, they have been addressed as “listener feedback” or “supportive feedback” (Stubbe, 1998), “listener responses” (Dittmann and Llewellyn, 1968), “minimal responses” (Fellegly, 1995), “continuers” (Schegloff, 1982), “acknowledgment tokens” (Jefferson, 1984), “accompaniment signals” (Kendon, 1967) and “reactive tokens” (Clancy et al., 1996), *inter alia*.

Further labels for various types of backchannels have been assigned based on their perceived function within the interaction, even though it is professedly problematic to pinpoint this for each individual token, considering the vast range of purposes for which backchannels can be used (e.g., Gardner, 1997; Heritage, 1984). For instance, in her corpus analysis, Knight (2011, p. 90) adopts O’Keeffe and Adolphs (2008)’s approach, and labels different backchannels as “Continuers”, whose main functions are inviting the speaker to proceed and maintaining the flow of discourse; “Convergence Tokens”, reinforcing rapport between interlocutors; “Engaged Response Tokens”, signaling an emotive reaction; and “Information Receipt Tokens”, marking the close or shift of topic. Heinz (2003) provides a list of the different functions that have been assigned to backchannels in the literature: they can be used to display attention, agreement, involvement, acknowledgment, or also express dis-

agreement and lack of understanding. Backchannels can be also used by the listener to refuse the opportunity to speak when the floor is yielded to them by the speaker (Cutrone, 2005; Ward and Tsukahara, 2000). Backchannels with this particular function have been referred to as “perverse passive”, (Ward and Tsukahara, 2000, p. 1184; cf. Jefferson, 1984). Similarly, listeners have been observed to utilize backchannels to contribute to the conversation without assuming primary speaker-ship (Cutrone, 2005).

The next chapters will also refer to backchannels as listener responses, listener feedback, or simply as feedback. Since the identification and analysis of backchannels’ function is outside the scope of this dissertation, those labels will be treated as synonymous, without reference to any specific features or perceived purposes of the backchannel in the conversation.

In the verbal domain, the existence of these signals as a group of words with a function of their own in conversation, or, more specifically “whose main business was not at the level of sentence but at the level of conversational interaction” (Dingemanse, 2024, p. 261; Knudsen et al., 2020), was first reported in a corpus study on English by Fries (1952), that noticed that tokens such as “yes”, “yeah”, “good”, were used by the listener as “signals of continued attention” (1952, p. 49) without interrupting the speech by the current speaker. However, the term “backchannel” to describe these particular utterances in interaction was first used by Yngve (1970), in order to underline the contribution of the non-speaking party to the conversation. Specifically, Yngve (1970) claims that both interlocutors in the conversation are, at the same time, both engaged in the act of speaking and listening thanks to the existence of the “backchannel” of communication (p. 568), where listeners produce their responses, while the main channel is occupied by the current speaker’s turn. Similarly, Markel (1975) operationalize the conversational turn as the act of “solo talking” (p. 190) by one interlocutor. During this phase, the other interlocutor may engage in simultaneous speech, offering positive reinforcement without assuming control of the conversation from the current speaker, due to the presence of a “concurrent listening turn” (p. 190). Clark and Krych (2004) characterize conversations as an exchange where speakers attentively monitor both their own speech and the behavior of the interlocutor. On the other end, listeners “try to keep speakers informed of their current state of understanding” (Yngve, 1970, p. 66).

Therefore, despite backchannel production being by definition optional and not actively acknowledged by the speaking interlocutor, the listener's active contribution to the structure of turn-taking is fundamental for the smooth development of the conversation. The form, frequency, and placement of feedback within the current turn are not random and are significant for maintaining the flow of the exchange. Bavelas et al. (2000) report that in dyadic interactions, when the listener used fewer backchannels, especially those with more specific lexical forms directly addressing the content of the current turn, the speaker's storytelling was less effective and efficient, and their speech was less fluent as a result. Studies primarily focusing on backchannel production in L2s suggest that, although feedback is typically not consciously acknowledged by the current speaker, it becomes noticeable when either too frequent or too scarce, and it may even irritate the current speaker (e.g., Heinz, 2003; Krause-Ono, 2004; Mizutani, 1982). This tendency is attributed to the different frequency baselines that languages have regarding backchanneling in their cultural norms related to listener behavior in conversation, a topic that will be addressed in Chapter 4. Poppe et al. (2011) conducted a perception study controlling the frequency and placement of multimodal feedback by a virtual agent. They find that backchannel behavior is perceived as less human-like when backchannels are too frequent or too rare, as well as when they occur randomly within the current speaking turn, particularly concerning verbal backchannels. Therefore, what clearly emerges from the literature is that both speakers and listeners are very sensitive to the precise realization of backchannels, and their appropriate distribution is fundamental for the conversation (Wehrle et al., 2023).

As previously mentioned, listener behavior is multimodal and backchannels across modalities have different distributional properties. Feedback produced in the vocal modality can either comprise lexical items or short phrases (e.g., "ja", "genau", or "das stimmt" in German), or it can be constituted by non-lexical utterances such as "mh-hm". Wong and Peters (2007) elaborate a classification of the different lexical forms that were found for English, and identify three categories: minimal, lexical, and grammatical. Minimal backchannels are the simplest construction among the three, consisting of non-lexical items like "uhuh," "mhm," "oh," "mm," as well as interjections like "yes," "no," "yeah," and "nope," primarily conveying polarity. Lexical backchannels comprise single-word feedback, such

as “really,” “right,” and “good,” carrying more semantic weight. Finally, grammatical constructions involve more complex forms, including short fixed phrases like “I see” and “that’s true” (Wong and Peters, 2007). While these constructions can occur on their own as a backchannel response, often they appear combined with each other, or in clusters of repetitions of the same form (Tao and Thompson, 1991; Tottie, 1991). For instance, Tottie (1991) identifies doubled forms such as “mhm mhm”, “yeah yeah”, or complex forms like “mhm gee, yeah”, and “that’s right yeah yeah”.

Verbal utterances with a feedback function tend to be short and non-interrupting, comprising mono- or disyllabic expressions (Gardner, 2001; Müller, 1996). Languages across the world show a striking degree of regularity in the form of verbal backchannels. Liesenfeld and Dingemanse (2022) report that, in 16 languages from 8 different language families, a basic formal distinction appears to be between a minimal backchannel form, including a nasal, and more complex forms comprising vowels and consonants other than nasals (e.g., “mhm”, “aha” and “ja” for German; Liesenfeld and Dingemanse, 2022). Complex forms have been reported to be used as a signal that the listener, while not actively attempting to take the floor at that precise moment, intends to speak in the near future (Knight, 2011). They have been characterized as backchannels serving as turn requests (Dittmann and Llewellyn, 1968), or compared to “a raised hand in a classroom” (Oreström, 1983, as cited in Knight, 2011, p. 89). Both simple and complex forms usually appear within an interaction: since the repetition of the same verbal backchannels by the same listener over extended intervals in the conversation may indicate boredom and inattention, listeners tend to vary the forms of their verbal feedback in conversation (McCarthy, 2003; Schegloff, 1982).

In the vocal modality, feedback can also occur in the form of laughter (Gardner, 2001; Krepsz et al., 2022; Tottie, 1991; Trouvain and Truong, 2012). Laughs and speech-laughs are characterized by rhythmic bouts of syllables uttered during exhalations and separated by inhalation, and speech-laughs are further characterized by the simultaneous occurrence of speech (Ludusan and Wagner, 2022a). As reported in Provine (1993), contrary to the general belief, laughter in conversation occurs less frequently as a reaction to humorous attempts, but rather after generic utterances, indicating a wider range of social and interactive functions than simple amusement

signaling (Ludusan and Wagner, 2022b; Vettin and Todt, 2004). In line with the other vocal backchannel forms, laughter with a social function in conversation is used to signal attention and interest (Chapman and Foot, 1976), to build rapport between interlocutors and maintain a smooth flow of the conversation (Provine, 1993; Scott et al., 2014). Even though laughter in conversation may not be reduced in its prosodic form, as evidenced by higher F0 and intensity values compared to speech, and longer duration than other vocal backchannels (e.g., Krepsz et al., 2022; Ludusan and Wagner, 2022a; Vettin and Todt, 2004), perception studies frequently indicate that laughter in conversation often goes unnoticed by both the person producing it and the listener/rater (Vettin and Todt, 2004), as is often the case for other backchannels as well (Shelley and Gonzalez, 2013; Ward and Tsukahara, 2000).

In the visual modality, several types of listener behaviors have been identified as providing feedback to the interlocutor. These include shifts in posture (Bird-whistell, 1966), eyebrow movements, facial expressions, gaze patterns (Paggio and Navarretta, 2011), and eye blinking (Hömke et al., 2018). However, predominant among these visual cues is the occurrence of head movements (Cerrato, 2012; McClave, 2000; Paggio and Navarretta, 2011), either in isolation or concurrent with both verbal responses (Dittmann and Llewellyn, 1967) and laughter (Hadar et al., 1983). Head nods emerge as the most frequent head movement type produced by listeners (Malisz et al., 2016; Paggio and Navarretta, 2011), and are the most studied as well. However, several types of head movements have been described to occur either in conversation or, in general, in combination with speech. In their work providing a comprehensive paradigm for gesture annotation, Rohrer et al. (Rohrer et al., 2023; Rohrer et al., 2020), based on the previous work and classification provided by Wagner et al. (2014), report descriptions of five possible head movement forms that can be observed in speech: “nod”, indicating up and down movements of the head; “turn”, for left and right movement; “tilt”, for cases where the top of the head goes in one direction and the chin in the opposite one; “slide”, describing horizontal movements of the whole head left and right; and “protrusion”, indicating a movement of the whole head forward or backwards (Rohrer et al., 2023, pp. 22–23).

Similar to verbal tokens, head movements have been found to actively regulate turn-taking (McClave, 2000), to encourage the speaker to continue with their turn, as well as signaling understanding of what is being said (Cerrato and Skhiri, 2003).

Being produced on the visual channel, they are considered to be less disruptive than verbal responses for the current speaking turn (Dittmann and Llewellyn, 1968; Ferré and Renaudier, 2017). In fact, even when placed with higher degrees of random distribution, head movements used as feedback are more rarely deemed as inappropriate than verbal backchannels in Poppe et al. (2011)'s perception study. According to Hadar et al. (1983) and Knight (2011), the timing of the head movement with respect to the current speaker's speech is fundamental to disambiguate a movement with a feedback function and movements with other functions, such as conventionalized responses to polar questions for nods and turns, or floor-taking attempts. Moreover, Knight (2011), specifically referring to nods, indicates that their co-occurrence with a verbal form might also help identify a movement as a backchannel. Dittmann and Llewellyn (1968) found more backchannels in which the verbal element and the head movements were combined than they expected. They also observe that the onset of the gestural component usually precedes that of the vocal response, and the head movement often starts in anticipation of an upcoming TRPs, when the current speaker is still talking.

3.2.1 Backchannel distribution in conversation

Even though backchannels are less constrained than full-fledged turns within the turn-taking structure, and their occurrence in overlap with speech is not necessarily considered inappropriate (Roberts et al., 2015), listeners and speakers alike are highly sensitive to the distribution of feedback across modalities within the turn by the current speaker and in relation to its acoustic configuration. The majority of the studies that have taken into account the location of backchannels and the acoustic features in the preceding speech refer to such concepts as "backchannel relevance spaces" (Heldner et al., 2013) and "backchannel inviting cues" (Gravano and Hirschberg, 2009a, 2011; also "feedback inviting cues", e.g., Boudin et al., 2021; Skantze et al., 2014), respectively. In line with the concept of TRPs and the related turn-taking cues signaling the possibility of a speaker change, backchannel relevance spaces describe a location in speech where it would be appropriate for a listener to produce feedback, while backchannel inviting cues indicate the multimodal cues employed by the speaker to signal to the listener that the possibility of producing a response is approaching (Boudin et al., 2021; Heldner et al., 2013). While the

literature provides clear evidence, to be discussed below, that backchannels across modalities are carefully timed by listeners, and tend to follow specific configurations in conversation, the definitions of backchannel relevance spaces and backchannel inviting cues appear to move the focus from the listener's intentions and perceptions on to the current speakers action of, to some extent, requesting feedback while at the same time holding the turn and structuring their talk. However, it is fundamental to consider that backchannels are optional and not fully acknowledged by the speaker; they are produced at the discretion of the listener, based on their own perception of appropriateness in their location, as well as their frequency. Thus, in order to stress these properties of feedback, the present work will address the acoustic configurations that systematically precede the production and the patterns of distribution of backchannels (as well as their language-specific and/or gender-specific variation); however, here, like in Chapter 9, the specific terminology of "backchannel relevance spaces" and "backchannel inviting cues" will not be employed.

As for the distribution of feedback within the current speaking turn, Howes and Eshghi (2021) indicate that backchannels tend to occur at clause boundaries, after a syntactically complete unit. They can also arise before the boundary, in cases where the following content and the upcoming completion appears to be easily predictable (Howes and Eshghi, 2021). Dittmann and Llewellyn (1967) find that listeners in simulated phone conversations produce vocal feedback expressions after the conclusion of phonemic clauses (i.e., units of speech identifiable through pitch and stress units), rather than within them. In particular, they observe that the most backchannels occurred during pauses. In their following study, Dittmann and Llewellyn (1968) also include head nods and strongly confirm their former observation that the listeners time their responses to not interrupt the current speaker. Truong et al. (2011) analyze Dutch face-to-face dialogues and report that backchannels in the visual modality tend to be produced mostly in overlap with speech, while those in the verbal modality arise after the offset of speech, during a silence. Bimodal feedback (i.e., comprising both a verbal and a gestural component, referred to in the present dissertation as "multimodal") shares distributional features of both verbal and gestural backchannels, occurring both in overlap with speech and during silence (Truong et al., 2011). The distributional features of backchannels across modalities are confirmed by Ferré and Renaudier (2017) for English dia-

logues. Moreover they find that, for bimodal backchannels, the duration of the head movement component is longer than that of purely gestural ones, and that it tends to be initiated during the speech by the current speaker, while the vocal component arises after the offset of speech, in line with Dittmann and Llewellyn (1968)'s previous findings. Truong et al. (2011) also report that visual backchannels tend to be produced after high intensity profiles in the current speaker's turn, possibly corresponding to emphasized speech, while verbal ones generally follow a decreasing intensity profile, associated with a terminating utterance preceding a silence.

While Truong et al. (2011) do not find any configurations related to the variation of F0 preceding feedback, other studies attribute a central role to it. For instance, Levow and Duncan (2012) report that a "lowness" (p. 837) in the levels of both pitch and intensity precede vocal backchannels across languages. In particular, they consider Arabic, Mexican Spanish and American English. Koiso et al. (1998) indicate that for Japanese, the 100 milliseconds of speech preceding feedback are characterized by a flat-fall or a rise-fall pitch contour. Ward and Tsukahara (2000) describe in detail the role of the pitch in the speech preceding feedback in English and Japanese. They report that regions of low pitch lasting at least 110 milliseconds precede the production of backchannels by the listener. The communicative meaning associated with such regions of low pitch is the finalized transmission of information, after which it would feel natural to receive feedback from the listener. Ward and Tsukahara (2000) also report a series of other features that might accompany low pitch in the speech before feedback, conveying either turn-finality or the need for some confirmation signal from the listener. Rising intonation – or up-talk – might elicit feedback, as it might be perceived as an insistent request for a reaction or an acknowledgment by the listener; however, when up-talk is observed, it is also preceded by a low pitch region, to which the final rise adds the insistence nuance. Moreover, the completion of a grammatical clause might also be accompanied by low pitch, as well as vocal fry, or vowel lengthening (Ward and Tsukahara, 2000).

Boudin et al. (2021), on the other hand, find rising intonation contours in the speech preceding feedback, and associate it with the presenting of new information, to which the listener reacts. Gravano and Hirschberg (2009a) examine pitch slopes over the final 200 and 300 milliseconds before verbal backchannels, and observe patterns of rising intonation before these, in particular high-rise or low-rise

contours. These differ from the contours observed for turn holds, characterized by final plateau intonation, or speaker changes, where final falls, or also high-rises are observed (Gravano and Hirschberg, 2009a). Furthermore, they report that mean intensity and mean pitch levels tend to be higher over the last 500 milliseconds and 1 second in the speech preceding backchannels than before other turn-taking structures, while NHR values are lower. They also indicate that each subject in their dataset seems to time their backchannels based on an individual preference for particular cues, or combination of cues, highlighting a high degree of individual variation (Gravano and Hirschberg, 2009a, 2011).

In conclusion, what emerges from both the literature on speaker changes and turn holds, as well as on backchannel distribution, is that both the speaker and the listener put effort into coordinating turn-taking by either signaling their conversational intentions with syntactic and acoustic means, on one hand, and by providing carefully-timed backchannels demonstrating active listenership and grounding, on the other. The interactional efforts made by the two parties in a dialogue aim to achieve smooth exchanges. However, successful interactions in real life are not solely characterized by seamless turn-taking, where changes, holds, or backchannels occur without any silent gaps or speech overlaps in between. As noted by Heldner and Edlund (2010) and Heldner (2011), natural data consistently show naturally occurring overlaps or gaps that do not break down the progression of the conversation or are perceived as disruptive by interlocutors. In fact, in specific contexts with specific participants, the default or preferred option may involve overlapping with each other or, conversely, completely avoiding overlaps and carefully leaving silent gaps before producing a response. It is within these context-specific variations that sociolinguistic and language-specific differences, to be discussed in the following Chapter 4, come into play to shape the variations in turn-taking dynamics.

Chapter 4

Sociolinguistic variation of turn-taking

Informal face-to-face interaction is the primary environment of sociality in all languages, i.e., it is the location in which individuals negotiate meaning and construct their individual, social and cultural identities across turns at talk (Turnbull, 2003). The focus of research approaches such as social pragmatics and interactional sociolinguistics are aimed at understanding how individuals employ their conversational behaviors and acts in order to “co-construct images and representations of themselves, the interlocutor, their relationship, and their culture” (Agliati et al., 2005, p. 225). More generally, conversation within these frameworks is regarded as a universal system on which the social performance of the speakers is constructed across different situations and contexts (Agliati et al., 2005; Anolli, 2006).

The norms of turn-taking described by Sacks et al. (1974) and empirical cross-linguistic investigations, such as Stivers et al. (2009)’s, postulate a universal tendency of turn-taking to minimize silent gaps and speech overlaps in conversation, while accounting for a certain degree of language- or context-specific variability in the interactional tempo of conversation. In further empirical investigations on the quantification of gaps and overlaps in verbal interaction, Heldner and Edlund (2010) and Heldner (2011) challenge the idea of such a universal tendency to achieve smooth transitions by reporting that no-gap-no-overlap transitions are the least frequent across languages, and that the timing of turn-taking might be less precise than previously claimed. However, the notion of a universal mechanism and timing organization is crucial for assessing the social significance of variations within such system. While there is a universal tendency to adhere to turn-taking norms, deviations from them are not arbitrary or random; rather, they reflect the

social dimension of interaction, through which interlocutors express their individual identities within specific contexts.

Turn-taking is intrinsically social (Beňuš, 2009). How and when we take the floor or remain silent, i.e., our individual behavioral preferences within a system with a universal base, are reflections of our social roles and identities as speakers of a specific language, our adherence to particular cultural norms and our performance of sociolinguistic variables such as gender, age, or social class (Beňuš, 2009).

In exploring the social and cultural aspects, this work focuses on gender-related variations in turn-taking behavior and the use of turn-taking cues. As gender performance in conversation needs to be contextualized within a larger framework, including the language of the speakers, the analysis will also incorporate the aspect of cross-linguistic variation in conversational behavior.

4.1 Turn-taking dynamics across languages

The turn-taking system described by Sacks et al. (1974) presents a one-at-a-time speaker model, in which transitions involving simultaneous speech occur, but they are kept to a minimum by conversational participants. At the same time, efforts are made to avoid silent gaps and, according to Gardner and Mushin (2015), several Conversation Analysis studies following Sacks et al. (1974) associated delayed responses with problems within the interaction, such as word searches, speech impairments, disagreement or difficulties in understanding on the part of the recipient.

However, further studies in various languages have shown that both speech overlaps and silent gaps are present and not always negatively judged or actively avoided, and their distribution display a significant degree of cross-linguistic variation. Examples in the scientific literature illustrate varying degrees of tolerance for speech overlaps in different languages and communities: Finnish, for instance, exhibits a very low tolerance for simultaneous speech (Lehtonen and Sajavaara, 1985; Stubbe, 1998), while a study on Jewish Americans by Tannen (1984) reveals a prevalent use of supportive overlaps, which are perceived as interruptions by Californian American listeners. Additionally, results from Ulijn and Xiangling (1995), further supported by Deng (1998), indicate that Chinese speakers frequently overlap at potentially interrupting locations (i.e., non-TRPs) during the current speaker's turn,

both within their cultural context and when interacting with Finnish and Dutch speakers, suggesting language-specific dynamics grounded in cultural conventions. Hayashi (1988)'s analysis of conversations in Japanese and American English indicates that Japanese speakers use overlapping talk more frequently than English speakers. Agliati et al. (2005) compare dyadic conversation in Italian and Icelandic, and report that Icelandic participants displayed a higher rate of speaker changes than their Italian counterparts, resulting in shorter, non-overlapping turns. In contrast, Italian conversations exhibited a greater occurrence of simultaneous speech, where participants often began speaking before the interlocutor had finished, suggesting a strong sense of participation and engagement (Agliati et al., 2005). The authors explain that, while this dynamic may be perceived as chaotic or impolite in some cultures like German or Scandinavian, in Italian and French cultures it signals active involvement and contributes to a lively development of the conversation. Tannen (2012) reviews a series of studies illustrating that overlapping speech is common and positively viewed in many cultures, such as Antigua, Thai, Guyanese, and Samoan. While overlapping speech may appear chaotic to outside viewers, it is systematic and cooperative within the community, often serving as a form of joining-in rather than disruptive (Tannen, 2012).

Similarly, long silent gaps within and between turns might have different valence across languages. For instance, in examining the stereotype of "the silent Finn," Lehtonen and Sajavaara (1985) observe that Finnish speakers exhibit longer silent pauses compared to Southern Europeans or Anglo-Americans. However, these silences do not necessarily indicate disinterest or lack of commitment to interaction from the Finnish perspective (Carbaugh, 2005). Even though their findings were not quantitative, Carbaugh and Poutiainen (2000) additionally report that Finnish speakers exhibit longer silent pauses as a gesture of respect, especially during self-introductions, while American English speakers may be seen as overly talkative in Finnish contexts, trying to fill the silence left by their Finnish counterparts. In one of the earliest studies on silent pauses in intercultural communication, Scollon and Scollon (1981) claim that American English speakers dominate conversations with Athabaskan Native American speakers due to the longer switching pauses of the Athabaskan people, who tolerate silence for around 1.5 seconds, contrasting with studies suggesting tolerances of up to 1.0 second in native English speakers within

Western European and North American white culture (Nakane, 2012). Mushin and Gardner (2009) and Gardner and Mushin (2015) conducted empirical studies on inter-turn pauses among Garrwa language speakers in remote Australian Aboriginal communities, finding that longer gaps are frequent and not considered problematic for the interaction, particularly in relaxed and familiar contexts, which allow less pressure for continuous talk.

Similar observations on cross-linguistic dynamics of turn-taking in different languages were at the base of the empirical research conducted by Stivers et al. (2009). They examined the response offset – specifically the gap between polar questions and answers – in informal conversations across 10 languages from 5 continents, including regions such as Southeast Asia, Mexico, Namibia, and Papua New Guinea. Their findings suggest that the differences in response times stem from distinct cultural norms concerning conversational pace. The authors' argument is that these differences reflect minor adjustments in the implementation of a universal turn-taking system, where speakers strive to reduce pauses between turns, as well as overlaps. These adjustments may be influenced by specific cultural interactional rhythms or broader variations in the pace of social life. For instance, by reporting the mean response times for responses labeled as on-time in their data, they observe that, while the average silent gap between turns typically lasts around 200 milliseconds, the transition time varies significantly. This variation ranges from 7.29 milliseconds in Japanese exchanges to 468.88 milliseconds in Danish, highlighting significant differences in perceived appropriate turn-taking gaps among speakers of different languages. Additionally, Stivers et al. (2009) find that typological closeness does not fully account for these patterns. Even within Germanic languages like Danish, Dutch, and English, different response timings are observed, indicating that factors beyond language structure and typology contribute to these differences.

Since their patterns of variation are predominantly grounded in the ongoing interactive exchange (Stubbe, 1998), the frequency, distribution and modality of backchannels also display high rates of cross-linguistic differences, and have been the object of several studies comparing different cultures and communities between each other. Stubbe (1998) compared the use of backchannels by native Maoris and by Pakehas (New Zealanders of European descent) and found that Pakehas produced more listener feedback and expected it to be closely timed to the current

speaker's speech, whereas Maoris produced and expected less, in particular with respect to verbal backchannels, and delays were acceptable. Several studies have conducted comparisons involving Japanese and other languages, such as German (Krause-Ono, 2004) and American English (Cutrone, 2014; Maynard, 1990; Ward and Tsukahara, 2000; White, 1989), and Finnish with Central European languages or English (Lehtonen and Sajavaara, 1985; Stubbe, 1998). These comparisons reveal that Japanese listeners tend to be highly active in providing feedback to the current speaker, both visually and verbally, a trait that persists even when they speak a second language. This tendency aligns with the notion that backchanneling constitutes a core aspect of Japanese polite behavior, suggesting that Japanese individuals may find it challenging to deviate from this behavior beyond a certain extent (Krause-Ono, 2004). In contrast, for Finnish listeners, known for their minimal tolerance for speech overlaps and interruptions, feedback is sporadic and, when it is produced, listeners tend to rely on the visual modality, with verbal backchannels carrying negative connotations in their interactions (Lehtonen and Sajavaara, 1985; Stubbe, 1998, 2013).

Fant (1989) compares the turn-taking dynamics of Spanish and Swedish speakers, and report that, while Spanish displays more turn overlaps, Swedish listeners provide 15 times more verbal feedback than the Spanish. In fact, it seems that Spanish interlocutors rely more on gaze patterns to confirm attention to each other. Fant (1989) concludes that producing backchannels in the verbal modality suggests heightened consciousness and formality, indicating that Scandinavian cultures prioritize confirming attention in dialogue more than Hispanic culture, with Scandinavian backchanneling patterns also serving to regulate dialogue and align with norms for turn-taking and avoiding overlapping speech (Fant, 1989). Dideriksen et al. (2023) investigate the frequency of backchannels in Scandinavian languages by comparing Danish and Norwegian, two typologically related languages with similar societal and cultural backgrounds but differing sound structures. Their focus also includes further linguistic devices aimed at signaling and achieving grounding among interlocutors, i.e., repairs and conversational entrainment. Their findings showed that Danish speakers put more effort into displaying mutual understanding compared to Norwegian. Dideriksen et al. (2023) suggest that this could be due to Danish having more complex and opaque sound structure than Norwegian,

indicating that differences in language-specific features, combined with cognitive processing elements such as mutual understanding and context-driven inferences, contribute to variations in the use of conversational devices. Finally, comparing telephone interactions between American English speakers, and between German monolingual and bilingual (German-English) speakers, Heinz (2003) finds that, overall, German participants produce less backchannels than American English, and less often in overlap with the current speaker's speech. Furthermore, in German conversations, bilinguals produce a higher rate of backchannels, often in overlap, than German monolingual listeners (Heinz, 2003). The author provides a possible explanation for this behavior of bilinguals as being the result of a possible higher degree of identification of the individuals with the culture of their second language (Heinz, 2003).

In summary, studies on backchanneling behavior across languages underscore the presence of language-specific baselines for backchannel frequency (Stubbe, 1998; White, 1989), as well as specific expectations regarding their distribution.

4.2 Gender-specific variation in conversation

Within language, a fundamental factor of variation is provided by the gender of the interlocutors involved in the dialogue. As a sociolinguistic variable, gender is not the same as the anatomical sex of individuals. Specifically, as Kendall et al. (2023) state, the conceptual separation lies in that sex is a biological trait, while gender is a social construct defined by what an individual does with speech in order to perform, e.g., masculinity or femininity within a specific context or community. Rather than something that someone is, gender is something that someone does – or performs – through social acts, such as interaction (e.g., Cannon et al., 2019; Eckert and McConnell-Ginet, 2013; Kendall et al., 2023; Ochs, 1992; Stokoe and Smithson, 2001).

This framework for the analysis of gender is adopted by the dynamic approach (e.g., Coates, 2004), in which gender is studied as a contextualized practice interacting with further contextual variables related to the current interaction (Swann, 2009). Earlier studies involving gender and turn-taking behavior within Western societies had different approaches as they were developed within a gender difference

paradigm (Stubbe, 2013). Specifically, the deficit approach (e.g., Lakoff, 1973, 1975), the dominance approach (e.g., Zimmermann and West, 1975), or the difference approach (e.g., Maltz and Borker, 1982; Tannen, 1990) were mainly focused on identifying which conversational habits could be attributed to men or to women, and how the dichotomy could be explained. The deficit approach was concerned with identifying the reflections of women's inferior social status in society through language use; similarly, the dominance approach seeks confirmation of men's higher power status in society through their greater use of interactional practices such as interruptions. Finally, the difference approach, which, unlike the previous two, continues to be employed (Plug et al., 2021; Stokoe and Smithson, 2001) relies on Maltz and Borker (1982)'s view of gender as a stable attribute acquired during early socialization, and men and women belonging to two different subcultures, with different conversational habits. For instance, Tannen (1990) suggested that women's conversations prioritize closeness, collaboration, seeking confirmation and support, while men in interactions aim to establish dominance, competition and leadership. In general, all-female conversations were described as being characterized by a cooperative dynamic, where the conventional rule of one speaker at a time is replaced by a collaborative sharing of the conversation with overlapping questions, restating, repeating, and the use of backchannels, all of which demonstrate active participation (Coates, 2004; Edelsky, 1993). Conversely, interactions among men are said to typically follow a more traditional one-speaker-at-a-time model, commonly involving either lengthy, uninterrupted speaking turns or brief, rapid exchanges, with minimal overlap, which is generally seen as disruptive (Coates, 2004).

Recent investigations into language and gender have begun to challenge earlier descriptive studies that sought to detect predetermined gender differences, as they did not take into consideration the fundamental role of contextual factors, or the specific dynamics of conversational interaction (Murphy, 2011; Stubbe, 2013). In fact, it has been highlighted how local contextual variables may influence the performance of gender identity, also in relation to its salience in a specific communicative situation. For example, Leaper and Robnett (2011) synthesized findings from 29 studies involving 3,502 participants examining gender-specific usage of tentative language. They reported that the conversational activity influenced participants' gendered behavior, particularly noting greater differences between male and female

speakers in task-oriented interactions, with female speakers displaying a more cooperative behavior than male speakers. This suggests that certain tasks may activate more stereotypical gender behaviors. Similarly, Vinciarelli et al. (2015) observed variations in interactional phenomena such as laughter, overlapping speech, and silences during cooperative phone tasks. Despite no inherent status differences introduced by the task, male speakers tended to exhibit stereotypically masculine behaviors associated with dominance, such as speaking for longer periods and laughing less in the task-based activity. Furthermore, the gender of the conversational partner influences interaction dynamics. In mixed-gender dyads, participants' gender identity becomes more salient, leading to a higher frequency of gender-specific stereotypical behaviors compared to same-gender interactions (Hogg, 1985; Palomares, 2004; Reid et al., 2003). The topic being discussed might also influence turn-taking behavior. For instance, Pilkington (1998, cited by Stubbe, 2013) found that, when gossiping, both men and women adopted a high-involvement style comprising frequent backchannels and cooperative overlaps.

Thus, approaches viewing men and women as possessing two diametrically opposed conversational styles, and as being monolithic groups behaving the same, has been contested; instead, there is a shift towards seeing masculine and feminine styles as being part of a continuum from which individuals can draw to different extent based on the identity they wish to perform in a specific context (Coates, 2004; Stubbe, 2013). Nevertheless, the results obtained from studies operating within previous frameworks should not be discarded; in fact, it is not their validity that is at question, as much as their interpretation (Stubbe, 2013). As claimed by, e.g., Cameron (1995), Mullany (2010), Murphy (2011) and Stubbe (2013), with the caution of avoiding too simplified and dichotomous explanations, those early studies on the different habits of men and women in interaction serve as an important starting base for the analysis of turn-taking behaviors as part of gender performance.

Recent developments in gender theory offer a more nuanced understanding of language variation by addressing individual differences such as class, race, and age, that can be simultaneously relevant within specific interactional contexts (Labov, 1990; Murphy, 2011). Moreover, research is increasingly acknowledging that gender is not binary and it is including transgender and non-binary speakers, recognizing the diverse ways in which gendered and sexual identities can be performed

(Kendall et al., 2023). However, the current Section, as well as the subsequent analyses, will focus on female and male speakers, as these are the two self-reported genders of the subjects in the datasets that are used in this work.

Empirical evidence from sociophonetics underscores the distinction between biological sex and gender as an identity and social trait. In fact, male and female speakers exhibit anatomical differences that are relatively constant across populations, and influence specific acoustic speech values. For instance, since men's vocal folds are generally longer and thicker than women's, they tend to vibrate more slowly and thus produce lower F0 values, i.e., a lower perceived pitch (Fant, 1966; Simpson, 2009; Titze, 1989). Voice quality also seems to be affected by the thickness of the vocal folds. In female voices, thinner vocal folds may result in incomplete closure, potentially contributing to the more frequent breathy phonation observed in female speakers (Simpson, 2009; Titze, 1989). Moreover, the higher formant values observed for female speakers can be partly attributed to the differences in vocal tract length between men and women (Fant, 1966; Weirich and Simpson, 2018).

However, sociophonetic cross-linguistic and cross-cultural comparisons reveal that these differences in speech values between men and women vary across languages and cultures. This indicates that speakers can adjust their speech to convey femininity or masculinity, or traits considered more feminine or masculine within their language or social group. Such adjustments are part of the performance of social roles in specific interactional contexts, indicating that the variation of acoustic features in speech is partly learned, and influenced by sociocultural factors (Simpson, 2009; Weirich and Simpson, 2018). For example, Weirich et al. (2019) and Weirich and Simpson (2023) investigated gender performance through phonetic detail in German and Swedish. Their investigations involved collecting phonetic data, such as F0 mean values and vowel space size, alongside gender identity data obtained through self-reported femininity scores by their subjects. These scores were derived from the GEPAQ_F+ questionnaire, which assesses positive personality traits traditionally associated with femininity, such as kindness, sensitivity, friendliness, and empathy. Weirich et al. (2019) noted a smaller difference in acoustic measurements between male and female speakers in Swedish, correlating with relatively similar gender identity scores among participants. Weirich and Simpson (2023) further observed a trend towards reduced gender-specificity

in Swedish, with female speakers within the population adopting characteristics more aligned with male values over time, e.g., lower F0. This shift was mirrored in lower femininity scores obtained from the questionnaire, which closely resembled the attributes self-identified by male participants. In contrast, the distinction between male and female speakers' F0 values and scores remained more pronounced in Germany, suggesting a stronger adherence to traditional gender norms in phonetic variation in speech and self-identification. Working on different varieties of German, Weirich and Zahner-Ritter (2023) found that speakers from Eastern Germany exhibited a tendency to express femininity through vowel space, unlike their subjects from Western Germany. The authors suggest that this difference could be linked to the perception of gender roles being more egalitarian in the Eastern region, thus favoring the performance of female identity through acoustic variations.

The difference of F0 values between genders in German was tested in both production and perception by Funk and Simpson (2023) in the speech of first-grade schoolchildren. They report that F0 values play a fundamental role in influencing the perceptual judgments regarding speaker's gender by adolescent and adult German listeners of the children's voices they heard. In particular, speakers with higher F0 tend to receive higher ratings for femininity. This suggests that, even without an anatomical basis in the prepubertal larynx or vocal tract, girls likely learn to adapt to the higher-pitched voices associated with women, while boys adopt the lower-pitched voices associated with men (Funk and Simpson, 2023). These results are in line with a previous study carried out by Ferrand and Bloom (1996) on American English schoolchildren. The authors find that the average F0 of boys decreased at 7-8 years of age, together with maximum F0, F0 range, and in the frequency of rises and falls, adopting the features of the adult male model; on the other hand, the F0 of girls of the same age did not display a similar decrease. With their results, Ferrand and Bloom (1996) highlight the ages between 7-10 years as a fundamental period for the acquisition of gender-specific pitch characteristics.

Further acoustic features have been targeted in the literature on gender-specific patterns of variation in speech, such as speech rate and intensity. For instance, Simpson (1998) reports that female German speakers generally produce vowels with longer durations compared to males. This was also observed in Swedish by Ericsson and Ericsson (2001), who found that female speakers generally exhib-

ited longer vowel durations compared to males, and that, in the same monosyllabic words, they also displayed larger differences in terms of duration between stressed and unstressed vowels. The results obtained on speech rate are not always consistent, sometimes finding faster rates in male and other times in female speakers (Hardy et al., 2020), while intensity is reported to be higher, as well as more variable, for male in conversational contexts (Gelfer and Young, 1997; Hardy et al., 2020).

In spite of these reported gender-specific differences of individual acoustic features in speech, little is known about their potential gender-specific variation as turn-taking cues in the context of TRPs. In fact, no study addresses specifically the turn-taking related variations of F0, intensity, duration and voice quality between male and female speakers. However, a few studies, despite focusing on different research questions or settings not involving interactions, indicate potential gender-specific patterns in the variation of acoustic turn-taking cues. Investigating the prosodic cues preceding speaker changes and turn-holds employed by Chinese learners of English as a Foreign Language, Wei and Lee (2019) mention that gender differences emerged from their results. In particular, they notice that, in the use of F0, intensity and speech rate, female learners appear to perform better in marking their conversational intentions. Another study indicating potential phonetic variation for interactional purposes was conducted by Whiteside (1995, 1996). It demonstrates that male and female speakers of British English in a reading task exhibit distinct tendencies in marking syntactic boundaries through prosodic patterns. Specifically, she observes that women tend to use more frequent silent pauses and exhibit final lengthening approaching boundaries, whereas men employ more frequent F0 shifts and tend to avoid pausing. The author hypothesizes that males' strategies for marking boundaries in a conversational setting would make it harder for the interlocutor to intervene and take up the next turn (Whiteside, 1995). While these studies point in the direction of possible gender-specific preferences in the acoustic marking of potential boundaries, investigations testing the influence of gender on the variation of turn-taking cues in detail are currently lacking. This thesis takes a first step towards investigating turn-taking cues from a sociophonetic perspective by quantifying the influence of the gender of the speaker and the gender of the listener on the acoustic configuration in the speech preceding turn transitions and backchannels.

Part II

Study Overview

Chapter 5

Studies and hypotheses

This dissertation comprises a series of analyses focused on turn-taking dynamics in face-to-face dialogues between German adult interlocutors. Swedish dialogues are analyzed as well, in order to carry out a cross-linguistic comparison. The studies that will be presented in the following chapters had the aims of exploring with a quantitative approach the influence of various aspects related to the social identities of the speakers and to contextual factors related to the communicative situations that, based on the previous literature, might influence the dynamics of interactions. However, the effect of these factors tend to be either marginally reported in studies focused on specific turn-taking aspects (e.g., gender and the acoustic organization of turn-taking), or to be addressed only in isolation rather than in interaction with further relevant variables (e.g., modality, gender and cross-linguistic variation of turn-taking dynamics).

Specifically, the studies that will be reported in the following chapters will take into account the effect of the gender of the interlocutors, the modality of the response and the native language (L1) of the subjects, as well as the interactions of the variables, on specific aspects of the turn-taking mechanism. The analyses will explore in detail the variation and the interaction of the linguistic and acoustic turn-taking cues employed to manage the floor by both speakers and listeners in conversation, as well as the frequency and the distribution of turn-taking structures (speaker change, turn keep and multimodal backchannels), both within the same language and cross-linguistically. Firstly, the influence of the variables of gender, modality and L1 on turn-taking dynamics will each be tested separately in three exploratory studies. Then, their various interaction will be analysed in two final in-depth analyses.

The central thesis of this work is that, despite all relying on the same system of turn-taking norms, real-life informal interaction often will not display a full adherence to such norms. Simultaneous speech, gaps and overlaps, as well as smooth transitions between speakership and listenership are observed as the result of the social background, the setting and the context of the conversation. This means that variables such as the gender of the speaker, that of the listener, and the modalities available to the interlocutors, in interaction with each other, have a fundamental role in shaping the turn-taking dynamics of the specific interaction. Thus, the deviations from a shared turn-taking system are social and contextual, and, as such, they are also language-specific, influenced by the diverse backgrounds of values and conversational preferences inherent to each language.

In particular, the analyses aim to address the following Research Questions (RQs):

- RQ1: Are the turn-taking-related variations of acoustic features and turn-taking dynamics influenced by the gender of the speaker and the gender of the listener in conversation?
- RQ2: How does the availability of multiple modalities (i.e., the visual and the vocal modalities) in the interaction affect the timing of backchannels in relation to the offset of speech by the current speaker and to the acoustic configuration of the current utterance?
- RQ3: Do the same patterns of variation occur across (typologically related) languages, or do social factors and the availability of multiple modalities influence conversation in language-specific ways, connected to the language-specific conversational habits, preferences, and values?

The first study (Chapter 6) reports the results of an analysis investigating the variation of acoustic turn-taking cues (F0, intensity and segmental duration) at different test time-points in the speech approaching TRPs (addressed in that specific analysis as Potential Transition Locations – PTLs – in order to underline the variable time window considered) followed by speaker changes and turn keeps. In parallel, the influence of the gender of the interlocutors on the variation of each value at the different time-points is tested through linear mixed-effect models. Given their documented acoustic differences in speech, as well as their possible specific preferences

in the attitude towards managing the conversational floor and turn-taking habits, the hypothesis at the base of this first study is the following:

Hypothesis 1 (H1): Speakers of German use different patterns of acoustic turn-taking variation to mark their conversational intentions in the speech approaching TRPs followed by speaker changes and turn-keeps influenced by their gender and the gender of their interlocutor.

The second study (Chapter 7) explores the frequency and the distribution of multimodal feedback within German face-to-face dialogues. Specifically, the study considers both verbal and non-verbal backchannels in the vocal modality, head movements in the visual modality, and combinations of vocal feedback and head movements. The distribution of backchannels is observed in relation to the speech produced by the current listener, thus feedback can occur either in overlap with it, after a silent gap, or smoothly at the offset of speech, without any audible silent gap or speech overlap. Different types of head movements are considered, and their temporal relation with the vocal component of multimodal listener responses will be explored as well. Specifically, the study is aimed at testing the following hypothesis:

Hypothesis 2 (H2): Listeners in conversation are sensitive to the placement of backchannels within the speaking turn and, when multiple modalities are available, they tend to rely on multimodal alternation to provide feedback that is the least interrupting as possible for the speech by the current speaker.

In the third study (Chapter 8), the analysis explores the extent of cross-linguistic variation in the use of turn-taking cues by comparing German dialogues with Swedish ones. The focus will be specifically on testing the variation of F0 values at different temporal locations towards TRPs (addressed in that specific analysis as Potential Turn Boundaries – PTBs – in line with the terminology adopted by Feindt et al., 2021). Given the varied results reported in the literature about the acoustic cues that are relevant for turn-taking in the two languages, as well as their different prosodic phonologies, the analysis investigate the variation of F0 towards turn ends, and explores to what extent it can be used to express individual patterns of

sociophonetic variation. The specific hypothesis that will be tested in this study is the following:

Hypothesis 3 (H3): German and Swedish speakers exhibit distinct patterns of variation in F0 as a turn-taking cue in the speech preceding TRPs, with German speakers displaying greater flexibility in marking conversational intentions and employing F0 for social dynamics, while the use of F0 at turn boundaries is expected to be more constrained for Swedish speakers.

The fourth and fifth studies include analyses that observe the effect of more factors together, in order to get a more comprehensive and multifaceted view of the interaction. Specifically, the fourth study (Chapter 9) focuses on the timing and on the acoustic configuration of speech preceding the occurrence of multimodal backchannels in German dialogues. The study will test whether male and female listeners producing feedback do so after different acoustic cues in same-gender and in mixed-gender dyads. Since previous literature described gender-specific preference in the use of supportive overlaps and feedback distribution, the analysis aims at testing empirically whether, based on their placements in specific acoustic locations within the turn, male and female listeners display distinct preferences for the production of backchannels within the current stream of talk. The interaction with modality is also considered, to observe whether male and female listeners produce gestural, verbal or multimodal backchannels at different moments during the current turn. Besides these contextual variables, the study also aims at addressing a research gap on the acoustic configurations of the speech preceding different types of backchannels. In particular, the structure of the quantitative analysis has the objective of investigating the relative contribution of different acoustic turn-taking cues for the (male or female) listener's local decision to place a specific backchannel at a given location during the current turn.

Hypothesis 4 (H4): The distribution and the placement by the listener of multimodal backchannels within the current speaking turn in German dialogues display gender-specific patterns of variation, also based on the gender of the current speaker and the modality of the feedback.

The last study (Chapter 10) takes into account the distribution and temporal dynamics of the different sequential structures (speaker change, turn keep and backchannels occurring in different modalities). Here, the effect of the gender of the speakers is observed both in interaction with that of the interlocutor, and from a cross-linguistic perspective. In this study, face-to-face dialogues in German will be compared to face-to-face dialogues in Swedish, in order to observe the influence of the specific language of the interaction on turn-taking dynamics. The cross-linguistic observation of the effect of gender on turn-taking dynamics will provide further insight into the extent to which gender performance influences conversational behavior in German compared to Swedish. Given the previous differences observed between these two specific languages, the hypothesis that will be tested in the analysis is the following:

Hypothesis 5 (H5): The distribution and the timing of speaker changes, turn keeps and multimodal backchannels display language-specific patterns in German and Swedish; the gender-specific turn-taking behavior displayed by male and female speakers in same-gender and mixed-gender dyads display language-specific dynamics as well.

The final chapters of the work (Chapter 11 through 13) provide a summary of the studies and their main findings, and the above outlined hypotheses will be discussed in light of the results obtained. Limitations to the findings and potential further development of the research will also be addressed.

Part III

Analyses

Chapter 6

Gender-specific turn-taking cues in German

The current chapter was published as: Rossi, M. (2022). Gender influence on phonetic turn-taking cues at potential transition locations in German. In R. Orrico & L. Schettino (Eds.), *The position of the speaker in interaction: Attitudes, intentions, and emotions in verbal communication* (pp. 127–147, Vol. 9). Officinaventuno: Milan. <https://doi.org/10.17469/O2109AISV000007>

This is a pre-print version.

The full summaries of the statistical models in the analysis are reported in Appendix 1 (p. i).

6.1 Introduction

Conversational interactions are a fundamental part of human social behavior. On a daily basis, individuals are involved in face to face or mediated conversations, producing about 1200 talk spurts amounting to around 2 to 3 hours of speech (Levinson and Torreira, 2015; Mehl et al., 2007). It is clear from even the most casual exchanges that people in conversation do not speak at random: in general, it can be expected that there will be one person talking at a time and that the interlocutor(s) will wait for them to finish speaking to launch their turn, usually trying to keep silent gaps and speech overlaps reduced to a minimum (Sacks et al., 1974).

In order to achieve a smooth exchange of turns, speakers seem to be able to predict the approach of a potential transition location (PTL from now on), i.e., a point in the current speakers' turn when speaker change becomes a possibility (Schiffrin, 1987;

Transition Relevance Places, Sacks et al., 1974; Potential Turn Boundaries, Zellers, 2017), which allows them to start planning their next conversational move. This can either be to take up the next turn, to remain silent and let the speaker continue, or to produce a non-interrupting backchannel to signal attention. In fact, interlocutors do not wait for the current speaker to finish with their turn before deciding what to do next; they start encoding their following turn while the current one is still ongoing (Levinson and Torreira, 2015), so that they are able to respond appropriately at the appropriate time. Evidence in favor of early planning within the current turn is given by the discrepancy between the language production system's latencies for the encoding of a single word and of a simple short clause, which are respectively of around 600 milliseconds (ms) (Indefrey, 2011; Indefrey and Levelt, 2004; Schnur et al., 2006) and 1500 ms (Griffin and Bock, 2000), and the average duration of a silent gap in between turns across several languages, which amounts to around 200 ms (Heldner and Edlund, 2010; Stivers et al., 2009). Predictions about the approach of a PTL and about what will come after it are made by interlocutors on the basis of turn-taking cues in the current speaker's turn.

Previous studies have identified several communicative means that are used by speakers to signal their conversational intentions for the next turn: gestural (Edlund and Beskow, 2007, 2009; Hadar et al., 1984; Mondada, 2007; Paggio and Navarretta, 2011; Zellers et al., 2019), linguistic (Ford et al., 1996; Levinson, 2013; Levinson and Torreira, 2015; Local and Walker, 2012) and phonetic (Gravano and Hirschberg, 2009b, 2011; Hjalmarsson, 2011; Local et al., 1986). Among the phonetic cues, the variation of F0, intensity and segmental duration have been found to significantly contribute to the turn-taking mechanism in several languages (e.g., Brusco et al., 2020; Gravano and Hirschberg, 2009b, 2011; Ishimoto et al., 2017; Zellers, 2017), including German (Dombrowski and Niebuhr, 2005; Kohler, 1983; Niebuhr et al., 2013; Peters, 2006; Selting, 1996). In particular, PTLs followed by a speaker change tend to be marked by either a rising or falling intonation and by a decrease in intensity, while turn holds are preceded by level F0 contours and higher intensity profiles.

Backchannels also seem to be preceded by determinate sets of prosodic cues, such as regions of low pitch and some cases of up-talk in English and Japanese spontaneous conversations (Ward and Tsukahara, 2000), and final rising intonation

(Skantze and Schlangen, 2009), together with higher intensity values (Gravano and Hirschberg, 2009a) in task-oriented interactions in Swedish and in English. Mixed patterns of variation towards PTLs are found for segmental duration. While some studies give evidence for pre-boundary lengthening before turn yields (e.g., Local et al., 1986, for British English; Gravano and Hirschberg, 2009b, 2011, for American English; Niebuhr et al., 2013 for German), others find increased segmental duration before turn holds and faster speech rate before speaker changes (e.g., Brusco et al., 2020; Koiso et al., 1998; Zellers, 2017).

6.2 Gender variation in conversational behavior

Sociolinguistic research on verbal interactions has revealed the existence of social variation in the way turn-taking takes place: in particular, the interlocutors' genders appear to have a strong role in conversational behavior. Tannen (1994, 1998) proposed that, as a result of their different social and cultural background, men and women tend to use and interpret linguistic strategies such as interruptions, taciturnity and indirectness in contrasting ways. For instance, all-female talk has been observed to be often characterized by a collaborative floor with co-constructed utterances (Edelsky, 1993), cooperative overlaps and frequent minimal responses (Menz and Al-Roubaie, 2008; Stubbe, 2013), while men interacting with each other tend to stick to the one speaker-at-a-time model, with either long uninterrupted turns or short rapid ones, where overlaps are rare and perceived as deviant (Coates, 2004). Thus, while in some contexts conversational strategies like interruptions might be used to show support and achieve cooperation, in others they can also be employed to reinforce asymmetry and establish dominance over the other speaker. Asymmetrical interactional patterns have been often observed in mixed gender conversations, with one speaker saying too much or too little, interrupting the other or virtually withdrawing from the conversation (e.g., Coates, 2004; Waara and Shaw, 2006; Zimmermann and West, 1975). The configuration of same-gender and mixed-gender interactions, however, does not always follow these exact patterns: results from different studies give oftentimes seemingly inconsistent results, since the influence of speakers' genders on their turn-taking behavior seems to be mediated by other so-

cial and contextual factors, such as the conversational goal and interactional setting, institutionalized roles and gender identity salience (Plug et al., 2021).

The majority of the scientific literature on gender variation in turn-taking has strongly focused on pragmatics, addressing, for example, talkativeness, the use of tentative language, and interruptions. It is not clear, however, if the similarities and the differences between genders also extend to the use they make of turn-taking cues, such as phonetic ones, to signal to the interlocutor their conversational intentions of offering the next turn or continuing to speak. We are aware of the overall differences in F0, intensity and segmental duration in men's and women's speech (e.g., women tend to have a higher mean F0 than men [Weirich and Simpson, 2018], men have higher conversational intensity levels than women [Gelfer and Young, 1997] and slightly lower durational values as well [Pépiot, 2014]). However, the possible influence of gender on the phonetic variation of these features as turn-taking cues has, to the best of the author's knowledge, never been empirically investigated in any language. Thus, the analysis presented in the current paper aims at proposing a first sociophonetic exploration of turn-taking by testing if, and how, social variables such as the gender of both interlocutors in a dialogue might have an effect on the acoustic cues that speakers use to signal their intentions for the next conversational turn.

6.3 Methods

In order to offer more insight into the variation of phonetic turn-taking cues in spontaneous interactions in German and to investigate the possible influence of gender on such variation, a dataset composed of two-party conversations between German native speakers, balanced for the gender of the interlocutors, was annotated, and values for F0, intensity and segmental duration were extracted at different test locations approaching a PTL.

6.3.1 Dataset

The two-party conversations analyzed for this study are part of the German sub-corpus of the DUEL Multi-lingual Multimodal Dialogue Corpus (Hough et al., 2016). The DUEL corpus consists of naturalistic, face-to-face conversations based

off loosely-directed tasks assigned to the participants. The tasks were specifically designed to provide subjects with a theme to discuss without, at the same time, being constrained in how to develop their interactions (*ibid*). In particular, in the portions of dialogues analyzed for the present research, the speaker pairs dealt with either the “Dream Apartment” or the “Film Script” tasks, in which participants had to imagine and plan out the organization, furniture and decoration of a hypothetical shared apartment or imagine and describe a movie scene centered on an embarrassing situation, respectively. The two interlocutors in each dyad were recorded on separate channels, which allows the phonetic analysis of speech even when the participants talk in overlap with each other. For the present study, 6 dyads have been annotated and analyzed, for a total of 12 different speakers, all German native speakers and (at the time of the recording) students at the University of Bielefeld, aged 21-28. The dyads were selected with the aim of having both speakers of the same gender in conversation with each other, as well as subjects of different genders interacting with each other. The dyads thus included in the dataset for the study are 2 male-to-male (MM) conversations, 2 female-to-female (FF) conversations and 2 mixed-gender (MF) conversations (see Tab. 6.1). The first 5 minutes of the tasks have been taken into consideration for this analysis.

6.3.2 Annotation

The annotation was carried out in Praat (Boersma and Weenink, 2023) and it started with the manual annotation of PTLs (Schiffrin, 1987). As described by Sacks et al. (1974), such locations in conversations, defined by them as Transition Relevance Spaces, arise at the end of Turn Constructional Units, when speaker change becomes possible, though not mandatory. Following Zellers (2017) classification of Potential Turn Boundaries, these locations were identified not only after syntactically complete utterances, but also after syntactically incomplete utterances which functioned as semantically or pragmatically complete in the local context of the interaction. Such locations were identified in the data using the right boundaries of the “X-utts” tier provided in the DUEL corpus, which contained the segmentation and the orthographic transcription of speakers’ utterances within each turn (annotated in the “X-turns” tier).

The guidelines for the segmentation of the utterances given to the annotators of

the DUEL corpus follow the notion of “slash unit” described by Meteer et al. (1995), in which an utterance should be comprised by maximally a complete sentence or a smaller unit, which may not be syntactically complete but is judged as complete in context by the annotators (Hough et al., 2016). Turns in the “X-turns” tier, instead, included all continuous stretches of speech by one speaker until the other one takes up the floor, or up until a silent gap after which either the current speaker continues talking, or the interlocutor takes up the next turn. By assuming the right boundaries of the “X-utts” tier as PTLs, it was possible to include in the analysis instances considered by the annotators as complete in context, but that were not necessarily followed by a speaker change or by a silent gap.

Once PTLs were identified, they were annotated using a set of labels (Feindt et al., 2021) describing the utterance by the current speaker and what came after it (see Fig. 6.1):

- Completeness label: indicated the syntactic/semantic completion of the utterance in context (“yes” for complete utterances and “no” for incomplete utterances);
- Sentence Type label: described the form of the utterance (“d” for declaratives, “q” for questions and “t” for tag questions);
- Sequential Structure label: indicated the conversational action by the current speaker or by the other participant which followed the PTL, i.e., the other participant took up the next full turn (speaker changes, labelled with “c”, also referred to as turn yields), the current speaker held the floor (keeps, labelled with “k”, also referred to as turn holds), the other participant produced a minimal, non-interrupting response (backchannels, labelled with “b”);
- Transition Type label: described the way in which the conversational action following the PTL took place, i.e. with a silent gap (“g”), with a speech overlap (“o”), in a smooth way, without any perceivable silences or speech overlaps (nogap-no-overlap, “n”); in particular, only those silences longer than 120ms were annotated as gaps, and only those stretches of overlapped speech longer than 120ms were labelled as overlaps, as it has been shown that gaps and overlaps shorter than 120ms are not perceived as such by listeners (Held-

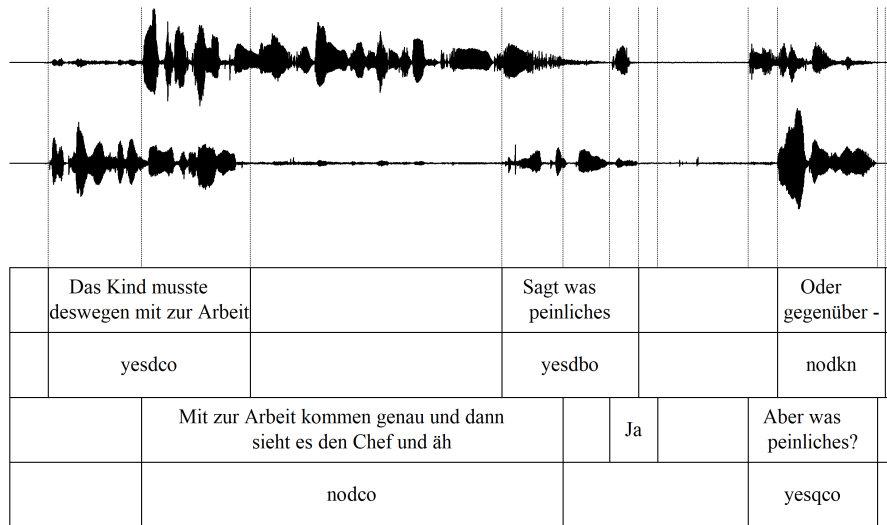


Figure 6.1: Example of the annotation of PTLs in Praat, from one of the dialogues in the dataset. The first part of the label(s) refers to the completeness in context (“yes” or “no”), the following letter refers to the sentence type (“d” for declarative, “q” for questions, “t” for tag questions), the third part of the label refers to the sequential structure that followed (“c” for speaker change, “k” for keep and “b” for backchannel), and the final letter describes the transition type (“g” for gaps, “o” for overlaps, “n” for no-gap-no-overlaps).

ner, 2011). Transitions occurring with possible silences or overlaps shorter than 120ms were considered as no-gap-no-overlaps.

For each utterance, words, syllables and segments were also annotated.

Due to space constraints, the analysis involving transition types (i.e., gaps, overlap and smooth transitions) will not be reported, and, for the purposes of this study, sentence types such as questions and tag questions are momentarily set aside, in an attempt to limit the sources of prosodic variation. Finally, the exclusion of TRPs preceding backchannels is due to limitations in the current annotation scheme and analysis structure. In fact, several factors have been found to influence backchannel’s placement (e.g., the lexical or non-lexical content of the backchannel, or its modality, cf. Ferré and Renaudier, 2017; Truong et al., 2011), and would thus have to be considered when targeting their distribution and the cues preceding them. Since such factors are not present in this annotation scheme, nor they would fit into the rest of the analysis, the description that would have resulted would have been biased and ambiguous. Further developments of this research, though, include a more precise annotation of backchannels, as well as their analysis.

6.3.3 Data extraction

As a part of a wider research project aimed at investigating the location and the extension of the transition space in conversation and testing the hypotheses related to the relevance for turn-taking of time windows against that of phonological categories (issues that will not be discussed in the current paper), the present analysis observes the variation of the phonetic parameters using three different time-points approaching a PTL as reference (Feindt et al., 2021). The phonetic parameters investigated are F0, intensity and segmental duration, and the three test time-points approaching a PTL are located at 500 ms before the end of the utterance, at 200 ms before the end of the utterance, and at the end of the utterance. The phonetic values were extracted automatically using a Praat script from utterances with a duration of 1 second and up, so that it was possible to extract data from all three test locations without them being too close to the start of the utterance (Feindt et al., 2021).

F0 readings, extracted using Praat with the settings for semitones (st) above 1Hz, were then normalized with the individual speaker's baseline in order to exclude the influence of physiological factors. Using a sample of F0 datapoints in semitones extracted from 2 minutes of clear speech for each subject, the individual baseline was calculated as 0.75 times the first quartile of the data, following Zellers and Schweitzer (2017) and Zellers (2021). Intensity was extracted in decibels (dB) and normalized with the speakers' mean (Ludusan and Dupoux, 2015). Values for segmental duration were also extracted at three different test intervals approaching the PTL: the average segmental duration, in milliseconds (ms), was extracted over the last 500 ms and over the last 200 ms (i.e., from the offset of speech to 500 ms before that and from the offset to 200 ms before that) and, finally, the duration of the last segment at the end of the utterance was extracted. This way, it was possible to observe how average segmental duration varied towards PTLs. No normalization was carried out at this stage for segmental duration; however, the randomness of the segments included in the analysis allows us to make general preliminary observations about the variation of segmental duration approaching PTLs, excluding the potential influence of the different segment types.

6.4 Analysis and results

A total of 489 PTLs with the related phonetic values at the three test locations were extracted. The qualitative and quantitative analysis of the data was carried out in R (R Core Team, 2022). Linear mixed effects models with the subject as a random factor were used for the quantitative analysis of the data using the R package *lmerTest* (Kuznetsova et al., 2017).

From a first qualitative exploration of the data, differences between the behavior of the male and female speakers in conversation appear. For instance, the numbers of speaker changes and keeps in FF conversation were equal: female speakers tend to yield the turn as much as they hold it (78 changes and 70 keeps); on the contrary, in MM conversation there is a bigger disproportion between turn holds and speaker changes, with male speakers keeping the floor for longer, and more frequently than they are ceding it (44 changes and 92 keeps). Moreover, it was observed that smooth transitions (i.e. no gap no overlap) occurred more frequently in same-gender conversation, thus 26% of the transitions are smooth in MM conversation and 34% in FF conversations, while only 16% of transitions were smooth in mixed-gender conversations (see Tab. 6.1). Silent gaps occur more frequently in MM conversations than in FF and MF ones, while speech overlaps are observed more frequently in FF dyads (see Tab. 6.1).

Table 6.1: Transition type (gaps, overlaps, no-gap-no-overlaps) distribution in same-gender (“MM” for male-to-male, “FF” for female-to-female) and mixed-gender (“FM” for female-to-male) dyads

Gender	Transition Type		
	Gaps	Overlaps	No-gap-no-overlaps
MM	61% (97)	13% (20)	26% (41)
FF	39% (64)	28% (46)	34% (56)
MF	56% (93)	27% (45)	16% (27)

The analysis reported below focuses on the pre-final phonetic marking of syntactically/semantically complete utterances in declarative form preceding speaker change and keep cases, and how the gender of the interlocutors influences that.¹ The PTLs thus considered for the following analysis are 239, including 131 followed by speaker change cases and 108 by turn holds, constituted by syntactically/semantically complete utterances in declarative form.

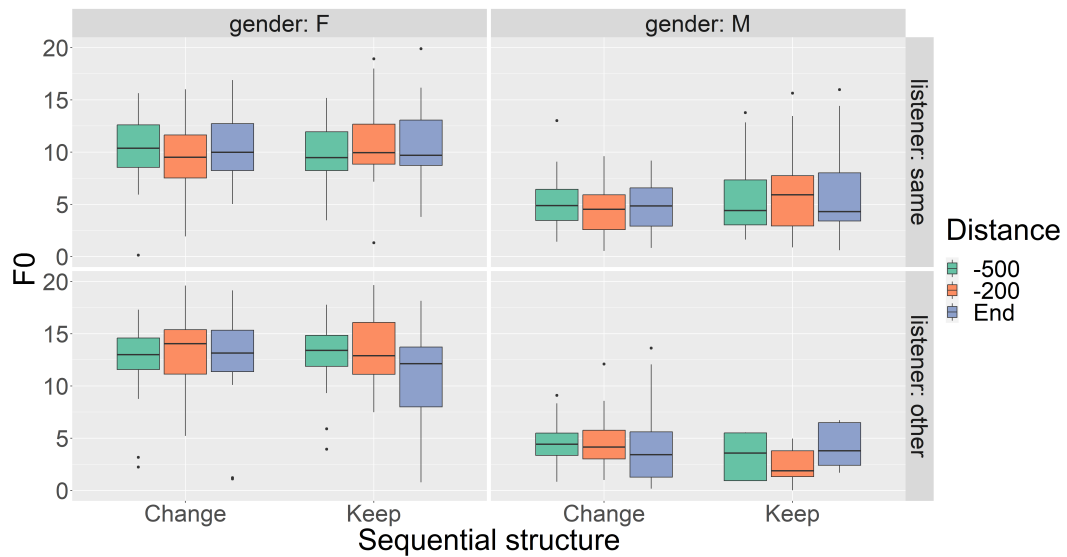


Figure 6.2: F0 datapoints distribution (in st, above the speaker's baseline) for speaker change and keep cases at the three test locations ("Distance": -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other).

6.4.1 F0 values

The variation of the normalized F0 values towards PTLs revealed significant differences both between the speaker changes and keeps, and between male and female interlocutors and the way in which they approach different sequential structures. Male speakers' values at the three test locations analyzed (at the end of the utterance, at 200 ms from the boundary and at 500 ms from the end of the utterance) are closer to their baseline, at around 5 st, while the average values for female speakers are higher, mostly above 10 st.

Considering both the gender of the speaker and the gender of the interlocutor, in speaker change cases the values of male speakers are closer to their baseline, both in MM and MF conversations. In all conditions, the contour created by the three datapoints tends to be flat, though with a certain degree of variability, with the exception of male speakers in MF conversations, where the values of F0 at the End location fall closer to the speakers' baselines, at 2.5 st, while for female speakers they tend to remain higher, closer to the previous ones, at 12.5 st (see Fig. 6.2). Moreover, in both speaker change cases and keep cases (see Fig. 6.2), it appears that female speakers are always higher than males, but they are even higher when

in conversation with a male speaker; similarly, male speakers in MM conversation show low values, but males in conversation with females are even lower.

Table 6.2: Summary of the linear mixed model for F0 in speaker change cases. A significant three-way interaction is found between the value of the final F0 test location (End), the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(F0 \sim distance * gender * listener + (1 | speaker))$

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	11.0322	1.6712	9.2173	6.601	<0.0001
-200: Male	2.4714	0.8134	1465.0749	3.038	0.002420
End: Male	4.4082	0.8134	1465.0749	5.420	<0.0001
-200 other	1.9198	0.8134	1465.0749	2.360	0.018391
End: other	4.5724	0.8134	1465.0749	5.621	<0.0001
Male: other	-2.2018	4.0219	8.5934	-0.54	0.597998
-200*Male: other	-1.5617	1.2932	1465.0749	-1.20	0.227398
End*Male: other	-7.1274	1.2932	1465.0749	-5.51	<0.0001

A linear mixed model shows that the value of F0 at the PTL differ significantly for male speakers in mixed-gender conversations (see Tab. 6.2). A significant three-way interaction between the distance from the PTL, the gender of the speaker and the gender of the interlocutor is found for keep cases, too: the height of F0 at the -200 datapoint differ significantly from the intercept for male speakers in mixed-gender conversations (see Tab. 6.3). In these interactions, F0 is rising at the end of the utterances for male speakers, while it is falling for female speakers (see Fig. 6.2).

Table 6.3: Summary of the linear mixed model for F0 in turn hold cases. A significant three-way interaction is found between the value of the penultimate F0 test location (-200), the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(F0 \sim distance * gender * listener + (1 | speaker))$

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	9.1696	1.2859	7.3667	7.13	0.000148
-200: Male	-0.8469	0.4724	2896.9797	-1.79	0.073108
End: Male	-1.4471	0.4724	2896.9797	-3.06	0.002209
-200: other	-0.3076	0.5112	2896.9797	-0.60	0.547383
End: other	-2.3349	0.5112	2896.9797	-4.56	<0.0001
Male: other	-2.5707	3.6319	7.3266	-0.70	0.500957
-200*Male: other	-3.6392	0.9743	2896.9797	-3.73	0.000191
End*Male: other	1.3532	0.9743	2896.9797	1.389	0.164965

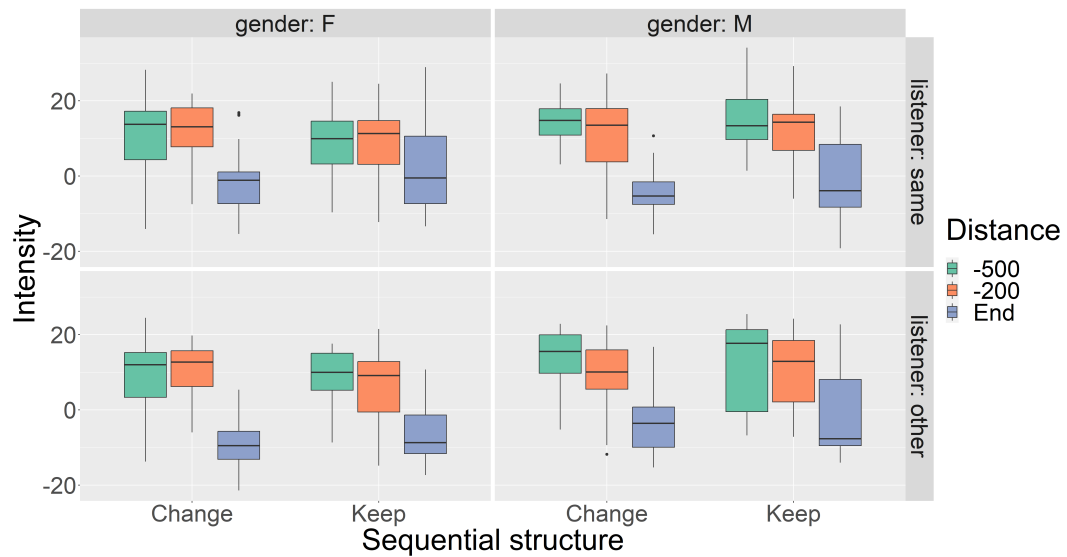


Figure 6.3: Intensity datapoints distribution (in dB, normalized to the speaker's mean) for speaker change and keep cases at the three test locations ("Distance": -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other).

6.4.2 Intensity

For each of the sequential structures analyzed, intensity decreases approaching the end of the utterance, with the values of the last test location dropping below the speakers' means. The decrease is higher for PTLs preceding speaker change cases, where it appears that speakers lower their voices much more than before keeps. In speaker change cases, male speakers appear to be louder than females, except at the end of the turn in MM conversations, where their values drop at -5 db below their means, while the final values for females in FF conversations are closer to their means. On the contrary, female speakers talking to a male conversational partner tend to lower their intensity much more approaching the end of the turn, while males in conversation with a female partner end their turns with a higher intensity, closer to their means (see Fig. 6.3).

For keep cases, when the speaker held the floor, intensity decreases less towards the PTL. In fact, in these cases the intensity values at the last test location fall below the speakers' means, but they remain closer to it than in speaker change cases. Considering both the speakers' and the interlocutors' gender, however, it is possible to detect an exception in this trend: intensity values at the end of the utterance for male speakers in MF dyads decrease well below their mean, more than it does for

female speakers in MF dyads. For both speaker change cases and keep cases, two separate linear mixed effects models show a three-way interaction between the gender of the speaker, the gender of the interlocutor and the intensity at the final test location (see Tab. 6.4 and Tab. 6.5).

Table 6.4: Summary of the linear mixed model for intensity in speaker change cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(intensity \sim distance * gender * listener + (1 | speaker))$

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	10.1865	1.062	9.7709	9.58	<0.0001
-200: Male	-3.2728	0.858	3517.058	-3.81	<0.0001
End: Male	-6.6951	0.858	3517.058	-7.80	<0.0001
-200: other	0.1572	0.850	3517.058	0.18	0.853326
End: other	-6.7422	0.850	3517.058	-7.92	<0.0001
M: other	1.0235	2.576	9.3741	0.37	0.700063
-200*Male: other	-1.8130	1.274	3517.058	-1.42	0.154837
End*Male: other	8.8455	1.274	3517.058	6.94	<0.0001

Table 6.5: Summary of the linear mixed model for intensity in turn hold cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $lmer(intensity \sim distance * gender * listener + (1 | speaker))$

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	10.3673	2.121	7.5234	4.88	0.00145
-200: Male	-2.6697	0.943	2896.948	-2.83	0.00468
End: Male	-7.3551	0.943	2896.948	-7.79	<0.0001
-200: other	-2.7057	1.020	2896.948	-2.65	0.00807
End: other	-6.9076	1.020	2896.948	-6.76	<0.0001
Male: other	-0.4589	5.988	7.4643	-0.07	0.94093
-200*Male: other	3.8497	1.945	2896.948	1.97	0.04790
End*Male: other	8.0387	1.945	2896.948	4.13	<0.0001

6.4.3 Segmental duration

Segmental duration is significantly influenced by the distance from the PTL and the different sequential structures that follow the PTL. Duration increases over the three test locations up until the end of the utterance for PTLs preceding speaker change cases. Average final segmental duration is higher also for turn holds, but it remains closer to the average of the previous test location (-200). This indicates that speakers

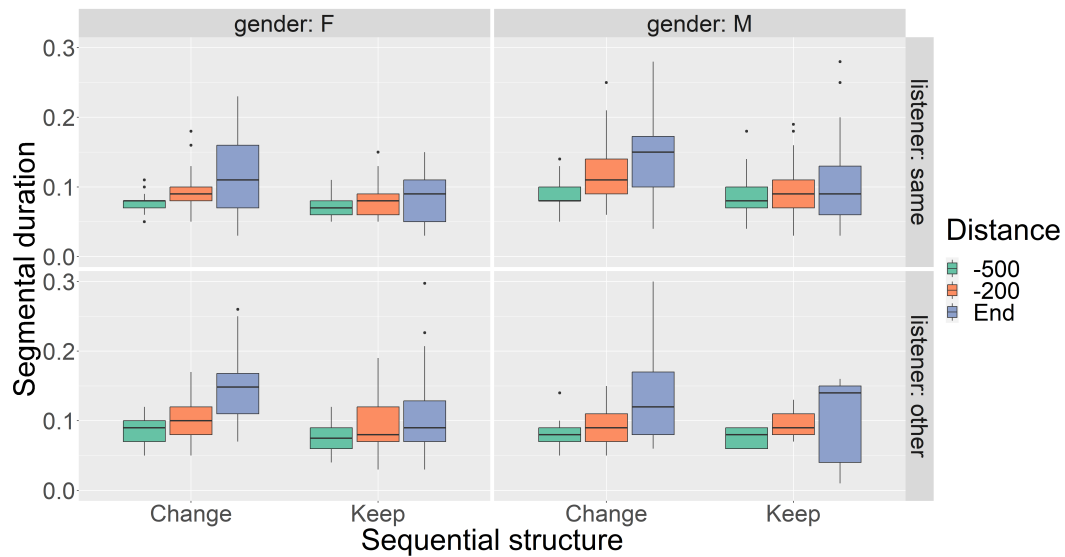


Figure 6.4: Average segmental duration in ms for speaker change and keep cases over the three test locations (“Distance”: -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other).

do slow down their speech when approaching a PTL before a keep, but not as much as they do before a speaker change (see Fig. 6.4).

For speaker change cases, final segmental duration increases to a higher degree for male speakers in MM conversations than for female speakers in FF conversations. In MF dyads, on the contrary, duration at the final test location increases for female speakers, while its average remains closer to the previous locations for male speakers (see Fig. 6.4). A linear mixed effects model shows a three-way interaction between the gender of the speaker, the gender of the interlocutor and the duration values at the final test location (Tab. 6.6).

In keep cases, as already mentioned, final duration does not increase as much as in speaker change instances. However, this does not seem to be true for male speakers in MF dialogues. In fact, they seem to reach the potential turn end with an increased segmental duration also before turn holds (see Fig. 6.4). Even if it is possible to detect qualitative differences, in this case, they did not reach statistical significance.

6.4.4 Summary of the results

The phonetic cues preceding speaker changes and those preceding keep cases show different patterns of variation. Intensity tends to decrease towards the end of the

Table 6.6: Summary of the linear mixed model for segmental duration in speaker change cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: $\text{lmer}(\text{duration} \sim \text{distance} * \text{gender} * \text{listener} + (1 | \text{speaker}))$

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	0.077	0.00645	9.551	11.974	<0.0001
-200:Male	0.015	0.00513	3517	2.881	0.0039
End:Male	0.047	0.00513	3517	9.235	<0.0001
-200:other	0.00418	0.00509	3517	0.822	0.4112
End:other	0.0302	0.00509	3517	5.932	<0.0001
Male:other	-0.0156	0.0156	9.173	-1.003	0.3417
-200Male:other	-0.0197	0.00762	3517	-2.594	0.0095
EndMale:other	-0.058	0.00762	3517	-7.609	<0.0001

utterance, but to a lesser extent when the current speaker wants to continue talking. Segmental duration increases towards the end, but to a lesser extent, again, when the current speaker has the intention of continuing. F0 values show a high degree of variation, so it is harder to get a clear pattern of variation related to the turn-taking structure. As for the sociophonetic variation of turn-taking cues, there are significant differences in the way male and female speakers in same-gender and mixed-gender conversations mark the test locations for both keep and speaker change cases. In particular, the results obtained using linear mixed models, indicate that, in speaker changes, the interaction between the gender of the speaker, the gender of the interlocutor and the final datapoint, i.e., the end of the utterance, was significant for F0 values (β : -7.1274; t value: -5.511; $p < .0001$) for intensity (β : 8.8455; t value: 6.942; $p < .0001$) and segmental duration (β : -0.05805; t value: -7.609; $p < .0001$). For segmental duration, the interaction of the speaker's and the interlocutor's genders with the "-200" location was significant as well (β : -0.01979; t value: -2.594; $p < .01$). For keep cases, the variation of F0 values is influenced by the interaction of the gender of the interlocutors and the "-200" location (β : 1.3532; t value: 1.389; $p < .0001$), and, for intensity, by the gender of the interlocutors and the "-200" (β : 3.8497; t value: 1.979; $p < .05$) and "End" (β : 8.0387; t value: 4.133; $p < .0001$) locations.

6.5 Discussion

The results related to the phonetic variation of turn-taking cues in this dataset show patterns that had already been observed in previous research, i.e., speakers tend to maintain their intensity higher (Gravano and Hirschberg, 2009b, 2011) and their speech rate faster (Local et al., 1986; Niebuhr et al., 2013) in order to hold the floor. This configuration of the phonetic turn-taking cues approaching PTLs is in line with the idea that speakers might put more effort into marking turn holds than into signaling turn-yielding intentions because speaker change might be “a kind of default option” (Zellers, 2017: 13) in spontaneous conversation. This might be especially relevant when the speaker’s utterance is syntactically/semantically complete, as was the case for the utterances considered in this study. In fact, in the presence of syntactic/semantic completion, it might be more likely that the interlocutor will assume that the current speaker has finished with their turn, so a more prominent phonetic marking is needed. Patterns for F0 variation related to turn-taking are less clear. In fact, no recurrent or systematic patterns emerged from the qualitative analysis of the data. This observation is in line with the results obtained by Feindt et al. (2021) who suggest that, since F0 movements in German are not restricted by any phonological function, as it is the case for other languages, e.g., Swedish, speakers in conversation tend to use F0 more freely towards turn ends and make use of their entire F0 range. Moreover, the fact that only syntactically/ semantically complete utterances were taken into consideration for the present study gives a further explanation to the high degree of variation obtained from this data sample: in fact, it has been observed that when the upcoming completion of an utterance is signaled by syntactic, semantic or pragmatic means, F0 patterns are less restricted for turn-taking function, and speakers tend to show a higher degree of variability towards the end of their turn (Feindt et al., 2021; Rossi et al., 2022a; Selting, 1996).

As for gender variation, a few differences were observed in the phonetic patterns of the parameters analyzed, which were seemingly influenced not only by the gender of the speaker, but also by the gender of the interlocutor. First of all, the data-points extracted for F0 (in semitones and normalized with the speakers’ baseline to exclude the influence of physiological factors) were always higher for female speakers and closer to the baseline for male speakers, a result consistent with other studies

on German and other languages (Andreeva et al., 2017; Pépiot and Arnold, 2021; Pépiot, 2014; Weirich et al., 2019). The difference in F0 height was particularly striking in MF conversations, where F0 was closer to the baseline for male speakers than in MM conversations, while for female subjects it was higher in MF than in FF conversations. Moreover, more rises and falls in F0 were observed for male speakers than for female speakers and, in a few occasions, F0 even showed a more consistent patterning for them, as for example in the keep cases in MF dyads (see Fig. 6.2).

Variation related to turn-taking for intensity and segmental duration, instead, seemed to be employed more consistently by female speakers in both FF and MF conversations, and by male speakers in MF conversations. Interestingly, in fact, male speakers in MF dyads seemed to use these values to a smaller extent to signal their conversational intentions, e.g., they tend to not lower their final intensity as much as it happens in other conditions (e.g., in MM dyads), as well as do not vary their speech rate in the direction it would be expected, or as much as their female conversational partners. On the contrary, male speakers make use of these cues in line with the expectations in MM dyads and, in these cases, their variation related to turn-taking is much more striking than it is for female speakers in FF dyads. Thus, in general, it appears from these observations that female speakers make more use of duration and intensity as turn-taking cues, while it is possible that male speakers may vary more with F0 towards turn ends to signal their intentions. This observation would be in line with the results obtained by Whiteside (1995, 1996) for British English. She finds that, in a reading task, women mark syntactic boundaries by pausing and using phrase-final lengthening, while male subjects rely more on F0 shifts and pause less frequently, which, in a conversational setting, would make it harder for the interlocutor to take up the next turn (Whiteside, 1995). Moreover, their more consistent use of duration and intensity in MF dialogues may suggest that female speakers in this data sample are doing more work in terms of signaling their conversational intentions than male speakers do. However, similar patterns of variation, even more accentuated, are observed for male speakers in MM conversations. So, in general, it appears that all subjects put more effort in terms of signaling their intentions to either cede or hold the floor when the other interlocutor was a male.

As previously mentioned (see Section 6.2), the interactional setting should be

taken into consideration when discussing gender influence on conversational behavior. Previous studies on talkativeness and assertiveness and their relation to gender identity reported that, in mixed-gender dyads, gender identity was salient for the participants, since the effects of controlled factors were significantly bigger than in same-gender ones (Hannah and Murachver, 2007; Leaper and Ayres, 2007). Applying this to our results may suggest that female speakers in mixed conversations, where gender might have been more salient, tended to take up a more cooperative and facilitating role (Holmes, 1995), focused on keeping the flow of the conversations as smooth as possible by making their turn-taking intentions clear. However, the structure of the exchanges of same-gender dyads suggest some influence of the gender of the two interlocutors, too. For example, a greater number of overlaps were found in female-to-female speech, while a lower number of turn yielding cases accompanied by a higher number of holds was observed for male-to-male dyads, which is similar to the descriptions of same-sex conversations provided by the literature (see Section 6.2). In fact, it has also been indicated how also experimental activities seem to interact differently with gender, with task-oriented dialogues making interlocutors' genders more relevant than non-structured conversations (Leaper and Ayres, 2007; Leaper and Robnett, 2011).

Finally, in this context, the result of male speakers doing more signaling work than female speakers in MM conversation is interesting and unexpected. A possible explanation could be that the goal of the conversation, i.e., discuss with each other to complete a cooperative task, was salient and thus the interlocutors might have focused on that, and tried to cooperate as much as possible with each other by facilitating turn-taking and putting more effort into signaling their conversational intentions.

6.6 Conclusions and outlook

From this first exploration of the influence of gender on the variation of turn-taking cues, it appears that male and female speakers in this data sample share both commonalities and differences. All the speakers display the characteristics of potential turn ends already investigated for German, with certain patterns of variation of F0, intensity and segmental duration before turn holds and speaker changes. A few

differences were observed, however, in the degrees of variation of such cues, especially when the gender of the interlocutor was taken into consideration as well. Even if quantitatively significant, it must be kept in mind that these results concern a relatively small set of data and speaker sample. Moreover, since it seems that no previous study has tested how social and individual variables influence the variation of phonetic turn-taking cues, it is not possible to generalize or make broader claims on the issue.

However, these results provide some preliminary evidence for a possible influence of gender on the way male and female speakers mark PTLs in spontaneous interactions in German. Thus, it can be a matter worth of being investigated further, since it might give us some more insight into what contributes to produce the different structures that same-gender and mixed interactions sometimes seem to assume. Further developments of this research will focus on a larger set of data and will include more phonetic parameters in the analysis, such as voice quality and segmental reduction measures, in order to observe their variation towards PTLs in German and to test the influence of the interlocutors' genders on their use. Then, a better tailored annotation and analysis will focus on backchannels, both vocal and gestural (in the form of head movements) and on the phonetic marking of the speech preceding them. In line with the current one, the investigation of backchannels in German will deal with their form, their distribution in the conversation, and the extent to which the gender of both interlocutors might have an effect on backchannel production and the phonetic cues preceding it.

Chapter 7

Multimodal backchannels in German

The current chapter was published as: Rossi, M., Schröer, M., Ludusan, B., & Zellers, M. (2023). A multimodal account of listener feedback in face-to-face interactions. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2023)*, 4121–4125.

This is a pre-print version.

The full summaries of the statistical models in the analysis are reported in Appendix 2 (p. ix).

7.1 Introduction

In conversation, listeners tend to provide feedback to their conversational partner, i.e., they signal their active attention or react to what is being said, without interrupting the speaker's stream of talk (Duncan, 1974; Kjellmer, 2009; Truong et al., 2011; Yngve, 1970). In face-to-face interactions, listener feedback phenomena, also called “backchannels” (Yngve, 1970) or “listener responses” (Maynard, 1990), are observed in both the vocal and the visual modalities, i.e., through verbal and non-verbal vocalizations, and through gestures.

Verbal feedback is constituted by lexical items or phrases with a very short constituent size (Müller, 1996). In German, typical lexical feedback expressions are, e.g., “ja” “genau” “okay” “eben” “achso” (Knudsen et al., 2020). Moreover, non-lexical vocalizations involving nasals, such as “mmm” or “mhm”, are often also candidates for backchannels, in German as well as cross-linguistically (Liesenfeld and Dingemanse, 2022). Laughter is another non-verbal vocalization employed for backchanneling purposes, which often occurs in spontaneous interactions (Trouvain and Truong, 2012). It has several functions in conversation, including linguis-

tic ones (e.g., for discourse organization Ludusan and Schuppler, 2022). Previous work has shown that laughter is used as a feedback signal across languages, with the proportion of laughter out of the total backchannels varying between a few percentage points and almost a quarter of all backchannels (Cutrone, 2014; Krepsz et al., 2022; Pammi and Schröder, 2009).

Finally, head movements, such as head nods, are another feedback type occurring in face-to-face conversations (McClave, 2000; Yngve, 1970). Being communicated via the visual channel, head movements are even less disruptive than vocal feedback and can potentially occur more often and in more locations than vocal ones without being judged inappropriate (Ferré and Renaudier, 2017). They can occur on their own, but are very often observed in co-occurrence with both verbal and non-verbal feedback expressions, or in their vicinity, often preceding vocalizations (Dittmann and Llewellyn, 1968; Maynard, 1990).

Previous experimental research on listener responses in conversation has focused mostly on the identification and description of feedback inviting cues (e.g., Gravano and Hirschberg, 2011; Ward and Tsukahara, 2000), while less attention has been put on the phonetic features of feedback items and how they vary depending on their type and their placement in the ongoing talk (Krepsz et al., 2022). In particular, while there is a growing body of research targeting multimodal backchannels (e.g., Ferré and Renaudier, 2017; Truong et al., 2011), non-verbal vocalizations tend not to be included in the investigations, so that not much is known about how and to what extent they differ from each other and from other types of feedback.

7.2 Methods

7.2.1 Dataset

We investigated face-to-face dialogues between German native speakers using audio and video files taken from the German subset of the multimodal DUEL corpus (Hough et al., 2016). We analyzed 12 dyads involving 24 different speakers for a total of 2 hours and 15 minutes of dialogue. The recording setup for the DUEL corpus consisted of lapel microphones in front of each subject and two video cameras capturing the gesture space and the faces of the subjects. Participants were involved in different interaction scenarios designed to have them start talking without having

to select a subject, while also allowing a free discussion, and to ensure the presence of laughter in different contexts. The scenarios analyzed for this study were “Film script” and “Dream apartment”. The corpus includes the full transcription of the dialogues, their subdivision in conversational turns and utterances, as well as laughter occurrences.

7.2.2 Annotation

Listener feedback was identified in the dyads using the corpus’ orthographic transcription, the audio and the video files. We consider as listener feedback all those lexical and non-lexical backchannels, laughter and head movements produced/performed by one of the interlocutors as an optional response to what the current speaker is saying (e.g., answers to questions are not considered as listener feedback; Ward and Tsukahara, 2000), and which are not part of a full-fledged turn (e.g., turn-opening, turn-closing items or discourse markers are not considered as listener feedback).

In Praat (Boersma and Weenink, 2023) we isolated lexical and non-lexical backchannels, and carried out a further annotation of laughter, in order to consider only those instances which corresponded to our feedback identification guidelines. The category of verbal feedback includes all the items that functioned as feedback, e.g., “ja”, “genau”, “okay”, “das stimmt”, while non-lexical feedback includes items such as “mhm”. No form distinction will be used for the purposes of the current study. For the annotation of gestural feedback, we used the software ELAN (Wittenburg et al., 2006) and the guidelines provided by the M3D annotation scheme (Rohrer et al., 2023) for head movements. First, all head movements were annotated using the video without the audio; then, in Praat, additionally considering the corresponding interlocutor’s turns, a further annotation was carried out to isolate the movements which occurred as a stand-alone listener response, or in connection with a verbal, non-verbal or laughter type feedback.

Moreover, to investigate feedback distribution, we observed their location with respect to the current speaker’s turn. If the feedback occurred completely in overlap with the other interlocutor’s speech, it is considered “turn-internal”, while if it occurred right after the interlocutor’s turn end or it finished outside of it, followed by a silence, it is considered “turn-external” (Kjellmer, 2009). Finally, our study

included a further analysis of turn-external feedback by categorizing the type of transition with which it occurred. Specifically, we labelled the transition type of feedback that occurred during an overlap with the other speaker's turn as "overlap," and feedback that occurred after a silent gap as "gap" (Heldner, 2011). If the silent gap or portion of overlapped speech was less than 120 ms, the transition was considered "no-gap-no-overlap," as previous research has established 120 ms as the detection threshold for silences and overlaps (Heldner, 2011).

7.2.3 Phonetic features

To investigate the phonetic variation of both laughter and verbal feedback we focus on the features of loudness and voice quality, using the values of maximum intensity (intmax) and the mean of the cepstral peak prominence (CPP). In contrast with other features, such as F0 and duration, intensity and voice quality have yet to be well investigated in feedback expressions. For lexical and non-lexical backchannels, Beňuš et al. (2007) and Beňuš (2016) report that backchannels tend to be louder than the other affirmative words in American English, and that non-lexical items tend to be significantly less loud than short lexical expressions in Slovak. Using the parameter of jitter, Zellers (2021) studied voice quality in backchannels in Ruuli/Lunyala, finding higher values, correlated with lower periodicity, in longer lexical items, while non-lexical items displayed lower jitter. Intensity and voice quality have been studied also for laughter (e.g., Ludusan and Wagner, 2022a), with a higher intensity and a less modal production being found for laughter, compared to speech. For the current study, intmax and CPP were both extracted from the entire interval constituting either the feedback item (lexical, non-lexical, laughter) using a Praat script. Intensity values were normalized with the individual speaker's mean.

7.3 Results

A total of 771 feedback items were extracted from the 12 dialogues. In terms of frequency, lexical items and head movements (without speech) are the most frequent feedback types in our data (with, respectively, 294 and 240 occurrences), followed by laughter (n=158), and lastly by non-lexical backchannels (n=79). Statistical analyses were carried out using R (R Core Team, 2022).

7.3.1 Head movements

A total of 462 head movements were extracted, including both those cases which constituted a feedback item on their own and those which accompanied vocalizations. The most common type of head movement with a feedback function is the nod ($n=383$, 83%), followed by head tilts ($n=37$, 8%). We also extracted a few occurrences of protrusions ($n=32$, 7%), which mostly accompany laughter, and head turns ($n=10$, 2%). Out of the 531 vocal feedback items, the 42% of the lexical backchannels are accompanied by a head movement ($n=136$), as well as the 56% of the non-lexical ones ($n=44$) and 27% of the laughs ($n=42$). The onset of the head movement tends to occur mostly before the vocalization, specifically 84 ms before lexical backchannels, 200 ms before non-lexical items and 140 ms before laughs, while the movement offset is observed 340 ms before the offset for laughter, and 140 ms and 172 ms after the end of the vocalization for lexical and non-lexical feedback respectively. Linear mixed models, with the speaker as a random effect, and subsequent Tukey post-hoc pairwise comparisons indicate a significant difference between the timing of the head offsets for laughter and lexical feedback ($\beta = -.4691$, t ratio = -4.696 $p < .0001$) and laughter and non-lexical feedback ($\beta = -.5110$, t ratio = -4.219 , $p = .0001$), while the difference between the head offset timing for lexical and non-lexical item is not significant ($\beta = -.0419$, t ratio = $-.418$, $p = .9084$).

Table 7.1: Count of observations (and residuals) of the χ^2 test for the transition type relative to the feedback item (lexical, non-lexical, laughter, head movement).

item	gap	nogap	overlap
<i>lex</i>	89 (2.17)	25 (0.44)	68 (−2.17)
<i>non</i>	18 (0.62)	10 (2.22)	12 (−1.68)
<i>laugh</i>	23 (−2.12)	9 (−0.75)	60 (2.28)
<i>head</i>	37 (1.19)	10 (−1.19)	69 (1.68)

7.3.2 Distribution

With respect to the other interlocutor's speech, feedback constituted only by a head movement tend to occur mostly turn-internally, while all the other types of feedback are mostly observed turn-externally, with the exception of non-lexical backchannels, which are observed in almost equal measure turn-internally and turn-externally

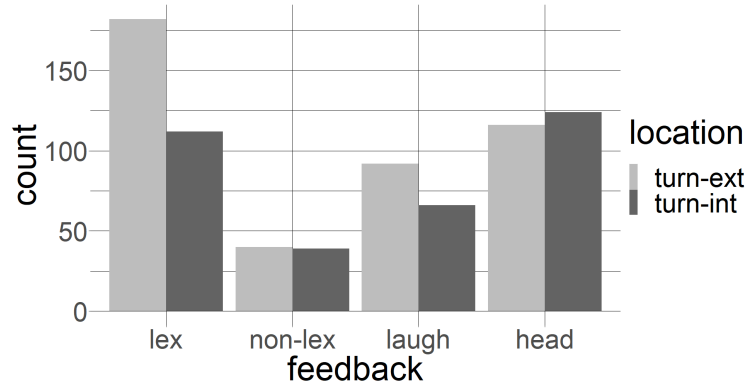


Figure 7.1: Distribution of feedback items (lexical, non-lexical, laughter, head movements) with respect to the interlocutor's current speaking turn (turn-external, turn-internal).

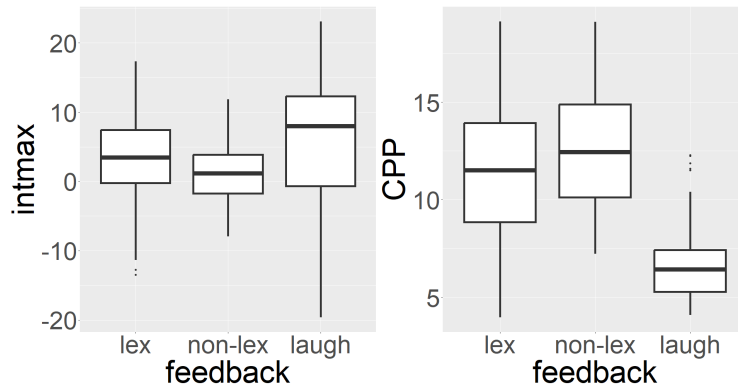


Figure 7.2: Maximum intensity (normalized with speaker's mean) and CPP values for vocal feedback items (non-lexical, lexical, laughter).

(see Fig.7.1). A χ^2 test confirms the existence of a relationship between the feedback type and the location in which it is produced by the listener ($\chi^2 = 11.099$, $df = 3$, $p = .0112$). There is also a relationship between the feedback type and the transition with which it occurs considering the final boundary of the other interlocutor's turn, i.e., with a gap, an overlap or a no-gap-no-overlap ($\chi^2 = 33.869$, $df = 6$, $p < .0001$). The Pearson residuals for the test indicate that lexical feedback occurs significantly more frequently with a gap, laughter more frequently with an overlap, and non-lexical backchannels with no-gap-no-overlap transition (see Tab.7.1).

7.3.3 Phonetic variation

As for the phonetic features of the vocal feedback items, we find that laughter tends to have a higher intmax than the other types of backchannels, followed by lexical items and non-lexical items (see Fig.7.2). Conversely, mean CPP values are higher

for non-lexical items, followed by lexical feedback and laughter (see Fig.7.2). Linear mixed models and Tukey post-hoc tests show that intensity values are significantly different between lexical feedback and laughter ($\beta = -2.75$, t ratio = -4.062 , $p = .0002$) as well as between non-lexical items and laughter ($\beta = -4.59$, t ratio = -5.132 , $p < .0001$), while the difference is not significant between lexical and non-lexical items ($\beta = 1.83$, t ratio = 2.191 , $p = .0737$). CPP values are significantly different between each pair in the post-hoc comparisons (lexical-non-lexical: $\beta = -1.26$, t ratio = -3.584 , $p = .0011$; lexical-laughter: $\beta = 4.94$, t ratio = 17.305 , $p < .0001$; non-lexical-laughter: $\beta = 6.20$, t ratio = 16.440 , $p < .0001$).

Additionally, we tested if the presence of a head movement might also correlate with variation in the phonetic features of vocalizations. Except from some qualitative differences, feedback items with a co-occurring head movement do not have significantly distinct features from those where the movement is not present, with the exception of the intensity values for lexical items. In these, lexical items with a co-occurring head movement seem to be significantly louder than the unimodal ones, i.e., constituted only by the vocal element ($\beta = -2.422$, t ratio = -3.149 , $p = .0214$).

7.4 Discussion

We investigated the distribution and the acoustic realization of multimodal feedback in German face-to-face dialogues by observing lexical backchannels, head movements, laughter and non-lexical backchannels. Lexical items are the most frequent type of feedback in our dataset, followed by head movements. In line with previous studies (e.g., Truong et al., 2011 for Dutch and Ferré and Renaudier, 2017 for British English), lexical backchannels tend to occur mostly turn-externally, and after a gap, while head movements, being the least disruptive type of feedback, occur more often turn-internally. Lexical items tend to be quieter than laughter, as previously observed for German by Ludusan and Wagner (2022a), but they are generally louder than non-lexical items. In terms of voice quality, they appear to be less periodic than non-lexical items, which could be motivated by their segmental content including voiceless segments or even short silences in complex lexical phrases.

Our results on head movements' temporal alignment with speech in multimodal feedback items show that the onset of the head movements precedes the vocalization by around 141 ms, which, as hypothesized by Dittmann and Llewellyn (1968), is done by the listener to anticipate the start of a feedback response without interrupting the interlocutor. The alignment of the head movement with laughter is significantly different when compared to the other feedback types, since the offset of the movement occurs well before the laugh's end. This might be due to the fact that laughter is the type of feedback where the most protrusions are observed. A head protrusion would generally consist of a single movement of the head forward or backwards (and not in repeated back-to-back movements, as in e.g. head nods), so the head might tend to return to its resting position before the laughter is finished.

Moreover, we observe that multimodal lexical backchannels in our data appear to be louder than unimodal lexical ones. This might suggest a relation between the phonetic features of feedback items and head movements. For instance, Tiede et al. (2019) find that, for articulation rate and nodding, increased effort in speech production is accompanied by increased head movement. Further tests (e.g. logistic regression models) will be carried out on our data to find out whether this is the case for multimodal feedback as well.

Finally, we find that the two types of non-verbal vocalizations present some substantial differences from each other, both in terms of their distribution and their phonetic features. While non-lexical backchannels tend to occur at the same rate turn-internally and turn-externally, and are produced after a smooth transition at the end of the other interlocutor's turn, laughs mostly occur turn-externally, and often after a portion of overlapped speech at the end of the current speaker's turn.

7.5 Conclusions

With the current analysis we provided a description of the distributional features of different types of feedback, both unimodal and multimodal, and we tested how different factors, such as the feedback's form, location, its transition type and the presence of a co-occurring head movement, influence the acoustic realization of vocal and multimodal listener responses. Future research should include, e.g., the lexical and segmental content of the verbal backchannels to provide a more detailed de-

scription of their variation, and the phonetic features of the current speaker's turn corresponding to a listener response, in order to observe if determinate cues would elicit a specific type of feedback.

Chapter 8

Language-specific variation of turn-taking cues

The current chapter was published as: Rossi, M., Feindt, K., & Zellers, M. (2022). Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish. *Proceedings of Speech Prosody*, 185–189. <https://doi.org/10.21437/SpeechProsody.2022-38>.

This is a pre-print version.

8.1 Introduction

In conversation, speakers produce acoustic-prosodic cues, along with morphosyntactic and gestural ones, to signal to listeners their intention to either cede the floor or continue speaking. In order for the listener to launch their turn in time or to remain silent, and thus to comply with the conversational expectations of keeping silent gaps and overlapped speech portions reduced to a minimum (Levinson, 1986; Sacks et al., 1974), the variation of phonetic cues related to turn-taking is hypothesized to be initiated early within the current turn, from half a second (Levinson and Torreira, 2015) up to one full second before the end of the turn (Zellers, 2017). The precise location of the transition space, however, is yet to be defined, as well as the relative contribution to turn transition of several proposed phonetic cues across languages. The current study adds to the ongoing discussion about the turn-taking mechanisms by observing the variation of a phonetic parameter which has been assigned a central role by research in signaling speakers' conversational intentions, i.e., F0, towards potential turn boundaries (PTBs) (Zellers, 2017).

To test the extent of F0 variation as a turn-taking cue, we adopted a cross-linguistic

perspective and analyzed fundamental frequency contours in German and Swedish. The comparison of these two languages is motivated by the fact that, even though related, they differ significantly in their prosodic phonology. German is an intonation language in which pitch accents mark prominent syllables and boundary tones mark the end of an intonational phrase, whereas Swedish is a pitch-accent language (Gårding, 1989) in which a two-way lexical pitch accent contrast distinguishes word pairs and focus is signaled with an additional high tone (Bruce, 1977; Gårding, 1989). Moreover, boundary tones in Swedish rarely show a rising pitch contour; even questions end mostly with a falling contour (House, 2005; House, 2004).

The objectives of this study are thus to investigate the differences in F0 marking of PTBs before both speaker change and floor holds and, furthermore, to observe how speakers of German and Swedish use F0 variation to signal their intentions for the following turn in the ongoing conversation.

8.2 Methodology

8.2.1 Dataset

The production data analyzed for this study consists of two-party spontaneous conversations in German and Swedish. The German data comes from the Lindenstraße task in the Kiel Corpus of Spoken German (Kohler et al., 2017), while the Swedish conversations come from the Spontal Corpus (Edlund et al., 2010). Subjects from the two corpora are native speakers of German and Swedish respectively. In both corpora, speakers were recorded in separate channels, allowing for a phonetic analysis of speech even when it was produced in overlap. Thus far we have annotated and analyzed a total of 4 conversations involving 8 different speakers. Each subject has been given an alphanumerical identification code (e.g. SG01A or SS02B) indicating their L1 being either German (SG) or Swedish (SS) and the conversation in which they took part (01 or 02, with two conversations per language). The letter “A” or “B” at the end of the code identify the two participants in the same conversation.

The current subject sample was not balanced for the speakers’ gender: all German speakers are female while 3 Swedish speakers are male and 1 is female. This issue and its possible consequences on our results and their interpretation will be

addressed in Section 8.4; however, to exclude the variation related to physiological factors, F0 values have been normalized with the procedure reported in 8.2.3. The dataset analyzed for the present study is comprised of 521 potential turn boundaries in declarative form. All annotations were carried out using Praat (Boersma and Weenink, 2023).

8.2.2 PTBs annotation

PTBs are defined as locations in conversation where it is possible, although not obligatory, for one speaker's turn to end; they are roughly comparable to transition relevance places (Sacks et al., 1974) or SYNCOMPs (Local and Walker, 2012). For the current study, PTBs were identified following the automatic detection of silent pauses in the individual audio tracks. They were later manually annotated for syntactic/pragmatic completeness, sentence type (declarative, question or tag question; however, the current study focuses only on declarative utterances), sequential structure and transition type.

On the basis of what followed the PTB, one of the following sequential structure labels was assigned:

- “c” for change: the other conversational participant takes up the next full turn;
- “k” for keep: the current speaker holds the floor by continuing his/her turn;
- “b” for backchannel: the other conversational participant produces a minimal response and the first speaker continues his/her turn.

The transition type label was applied to describe how the transition between the above-mentioned sequential structures occurred:

- “g” for gap: between the turns there is a marked silent pause, i.e., longer than 120 ms (Heldner, 2011);
- “o” for overlap: between the turn there is a marked portion of overlapped speech in which both speaker talk simultaneously for more than 120 ms (Heldner, 2011);
- “n” for no-gap-no-overlap: the transition between turns occurs smoothly, with potential gaps or overlaps shorter than 120 ms.

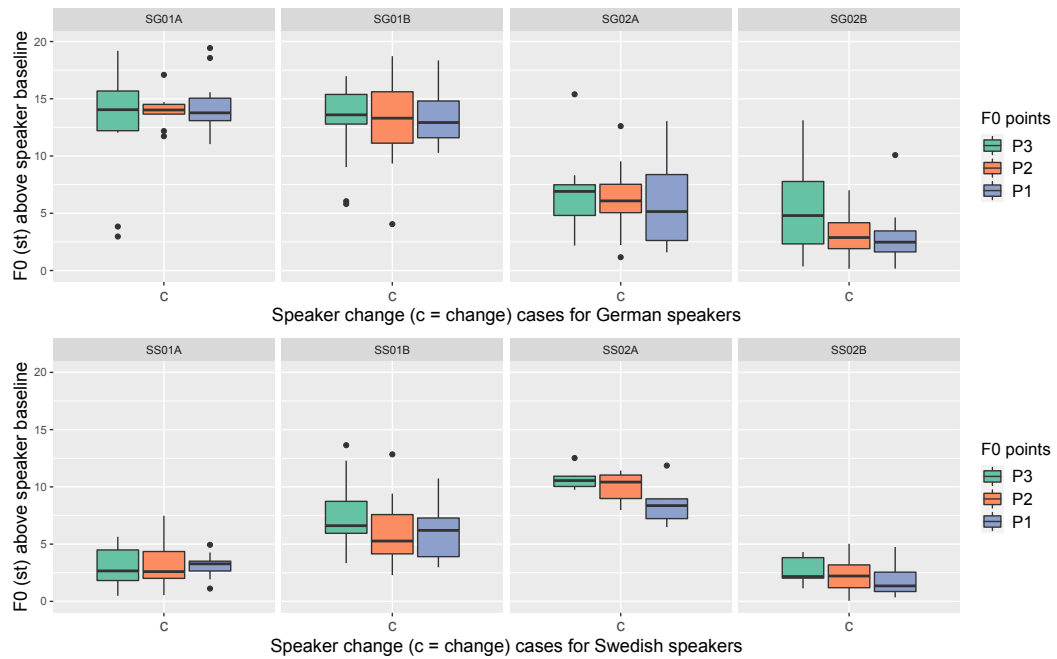


Figure 8.1: Boxplots of normalized F0 for German speakers (SG01A, SG01B, SG02A, SG02B) and Swedish speakers (SS01A, SS01B, SS02A, SS02B) preceding a speaker change. Numbers “01” and “02” in the German and Swedish participants’ identification codes indicate the conversation in which they took part.

8.2.3 F0 points

To observe the F0 movements leading up to the turn boundary and to investigate the extension of the transition space, F0 values were automatically extracted at three test locations: P1, i.e., at the offset of speech; P2, i.e., at 200 ms preceding the boundary; P3, i.e., at 500 ms preceding the boundary. The extraction was carried out using Praat’s setting for semitones above 1 Hz. The values obtained were then normalized for each speaker by calculating a baseline F0 value for the speaker and subtracting it from the measured values in the data. The baseline F0 was calculated using a similar method to Zellers, 2017, using an automatic extraction of values and assigning the value at the first percentile as the baseline F0 value for the speaker. Thus all reported F0 values are measured in semitones (st) relative to the speaker’s baseline F0.

8.3 Analysis and Results

Results from an exploratory analysis of the data give evidence for a different use of F0 between the two groups of speakers, with a few cross-linguistic similarities. The first main difference between the two languages is the overall F0 span of data

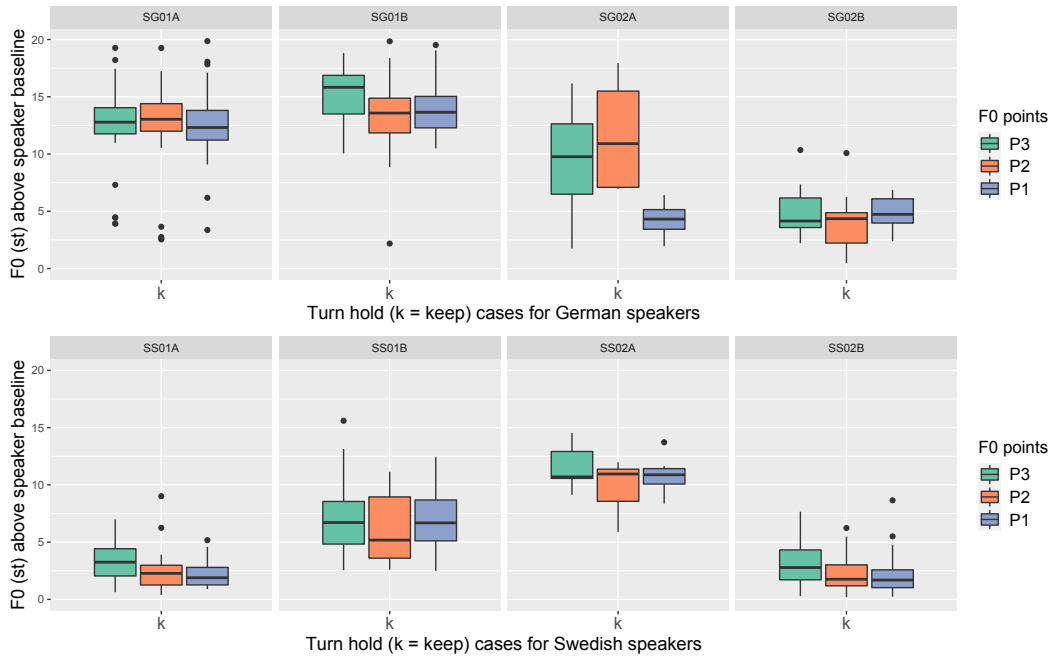


Figure 8.2: Boxplots of normalized F0 for German speakers (SG01A, SG01B, SG02A, SG02B) and Swedish speakers (SS01A, SS01B, SS02A, SS02B) preceding a keep. Numbers “01” and “02” in the German and Swedish participants’ identification codes indicate the conversation in which they took part.

points at the test locations. Where German speakers’ F0 range of values at turn ends extends from 3 st to around 20 st, Swedish F0 span of data points at turn ends remains overall closer to the speakers’ baseline, going from 2 st to 8 st: the range of F0 values observed for the Swedish speakers appears to be much narrower than that used by the German speakers. In speaker change cases, the variation of the P1 (F0 values at the boundary) shows that German speakers end their utterances at around 10 st above their baseline, while they end higher for keeps, i.e., at around 14 st. On the other hand, Swedish speakers end higher for speaker change, at around 5 st, and lower for keep, at around 3 st.

A recurrent F0 contour pattern for changes and keeps, however, did not clearly emerge from our dataset, due to the very high degree of inter-speaker variability observed for German subjects and, to a lesser extent, for Swedish ones (see Fig. 8.1). For example, in the German data, where subject SG01A ends at 15 st above the baseline, subject SG02B ends much lower at 2.5 st. These are, however, speakers from two different conversations. If we observe German speakers interacting with each other in the same conversation, the two pairs appear to behave in a similar way, seemingly accommodating their F0 movements to each other. The F0 values at the

three test locations average at 15 st in conversation 01, while they are closer to the speakers' baselines in conversation 02, at 5 st, with some variability. Local similarity for the German speaker pairs appears in the data for keeps as well (see Fig. 8.2). Speaker SG02A displays a high degree of variability for the two F0 points (P3, P2) preceding the turn final one (P1), which falls however to the same average of the other conversational participants, i.e., 5 st above the baseline. In conversation 01, both speakers behave once again very similarly and maintain the F0 at 15 st above their baseline.

On the other hand, in the Swedish data there is no evidence for possible F0 accommodation between speaker pairs. For instance, in speaker change cases (Fig. 8.1), participants of conversation 01 fall to 2.5 st and 5 st respectively; in conversation 02, an almost divergent behavior can be observed, where one speaker rises to 8 st while the other falls very close to his baseline.

In spite of the high degree of variability in the data, we were able to make some general observations about the two groups' turn-taking behavior. For both languages, in speaker change cases followed by a smooth transition, the F0 values showed a wide span of variability. F0 in speaker changes followed by a gap, instead, shows a more consistent patterning towards the boundary, with the German P1 averaging at around 13 st and the Swedish P1 at around 2.5 st (higher than the other averages for German, lower than the others for Swedish). The fact that such speaker change cases are followed by a gap offers a possible explanation for the situation, i.e., the next speaker interpreted the final rise or the final fall, respectively for German and for Swedish, as a signal that the current speaker wanted to continue talking and take up the following turn, too. These values, in fact, resembles those from the keep condition, with the German speakers' F0 at around 14 st above the baseline and the Swedish at 3 st. Thus, ending higher would be a signal for German speakers to hold the floor, while for Swedish the same intention would be signaled by a lower F0.

Sentence completeness gives more insight into the high F0 variability for speaker change cases. In the no-gap-no-overlap transitions, as well as for overlaps, we observed a high degree of F0 variability in sentences labeled as complete: it appears that, when the syntax or pragmatics signal the completeness (or the forthcoming completeness) of the sentence, then F0 variation is more free and less restricted.

8.4 Discussion

We analyzed the F0 marking of turn-taking in two German speaker pairs and two Swedish speaker pairs (four different speakers for each language) over the last 500 ms of speech in conversation. What was immediately apparent from our data was the high degree of inter-speaker variability in the F0 movements over our three test locations. In the German data, however, we noticed that the four subjects in the two conversations analyzed showed a convergent behavior in F0 variation in both change and keep cases, i.e., the speakers in the same conversations displayed a very similar range of values of F0 at the three locations analyzed. We hypothesize that this could potentially be evidence for a possible accommodating behavior between German conversational participants, who may entrain with each other in F0 and its variation when signalling different turn-taking intentions.

Entrainment of acoustic-prosodic parameters between conversational participants was observed by Levitan et al. (2011), who reported local F0 accommodation at the turn level, as well as Lubold and Pon-Barry (2014), with F0 being one of the two most significant features that speakers in collaborative dialogues entrained when there was rapport between participants. At the current stage of this study, our data are still limited to a relatively small set of speakers and the results obtained are descriptive. In fact, our observation is based on an seemingly convergent qualitative distribution of F0 values in the conversations of two speaker pairs in each language, and convergence or proximity measures have not been taken into consideration yet. However, the fact that our data are normalized to the speakers' baseline allows us to exclude the possibility of a similarity based on physiological factors, so we hypothesize the presence of a local entrainment in F0 between the subjects.

Moreover, as we have already mentioned in 8.2.1, our corpus is not balanced for subjects' gender: speakers in the German group are all females, while all except one speaker (i.e., SS02B) in the Swedish group are male. It could be argued that this gender difference could be the explanation for the observed convergent behavior in the German subset, since female speakers have been observed to accommodate more frequently in conversation (Cabarrão et al., 2016; Reichel et al., 2018). However, studies on prosodic entrainment have reported irregular results on the matter and, after their systematic analysis of two large corpora, Weise et al.

(2019) claimed that gender does not have a significant effect on the accommodation of acoustic-prosodic features, including F0. It is thus still possible to attribute the different behavior observed between the two groups to cross-linguistic differences in the two languages' intonational phonology, particularly since the differences we find are consistent with what could be predicted on the basis of the phonological systems. As discussed in Section 8.1, Central Swedish has a two-way lexical pitch accent system in which pitch distinguishes segmentally-identical word pairs, and focused content words and phrase-final words are marked with an additional high tone (Bruce, 1977; Gårding, 1989).

Our results are in line with Zellers (2013) and Zellers (2014, 2017)'s hypothesis that Swedish pitch could be already saturated as a signaling tool, thus speakers tend to rely more consistently on other phonetic features, such as segmental duration. On the other hand, German speakers are able to vary pitch more without impact on the lexical content, which leads to them having more flexibility and thus the space to entrain with each others' F0 in conversation. This does not exclude the possibility of entrainment in Swedish: conversational participants may accommodate to each other in other phonetic features, such as e.g., segmental duration, intensity or voice quality. Further research will investigate this hypothesis, particularly focusing on clusters or constellations of phonetic features.

The observation that Swedish F0 may be already saturated as a signaling cue could also explain the narrower F0 range observed for Swedish in comparison to German: Swedish boundary tones are more constrained by the lexical-phonological function of the pitch, and thus F0 variation range at the boundary will be smaller when compared to that of an intonational language, such as German, although variations in boundary tone height may still have a subtler pragmatic function (Sara Myrberg, personal communication).

Although we have been able to make some general observations about the F0 marking of turn-taking, our analysis revealed a very high degree of variability of F0 approaching PTBs, especially, as already mentioned, for German speakers. A possible reason for this could be that F0 movements were the only parameter taken into account for our the present corpus study. Clearer patterns of variation may emerge if other acoustic features were to be taken into account, such as intensity, voice quality, and segmental duration (Niebuhr et al., 2013; Zellers, 2017). Our results thus

support the hypothesis that phonetic cues related to turn-taking should not be analyzed in isolation, but rather as part of a gestalt of features (Couper-Kuhlen and Selting, 1996; Selting, 1992).

8.5 Conclusion

We analyzed F0 variation at three time points towards turn boundaries in spontaneous conversation in German and Swedish. These are two related languages with different intonational phonologies, which appear to lead to cross-linguistic differences in turn-taking signaling. Individual variation was particularly evident in our sample, and it reflects to some extent the way in which German and Swedish deal with F0 as a turn-taking cue. In fact, German speakers display a higher degree of variability, along with entrainment tendencies among speaker pairs, while, on the other hand, we find that Swedish speakers did not show any convergent behavior for F0, and utilize a smaller F0 range when approaching the boundary. The fact that Swedish F0 appears to be less flexible at PTBs may suggest that it is less available to speakers as an intention signalling cue in conversation, while its wider variability range in German, as well as its use as a feature for entrainment, may indicate a greater contribution to the turn-taking mechanism regulating interactions. Further research on these two languages, including a larger sample of speakers, will investigate further how F0, along with other phonetic features, varies to signal turn-taking intentions, both cross-linguistically and between individual speakers.

Chapter 9

Gender-specific and multimodal variation in backchannel distribution

The full summaries of the statistical models in the analysis are reported in Appendix 3 (p. xvii).

9.1 Introduction

Smooth and successful conversational exchanges are the result of the coordination and the synergistic efforts of the interlocutors involved. In everyday conversations, both the current speaker and the listener tend to do interactional work to ensure that the conversation will flow as seamlessly as possible. The speaker, on his/her end, uses a set of verbal and non-verbal turn-taking cues to signal his/her intentions to either cede the conversational floor to the listener, or to keep it further and continue speaking; on the other end, the listener, uses a set of vocal and non-vocal means to display his/her active attention and understanding of what is being said without interrupting or claiming the turn. Such verbal and gestural signals employed by the listeners are referred to in the literature as backchannels (Yngve, 1970), listener feedback (Allwood et al., 1992), minimal responses (Coates, 2004) or listener responses (Dittmann and Llewellyn, 1968), and they play a fundamental role in the organization of face-to-face interactions. Even though backchannels “do not appear to be subject to the usual turn-taking constraints (i.e., avoidance of overlaps and long gaps) and may appear in overlap or be overlapped more frequently than other types of utterances” (Roberts et al., 2015: 120), they are not produced randomly by the listener in conversation. Instead, they seem to be locally timed to the current speaker’s

speech by monitoring linguistic and phonetic turn-taking cues (e.g., Gravano and Hirschberg, 2009a, 2011), and to follow specific patterns which are related to the modality in which they can occur in face-to-face interactions, i.e., visual or verbal (e.g., Truong et al., 2011), and to the social background of the speakers, in particular their gender (e.g., Zimmermann and West, 1975).

The current work aims to explore the extent to which the gender of both the listener and the current speaker play a role in multimodal backchannel timing and planning in face-to-face spontaneous dialogues in German. More specifically, the study tests whether male and female listeners time their feedback differently in relation to the current speaker's turn and, by analyzing the phonetic features occurring in the preceding speech, whether the gender of both interlocutors influences which turn-taking cues and which patterns of phonetic variation are used to locally guide the planning and production of backchannels across modalities. More broadly, this investigation seeks to test the hypothesis advanced in Rossi (2022), who suggests that gender is indexed not only in the acoustic features of speech taken in isolation, but also in how interlocutors employ and perceived turn-taking-related variations of these cues. Ultimately, the study aims at informing theories on gender-related conversational dynamics by providing an empirical and quantitative account of the influence of gender on multimodal listener behavior.

9.2 Feedback in face-to-face conversation

Backchannels are defined as brief and unobtrusive contributions provided by the listener while the current speaker is talking in order to display an active role in conversation without claiming the turn (Ward and Tsukahara, 2000). The term “backchannel” was first employed by Yngve (1970), as such responses occur over the back channel of the interaction while the main channel is occupied by the current speaker's utterance. According to Yngve (1970), all conversational participants fulfill the role of speaker and listener simultaneously, and they do so by switching from one channel to the other, either producing full conversational turns or non-competitive minimal responses. Cutrone (2014), following Markel (1975)'s definition of turn-taking, adopts the view that, in interactions, along with the full-fledged

speaking turns, “a concurrent listening turn” also exists, “which is the behavior of one or more non-talking interlocutors present” (Cutrone, 2014, p. 84).

While backchannels’ main function is generally indicated to be that of signaling active listenership, reacting to what is being said by the speaker and encouraging the speaker to continue (Duncan, 1974; Schegloff, 1982; Ward and Tsukahara, 2000), they are used in many different ways. Notably, they can be used to refuse the conversational floor. For instance, Ward and Tsukahara (2000), as well as Cutrone (2005), describe cases where the speaker yields the conversational floor to the interlocutor and waits for them to start a new turn, but the interlocutor produces only a “perverse passive” backchannel to refuse the opportunity to speak (Ward and Tsukahara, 2000, p. 1184; cf. Jefferson, 1984). Similarly, cases where subjects use backchannels as a way to contribute without having to take primary speakership are also reported in the literature (Cutrone, 2005).

In face-to-face conversations, interlocutors rely on the different modalities available to them to provide feedback, specifically, the vocal and the visual modality. In fact, in this type of interactions backchannels can be (at least) either verbal, gestural or multimodal (i.e., constituted by both a verbal and a gestural component occurring simultaneously).

Verbal backchannels tend to have a short constituent size and to be prosodically reduced in comparison to surrounding speech (Gardner, 2001; Müller, 1996). In particular, they tend to be short, non-interrupting utterances, comprising either one or a few syllables, and can have the form of single-word lexical expressions, short phrases, or non-lexical vocalizations (Liesenfeld and Dingemanse, 2022; Mulac et al., 1998; Zellers, 2021). Some examples of listener responses in German are “ja” (‘yes’), “genau” (‘exactly’), “stimmt” (‘true’), “mhmm”. Backchannels, however, can also have longer forms, as they can be constituted by complex phrases or by combinations of multiple lexical and/or non lexical units (Krepsz et al., 2022; Totte, 1991). Some studies (Cutrone, 2011, 2014; Stubbe, 1998) have suggested the existence of a continuum between forms, which is also associated with the context and the level of involvement of the listener in the conversation, with shorter and reduced forms on one end, suggesting lower support and involvement, and longer and more salient forms at the other end of the continuum, which seem to display higher engagement levels.

In the visual modality, several types of movements have been described as part of the listener behavior. For instance, changes in posture (Birdwhistell, 1966), eyebrow movements, facial expressions, gaze patterns (Paggio and Navarretta, 2011) and eye blinks (Hömke et al., 2018) are all visual means employed by listeners to provide feedback to the current speaker. However, the most frequent type of backchannel in the visual modality has been found to be head movements (Cerrato, 2012; McClave, 2000; Paggio and Navarretta, 2011), both on their own or in co-occurrence with verbal feedback (Dittmann and Llewellyn, 1968). Head nods are the most common type of movement used as a backchannel (Paggio and Navarretta, 2011), as well as the most studied one, but other types of head movement have also been observed. For instance, Rossi et al. (2023) also identified protrusions, tilts, slides and turns in German dialogues, occurring as gestural feedback, or as part of a multimodal unit with verbal and non-verbal vocalizations (i.e., non-lexical backchannels and laughter).

As mentioned in Section 9.1, backchannels do not appear to follow the common turn-taking constraints associated with full-fledged turns, and frequently occur in overlap with the current speaker's speech (Roberts et al., 2015). However, results from previous research suggest that, generally, listeners rely on multimodal alternation in order to cause the least disruption for the developing turn (Heldner et al., 2010). For instance, Truong et al. (2011) and Ferré and Renaudier (2017), investigating Dutch and British English interactions respectively, find that verbal backchannels occur mostly during silent pauses while head movements are more frequently observed in overlap with the speech by the current speaker. This pattern has also recently been reported by Rossi et al. (2023) for German dialogues. Additionally, Rossi et al. (2023) find that, in multimodal feedback responses, the onset of the head movement tends to anticipate the onset of the vocal component by 84 milliseconds on average for lexical feedback items, and by 200 milliseconds for non-lexical forms. This result, in line with a previous finding by Dittmann and Llewellyn (1968), suggests that the less disruptive feedback type (gestural) would anticipate the more disruptive one (verbal) in order to prepare for the upcoming verbal production, in an effort to be less disruptive or competitive for the current turn (Dittmann and Llewellyn, 1968; Rossi et al., 2023).

9.3 Feedback and turn-taking cues

Interlocutors involved in a conversation do not intervene at random, but rather time their responses based on a complex set of linguistic and phonetic turn-taking cues which projects Transition Relevance Places (TRPs), i.e., locations in the current speaker's turn where speaker change becomes a possibility (Sacks et al., 1974; Potential Transition Locations, Schiffrin, 1987; Potential Turn Boundaries, Zellers, 2017), and, generally, try to minimize silent gaps and speech overlaps (Sacks et al., 1974). As a TRP arises in the speech by the current speaker, the listener can either launch his/her speech for a next turn at an appropriate time, or remain silent in case the cues in the interlocutor's turn project more talk. A number of studies have investigated the variation of phonetic turn-taking cues towards PTLs. In general, they find that, in combination with the syntactic information of the current utterance, the cues that project a turn change are: a rising or falling F0, a significant decrease in intensity levels and a slower speech rate (some studies, however, find an increased speech rate to precede turn changes and a slower speech rate before turn holds, e.g., Brusco et al., 2020; Koiso et al., 1998; Zellers, 2017); on the other hand, the cues signaling turn-holding intentions are: lack of syntactic/semantic completeness, a flatter F0 contour, louder speech and a more modal phonation quality (Gravano and Hirschberg, 2011; Koiso et al., 1998; Local and Walker, 2012; Local et al., 1986; Niebuhr et al., 2013; Rossi, 2022; Włodarczak and Heldner, 2022; Włodarczak et al., 2023; Zellers, 2021).

Over the course of the current turn, not necessarily only in co-occurrence with TRPs marked by turn-hold cues, a listener might produce a non-interrupting backchannel, in order to display participation and attention without claiming the floor. Since backchannel placement and frequency are very important so as not to disrupt the flow of the interaction and to maintain rapport between speakers (Krause-Ono, 2004; Poppe et al., 2011), it has been hypothesized that, just as they would mark a TRP using specific turn-taking cues, current speakers could also use turn-taking cues to mark specific locations in their turn where it would be appropriate for the listener(s) to produce a backchannel. Such locations in conversation have been defined as “backchannel relevance spaces” (Heldner et al., 2013) or “feedback relevance spaces” (Howes and Eshghi, 2021), while the cues used by the cur-

rent speakers to signal them have been termed “backchannel inviting cues” (Gravano and Hirschberg, 2009a, 2011). Feedback relevance spaces have been found to occur mostly at clause boundaries after a semantically complete unit, or also earlier in cases where the following lexical content of the utterance is easily predictable (Howes and Eshghi, 2021). The phonetic cues that have been identified as “backchannel inviting” for the listener are a 110ms region of low pitch (Ward and Tsukahara, 2000), a rising final intonation together with higher pitch levels, higher intensity levels and lower noise-to-harmonic ratio values (Gravano and Hirschberg, 2011), or a rising intensity for visual backchannels and a decreasing intensity profile for verbal ones (Truong et al., 2011).

However, identifying the current speaker as the one inviting, or, somehow, regulating backchannel distribution, and thus adopting the concepts of “feedback relevance spaces” and “backchannel inviting cues”, may not be entirely fitting to provide a realistic account of listener’s behavior and how different linguistic and sociolinguistic factors might have an impact on it. In fact, even though fundamental for the development of a conversation, it is an intrinsic feature of backchannels to be optional during the current speaking turn (Pope et al., 2011, 2013; Ward and Tsukahara, 2000). Thus, it is the listener that locally decides if, when and where to produce a backchannel. Such a decision could be highly dependent on several different factors, including culturally and socially defined ones, as highlighted in Section 9.4.

The current study will thus adopt a slightly different perspective and move the focus onto the listener. More specifically, this study will consider as Feedback Locations (FLs) only those points in the current turn where a backchannel is actually produced by the listener, and will not pay attention to other locations where a backchannel could plausibly have been produced but, in actuality, was not (c.f. Heldner et al., 2013). With this approach, we aim to provide an ecologically valid account of when, where and, possibly, why particular listeners, paired with particular conversational partners, produce a backchannel during the current speaking turn. In other words, by adopting the listener’s perspective, we assume that it is not the speakers who offer the chance to produce a backchannel using specific set of cues; rather, that it might be dependent on the listener’s interactional attitudes to locally determine the most appropriate timing and location to produce feedback.

9.4 Backchannels and sociolinguistic variation

In sociolinguistic research, the use of feedback in conversation has been a source of great interest for the investigation of culturally defined interactional behaviors. In fact, since backchannels are not as constrained as other conversational moves, and are highly contextually variable, they can reveal some of the underlying dynamics upon which a conversation is built (e.g., Coates, 2004; Dingemanse et al., 2022; Kraaz and Bernaisch, 2022; Stubbe, 1998; Vinciarelli et al., 2015; Zimmermann and West, 1975). Cross-linguistic analyses of feedback use have found differences in listeners' baseline for backchannel frequency, which has been connected, for example, to the cultures' tolerance for interruptions and silences, or the importance given to the overt display of attentiveness and support in conversation (Cutrone, 2011; Stubbe, 1998, 2014). In fact, the norms governing interactional behavior often exhibit social and cultural variability and, even though backchannels in conversations play a comparable fundamental role across cultures, particular features of the feedback mechanism adopted by the listeners can display significant differences depending on the language and on the societal context (Stubbe, 1998).

Within a given language, gender has been identified as an important source of variation for conversational behavior and communication styles (Coates, 2004; Tannen, 1990), which also include the use of listener feedback in terms of form, frequency and distribution within the turn in interactions (Feke, 2003; Helweg-Larsen et al., 2004; Krepsz et al., 2022; Tannen, 1990; Zimmermann and West, 1975). As for the general conversational behavior stereotypically associated with either female or male speakers in Western contexts, a more supportive and tentative speech, with a higher frequency of overlapped utterances and co-constructed turns, is associated with the conversational style employed by female interlocutors (Edelsky, 1993; Menz and Al-Roubaie, 2008; Stubbe, 2013), while a higher amount of silent gaps, longer turns or rapid exchanges of short ones, and less overlapped portions of speech are features generally connected to male speakers' conversational styles (Brescoll, 2011; Coates, 2004).

As far as backchannels are concerned, a general tendency in the results of studies on different languages points towards female speakers backchanneling more frequently than male speakers, with both verbal and gestural resources (e.g., Ha-

balet, 2019; Helweg-Larsen et al., 2004; Kjellmer, 2009; Ueno, 2004) or timing their backchannels in order to leave fewer or shorter silent gaps (Beňuš et al., 2007; Kreps et al., 2022). This is, however, not always the case. In fact, a number of investigations do not report a significant difference in backchannel use between male and female speakers, in either frequency or distribution (e.g., Jiahong Yuan, 2007; Zellers, 2021), or even claim that their findings seem to be in contradiction with what male or female behavior is stereotypically supposed to look like, with male listeners backchanneling more frequently than female listeners in conversation with women (e.g., Dixon and Foster, 1998). More broadly, results from various studies concerned with how men and women speak have not consistently reproduced previous findings, leading to skepticism and questioning of long-standing assumptions about gender-specific language usage patterns (Plug et al., 2021).

The apparent inconsistency between results from different studies highlights the need for a less static approach to the observation of gender variation in sociolinguistic studies. In fact, in the past few decades, the focus has shifted away from research aimed at defining the ways in which men versus women talk as a result of societal power imbalance (e.g., Lakoff, 1973, 1975), or different socialization patterns and cultural backgrounds (e.g., Tannen, 1990; Uchida, 1992). More recent investigations involving gender instead adopt what has been defined as the dynamic approach (Coates, 2004), which considers gender not as a fixed category tied to biological sex, but rather as a cultural behavior, intertwined with other social variables such as social status, age and ethnicity. According to this theoretical framework, men and women 'perform gender' (Butler, 1993), 'do gender' (Goffman, 1976; Speer and Stokoe, 2011; West and Zimmerman, 1987), or 'index gender' (Ochs, 1992), i.e., they choose to adopt culturally shaped masculine or feminine language patterns to different extents based on the communicative situation in which they are involved and on their conversational goals (Plug et al., 2021). In the dynamic framework, the context of the interaction plays a fundamental role, in that the speakers might or might not perform different gendered behaviors to based on the perceived relevance of their gender to the specific situation (Coates, 2004; Marche and Peterson, 1993; Plug et al., 2021). For instance, Leaper and Ayres (2007) and Leaper and Robnett (2011), investigating assertiveness and tentative language respectively, indicate that in task-oriented dialogues gender-specific language use tends to be more prominent

than in unstructured dialogues. To explain this, Leaper and Robnett (2011) suggest that activities like debating, negotiating, or problem-solving, being "masculine gender-typed contexts" (p. 138), might make gender roles more salient, leading some men to prioritize authority and some women to emphasize cooperation and sensitivity. Vinciarelli et al. (2015)'s investigation further emphasizes the activation of gender stereotypes in a cooperative task. Additionally, Leaper and Ayres (2007) and Hannah and Murachver (2007) note that gender assumes greater importance in mixed-gender dialogues. Notably, Hannah and Murachver (2007) also propose that speakers tend to exhibit accommodating tendencies in unfamiliar contexts, gradually adopting more gendered speech patterns as they become more acquainted with one another.

Besides speech acts, gender is also performed through phonetic detail. While biological sex is at the physiological base of some differences in the way men and women speak, for instance, women use a higher pitch (Mennen et al., 2012; Weirich et al., 2019) and a breathier voice than men (Mendoza et al., 1996; Van Borsel et al., 2009), while men tend to be louder (Gelfer and Young, 1997) and to speak faster than women (Pépiot, 2014), a number of cross-linguistic investigations indicate that the speaker's gender and cultural background play a significant role in shaping the size of said differences. For instance, Van Bezooijen (1995) observed that the distinctions in F0 values between male and female speakers tend to be more pronounced in Japanese compared to Dutch subjects, reflecting varying perceptions of attractiveness attributes in the respective cultures; Weirich et al. (2019) noted larger differences in F0 and vowel space among German men and women in contrast to Swedish, together lower self-ascribed femininity scores reported by Swedish women. Within the same language, Weirich and Zahner-Ritter (2023) discovered that speakers from Eastern Germany indexed femininity through vowel space, a trend not observed in speakers from Western Germany, a discrepancy which the authors hypothesize may be connected to the more egalitarian perception of gender roles in the East, encouraging a performance of female identity through acoustic variation.

Thus, phonetic detail together with conversational practices and behaviors are used by speakers to convey identity and social meaning. However, studies bringing together these two aspects and investigating how they are connected to each other are currently lacking, and gender-specific phonetic variation in marking conversa-

tional intentions and regulating turn-taking in conversation currently represents a gap in the literature. Based on the findings reported above, it can be hypothesized that gender performance might also be observed in the way speakers of a particular language produce the different turn-taking cues which regulate the timing of turn changes and turn keeps in natural interactions. Additionally, and most importantly, this might extend to the way listeners in conversation respond to those cues, influencing the timing and the distribution of feedback responses in alignment with what they consider more appropriate for the given context.

A first example in this direction is given by Rossi (2022)'s study, which reports significant gender differences in the turn-taking related acoustic variation approaching TRPs (referred to as Potential Transition Locations in that study) in German spontaneous dialogues. In particular, she found that both male and female speakers put more effort into signalling their conversational intentions to either cede the floor or to keep it when talking to a male conversational partner. More specifically, in conversation with a male interlocutor, speakers displayed significantly lower intensity and slower speech rate to yield the turn, or used higher intensity and faster speech rate to keep the floor than when in conversation with a female conversation partner. Another study which suggests the possibility of phonetic variation for interactional purposes was carried out by Whiteside (1995, 1996), who show that male and female speakers of British English have a tendency to mark syntactic boundaries using different prosodic patterns. She finds that women tend to pause more and produce final lengthening approaching boundaries, while men use more F0 shifts and tend not to pause, which could require extra effort for a conversational partner to take the floor.

9.5 Objectives

The present study aims at investigating whether, and to what extent, the gender of the listener and the gender of the current speaker involved in a dialogue influence the timing of backchannels in spontaneous cooperative interactions. Since in the conversations investigated the interlocutors were face-to-face with each other, the modality of the feedback will be considered as well. Moreover, we take into account phonetic cues which are relevant for the turn-taking mechanism and we

explore whether the gender of both interlocutors has an effect on their patterns of variation in the speech preceding feedback, in order to assess if different listeners rely more or less on specific cues to locally plan and produce backchannels in different modalities. In particular, our research questions are the following:

1. To what extent does the gender of the listener and the gender of the current speaker influence the timing of backchannels in face-to-face dialogues?
2. Do phonetic cues such as F0 values, intensity, speech rate, and voice quality vary in the speech preceding feedback, and does the gender of both interlocutors play a role in how these cues are utilized for planning and producing backchannels?
3. How does the modality of the feedback (verbal, gestural or multimodal) factor into the timing and placement of backchannels and in the way they are used by different listeners?

Given the existence of gender-related preferences and tendencies, as reported in the literature on men's and women's conversational habits and dynamics (see Section 9.4), we expect to discover different patterns of variation for both backchannel timing and planning. For instance, the higher tolerance for overlapped speech reported for women's style and the higher frequency of silent gaps in men's talk might translate into different timing patterns for backchannels, with female listeners potentially providing feedback more frequently in overlap, while male listeners tending to do so more often during silent pauses. This might be relevant for preceding turn-taking cues as well. As a consequence of such possible preferred dynamics, it is likely that female listeners might time their feedback with phonetic cues resemble those marking turn keeps, while male listeners could react to cues more similar to the turn-yielding ones. These cues may involve either one or more of the acoustic cues considered, i.e., F0, intensity, speech rate, and phonation variations. An effect of the gender of the current speaker might also be possible: based on previous research (e.g., Hannah and Murachver, 2007; Leaper and Ayres, 2007), we might expect to find more striking differences between male and female listeners' behavior in mixed-gender dyads, where gender might be perceived as more salient. However, we do not exclude the possibility of observing accommodation tendencies between interlocutors. Finally, the modality of the feedback should also be relevant for

backchannel distribution. Given the reports provided by a previous analysis on the same corpus that is employed in the current study (Rossi et al., 2023), we anticipate a general trend of backchannels in different modalities to be produced where they might potentially cause the least disruption for the speech by the current speaker.

9.6 Methods

We analyzed 12 face-to-face dialogues involving 24 adult native speakers of German, for a total of 1235 utterances. The dialogues come from the German sub-corpus of the DUEL Multi-lingual Multimodal Dialogue Corpus (Hough et al., 2016). The experimental setting of the corpus data collection had the two dialogue participants sitting across from each other, each with close lapel microphones to capture the speech of the two participants on separate channels, allowing for the analysis of both speech signals even within portions of overlapped speech. Each dyad was assigned one of two interactional scenarios, i.e., either the “Dream Apartment” or “Film Script” scenarios, which were designed to have the two participants start talking without having to spend time to find a topic of discussion, but allowing them to speak freely and develop spontaneous exchanges of turns and the development of spontaneous dialogue phenomena and structures. In the “Dream Apartment” scenario, participants were told to discuss the plan and the furnishing of a shared apartment with a large sum of money at their disposal (Hough et al., 2016; Kousidis et al., 2013), while in the “Film Script” scenario interlocutors were asked to come up together with movie scene involving an embarrassing situation, which could also be based off of their personal experience (Hough et al., 2016).

The dataset analyzed for the current study is balanced for both interlocutors’ genders. The participants’ self-reported genders in the corpus were “Male” and “Female”, so the dyads were selected in order to have the same number of males in conversation with males and with females, and the same number of females in conversation with females and with males. Thus we have annotated and analyzed three female-female, three male-male and six female-male conversations. To explore the possible variability in listener behavior when in conversation with a male or a female speaker, we coded the gender of the listener as “Male” or “Female” (as self-reported in the background questionnaire filled out by the subjects recorded for the DUEL

corpus) and the gender of the interlocutor as either “Same” to describe same-gender dyads or “Other” for mixed-gender dialogues.

Speech transcription and segmentation into turns and utterances was provided with the corpus in Praat TextGrids (Boersma and Weenink, 2023). In particular, as per the corpus annotation guidelines (Hough et al., 2016), turns are defined as uninterrupted stretches of speech by the same speaker, while utterances are segmented following Meteor et al. (1995)’s definition of “slash units”. Based on this definition, utterances can be full syntactic sentences or units below the sentence level which “correspond to those parts of the narrative which are not sentential but which the annotator interprets as complete” (Meteor et al., 1995 cited in Hough et al., 2016, p. 1786). Word, syllable and phonetic segmentation was carried out using the WebMAUS pipeline without Automatic Speech Recognition (Kisler et al., 2017; Schiel, 1999) and manually corrected in Praat.

9.6.1 Feedback annotation

All instances of head movements were manually identified and annotated using the video files in ELAN (Wittenburg et al., 2006), without the audio track, using the movement type tier provided by the Multimodal and M3D Multidimensional Labelling Scheme (Rohrer et al., 2023; Rohrer et al., 2020), based on Wagner et al. (2014)’s head movement classification. The left boundary for the gesture was placed at the beginning of the head movement, i.e., as soon as the head was displaced from the rest position, while the final boundary of the movement was placed when the head came back to its rest position and was completely still. In line with the annotation carried out on the Active Listening Corpus (ALICO) presented by Malisz et al. (2016), uninterrupted movements constituted a single annotation unit. This means that gestures with multiple cycles (e.g., repeated uninterrupted head nods) were included in the same unit. However, unlike Malisz et al. (2016), in case of repeated movements, the number of gesture cycles was not labeled at this stage of annotation, as it fell outside the scope of the current study. The annotation process paid attention to identifying and including only those head movements which were intentionally produced by the subject and not caused by a general movement of the upper body or a change in posture. The annotation was then transferred from ELAN into Praat.

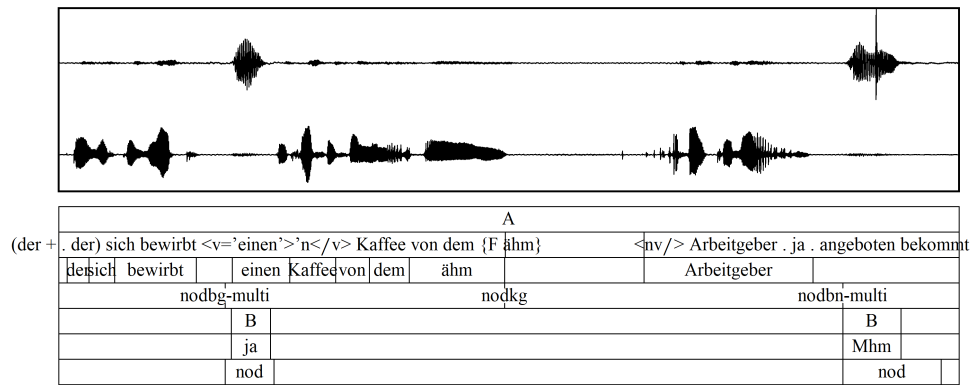


Figure 9.1: Example of the FL (Feedback Location) annotation in Praat. In this example, speaker A has the floor. The 2nd tier contains the transcription of the utterance, and the 3rd tier the segmentation of the utterance into words. B is listening and producing backchannels in both the verbal and the visual modality, annotated on the 6th and 7th tier respectively. FL are labelled on the 4th tier, using a point label placed at the left boundary of the backchannel, in correspondence with the current utterance. The reported annotation does not include the syllable and segment tiers, nor the head movement annotation for speaker A.

After that, all feedback instances, both verbal and gestural, were identified from the existing annotations. The guidelines used to identify which short utterances and which head movements were considered as feedback are the same used by Ward and Tsukahara (2000) and already applied to the DUEL Corpus in Rossi et al. (2023). Accordingly, we considered as backchannels all the short utterances or head movements that were (1) a direct reaction to the content of the current speaker's utterance, (2) optional, and (3) not acknowledged by the current speaker. Thus, head movements or verbal discourse markers occurring as the start or at the end of a full-fledged turn, answers to questions, and short clarification requests were not considered as backchannels.

Two tiers in Praat contained the annotations of the verbal backchannels and the head movements separately. A point label was then added below the current speaker's turn corresponding with the starting point of the backchannel item, i.e., the FL, which was then manually labelled using the set of labels already employed by Rossi (2022) indicating:

1. Completeness status, indicating whether the FL had reached a point of syntactic/semantic completion in context or not (*yes* or *no*);
2. Sentence type, indicating the form of the utterance preceding the FL, i.e., ei-

ther a declarative (*d*) or a question (*q*); based on our feedback identification guidelines, only the FLs labeled as declarative were included in the final dataset;

3. Sequential structure, indicating what type of conversational move followed the location, i.e., a turn keep (*k*) by the current speaker, a speaker change (*c*), or a backchannel (*b*) by the current listener; for the current study, only the location marked as backchannels were considered;
4. Transition type, describing the temporal relationship between the FL and the speech by the current speaker. Following Heldner (2011), who reported that the average detection threshold for gaps, overlaps and smooth transitions in conversations is 120 milliseconds, the transition type was labeled as gap (*g*) when the FL occurred after the offset of speech with an audible silence, i.e., longer than 120 milliseconds; it was labeled as overlap (*o*) if it occurred while the current speaker was still talking and the backchannel overlapped the speech for more than 120 milliseconds; finally, the transition was labelled as no-gap-no-overlap (*n*) if the FL occurred either smoothly after the offset of speech (without any gaps or with a gap inferior to 120 milliseconds), or in overlap with the current turn for less than 120 milliseconds (Heldner, 2011);
5. Modality, describing the modality in which the backchannel was produced, i.e., *verbal*, when the feedback was constituted by a lexical or a non-lexical utterance, *gestural*, when it was constituted by a head movement, or *multimodal*, for cases where both the visual and the verbal element occurred simultaneously.

An example of the annotations in Praat is shown in Figure 9.1.

9.6.2 Extraction of turn-taking cues

In order to analyze the variation of the phonetic cues in the speech preceding a FL, we extracted F0 values, intensity, speech rate and voice quality measurements from the syllable nuclei in current speaker's turn. Following findings by previous studies on the turn-taking related variation of phonetic cues and on the cues preceding

backchannels (e.g., Gravano and Hirschberg, 2011; Zellers, 2017), values were extracted starting from the syllable immediately preceding the FL up until one full second before it (syllables that were only partially included in this interval, e.g., started outside the 1 second interval but ended within it, were not used for this analysis). We also extracted the distance in milliseconds of the syllable from the FL or from the offset of speech for backchannels occurring after silent gaps. This information was stored as the variable *distance (from the FL)* to observe the influence of the distance from the FL on the variation of the acoustic values.

The extraction was done automatically using a Praat script written by the authors for the present study. Following a procedure similar to that proposed by de Jong and Wempe (2009), the script detected the loudest part of the syllable and then extracted the relevant phonetic features from the corresponding segment labeled on a separate tier (see Section 9.6), together with the segment's percentage of voiceless frames. After the extraction, all the segments with a voiceless section percentage below 30% were excluded from the analysis since, from our data, we assessed that this threshold was the most reliable to discriminate actual vowels from intense consonants, such as sibilant fricatives.

F0 values were extracted with the standard Praat settings for semitones above 1 Hz and were normalized using the individual speaker's mean in order to enable a direct comparison regardless of biological influences on F0. Intensity was extracted in dB and also normalized using the individual speaker's mean. The individual mean for both F0 and intensity was calculated based over the values extracted from speech intervals (excluding silences) for each recording. The individual mean was then subtracted from the values extracted with the script for the test segments. Speech rate, which was operationalized as syllable duration (Plug et al., 2022a), was then normalized for each speaker by subtracting their respective mean syllable duration. This mean was calculated by dividing the total number of syllables by the overall duration of the subject's speaking turns during the dialogue. Finally, voice quality was analyzed observing the variation of the Cepstral Peak Prominence (CPP), and it was normalized using the individual speaker's values of fully voiced and modal /a/ vowels (Murton et al., 2020) in stressed syllables, allowing us to visualize and account for the degree to which the extracted values would approximate or deviate from the fully modal value.

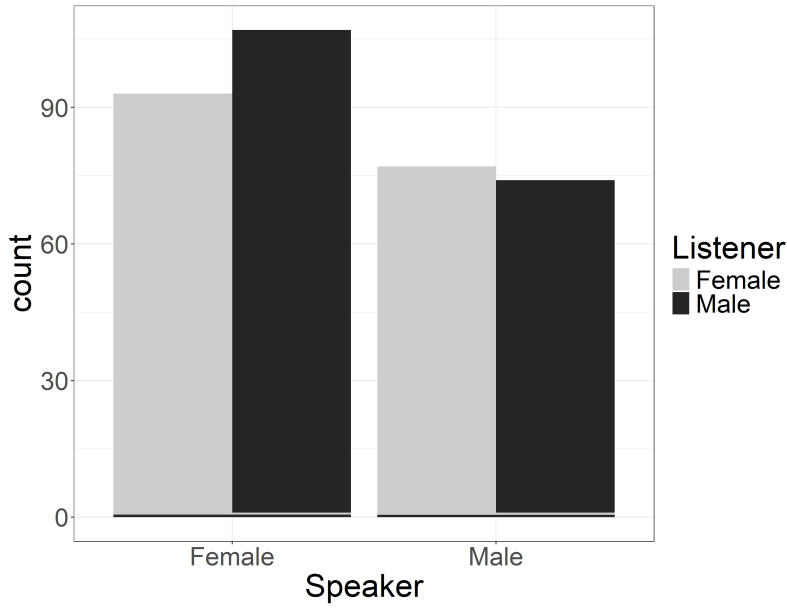


Figure 9.2: Distribution of backchannels produced by either male or female listeners (*Listener*) in conversation with either male or female conversational partners (*Speaker*).

9.7 Data and analysis

We extracted 351 FLs, i.e., locations where the listener started to produce feedback, occurring over 300 utterances by the current speaker (24% of the total number of the speakers' utterances in the dataset). Out of 351 FLs, 170 correspond to backchannels produced by a female listener, while 181 by a male listener and, from a first exploration, we observed that both male and female listeners produce more backchannels when the current speaker is female (see Figure 9.2). As for the modality of the response, we extracted 152 locations where the listeners produced verbal backchannels, 101 multimodal and 98 gestural ones. As it can be seen in 9.3, female and male listeners use feedback in different modalities with similar frequency.

The quantitative analysis of the data was carried out in R (R Core Team, 2022). First, using the package `nnet` (Venables and Ripley, 2002) we calculated a multinomial regression model to investigate the timing of multimodal backchannels by considering the transition type in relation to the current speaker's utterance. Secondly, using the acoustic values extracted from the current speaker's utterance, we analyzed which factors have an influence on the variation of the preceding phonetic turn-taking cues using linear mixed models with the package `lme4` (Bartoń, 2022) and post-hoc comparisons using the package `emmeans` (Lenth, 2022). Results are

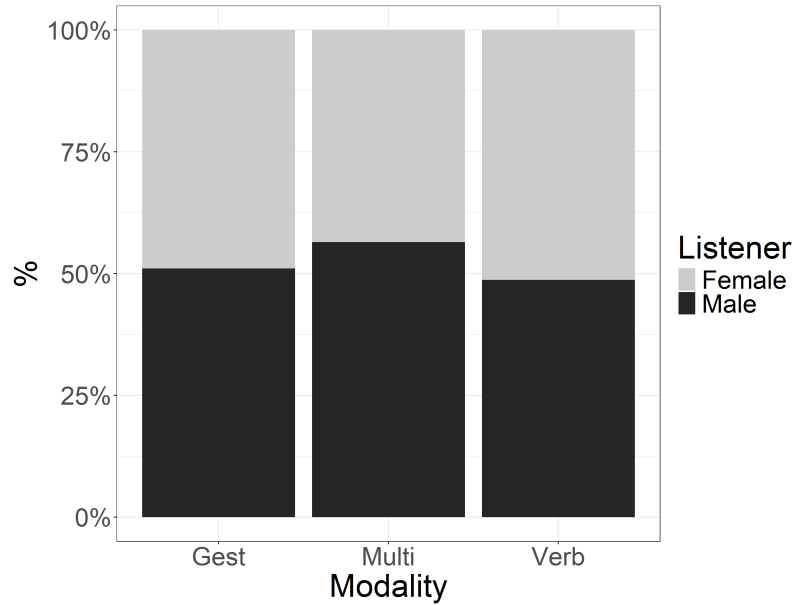


Figure 9.3: Frequency of listener responses in the different modalities (*Gestural*, *Verbal*, *Multimodal*) in our dataset produced either by a female listener (*Female*, *Gestural* = n. 48; *Verbal* = n. 78; *Multimodal* = n. 44) or a male listener (*Male*, *Gestural* = n. 50; *Verbal* = n. 74; *Multimodal* = n. 57).

visualized using the *emmeans*, the *sjPlot* (Lüdtke, 2022), the *Effects* (Fox, 2003; Fox and Weisberg, 2019) and *ggplot2* (Wickham, 2016) packages.

9.8 Results: Backchannel timing

This phase of the analysis focused on determining the relevant factors influencing the timing of multimodal backchannels in German dialogues. We calculated a multinomial regression model with the three different transition types (*Gap*, *No-gap-no-overlap*, *Overlap*) as outcomes and the gender of the listener, the gender of the current speaker, feedback modality, and their interactions as predictors, using a multinomial logistic regression model in R's *nnet* package.

Significant main effects for the transition type with which the feedback was produced were found for syntactic completeness of the FL and the modality of the backchannel; additionally, a significant interaction was found between the gender of the listener and that of the current speaker (see Figure 9.4 for the summary of the model). In general, feedback is significantly less likely to occur in overlap with the current turn when the location is syntactically/semantically complete (β : 3.33668, z value: 8.925, $p < .0001$), and when it is produced as a verbal (β : -1.66991, z value:

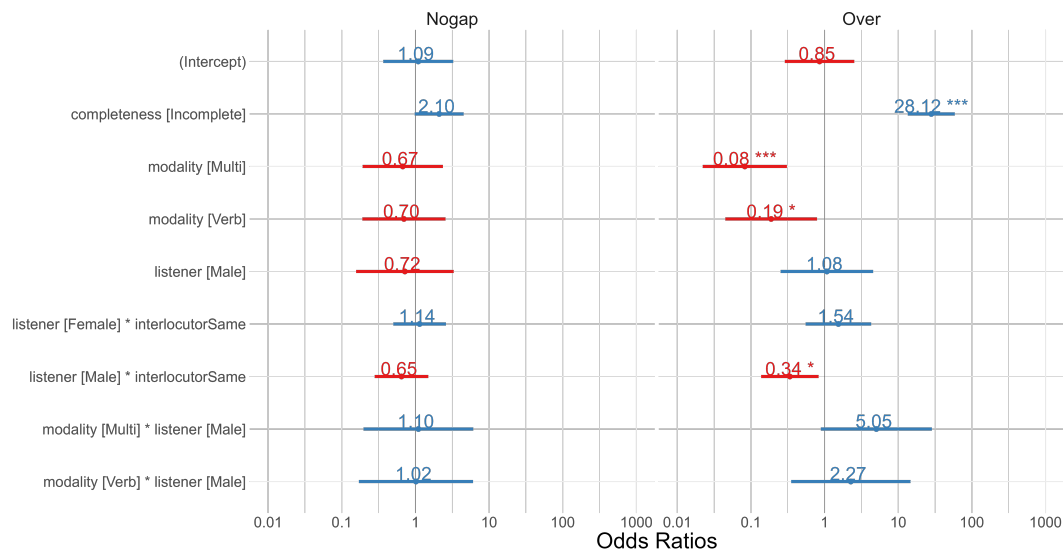


Figure 9.4: Summary of the multinomial model including the transition types, *No-gap-no-overlap*, *Overlap* and *Gap* as the categorical outcomes and the completeness status of the FL (*Complete* or *Incomplete*), the modality of the backchannel, the gender of the listener (*Male* or *Female*), the gender of the interlocutor (*Same* or *Other*), the interactions between the gender of the listener and the gender of the speaker and between the gender of the listener and the modality of the feedback (*Gestural*, *Multimodal* or *Verbal*). The Odds Ratios on the x-axis represent the likelihood of transitioning to the category *No-gap-no-overlap* or *Overlap* compared to the reference category *Gap* for a one-unit increase in the corresponding predictor variables or variable interactions.

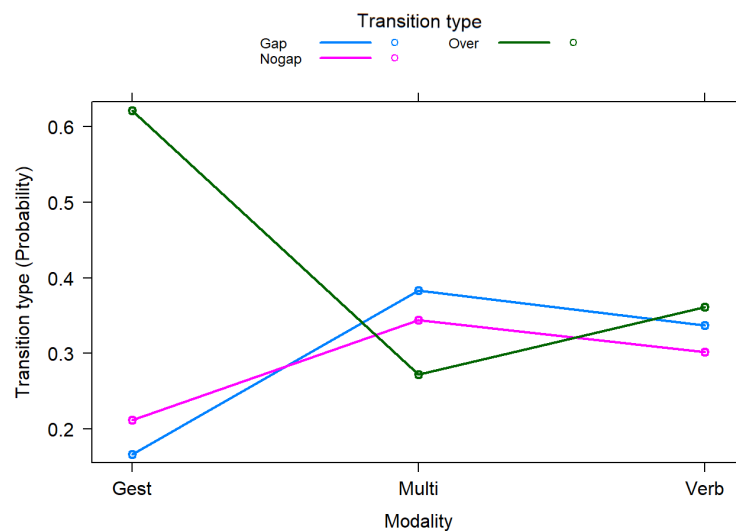


Figure 9.5: Effects plot of the multinomial model representing the likelihood of each transition type (*Gap*, *No-gap-no-overlap*, *Overlap*) occurring in relation to the modality of the backchannel (*Gestural*, *Multimodal*, *Verbal*).

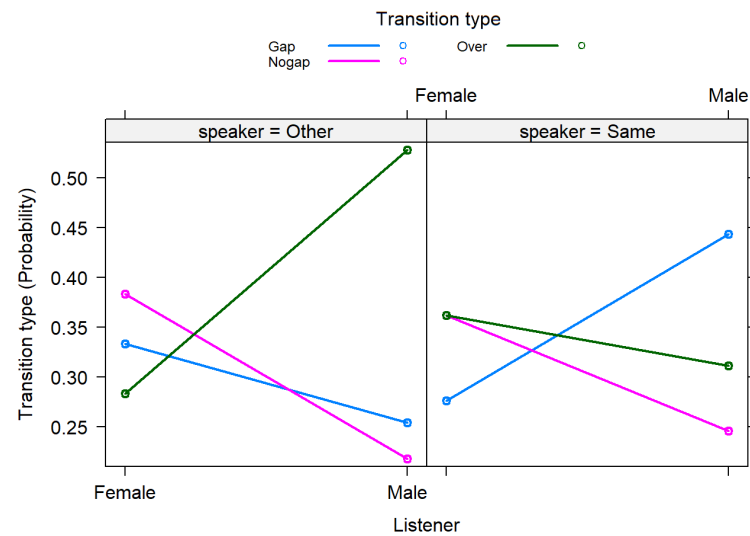


Figure 9.6: Effects plot of the multinomial model representing the likelihood of each transition type (*Gap*, *No-gap-no-overlap*, *Overlap*) occurring in relation to the interaction between the gender of the listener (*Male*, *Female*) and that of the interlocutor (*Same*, *Other*).

-2.289, $p < .05$) or a multimodal backchannel (β : -2.49318, z value: -3.725, $p < .001$), i.e., when it involves a vocal component, than when it is constituted by a gesture only (see Figure 9.5). Moreover, it appears that male listeners in conversation with male speakers tend to backchannel significantly less frequently in overlap (β : -1.08523, z value: -2.376, $p < .05$) than females do when listening to a male speaker. By taking a closer look at this interaction (see Figure 9.6), it emerges that in same-gender dyads, male listeners are more likely than females to produce feedback after silent gaps, while female listeners are more likely to backchannel in overlap or with a no-gap-no-overlap transition. On the other hand, in mixed-gender dyads, male listeners are more likely than females to produce backchannels in overlap, while female listeners are more likely to give feedback after a silent gap or with a no-gap-no-overlap transition.

From this first analysis, we observed that feedback timing in relation to the current speaker's speech can be predicted by the interaction of a set of linguistic and social factors, such as the completeness status of the FL, the modality used, and the gender of both interlocutors involved in the dialogue. In the following section, we will address whether, and to what extent, different combinations of such factors correlate with different patterns of phonetic variation in the speech preceding the production of a backchannel by the listener. In other words, we will test whether

listeners make use of specific patterns of linguistic and phonetic variation of turn-taking cues in order to locally plan when to produce feedback at an appropriate location in conversation.

9.9 Results: Acoustic configurations preceding feedback

From our dataset, we extracted 1131 syllables in the speech before backchannels and observed the variation of F0, intensity, speech rate and voice quality leading up to the FL. The variation of each acoustic parameter in relation to the variables extracted from the data and from the annotation, i.e., gender of the listener, gender of the current speaker, temporal distance from the FL, syntactic/semantic completeness of the FL, transition type, modality of the feedback, was tested using linear mixed effect models.

Since the interactions between these variables in the turn-taking structure can be multiple and multifaceted, we started this analysis by using R's step function for step-wise selection with models, for each phonetic parameter, which included all possible main effects and all possible two-way interactions between them in order to select the most appropriate set of predictor variables to fit the data. We also tested the contribution of possible random intercepts (the individual listener and the individual backchannel token) and random slopes (the individual listener and the transition type, the transition type and the modality). The only random factors that made a statistically significant contribution to each model were the individual listener and the individual backchannel token as random intercepts. We report below the models and the results obtained for each phonetic parameter.

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	-1.304	0.63	31.37	-2.070	0.047
listenerMale	0.825	0.88	32.39	0.933	0.358
Incomplete	0.622	0.29	337.96	2.113	0.035
Gap	-0.343	0.40	323.58	-0.851	0.395
Overlap	-1.066	0.42	336.70	-2.510	0.013
Male:Gap	0.600	0.60	336.24	1.001	0.318
Male:Overlap	1.607	0.56	308.11	2.866	0.004

Table 9.1: Summary of the linear mixed model analyzing the variation of F0 in the speech preceding feedback: $f0_{mean} \sim listener + completeness + transitiontype + (1|listener) + (1|token) + listener * transitiontype$.

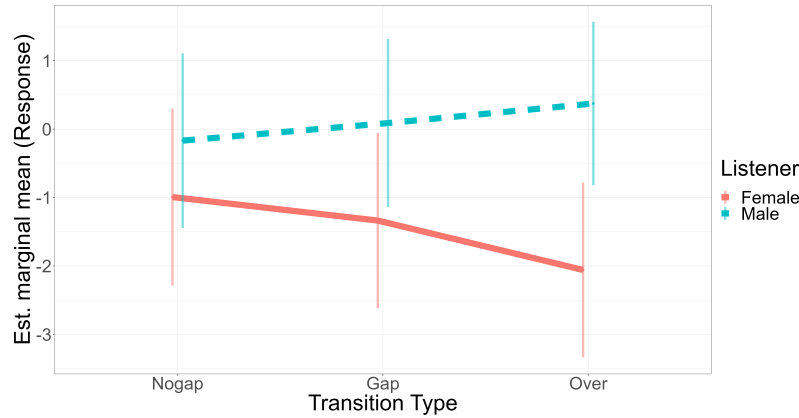


Figure 9.7: Estimated marginal means plot for the variation of F0 in the post-hoc contrast for the interaction between the gender of the listener (*Female*, *Male*) and the transition type of the feedback (*No-gap-no-overlap*, *Gap*, *Overlap*).

9.9.1 F0

The best-performing model found through the step-wise selection method for the variation of F0 in the speech preceding feedback included the main effects of the gender of the listener, the completeness status of the FL, the transition type of the feedback, as well as an interaction between the gender of the listener and the transition type. The summary of the model is reported in Table 9.1. There is a main effect of the syntactic/semantic completeness status of the location, with lower F0 values preceding backchannels produced after complete FLs and significantly higher F0 values preceding feedback at incomplete FLs (Incomplete-Complete: $\beta = 0.622$, t ratio: 2.113, $p < 0.05$).

The interaction between the gender of the listener and the transition type with which the feedback was produced was further tested using post-hoc contrasts with the *emmeans* package (Lenth, 2022). There is a significant difference between male and female listeners for backchannels produced in overlap with the ongoing speech of the current speaker. Irrespective of the gender of the speaker, a distinct pattern emerges in the behavior of male and female listeners. Specifically, male listeners appear to produce feedback in locations where the F0 preceding them is comparatively higher. On the other hand, in the same context, female speakers produce backchannels after lower F0 in overlap (Male-Female: $\beta: 2.433$, t ratio: 2.884, $p < 0.01$), see Figure 9.7.

9.9.2 Intensity

The most appropriate model for the analysis of intensity variation before FLs, suggested by the stepwise procedure, reveals a significant main effect of modality, of completeness, and of the interaction between the gender of the current speaker and the intensity contour as it approaches the FL. The summary of the model is reported in Table 9.2.

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	-2.052	0.46	118.49	-4.458	0.000
speakerOther	-1.005	0.53	71.09	-1.904	0.061
Gest	0.977	0.44	174.33	2.208	0.029
Multi	-0.457	0.41	321.84	-1.103	0.271
Incomplete	1.181	0.40	396.34	2.946	0.003
distance	-1.203	0.68	1119.03	-1.765	0.078
speakerOther:distance	-1.578	0.72	967.85	-2.196	0.028
distance:Gap	-1.111	0.62	660.00	-1.784	0.075
distance:Overlap	-0.350	0.64	873.38	-0.544	0.587

Table 9.2: Summary of the linear mixed model analyzing the variation of intensity in the speech preceding feedback: $intmean \sim speaker + modality + completeness + distance + (1|listener) + (1|token) + speaker * distance + transitiontype * distance$.

As for the main effect of the modality of the backchannel, the model reveals a significant difference in the intensity values preceding verbal and gestural tokens (Gestural-Verbal: β : 0.977, t ratio: 2.208, $p < 0.05$). In particular, gestural feedback is produced by listeners after speech with higher intensity values than verbal backchannels. On the other hand, the intensity values preceding multimodal items are lower than verbal backchannels, but this difference is not significant. The mean intensity values in the speech preceding FLs in different modalities are represented in Figure 9.8.

The interaction between the gender of the current speaker and the distance from the FL indicates that intensity values assumes different patterns of variation approaching the FL for male and female speakers. These are represented in Figure 9.9. Independent of their gender, listeners appear to backchannel after a decreasing intensity profile, i.e., at quieter locations, when in conversation with a speaker of a different gender from theirs. On the other hand, when in conversation with a speaker of their same gender, listeners produce feedback at comparatively louder

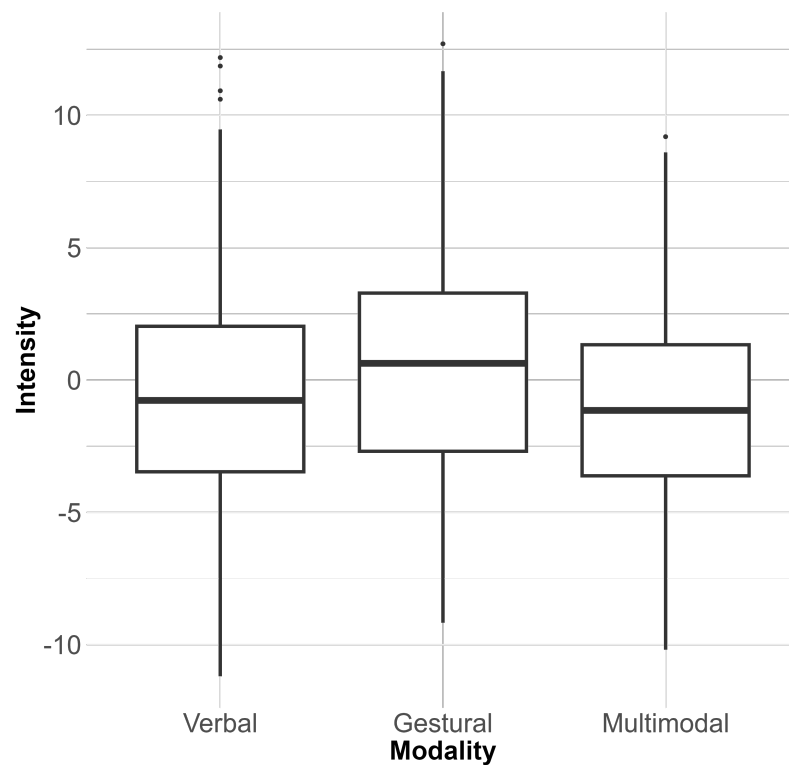


Figure 9.8: Mean intensity values (in dB, normalized with the speaker's individual mean) in the speech preceding backchannels across modalities (*Verbal, Gestural, Multimodal*).

locations. In these cases, in fact, intensity does not decrease approaching the FL (see Figure 9.9).

Even though it does not reach statistical significance in the model (see Table 9.2), the interaction between the distance from the FL and the transition types reveal interesting insights into how different intensity profiles could influence the timing of listener responses in conversation. In Figure 9.10, the graphical representation of intensity profiles preceding feedback suggest that listeners produce backchannels with different timings based on the varying intensity values. For instance, backchannels occurring in overlap are preceded by a higher intensity values. Backchannels produced after a silent gap are preceded by a declining intensity profiles, while backchannels produced with a no-gap-no-overlap transition are preceded by a declining intensity, but a rise in the final 250 milliseconds before the offset of speech. This difference between no-gap-no-overlaps and gaps approaches significance but does not reach it (Gap-No-gap: β : -1.111, t ratio: -1.784, $p = .07$).

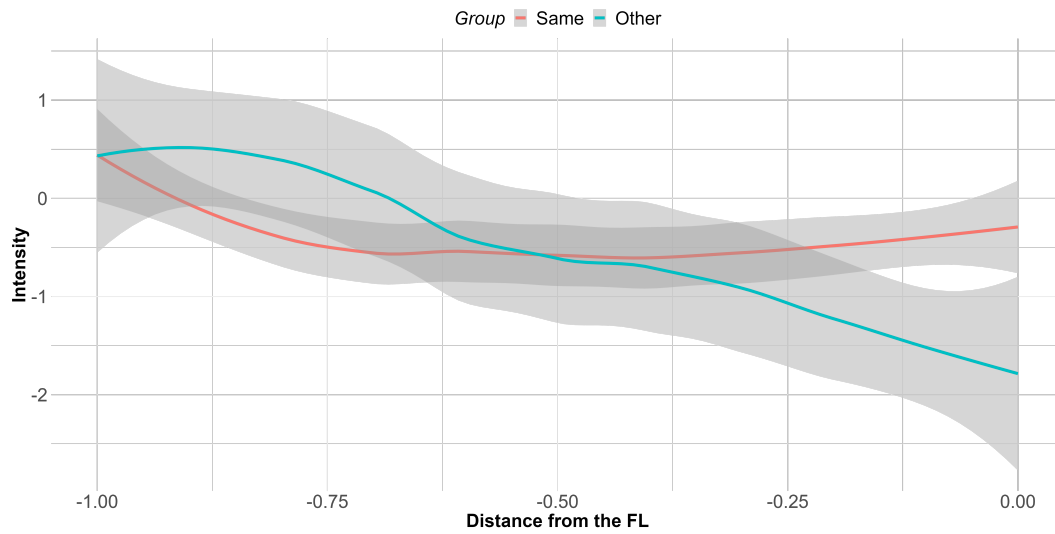


Figure 9.9: Intensity values' (in dB, normalized with the speaker's individual mean) variation over time approaching the FL in same-gender (*Same*) and mixed-gender (*Other*) dialogues.

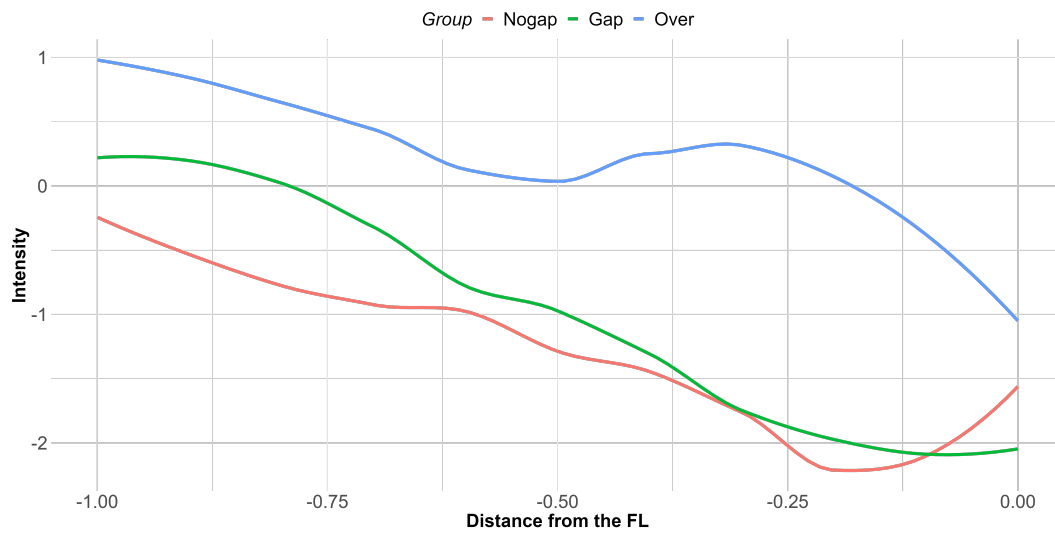


Figure 9.10: Intensity values' (in dB, normalized with the speaker's individual mean) variation over time approaching FLs with different timings in relation to the current speaker's utterance (*No-gap-no-overlap*, *Gap*, *Overlap*).

9.9.3 Speech rate

The model for the analysis of speech rate includes gender of the listener, completeness status of the FL, transition type of the feedback and distance from the FL as main effects, as well as the interactions between gender of the listener and transition type, transition type and distance from the FL, and completeness and distance from the FL. The summary of the model (see 9.9.3) reveals a significant interaction between the gender of the listener and the transition type with which the subject produced the backchannel. This interaction was further explored using post-hoc contrasts, which revealed male listeners' different backchannel timings in relation to varying speech rates in the current speaker's utterance. In fact, male listeners backchannel earlier (with a speech overlap or with a no-gap-no-overlap transition) after a faster speech rate by the current speaker, while they backchannel after a silent gap if the speech rate is slower (Gap-Nogap: β : 0.0319, t ratio: -2.826, $p < .01$; Overlap-Gap: β : -0.0261, t ratio: -2.469, $p < .05$). On the other hand, female listeners do not seem to rely on speech rate variations for the timing of their backchannels.

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	-0.041	0.03	41.35	-1.621	0.113
listenerMale	-0.042	0.03	26.61	-1.319	0.198
Incomplete	-0.013	0.01	1021.06	-0.848	0.397
Gap	-0.007	0.02	847.09	-0.375	0.708
Overlap	-0.041	0.02	894.46	-2.220	0.027
distance	0.086	0.02	1048.15	4.484	0.000
Male:Gap	0.040	0.02	190.66	2.665	0.008
Male:Overlap	0.015	0.01	161.41	1.078	0.283
Incomplete:distance	-0.047	0.02	1026.15	-1.990	0.047
Gap:distance	0.003	0.03	1062.93	0.110	0.912
Overlap:distance	-0.059	0.03	1023.43	-2.189	0.029

Table 9.3: Summary of the linear mixed model analyzing the variation of speech rate in the speech preceding feedback: $\text{duration} \sim \text{listener} + \text{completeness} + \text{transitiontype} + \text{distance} + (1|\text{listener}) + (1|\text{token}) + \text{listener} * \text{transitiontype} + \text{completeness} * \text{distance} + \text{transitiontype} * \text{distance}$.

Transition type is also involved in a second significant interaction for the speech rate model, i.e., with distance from the FL. In fact, as it can be seen from the model summary (see Table 9.9.3), speech rate varies significantly approaching the FL between the no-gap-no-overlap and the gap transition types. Since we previously observed a significant interaction between the gender of the listener and the transition

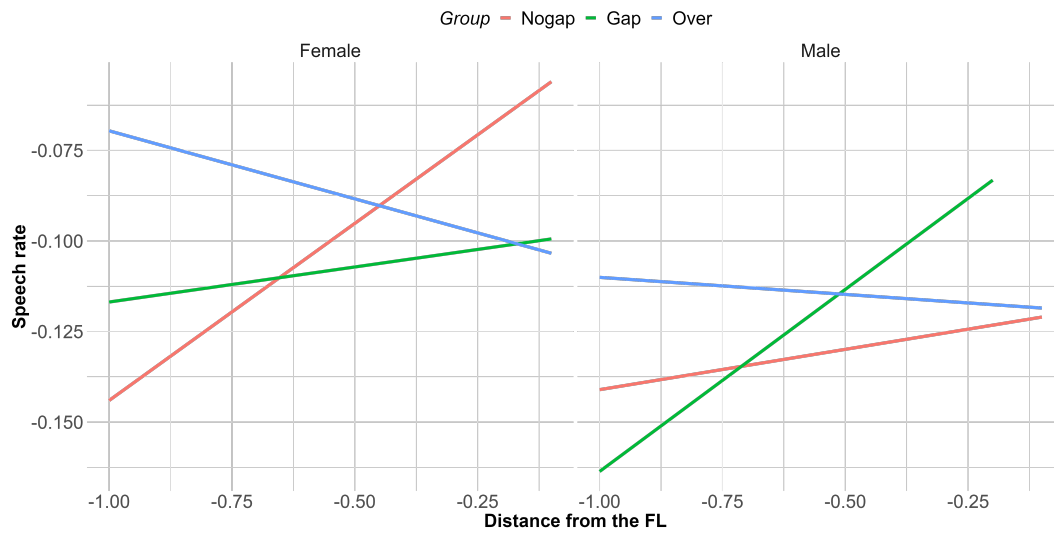


Figure 9.11: The plot visualizes the relationship between the temporal distance from the FL and the variation of speech rate (operationalized as syllable duration in ms, normalized with the individual speaker's mean) before the different transition types (*No-gap-no-overlap*, *Gap*, *Overlap*) in the male and female listener groups. The lines are the result of a linear regression model (lm) fit to the data visualizing the overall trend in the data.

type, we visualize speech rate variation over time separately for male and female listeners, in order to observe how they react to different speech rate variations in terms of backchannel timing. From Figure 9.11, it appears that, with a decrease in speech rate (i.e., syllable duration closer to the individual speakers' mean) towards the FL, female listeners produce feedback promptly, with a no-gap-no-overlap transition, while male speakers do so after a silent gap. On the other hand, a faster speech rate, with only a slight decrease towards the FL, precedes male listeners' backchannels with a no-gap-no-overlap transition.

The last significant interaction in the model for speech rate is between the completeness status of the FL and the distance from the FL, with syntactically/semantically complete or incomplete FLs being preceded by varying speech rate patterns over the last second of speech. More specifically, speech rate tends to be slower before incomplete FLs compared to complete FLs. This slower rate persists until around 300 milliseconds before the FL, where, after a decrease, speech rate appears to be slower before backchannels produced at syntactically/semantically complete locations (see Figure 9.12).

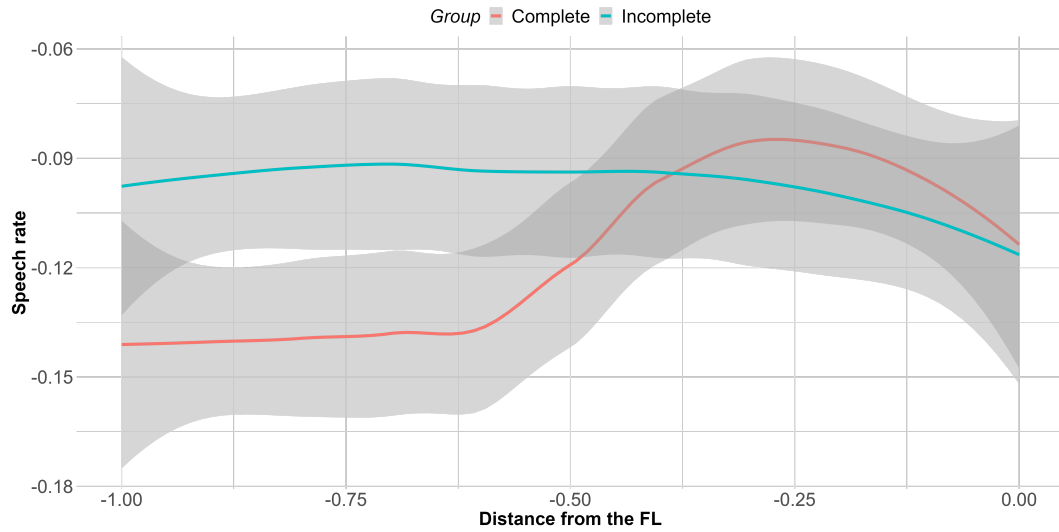


Figure 9.12: Speech rate's (operationalized as syllable duration in ms, normalized with the individual speaker's mean) variation over time approaching FLs with different syntactic/semantic completeness status (*Complete*, *Incomplete*).

9.9.4 Voice quality

To investigate variation in voice quality, measured through normalized CPP mean values, the model included only the main effect of completeness of the feedback location. Voice quality does not appear to be a significant cue influencing gender-specific or multimodal behavior in the placement of backchannels in the conversations of speakers within this dataset. It appears that the feedback locations where there is no syntactic completeness are the more modal ones, while the complete ones display significantly less modal values (Incomplete-Complete: β : 0.6059, t ratio: 2.385, $p < 0.05$) (see Table 9.9.4).

	Estimate	Std. Error	DF	t-value	Pr(> t)
(Intercept)	-0.303	0.19	45.40	-1.636	0.109
Incomplete	0.606	0.25	291.29	2.385	0.018

Table 9.4: Summary of the linear mixed model analyzing the variation of voice quality in the speech preceding feedback: $c_{pp} \sim completeness + (1|listener) + (1|token)$.

9.10 Discussion

We investigated the temporal relationship between backchannels and the ongoing speech by the current speaker in German face-to-face dialogues. Firstly, the results

from the quantitative analysis suggest that backchannel timing appears to be influenced by several different factors, such as the syntactic/semantic completeness of the location where the feedback item occurred, the modality in which it is produced, as well as the gender of the listener and the gender of the speaker. Listener feedback tends to occur after the offset of speech when the utterance has reached a syntactic/semantic complete point, or when the feedback involves a vocal component (i.e., verbal or a multimodal backchannels). Moreover, backchannels in overlap with the current utterance are more frequent when the interlocutor speaking is female, while backchannels after the offset of speech are more frequent when the current speaker is male.

After observing the differences in timing, we investigated the phonetic cues preceding feedback. It was highlighted that backchannels produced by female listeners tend to be preceded by lower F0 values in overlap with the current turn. Male listeners, on the other hand, produce backchannels after higher F0 values in the speech of their interlocutor. The modality of the feedback had a significant effect on intensity variation, with gestural backchannels being produced at the loudest FLs. The dyad composition, that is, the gender of the current speaker (either “Same” for same-gender dyads or “Other” for mixed-gender dialogues) in interaction with the distance from the FL had a significant effect on the variation of intensity, too. Specifically, in mixed-gender dyads, listeners backchannel at quieter locations after a declining intensity profile, while in same-gender dyads they tend to produce feedback after a steadier and louder intensity. Different speech rate patterns mark the speech preceding feedback depending on the gender of the listener, the transition type with which the feedback was produced and the distance from the FL. In particular, female listeners are faster to backchannel (i.e., backchannel with a no-gap-no-overlap transition) after slower speech, and males after faster speech. Finally, from our results, variations in voice quality appear to be significantly influenced by the syntactic/semantic completeness of the FLs, with higher CPP values being connected to syntactically/semantically complete locations preceding backchannels, and lower values with incomplete locations.

In line with previous research (e.g., Krepsz et al., 2022; Poppe et al., 2010; Ward and Tsukahara, 2000), backchannels produced by the listeners in our dataset appear to be locally planned by the listener to mostly occur at locations during the

current turn where they would potentially cause the least disruption. Multimodal variation plays a fundamental role in this. In fact, we observed head movements to be frequently produced in overlap with speech, while feedback involving a verbal component, either on its own or in combination with a gestural backchannel, tends to be preferred after the offset of speech. The availability of both the verbal and the visual modality in face-to-face conversation for the production of feedback is again confirmed to play a significant role in its distribution in relation to the current turn (cf. Ferré and Renaudier, 2017; Rossi et al., 2023; Truong et al., 2011).

The results highlight how several linguistic, phonetic and contextual cues are involved, to different extents, in the planning of the next conversational move by the interlocutors involved in a conversation. From our analysis, it appears that there might not be a predetermined set of cues which serve as inviting the production of a backchannel. Instead, different cues might become more or less relevant, or might assume one pattern or another, depending on the specific features of the approaching location, as well as the social identity of the listener, for whom different expectations on conversational behavior might be more or less valid (Engwall et al., 2023). In other words, the findings suggest that the timing and preceding cues for various backchannels are to be interpreted not only as the result of the influence of single main factors, but also in light of the diverse interactions between those. In line with the holistic approach to the analysis of conversational turns adopted, for example, by Couper-Kuhlen and Selting (1996), Ford et al. (1996), Local and Walker (2005), and Ogden (2001), which assumes that different combinations of turn-taking cues and factors project conversational intentions and marks different interactional structures, and thus no relevant variable should be excluded a priori, our results underscore the crucial impact of the broader conversational context on backchannel production.

This study specifically took into consideration the gender of both conversational participants in order to investigate the social dynamics related to multimodal feedback, and our results point towards the existence of gender-specific patterns. In terms of frequency and distribution, female speakers in mixed-gender dyads are more likely to receive backchannels in overlap with their current speech, while male speakers are more likely to receive backchannels after no-gap-no-overlap transitions or after silent gaps. This might point towards the adoption of a backchanneling

pattern related to both the perceived gender of the speaker and the listener accommodating to it, with women's communication style being associated with feedback produced while the ongoing turn is still unfolding, demonstrating a form of supportive overlap that is more rarely found in men's conversational behavior (Coates, 2004; Tannen, 1994). Such a tendency was already observed by Feke (2003). In her dataset, however, only female speakers seemed to accommodate to the backchanneling style of male subjects (Feke, 2003), while in our study both female and male listeners appear to do so.

Even though striking differences between same- and mixed-gender dyads did not emerge from our quantitative analysis of the preceding turn-taking cues, another case where the dyad composition was found to be relevant for backchannel production is the variation of intensity values in the speech preceding feedback. In fact, in this case, the intensity patterns observed in the last second of speech preceding the production of a backchannel varied based on whether the listeners were engaged in conversation with someone of the same or a different gender. This indicates that listeners tended to place their backchannels after different intensity values depending on the gender of their conversation partner. Specifically, in mixed-gender dyads (or with "Other" speakers), the speech preceding the FL was characterized by lower intensity values compared to same-gender dyads, where higher intensity values were observed. This observation suggests a potential tendency to minimize disruption in conversations with individuals of a different gender, and a lesser concern for disruption when engaging in conversations with individuals of the same gender.

Other studies have also highlighted the significance of intensity variations in connection to the gender of dialogue partners, not only in the context of backchannel production but also in shaping overall conversational dynamics. Together with variations in F0 levels, intensity is the cue for which Levitan et al. (2011) find consistent levels of local accommodation between dialogue partners in the speech preceding backchannels: listeners tend to produce feedback expressions after intensity levels that are increasingly similar, as the conversation progresses, to those after which their interlocutor also produces backchannels. Variations in conversational intensity connected to the gender of both interlocutors also emerged in the findings reported by Jurafsky et al. (2009). In their study on a sample of American English speakers in a speed dating setting, they find that, when male speakers talk to a female inter-

locutor who they perceive as friendly, they become more sensitive to variation in intensity and display a larger range in their own speech. While we did not specifically detect accommodation patterns in the use of intensity preceding feedback, our results suggest that intensity might be employed differently by listeners to time backchannels based on the gender of the current speaker and the gender of the listener.

Notably, intensity was the only phonetic parameter in our results on which variation modality had a significant effect in the quantitative analysis. In fact, in line with previous research on the speech preceding feedback in face-to-face interactions, specifically with the results reported by Truong et al. (2011), gestural backchannels in our dataset are consistently preceded by higher intensity values than backchannels involving a verbal component. As Truong et al. (2011) suggest, this result can be connected to the tendency of visual backchannels to occur mostly in overlap with speech, while verbal backchannels arise during silent pauses. This indicates a significant role played by intensity in the planning of listeners' backchannels across modalities during the current speaker's utterance. Our results also suggest that the variation of intensity over time has the possibility to influence, to some extent, the timing of the backchannel in relation to the current speaker's utterance. Listeners seem to react differently to various intensity profiles in the speech preceding the FL. For instance, a decrease in intensity, commonly associated with signaling turn-yielding intentions in conversation (e.g., Gravano and Hirschberg, 2011; Rossi, 2022 for the same corpus), is followed by a tendency among listeners to produce backchannels after an audible silent gap. Although further analysis is required, it can be hypothesized that some of these backchannels, produced in such contexts, may serve as a means for listeners to decline the offer of the next turn from the current speaker by using a "perverse passive" backchannel (see Section 9.2). On the other hand, a decrease followed by a rise in the intensity profile before the FL was associated with backchannels promptly produced at the offset of speech by the listener, without any perceivable silent gap or speech overlap (i.e., with a no-gap-no-overlap transition). Overall, these results point towards intensity playing a fundamental role for the listener in both the planning and production of multimodal backchannels, together with a possible relevance linked to the gender of the interlocutors in the dialogue.

Further differences between female and male listeners' use of phonetic turn-taking cues to plan backchannels emerged in our data for F0 and speech rate. Firstly, female listeners produce backchannels in overlap with the current speaking turn after speech intervals marked by low F0 values, below the current speaker's mean, while male listeners backchannel after a higher F0. Areas of low F0 in the current turn have been documented and described by Ward and Tsukahara (2000) as a feature which precedes the production of a backchannel by the listener, since they tend to come after a new information has been provided, in combination with the end of grammar clauses, or with disfluencies (Ward and Tsukahara, 2000; Ward, 1999). Thus, female listeners in our dataset paying seemingly more attention to this cue for feedback in overlap with the current speaking turn might be indicative of a higher attention to the timing of feedback at a location where it would not be perceived as disruptive.

As for speech rate, the different patterns of response to prosodic variation highlighted in the analysis seem to reflect gender-specific turn-taking strategies related to boundary marking observed in previous studies focused on the speaker (e.g., Rossi, 2022; Whiteside, 1995, 1996). In the current analysis, we observed that female listeners tend to backchannel more promptly after areas of slower speech rate in the current turn, while males do so after a faster speech rate. This may indicate that, for male listeners, a faster speech rate serves as a clearer indication that the current speaker intends to hold the turn, possibly making it more challenging for the interlocutor to take over, contrary to slower speech rate (Whiteside, 1995, 1996). In this dataset, it makes sense that male listeners would backchannel more promptly after this cue, as the current speaker may put more effort into maintain the floor when the listener is male, as previously observed by Rossi (2022) for the same data. This further substantiates Rossi (2022)'s finding by suggesting that, on the receiving end, male listeners possibly perceive such a cue as a turn-holding indication. The slower speech rate preceding backchannels for female listeners might also be related to possible hesitations in the speech of the current speaker, and the production of feedback might serve as an encouragement to continue and to confirm that the channel remains open.

Finally, the social and multimodal variables considered in this study were not significantly linked to the variation of CPP, used to operationalize voice quality in

the speech preceding feedback. Previous findings by Heldner et al. (2019) already provided ambiguous results on the contribution of CPP measurements for the prediction of backchannels. In our results, CPP does not seem to be used any differently in relation to the modality of the backchannel or the gender of the listeners and the current speakers. In line with previous research, however, we were able to highlight its contribution to the turn-taking mechanism. In fact, syntactically/semantically complete locations are followed by a feedback expression when the voice is more modal, that is, has higher CPP values, which have been found to function as a turn-hold cue in conversation (Włodarczak and Heldner, 2022; Włodarczak et al., 2023); on the other hand, locations followed by a backchannel without syntactic completeness, where thus it would be more intuitive that the current speaker would want to continue talking, are preceded by lower CPP values, indicating a less modal phonation.

These findings are relevant for the following key reasons. Firstly, they provide a more comprehensive understanding of how listeners employ phonetic cues in speech to plan and to produce backchannels across modalities in face-to-face dialogues, and how such cues heavily interact with other turn-taking and contextual variables.

Secondly, beyond the empirical distinctions in the acoustic features of speech between self-identified male and female speakers, which for German have been thoroughly described by Weirich and Simpson (2018, 2023) and Weirich et al. (2019), these results suggest the existence of gender differences in the way in which phonetic cues are utilized for turn-taking purposes during spontaneous interactions, an area which so far has not been empirically investigated. Specifically, as for the cues contributing to the timing of backchannels, the differences observed suggest the existence of fine phonetic variation in how listeners weight and respond to these features, locally influencing their decisions on when to provide feedback.

However, it is important to underline that these differences in the use and perception of turn-taking cues do not inherently imply competition or a disrupted and hard to manage conversational flow. In fact, the dialogues analyzed in this study are collaborative in nature and involve interactions between peers. As previously noted by Vinciarelli et al. (2015), “even if the scenario of the corpus does not introduce a status difference between subjects and there is no status difference between

male and female subjects” (p. 4), gender differences in backchannel timing and distribution seem to emerge nevertheless. In the current study, the experimental setting and tasks in which the subjects were involved might have played a role in it. Different activities have in fact been reported to interact with gender to different extents. For instance, task-oriented dialogues tend to highlight the relevance of speakers’ genders compared to unstructured conversations (Leaper and Ayres, 2007; Leaper and Robnett, 2011). Even though very loosely task-based, as described in 9.6, the fact that the interlocutors in this dataset had to talk to each other towards a conversational goal might have influenced the speakers’ behavior in this direction. Moreover, we observed that the gender of the current speaker was relevant for the variation of several factors, with the majority of differences being observed in male-female dyads, where the gender of both interlocutors has been observed to be more relevant than in same-gender interactions (Hannah and Murachver, 2007; Leaper and Ayres, 2007).

At the same time, our results for mixed-gender dialogues highlighted the use of accommodation and cooperative strategies, which are indicative of a fine-grained adaptability of gender performance and communication style to develop a collaborative, friendly and harmonious conversational exchange. These insights suggest that, even though listeners and speakers may hold distinct expectations and interpretations regarding the dynamics of listener behavior (Engwall et al., 2023), due also to expectations regarding gender or gender performance, they would all pursue the ultimate shared objective of achieving conversational success and effective communication (Valentine, 1986), by cooperating and naturally accommodating between each other.

9.11 Conclusions and outlook

This analysis of the timing of multimodal backchannels in face-to-face conversation and of the phonetic turn-taking cues preceding FLs revealed the existence of gender-specific patterns of listener behaviors, as well as a significant influence of the gender of the current speakers, which seems to induce different trends and accommodating tendencies between conversational partners. Importantly, our findings suggest that some of the patterns of listener behavior in men and women reported in the previ-

ous sociolinguistic literature on conversational dynamics may be influenced by how male and female listeners employ and perceive distinct linguistic and phonetic turn-taking cues. This, in turn, would inform their local decision-making regarding the production of feedback expressions in a specific modality during the current speaking turn, based on their judgments of appropriateness. Such results underscore the importance of including the information related to the gender of the interlocutors, when available, in studies investigating interactional dynamics and phonetic turn-taking cues. More broadly, our results add to research on how gender influences the production and the perception of acoustic features in speech by suggesting that, in spontaneous interactions, interlocutors make different use of the phonetic turn-taking cues available to plan their next conversational move. Based on the nature of the interaction, as well as possibly the age, the status, the social class or the education level of the interlocutors, these distinct preferences and tendencies might or might not result in smooth and balanced exchanges. Further research involving a wider range sociolinguistic factors is necessary to explore if and to what extent the individual background and characteristics influence the way in which turn-taking cues are produced and perceived in conversation.

The current study focused on dialogues in German. Future developments of this investigation will involve comparing the German data with analogous face-to-face dialogues in a language typologically related to German, namely Swedish. In particular, previous cross-linguistic comparison between German and Swedish have highlighted differences both in the way phonetic features are used for turn-taking purposes (Feindt et al., 2021; Rossi et al., 2022b) as well as in their speakers' distinct attitudes towards gender roles and dynamics in society and their reflection in acoustic values in speech (Weirich and Simpson, 2023; Weirich et al., 2019). This comparative approach will allow for a more comprehensive examination of the hypothesis that female and male interlocutors use turn-taking cues differently across languages. Additionally, it will allow for the exploration of potential variations in how interlocutors' genders influence conversational dynamics across languages.

Chapter 10

Cross-linguistic and gender-specific variation of turn-taking dynamics

The full summaries of the statistical models in the analysis are reported in Appendix 4 (p. xxiii).

10.1 Introduction

Daily verbal interactions represent an essential aspect of our social life. In conversations with others we continuously construct and reconstruct our identities as individuals belonging to a society, a community or a social group. In fact, the way we behave in interaction and our turn-taking habits, such as when we consider appropriate to take the floor, to remain silent, to produce a backchannel response, strongly convey several aspects of our cultural and social identities (Beňuš, 2009). While there are shared features related to conversational dynamics between the languages of the world, such as the avoidance of long silent gaps between turns and long stretches of simultaneous speech between interlocutors (Sacks et al., 1974; Stivers et al., 2009) or the fact that a non-speaking interlocutor will provide some form of feedback, or backchannel, to display attention and acknowledgement of what the current speaker is saying (e.g., Heinz, 2003), there is at the same time a significant deal of cross-linguistic variation in the way such turn-taking fundamentals are implemented by each language, even by typologically related ones. The observed cross-linguistic differences have been connected to the cultural values of the community in which the language is spoken, and to the performance of some aspects of the social identities of the speakers, such as their genders.

The dynamics of turn-taking behavior are fundamentally shaped by complex interactions of language-specific, social and contextual factors. To explore this assumption and to provide empirical and quantitative data in its support, the current study investigates language-specific turn-taking dynamics in German and in Swedish face-to-face dialogues. Previous cross-linguistic studies have focused on German and Swedish from multiple theoretical perspectives, and identified differences between these two languages. For instance, studies on turn-taking highlighted how German and Swedish speakers manage the conversational floor by resorting to different strategies and turn-taking cues as a possible result of the languages' phonologies (Feindt et al., in preparation; Rossi et al., 2022b; Zellers, 2017). Moreover, sociophonetic accounts have reported how German and Swedish male and female speakers do not index gender in speech in the same ways, pointing to the existence of socio-cultural differences between the two societies (Weirich and Simpson, 2023; Weirich et al., 2019). Building on these previous results, the present study brings together the two approaches, and investigates turn-taking dynamics in German and in Swedish by observing the frequency and the timing of speaker changes, turn keeps and backchannels in the two languages. Specifically, since the conversations took place face-to-face and the interlocutors could rely on both the vocal and the visual modality, backchannels will be analyzed from a multimodal perspective, i.e., by taking into account both verbal and gestural backchannels. Furthermore, the study analyzes how conversational variation might be used by the interlocutors involved in the dialogues to perform their gender, and to what extent gender performance might differ between the two languages and, consequently, their cultures.

10.2 Conversational dynamics across languages

The predominant use of speech and communicative gestures takes place in interaction. Daily informal conversations are characterized by a succession of fast and generally effortless exchanges of turns at talk (Levinson and Torreira, 2015; Sacks et al., 1974; Stivers et al., 2009), and brief and non-competitive multimodal backchannel signals displaying grounding and attention (Duncan, 1974; Schegloff, 1982; Ward and Tsukahara, 2000; Yngve, 1970). Smooth alternation of speakership and listenership between interlocutors is governed by a set of participant-managed turn-taking

norms determining who speaks and when is the most appropriate time to do so (Beňuš, 2009; Feindt et al., in preparation; Sacks et al., 1974). These turn-taking decisions are made by interlocutors at Transition Relevance Places (TRPs; Sacks et al., 1974; also defined by Schiffrin, 1987 as Potential Transition Locations, SYNCOMPs by Local and Walker, 2012, or by Zellers, 2017 as Potential Turn Boundaries), i.e., locations during the current turn where speaker change becomes a possibility. TRPs in speech are marked by a set of turn-taking cues through which the speaker conveys his/her conversational intentions for the upcoming turn and coordinates with the interlocutor about the next conversational action. (Levinson and Torreira, 2015). The syntactic/semantic completeness of the TRP, the preceding variation of acoustic cues such as fundamental frequency (F0) movements, intensity, speech rate and voice quality, together with gestures, facial expression or gaze patterns (e.g., Brusco et al., 2020; Ford et al., 1996; Gravano and Hirschberg, 2011; Kendon, 1995; Local and Walker, 2012; Local et al., 1986; Mondada, 2007; Niebuhr et al., 2013; Zellers, 2017) are all possible turn-taking cues that are used by the interlocutors to achieve close timing in conversation.

Specifically, in informal conversations, as outlined by Sacks et al. (1974), conversational participants conventionally take turns speaking and try to minimize silent gaps and speech overlaps. Stivers et al. (2009) set out to test empirically the cross-linguistic validity of these observations by investigating a sample of typologically related and unrelated languages of the world in search of potential turn-taking fundamentals. While they do find that, in general, languages display a tendency to minimize silences and overlapped turns, they also report a certain degree of cross-linguistic variability in the temporal dynamics of the turn-taking mechanism. More specifically, while the average silent gap between turns appears to have a duration of around 250 milliseconds (ms) (cf. Heldner and Edlund, 2010), the mean time for turn transition ranges from 7.29 ms in Japanese exchanges, to 468.88 ms in Danish. This indicates a considerable difference in what one language's speaker might perceive as an appropriate gap between turns, while to a speaker from another language, it might potentially sound like a considerable delay in turn allocation. Stivers et al. (2009) also report that the structure of languages, specifically their typological closeness, does not seem to be explanatory for the observed patterns. For instance, in their dataset, Germanic languages such as Danish, Dutch, and English exhibit

distinct response timings, indicating that language structure alone may not account for these differences.

As far as speech overlaps are concerned, the scientific literature presents examples of both languages with a low tolerance for overlaps, such as Finnish (Lehtonen and Sajavaara, 1985; Stubbe, 1998), and communities where simultaneous speech is more prevalent and accepted. For instance, Tannen (1984)'s study on the high-involvement style of Jewish Americans reports a pervasive use of supportive overlaps, that Californian American listeners perceive as interrupting (Tannen, 1984). Ulijn and Xiangling (1995), further supported by Deng (1998), observe that Chinese speakers make frequent use of overlaps at potentially interrupting locations during the current speaker's turn (specifically, at non-TRPs), both within their cultural context and when interacting with Finnish and with Dutch speakers, suggesting the influence of potential language-specific dynamics connected to cultural conventions (Deng, 1998). Hayashi's (1988) analysis of conversations in Japanese and English indicates that Japanese speakers employ overlapped talk more frequently, supported by quantitative analysis, and engage in more nonverbal interactive behaviors compared to their American counterparts.

Another turn-taking structure whose distribution and features offer a fundamental example of cross-linguistic and cross-cultural variation is backchannels, or listener feedback, i.e., responses provided by one interlocutor while the other is speaking. In face-to-face conversations, where interlocutors can rely on both the visual and the vocal channel to communicate with each other, listener feedback can be observed across modalities: it can be constituted by verbal, gestural or multimodal (i.e. both verbal and gestural) responses. Backchannels display active listenership and invite continuation, confirming that the channel remains open (e.g., Duncan, 1974; Ward and Tsukahara, 2000; Yngve, 1970). They tend to be unobtrusive and non-turn-competitive, and represent an important and pervasive turn-management device in interaction. Being less constrained than conversational turns with regards to their distribution within the turn-taking structure (Roberts et al., 2015), and being optional by definition (Truong et al., 2011; Ward and Tsukahara, 2000), backchannels are almost entirely rooted in the current interactive exchange, varying significantly based on the conversational practices of a specific linguistic community, as well as other social and contextual variables (Stubbe, 1998).

Investigations on backchanneling behavior of both native speakers and language learners describe significant cross-linguistic differences in the use of feedback. Examples of these studies have put Japanese in comparison with German (Krause-Ono, 2004) or with American English (Cutrone, 2014; Ward and Tsukahara, 2000; White, 1989); American English in comparison with French (Wieland, 1991) and Spanish (Berry, 1994); Finnish with Central European languages or English (Lehtonen and Sajavaara, 1985; Stubbe, 1998). From these accounts, Japanese listeners emerge as the most active in displaying feedback to the current speaker, both visually and verbally, as well as the ones transferring this tendency when speaking in an L2. "...[A]s backchanneling is one of the core points of Japanese polite behavior, it might be that Japanese cannot easily behave in a seemingly impolite way; ergo they cannot give up their backchanneling beyond a certain extent" (Krause-Ono, 2004, p. 78). On the other hand, Finnish listeners, with their very low tolerance for speech overlaps and interruptions in their conversations, could be placed on the other hand of the continuum, with their interactions being characterized by sparser feedback, mostly in the form of gestures, and with the use of verbal backchannels having a negative connotation (Lehtonen and Sajavaara, 1985; Stubbe, 1998, 2013). What emerges from these studies is the possible existence of language-specific and culture-specific baselines for backchannel frequency in conversation (Stubbe, 1998; White, 1989), as well as specific expectations regarding distribution. In fact, even if backchannels are, by definition, not openly acknowledged by the current speaker (e.g., Ward and Tsukahara, 2000), they are nevertheless taken into account to some extent: while they tend to go unnoticed, they become noticeable when there are too many of them or too little (Krause-Ono, 2004). They might also be considered as inappropriate when not well-timed within the current speaker's turn (Poppe et al., 2011). Previous studies suggest a preference for verbal backchannels to be produced after the offset of speech of the current speaker's turn, while gestural backchannels can also occur in overlap, as they are more likely to cause disruption (e.g., Poppe et al., 2011; Rossi, 2022; Truong et al., 2011).

Fant (1989) compares the turn-taking dynamics between Spanish and Swedish speakers and observes that Swedish listeners provide approximately 15 times more verbal feedback than Spanish subjects. In fact, it appears that Spanish interlocutors rely more on gaze patterns to signal attention to each other. Fant (1989) concludes

that the production of more backchannels verbally is indicative of the Swedish and, in general, Scandinavian communication style, which prioritizes overtly confirming attention and agreement in dialogue, more than Hispanic culture.

In general, these accounts indicate that, despite being founded on some shared rules, the timing and the overall dynamics of turn-taking structures vary across languages and speaking communities. However, as emphasized by Dideriksen et al. (2023), most investigations tend to compare languages that are typologically different and culturally distant from each other. As a result, the observed variation in the distribution of conversational devices remains confined to descriptive accounts and lacks more in-depth or fine-grained explanations. To address this limitation, Dideriksen et al. (2023) analyzed two related languages, Norwegian and Danish, which have very similar societies and cultures but differ in sound structure. The study specifically targeted backchannels, repairs, and conversational entertainment in dialogues, and their findings highlight that conversational devices aimed at demonstrating mutual understanding are more prevalent in Danish. The authors suggest that this discrepancy might be partly attributed to the relatively higher acoustic opacity of Danish compared to Norwegian. Consequently, Danish speakers might rely more on linguistic entertainment to display understanding and reduce the need for repairs. In conclusion, the authors propose that language-specific properties, linked with cognitive processing factors (e.g., mutual understanding and context-driven inferences), may underlie the variation in the use of conversational devices and dynamics in conversation.

In summary, while at the moment our understanding of how conversational dynamics systematically vary cross-linguistically is still limited (Dideriksen et al., 2023), possible reasons for this variation have been indicated in the “cultural interactional pace” (Stivers et al., 2009: 10590), in guidelines of appropriateness and politeness rooted in sociocultural values and widespread usage patterns (e.g., Kraaz and Bernaisch, 2022; Krause-Ono, 2004; Stubbe, 1998), or even in language-related properties and constraints which reflect in cross-linguistic differences in the use of conversational dynamics (Dideriksen et al., 2023).

10.3 Gender-related variation in speech

Some studies observing cross-linguistic variation in turn-taking patterns and conversational behaviors reported that the gender of the subjects in their data was at the base of some degree of within-language variation (e.g., Cutrone, 2005; Deng, 1998; Feke, 2003; Kraaz and Bernaisch, 2022; Stubbe, 1998). The influence of gender on interaction has been a source of interest for decades in sociolinguistics. Early studies were mainly aimed at identifying and describing the differences in the way men and women behave in conversation (Plug et al., 2021), as they saw language and interaction as reflecting the social imbalance and power inequalities (e.g., Lakoff, 1973; Zimmermann and West, 1975), or as the environment where the different cultural backgrounds in which men and women were educated would manifest and likely clash (e.g., Tannen, 1990). However, within more recent approaches, gender tends to be characterized, not as something an individual is, but rather as something an individual does (Cannon et al., 2019; West and Zimmerman, 1987), performs (Butler, 1993), and/or indexes (Ochs, 1992) through social acts and practices, such as speech and interaction. Studies within this framework do not necessarily look to outline specific boundaries between men and women in speech; instead, masculine and feminine styles are treated as being part of a continuum, and can be adopted by speakers based on the identity they wish to perform within a particular situation (Coates, 2004; Stubbe, 2013).

This view accounts for the fact that the observed differences in language use might not always be consistent from study to study, as other variables might interact with gender, such as the interactional role of the participants in the study, as well as its contextual settings (Plug et al., 2021). For instance, aggregating the results from 29 studies comprising 3,502 subjects investigating the gender-specific use of tentative language, Leaper and Robnett (2011) find an influence of the experimental task in the gendered behavior of participants. Specifically, they find a larger group difference between male and female speakers in task-based interactions, and hypothesize that such tasks might be responsible for the activation of more stereotypical gendered behaviors. Vinciarelli et al. (2015) also report differences in the occurrence of different interactional phenomena such as laughter, overlapped speech and silences. Subjects were involved in a cooperative task to be carried out over the

phone. In their results, Vinciarelli et al. (2015) report that, even if the task did not introduce a status difference, male speakers were found to adopt more stereotypically masculine behaviors associated with dominance, such as talking for longer and laughing less over the course of the task. Moreover, it has been observed that the gender of the interlocutor involved in the conversational activity has an influence on the dynamics involved. One of the tendencies observed is that, in mixed-gender dyads, the gender identity of the participants becomes more salient for them than in same-gender interactions and, “when a particular identity is salient, individuals prescribe to that which is most exemplary of the group, including communicative behaviors” (Palomares, 2004: 558). This means that a higher frequency of gender-specific stereotypical behaviors will be observed in conversations between male and female participants than in male-male or female-female ones (Hogg, 1985; Palomares, 2004; Plug et al., 2021; Reid et al., 2003).

In general, previous literature on gender and turn-taking carried out on men and women from Western countries, reports all-female conversations to be commonly marked by a cooperative approach, where the traditional norm of one speaker at a time is replaced by a shared conversational floor, overlapping questions, rephrasing, repetitions, and backchannels, indicating active engagement (Coates, 2004; Edelsky, 1993). On the other hand, interactions among men typically adhere more to the one-speaker-at-a-time model, featuring either extended uninterrupted speaking turns or brief and rapid exchanges, with minimal instances of overlap, considered disruptive (Coates, 2004). In their review of more recent studies on gender and conversational practices in dyadic interactions, Plug et al. (2021) highlight women’s tendency to exhibit a number of supporting turn-taking behaviors, expressing engagement in ongoing interactions across diverse contexts and roles by utilizing various communicative cues such as supportive interruptions, cooperative overlap, more frequent verbal backchannels and head nods than male subjects.

An important contribution to the empirical investigation of gender performance through speech has been given by studies in sociophonetics, particularly through cross-linguistic analyses that examine how male and female speech patterns may either converge or diverge in different languages. Starting from the observation that the speech of men and women, influenced by biological sex, inherently possesses distinct acoustic properties—male speakers’ F0 and formant frequencies are lower

than for female speakers (Coleman, 1976; Fant, 1966; Simpson, 2009), female voice is found to be breathier (Henton and Bladon, 1985; Klatt and Klatt, 1990) and to exhibit a larger vowel space (Henton, 1995)—sociophonetic research highlights how these biological differences are not uniformly reflected in speech across languages. In fact, the cultural construct of gender and its performance in a particular community (Kendall et al., 2023) and/or in a particular situation play a fundamental role in determining the extent or the size of the difference between males' and females' speech. That is, an important aspect of the acoustic variations of speech is learned (Simpson, 2009; Weirich and Simpson, 2018) and it is shaped through the social life of the individual. For instance, Simpson (2009) conducted a cross-linguistic review of various studies, including the findings reported by Traunmüller and Eriksson (1993), in which highlighted the minimal differences in F0 found in the Wü dialects of Chinese in comparison to the substantial variations observed between the F0 values of French and English men and women. Simpson (2009) comments that “it would be unreasonable to account for such large differences in terms of anatomical differences in the populations being investigated; part of the difference must be attributed to learned behaviors” (Simpson, 2009, p. 625). Van Bezooijen (1995) reports the differences in pitch values between Dutch and Japanese, and find that the higher pitch observed for Japanese women is connected to attributes considered feminine and attractive for Japanese society, while the mid-lower pitch is more connected to attractiveness in Dutch.

10.4 Previous comparisons of German and Swedish

German and Swedish—two Germanic languages, with societies that share important cultural ties and similarities (e.g., Weirich et al., 2019)—have been compared in terms of both turn-taking structures and gender performance through speech. Significant differences have emerged in both lines of investigation, laying important theoretical and empirical foundations for the current study's objective of assessing language-specific and gender-specific patterns of variation in conversational dynamics.

Firstly, it has been suggested that the prosodic phonologies of German and Swedish could be at the base of cross-linguistic differences in the acoustic signalling

of turn-taking intentions. German is an intonation language, marking prominent syllables with pitch accents that can carry information about the information structure and discourse relationship, while Central Swedish is a pitch accent language, where a two-way lexical pitch contrast is used to distinguish word pairs (Gårding, 1989; Gussenhoven and Bruce, 1999). Feindt et al. (2021, in preparation) and Rossi et al. (2022b) find that the variation of pitch (operationalized as F0 movements) towards TRPs is more constrained in Swedish compared to German, where speakers exhibit greater pitch variability and utilize it for turn-taking functions, as well as possible inter-speaker prosodic entrainment. These observations lead to the hypothesis that pitch might be already saturated with its lexical function in Swedish, resulting in reduced variation related to interactional and pragmatic functions, along with a preference for other acoustic turn-taking cues, such as segmental duration (Feindt et al., 2023, in preparation; Zellers, 2017).

Rossi et al. (2022c) also report a preliminary overview of the distribution of turn-taking structures, i.e., the frequency of speaker changes, turn-holds and backchannels across the two languages. Notably, they observe that Swedish speakers exhibited a higher rate of backchannels in their dialogues, while German speakers displayed a higher rate of speaker changes. However, they did not explore this cross-linguistic distributional difference any further. Moreover, the backchannels considered in Rossi et al. (2022b, 2022c) were only those produced in the verbal modality as, in the German dialogues that they use for their analysis, taken from the Kiel Corpus of Spoken German (Kohler et al., 2017), the interlocutors were not face-to-face, unlike the interlocutors of the Swedish dialogues. Therefore, while these findings suggest potential differences in conversational dynamics between the two languages, further analyses on more comparable data are essential to achieve a more comprehensive representation.

Cross-linguistic sociophonetic studies by Weirich and Simpson (2023) and Weirich et al. (2019) also reveal that German and Swedish speakers exhibit different patterns of gender-specific variation in speech. They collected both phonetic data (F0 mean values, vowel space size) and gender identity data with self-ascribed femininity scores in German and Swedish participant using the GEPAQ_F+ questionnaire (Runge et al., 1981), which comprises 8 bipolar adjective pairs characterizing positive personality traits conventionally associated with women, including

qualities like kindness, sensitivity, friendliness, and empathy, with higher scores indicating a positive association with these attributes. In Weirich et al. (2019), they observe that the difference in acoustic measurements between male and female speakers is smaller in Swedish, aligning with comparable values in gender identity scores, where female speakers displayed lower scores in Swedish. This finding was further confirmed in Weirich and Simpson (2023)'s investigation into the phonetic parameters of new parents in Germany and Swedish and their development over time. In Sweden, a reduced gender-specificity was once again observed, as female speakers shifted towards male values. This shift aligned with the lower femininity scores obtained from the questionnaire, which did not significantly differ from the self-ascribed attributes of male participants. In contrast, the differentiation between male and female speakers' values and scores was more evident in Germany.

10.5 Objectives

The current study has two main objectives. First, it aims to observe the distribution of turn-taking actions (speaker changes, turn keeps and backchannels) in German and in Swedish, in order to identify possible differences in the conversational dynamics between the two languages. The second objective of the study is to test the possible influence of the gender of the speaker and the gender of the listener on the distribution of the turn-taking structures. Specifically, since the self-ascribed genders of the subjects that are part of the datasets employed (described in Section 10.6) were male and female, the current study will adopt these binary gender labels for the following analysis, that will thus focus on men and women's conversational behavior.

The research questions of the current study are the following:

1. Are there differences in the distribution of speaker changes, keeps and backchannels between German and Swedish?
2. Does the gender of the interlocutors influence the distribution of turn-taking dynamics similarly in the two languages?

We hypothesize that significant differences in the distribution of conversational structures between German and Swedish will arise, consistent with the preliminary

observations made by Rossi et al. (2022b, 2022c). Their exploration indicates potential differences in the frequency of turn-taking structures, suggesting distinct conversational tendencies. For instance, it is possible that Swedish speakers, with potentially higher rates of backchannels, might place greater emphasis on active listenership. By incorporating multimodal backchannels and datasets from face-to-face dialogues in both German and Swedish, the objective is to present a more comprehensive analysis of the distribution and frequency of turn-taking structures in dialogues.

Furthermore, differences in the extent of gender influence between the two languages are expected. If Swedish exhibits a smaller degree of gender-specificity in speech than German, as suggested by Weirich et al. (2019)'s results, significantly different distribution patterns between male and female speakers in Swedish may not be observed, while more a more stereotypical differentiation might be found in German.

10.6 Methods

To test the influence of the language and the gender of the speakers on the distribution of conversational structures in German in comparison with Swedish, a set of 12 face-to-face dialogues between adult native speakers of German, for a total of 2 hours and 15 minutes, and to a set of 6 face-to-face dialogues between adult native speakers of Swedish, for a total of 30 minutes, were analyzed. The German data come from the DUEL Multilingual Multimodal Dialogue Corpus (Hough et al., 2016). The Swedish data come from the Spontal Corpus (Edlund et al., 2010). The dialogues analyzed for this study were balanced for the gender of both interlocutors, with the objective of having the same number of female and male speakers (encoded in the variables as "Female" and "Male") in conversation with an interlocutor of their same gender or of a different gender (encoded in the variables as "Same" and "Other" respectively). As previously addressed (Section 10.5), the binary labeling of gender is due to the fact that the available data only included male and female subjects, and no participants provided a known gender identity different from their biological sex.

The dyad composition is encoded by a combination of the two set of labels

(i.e., “Female-same” for same-gender conversations including two female speakers, “Female-other” for mixed-gender conversations including a female speaker and a male interlocutor, “Male-same” for same-gender conversations including two male speakers and “Male-other” for mixed-gender conversations including a male speaker and a female interlocutor). Two of the speakers in the Swedish dataset, one male and one female subject, separately took part in two dialogues in the dataset, once in a same-gender and once in a mixed-gender pair.

10.6.1 Datasets and comparability

The two datasets selected for the cross-linguistic comparison were chosen because they are comparable in terms of the types of interactions and conversational structures they comprise. For both corpora, the selected dialogues involve two participants who could see each other while engaged in conversation, sitting face-to-face. Moreover, in both the DUEL Corpus and the Spontal Corpus, participants wore head-mounted microphones to capture the speech of each participant on separate channels, which allowed for the analysis of portions of overlapped speech as well.

The two datasets present, however, a few differences, which, without undermining the overall comparability, might affect to different extents the results for this study’s specific research questions. One difference consists of the directions that participants were given about their interaction. While both corpora contain spontaneous exchanges, the dialogues in the DUEL Corpus were started with an input given by the researchers, and the conversational participants were not completely free in the choice of the overall interaction’s topic. Thus, the dialogues in German were “loosely task-directed” (Hough et al., 2016, p. 1784), as the researchers presented subjects with initial prompts, or “scenarios”, designed to initiate interaction without spending excessive time selecting a conversation subject.

Specifically, the dialogue portions from the German data analyzed for the current study are the ones corresponding to the “Dream Apartment” and the “Film Script” scenarios. For the “Dream Apartment” scenario, the dialogue partners were told that, with the prospect of becoming roommates, and with a large sum of money at their disposal, they should discuss their ideal apartment layout and furnishing. In the “Film Script” scenario, the pair was asked to imagine a movie scene which they thought would be very embarrassing, and could rely on their own personal

experiences or on movies they had watched. The DUEL corpus includes a third scenario called “Border Control,” involving a role-play interview at a desk, where one participant played the role of a traveler, while the other acted as a border control officer, directing the conversation with questions. Given the different nature of the interactions of this latter scenario, these dialogue portions were not considered for the analysis of the German data.

On the other hand, the Swedish dialogues retrieved from the Spontal Corpus were completely spontaneous, both in terms of the interactional dynamics as well as the topic of discussion, which came up naturally between the two subjects (Edlund et al., 2010). The difference between the two corpora thus lies in the conversational activity, which can be categorized following the classification by Leaper and Robnett (2011) as either structured (assigned topics of discussion) or unstructured (free choice). Since one corpus comprises entirely spontaneous interactions, while the other involves minimal task-oriented elements, potential differences might emerge in the interaction structure, since, as discussed earlier in Section 10.3, the task could influence how interlocutors’ gender affects speech. Although studying the task’s impact is beyond the scope of this study, it will be considered in the interpretation of the results.

Another difference between the datasets used for the study is the selection of the dialogue excerpts due to the data availability. In the DUEL Corpus, annotations were carried out starting from the beginning of the interactions, while for the Spontal Corpus the available files consist of extracted dialogue portions from longer conversations where speakers had already started interacting. This could potentially impact observed conversational patterns in different ways. In fact, the temporal dimension of conversations have been found to play a role in the distribution of interactional devices. For instance, backchannels have been reported to become more frequent over time (Fusaroli et al., 2017). The possible effect of global convergence (e.g., Weise and Levitan, 2018) might also bring two speakers to adopt increasingly similar behaviors the more they have been talking to each other. As this could potentially lead to the emergence of different patterns in the two languages which might not be directly connected to language-specific tendencies or the influence of gender, this dataset difference will also be considered in interpreting the results.

10.6.2 Annotation

Orthographic transcriptions of the dialogues, as well as the subdivision into speaking turns and utterances were provided with both the DUEL Corpus and the Spontal Corpus. Utterances were automatically divided into words, syllables and segments using the WebMAUS pipeline without Automatic Speech Recognition (Kisler et al., 2017; Schiel, 1999) and manually corrected in the relevant intervals by the author.

The annotation of turn-taking structures and timing was carried out in Praat (Boersma and Weenink, 2023). Point labels were placed corresponding with each speaker's utterances in locations where a speaker change or a backchannel by the interlocutor arose, or at the offset of speech if the current speaker kept the floor and continued speaking. The segmentation into words, syllables and segments was useful to identify potential utterance-internal silences or overlaps not entirely visible through the turn subdivision.

Following the approach used in Feindt et al. (2021, in preparation) and Rossi et al. (2022b), the turn-taking labels indicated:

1. Syntactic/semantic completeness in context of the speech preceding the label ('yes' for *Complete* and 'no' for *Incomplete*);
2. Syntactic form of the location, indicating if the speech preceding the label was in declarative form (labeled as 'd') or in the form of a question or a tag question (labeled as either 'q' or 't' respectively; given the availability of the data and the scope of the study, only the locations in declarative form were considered for the current analysis);
3. Sequential structure, or turn-taking structure, describing the conversational action that followed the label: if the floor passed from one speaker to the other, the sequential structure was labeled as *Change* (annotated with 'c'); if the same speaker held the floor and continued talking the label was *Keep* ('k'); if the interlocutor produced a backchannel while the current speaker was talking, or during a pause after which the current speaker continued keeping the floor, the structure was labeled as *Backchannel* ('b');
4. Transition type, indicating the timing with which the sequential structure

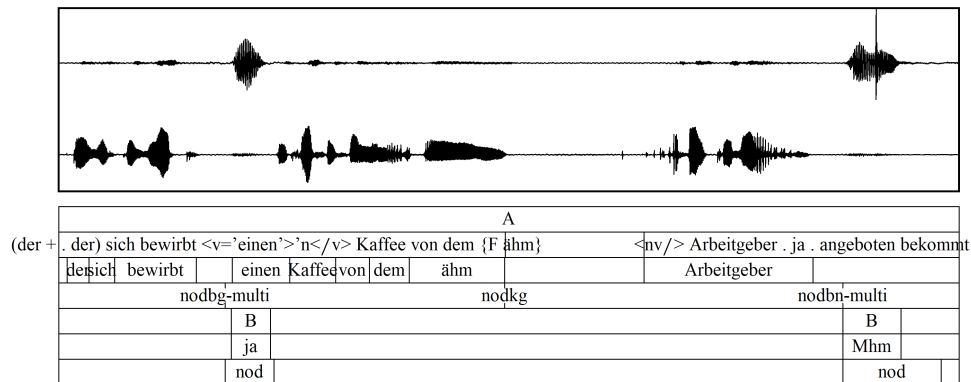


Figure 10.1: Example of the turn-taking annotation of speaker “A” in one of the German dialogue participants in Praat. In this portion of the dialogue, speaker A is talking (annotated on the first tier) and speaker B (fifth tier) is listening. The 2nd tier contains the transcription of A’s utterances, and the third tier the segmentation of the utterance into words (syllables and segments are not reported). B’s contributions in this portions are two multimodal backchannels, from which the verbal and the visual component are annotated on the sixth and seventh tier respectively. The turn-taking labels are placed on the fourth tier, using a point label, placed either at the offset of speech for turn keeps, or where backchannels (and speaker changes) arose.

arose: with a silent gap (labeled as *Gap*, ‘g’), with a speech overlap (labeled as *Overlap*, ‘o’) or without any audible gap or overlap, i.e., either arising smoothly at the offset of speech, or with possible speech overlaps or silent gaps under 120 ms (labeled as *No-gap-no-overlap*, ‘n’), following the detection threshold provided in Heldner (2011).

An example of the annotation is reported in Figure 10.1. All the labels were extracted using a Praat script written by the author.

10.6.3 Labels preceding backchannels

Specific attention was placed on identifying and labeling locations preceding backchannels in order to provide a comprehensive analysis of listener feedback dynamics.

As previously mentioned (see Section 10.2), a backchannel is a brief and unobtrusive signal in either the verbal, the visual or in both modality simultaneously, produced by the interlocutor who, in that precise moment of the conversation, has the role of the listener (i.e., does not have the conversational floor), while the current speaker is talking (Cutrone, 2014; Yngve, 1970). The annotation process considered

backchannels constituted only by a verbal element, by a gesture (more specifically, a movement of the head), and by both elements arising simultaneously.

Following the annotation guidelines of the same German dataset already presented in Rossi et al. (2023), the sequential structure was labeled as *Backchannel* if the feedback provided a direct reaction to the content of the current turn, was optional and was not acknowledged by the current speaker (Ward and Tsukahara, 2000). Thus, with this approach all instances of short utterances and/or head movements used as turn-openers or in correspondence with the end of a full-fledged turn, as well as answering questions or requesting clarifications are not considered as backchannels (Rossi et al., 2023; Ward and Tsukahara, 2000).

Following the annotation scheme presented in Rossi et al. (2023), the label preceding a backchannel also specified its modality (verbal, gestural or multimodal when a verbal item and a head movement occurred simultaneously), the lexical form of the item, if available, and the head movement type, if available. Verbal backchannels were identified in Praat, while head movements used as feedback were annotated in ELAN (Wittenburg et al., 2006) with their movement type, using the guidelines and the template provided by the the Multimodal and M3D Multidimensional Labeling Scheme (Rohrer et al., 2023; Rohrer et al., 2020), based on the head movement description provided by Wagner et al. (2014). Following Malisz et al. (2016), uninterrupted head movements, such as repeated nods, were treated as a single annotation unit. However, unlike in Malisz et al. (2016), where different head movement types were grouped together when continuous, each annotation in the present study contained only one type of gesture. This approach aimed to provide a more detailed representation of the temporal alignment between each head movement, the possible verbal component of the backchannel, and the speech of the current speaker. Labels were then extracted from ELAN and imported in Praat on a dedicated interval tier.

The labels used to annotate verbal feedback are: *mlx* for monosyllabic lexical items (e.g., “ja”, yes), *dlx* for disyllabic items (e.g., German “genau” and Swedish “precis”, i.e., exactly), *non* for non-lexical expressions (e.g., “mh” or “mhmh”), *cplx* for lexical phrases (such as “das stimmt” in German or “just det” in Swedish, i.e., that’s right), *rep* for repetition of the same lexical referent previously used by the current speaker, and *com* for combinations of the previous forms. This set of labels is based

on previous work by Rossi et al. (2023) and Zellers (2021). In line with Rossi et al. (2023), the head movement types identified in the datasets for gestural feedback are nods, tilts, turns, protrusions and slides (the gesture types and their distribution will not be further explored in the current study).

10.7 Analysis and Results

The analysis of the data was carried out in R (R Core Team, 2022). The packages `nnet` (Venables and Ripley, 2002) was used to build multinomial logistic regression models. The packages `ggplot2` (Wickham, 2016), `sjPlot` (Lüdtke, 2022) and `Effects` (Fox et al., 2022) were used to visualize the data.

10.7.1 Sequential structure distribution

The first part of the analysis aimed at observing the distribution of changes, backchannels and keeps, as well as their specific timing in relation to the speech by the interlocutor. Comparisons were carried out between German and Swedish, and within language, between male and female speakers. Given the different sizes of the two datasets, tests were conducted separately on each, and the results were then compared.

In total, the extracted data comprises 1,906 turn-taking locations, 1,325 from the German dataset and 581 from the Swedish dataset. A major cross-linguistic difference immediately emerges from the distribution of the different sequential structures in the two languages. In German, out of the total extracted turn-taking structures, 26% ($n=344$) are changes, 47% ($n=628$) are keeps and 27% ($n=353$) are backchannels; in Swedish, out of the total number of labels, 23% ($n=135$) are changes, 25% ($n=143$) keeps and 52% ($n=303$) backchannels. This indicates that in the German dialogues speakers kept the floor without a backchannel more frequently, while in the Swedish dataset the speakers produced a higher number of backchannels, in a quantity amounting to more than half of all the labels extracted. On the other hand, changes and keeps have almost the same frequency in Swedish.

Cross-linguistic differences connected to the timing of turn-taking are found as well. Chi-square tests including the sequential structures and the possible transition types indicate a significant association between the two variables in both languages

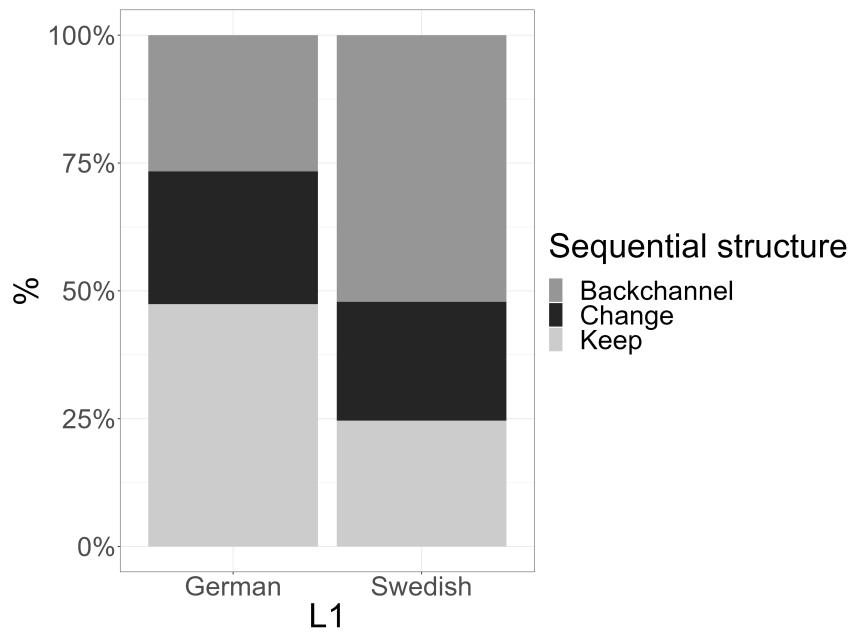


Figure 10.2: Distribution of the sequential structures (backchannels, speaker changes, turn keeps) in the German and Swedish speakers.

	Gap	Nogap	Overlap	Gap	Nogap	Overlap
	German			Swedish		
Backchannel	118 (-5.57)	87 (-0.26)	148 (9.81)	70 (-5.73)	61 (1.46)	172 (5.28)
Change	192 (0.06)	47 (-4.30)	105 (4.85)	55 (-0.78)	31 (1.78)	49 (-0.32)
Keep	426 (4.13)	202 (3.38)	NA	138 (9.10)	5 (-3.86)	NA

Table 10.1: Count of observations (and residuals) of the Chi-squared test for the transition type (*Gap*, *No-gap-no-overlap*, *Overlap*) relative to the sequential structure (*Backchannel*, *Change*, *Keep*) in German and in Swedish.

(German: $\chi^2 = 318.09$, $df = 4$, $p < .0001$; Swedish: $\chi^2 = 219.11$, $df = 4$, $p < .0001$). However, the Pearson's residuals indicate that keeps are significantly more likely to occur with a silence in Swedish. While the same can be observed in German as well, here keeps are also significantly associated with no-gap-no-overlap transitions. This indicates a one-to-one correspondence between silent pauses and keeps that seems to arise in Swedish, while this association is not observed in German (see Table 10.1).

The relationship between gender and both turn-taking structure distribution and timing reaches significance in both languages when the interlocutor is included in the tests, too, i.e., when the dyad composition is considered. In German and Swedish, the relationship between the dyad composition and the distribution of the sequential structures of the dialogues was found to be significant (German: $\chi^2 =$

	Backchannel	Change	Keep	Backchannel	Change	Keep
	German			Swedish		
Female_Other	107 (2.51)	86 (0.46)	122 (-2.23)	70 (-2.12)	45 (0.75)	58 (2.36)
Female_Same	95 (0.49)	71 (-1.81)	173 (0.97)	59 (1.10)	20 (1.30)	10 (-2.87)
Male_Other	77 (0.99)	74 (0.85)	107 (-1.38)	112 (0.08)	41 (-1.20)	60 (1.04)
Male_Same	74 (-3.43)	113 (0.55)	226 (2.16)	62 (1.60)	20 (-0.53)	15 (-1.81)

Table 10.2: Count of observations (and residuals) of the Chi-squared test for the sequential structure (*Backchannel*, *Change*, *Keep*) relative to the dyad composition (*Female_Other*, *Female_Same*, *Male_Other*, *Male_Same*) in German and in Swedish.

37.691, $df = 6$, $p < .0001$; Swedish: $\chi^2 = 79.22$, $df = 6$, $p < .0001$), even though with differences between the two languages highlighted by the Pearson's residuals (see Table 10.2).

In German, backchannels are observed to be more likely to occur with a female current speaker in mixed-gender conversations, and less likely between male speakers in same-gender dyads. On the other hand, in Swedish, female speakers in mixed-gender dyads were less likely to receive backchannels, while it is more likely for male speakers in same-gender interactions to receive them, even though this latter association does not reach statistical significance (see Table 10.2).

Moreover, in German, same-gender dyads between male speakers are more likely to include keeps, while keeps are significantly less likely to occur by female speakers in mixed-gender dialogues. Again, the picture is different for the Swedish data. In fact, female speakers in mixed-gender dyads are more less likely to receive backchannels and more likely to keep the floor. Moreover, keeps are negatively correlated with female-female conversations. There is also a negative association between keeps and male-male dyads in Swedish, even though this last association does not reach statistical significance (see Table 10.2).

The relationship between the dyad composition and the transition types was also found to be significant (German: $\chi^2 = 37.691$, $df = 6$, $p < .0001$; Swedish: $\chi^2 = 79.22$, $df = 6$, $p < .0001$) with cross-linguistic differences. In German, no-gap-no-overlap transitions are more likely to occur in female-female dialogues, while overlaps are less likely to occur in male-male dialogues. As for mixed-gender dyads, female speakers appear to be overlapped more frequently by male speakers (see Table 10.3).

On the other hand, female speakers in Swedish mixed-gender dialogues are less

	Gap	Nogap	Overlap	Gap	Nogap	Overlap
	German			Swedish		
Female_Other	170 (-0.37)	62 (-2.00)	83 (2.94)	113 (3.91)	28 (-0.16)	32 (-4.16)
Female_Same	165 (-1.69)	107 (2.26)	67 (0.28)	20 (-3.65)	12 (-1.07)	66 (4.70)
Male_Other	150 (0.55)	54 (-1.41)	54 (0.67)	101 (0.46)	36 (0.07)	76 (-0.55)
Male_Same	251 (1.42)	113 (0.80)	49 (3.36)	29 (-2.24)	21 (1.19)	47 (1.66)

Table 10.3: Count of observations (and residuals) of the Chi-squared test for the transition type (*Gap*, *No-gap-no-overlap*, *Overlap*) relative to the dyad composition (*Female_Other*, *Female_Same*, *Male_Other*, *Male_Same*) in German and in Swedish.

frequently overlapped, but silent gaps are more likely to occur these dyads. Moreover, in Swedish, gaps are negatively associated with same-gender dyads, for both male and female speakers, while overlaps are more likely to be observed in female-female dyads. Interestingly, they are positively associated with male-male dyads, too, even though the association does not reach statistical significance (see Table 10.3).

10.7.2 Multimodal backchannel distribution

Exploring more closely the distribution of backchannels, the analysis highlights both similarities as well as several cross-linguistic differences. As for the for of verbal backchannels (see Figure 10.3), in both German and Swedish the shortest ones, i.e., the monosyllabic and the non-lexical backchannels, are the most frequent. More in detail, non-lexicals are more frequent in Swedish, where they cover 49.77% of the data (while in German non-lexical backchannel represent 20.47% of the data), while monosyllabic forms are more frequent in German, in which they are 40.52% of all backchannels (vs. 30.41% in Swedish). The frequency of the occurrence of the different modalities appears very similar between the two languages. In German, out of the total amount of feedback, 43.05% (n=152) are verbal, 27.76% (n=98) are gestural and 29.17% (n=103) are multimodal; in Swedish, 42.90% (n=130) are verbal, 28.38% (n=86) gestural and 28.71% (n=87) multimodal. On the other hand, the timing of the feedback in different modalities has a different distribution pattern between German and Swedish (see Figure 10.4). In particular, in German gestural items occur with a higher frequency in overlap with the speech by the current speaker, while verbal and multimodal items are produced more frequently during silent pauses or right at the offset of speech. The same pattern is not reflected in

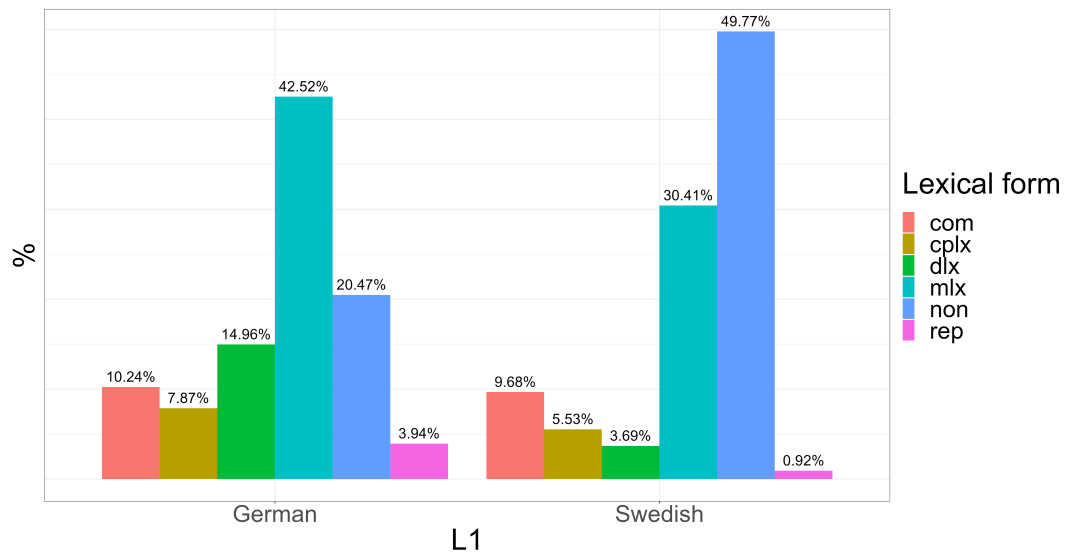


Figure 10.3: Frequency of the forms of verbal backchannels—combinations (*com*), complex lexical phrases (*cplx*), disyllabic lexical items (*dlx*), monosyllabic lexical items (*mlx*), non-lexical expressions (*non*) and repetitions (*rep*)—in the German and in the Swedish conversations.

Swedish, where feedback in all three modalities appears to be produced more frequently in overlap with the speech by the current speaker.

A separate multinomial logistic regression model was run for each language, including the transition type labels describing the timing (*Gap*, *No-gap-no-overlap*, *Overlap*) as outcomes, and the dyad composition, the completeness status of the location, and the modality of the feedback as predictors. The models revealed considerable cross-linguistic differences (see Figure 10.5 and Figure 10.6). In German, there are significant main effects of the completeness status of the location where the feedback occurs, the modality of the backchannel, as well as the dyad composition in terms of the gender of the interlocutor producing the backchannel and the current speaker. More in detail, German backchannels are significantly more likely to occur in overlap when the location is labeled as syntactically/semantically incomplete (β : 3.1983, z value: 8.799, $p < .0001$), while they are significantly less likely to occur in overlap when they include a vocal component, i.e., they are produced as verbal (β : -1.2506, z value: -2.926, $p < .01$) or as multimodal (β : -1.4422, z value: -3.186, $p < .01$) items. Moreover, they are significantly less likely to occur in overlap with male speakers' utterances in same-gender conversation than with female speakers' speech in mixed-gender dyads (β : -1.0838, z value: -2.391, $p < .05$).

On the other hand, in Swedish there is only a significant main effect of the com-

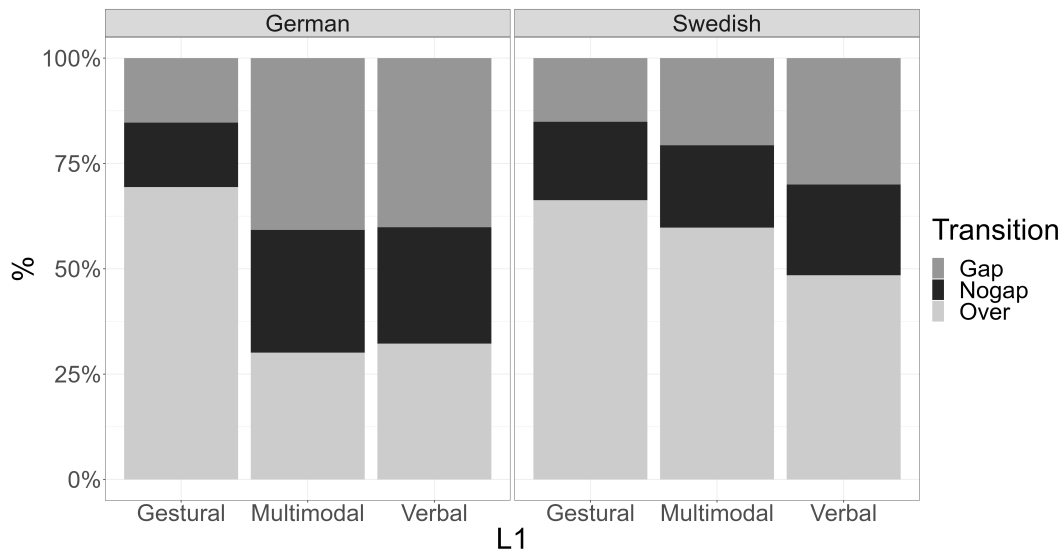


Figure 10.4: Distribution of gestural, multimodal and verbal backchannels based on their transition time (*Gap*, *No-gap-no-overlap* and *Overlap*) in the German and in the Swedish dialogues.

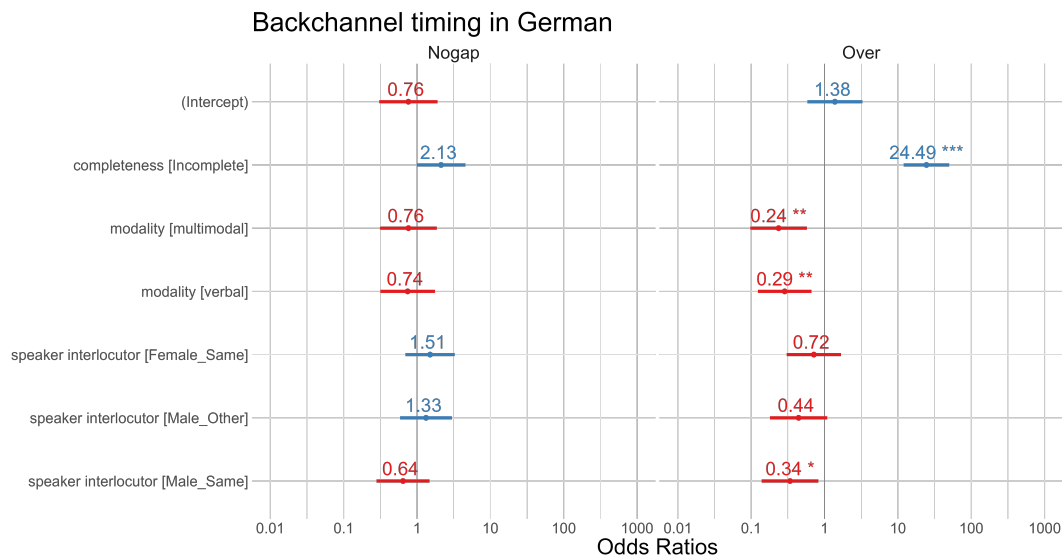


Figure 10.5: Summary of the multinomial model for German, including the transition types, *No-gap-no-overlap*, *Overlap* and *Gap* as the categorical outcomes and the completeness status of the label (*Complete* or *Incomplete*), the modality of the backchannel (*Gestural*, *Verbal*, *Multimodal*), and the dyad composition (*Female_Other*, *Female_Same*, *Male_Other*, *Male_Same*). The Odds Ratios on the x-axis indicate the likelihood of transitioning to the category *No-gap-no-overlap* or *Overlap* compared to the reference category *Gap* for a one-unit increase in the corresponding predictor variables.

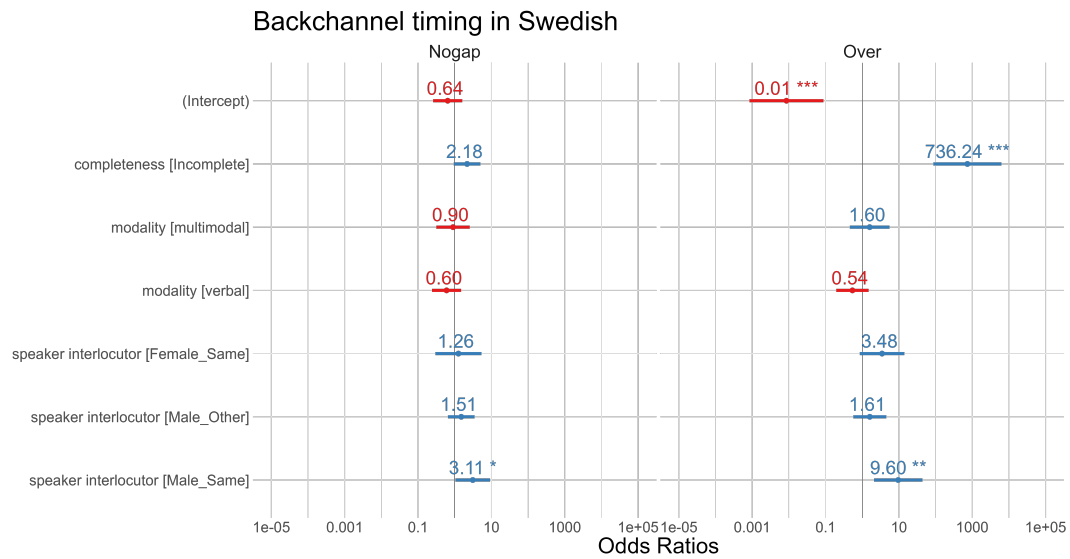


Figure 10.6: Summary of the multinomial model for Swedish, including the transition types, *No-gap-no-overlap*, *Overlap* and *Gap* as the categorical outcomes and the completeness status of the label (*Complete* or *Incomplete*), the modality of the backchannel (*Gestural*, *Verbal*, *Multimodal*), and the dyad composition (*Female_Other*, *Female_Same*, *Male_Other*, *Male_Same*). The Odds Ratios on the x-axis indicate the likelihood of transitioning to the category *No-gap-no-overlap* or *Overlap* compared to the reference category *Gap* for a one-unit increase in the corresponding predictor variables.

pleteness status of the location where the backchannel is produced. In fact, Swedish backchannels are significantly more likely to occur in overlap with the speakers' utterance at syntactically/semantically incomplete locations (β : 6.60149, z value: 6.077, $p < .0001$). The modality does not appear to be a significant predictor of backchannel timing in Swedish.

Finally, in Swedish, the dyad composition in terms of gender of the interlocutors is a significant predictor of backchannel distribution. Displaying once again a different image from German, Swedish male-male dyads are the most likely to contain backchannels in overlap with the current turn (β : 2.26165, z value: 2.923, $p < .01$) (see Figure 10.7).

10.8 Discussion

This analysis tested the distribution of turn-taking structures across German and Swedish and, intra-linguistically, across female and male speakers in same-gender and mixed-gender dyads. The main objectives were to observe potential language-specific patterns and to investigate the influence of gender on conversational dynamics. The results highlighted several cross-linguistic differences, both in the oc-

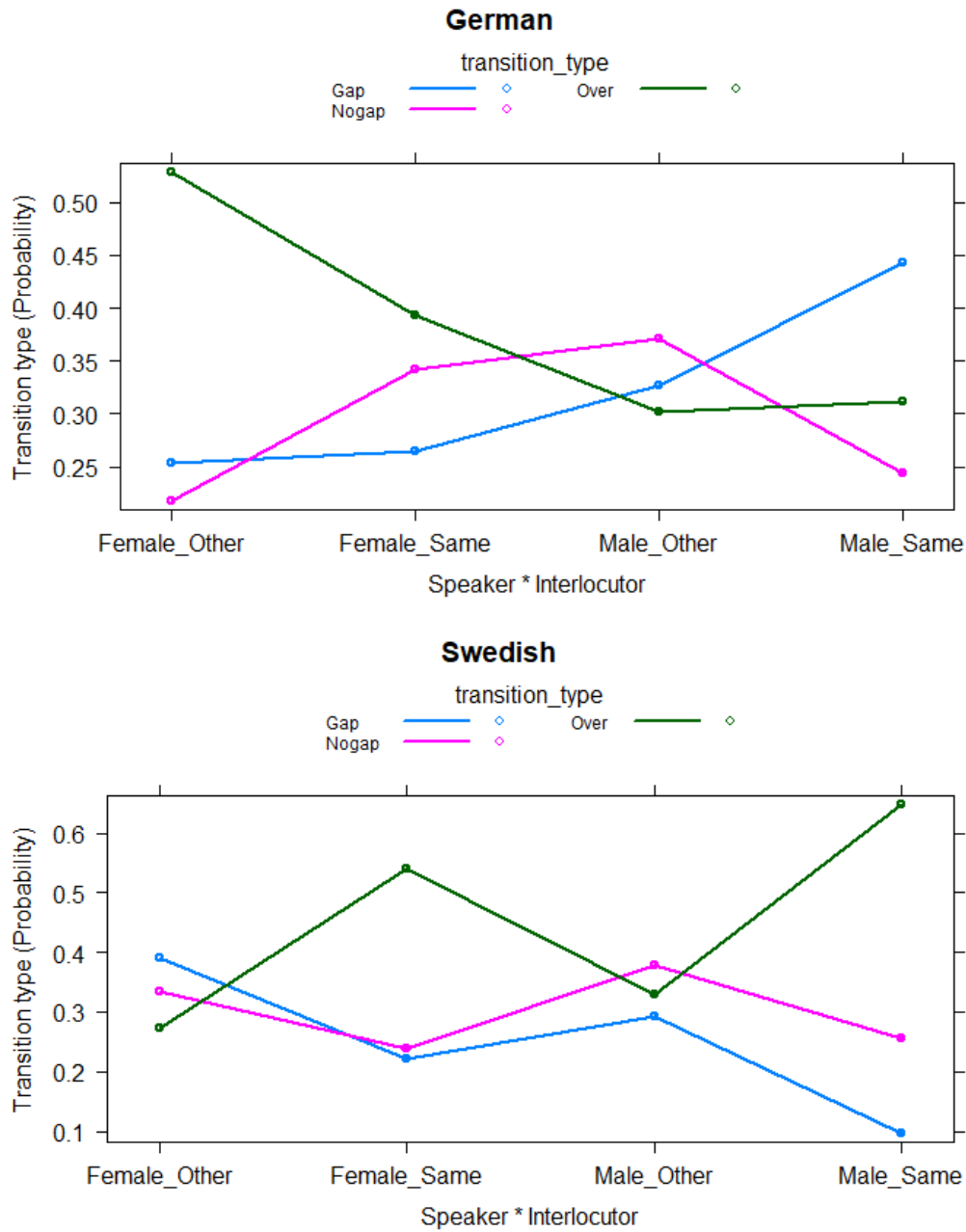


Figure 10.7: Effects plots of the multinomial models for German and for Swedish representing the likelihood of the transition types (*Gap*, *No-gap-no-overlap*, *Overlap*) of backchannels occurring in relation to the dyad composition (*Female_Other*, *Female_Same*, *Male_Other*, *Male_Same*).

currence of sequential structures and their timing in conversation, as well as distinct gender-specific dynamics in German and in Swedish, suggesting the possibility of language-specific dynamics at work in interaction.

10.8.1 Language-specific variation

The first striking cross-linguistic difference highlighted by the analysis is the notably higher frequency of backchannels in Swedish dialogues. In fact, 52% of the sequential structures extracted from the Swedish dataset were backchannels, vs. 27% in German. Both results are consistent with findings from previous studies. First, they confirm the higher backchannel frequency first emerged from the preliminary exploration provided in Rossi et al. (2022c). Furthermore, Knudsen et al. (2020) report 16% of verbal backchannels with respect to the full number of turns in their analysis of German data, and estimate a 15% to be the average of backchannel contributions to natural conversations in West Germanic languages, bringing together data for Dutch, English and German. On the other hand, studies on North Germanic languages report higher backchannel rates. For instance, Fusaroli et al. (2017) report for their Danish dataset that 54% of the overall number of turns are backchannel contributions. In their comparison of Danish and Norwegian, Dideriksen et al. (2023) report an average of 20-30% verbal backchannels in the overall datasets. For Swedish, Cerrato (2002) reports multimodal backchannels that comprise 50% of turns in her data on face-to-face exchanges, and 60% of turns in 10-minute dialogues are backchannels in Fant (1989)'s data.

Section 10.6.1 hypothesized a potential comparability issues regarding the different annotation intervals of the two datasets used for the current study, i.e., German dialogues were annotated from the beginning of the conversation, while in Swedish the two participants had already started interacting with each other. Considering the tendencies previously observed in the literature (Fusaroli et al., 2017), the higher percentage of backchannel in Swedish in comparison to German can be potentially connected, at least in part, to this dialogue excerpt difference. It might be possible that, if taken from the beginning of the dialogue, backchannels in Swedish might have been less frequent; on the other hand, with the conversation progressing, it could be possible that more backchannels could have been found in German. The fact that the results are in line with previous findings suggests that the observed backchannel frequencies might not only be an effect of the dialogue phase, but might be connected to more general language-specific tendencies. However, with the available data and current annotations, it is not possible, and outside the scope of this analysis, to verify the extent to which language-specific factors and dia-

logue progression might be interacting with each other, if at all. Since this potential interaction cannot be directly assessed within the current analysis, the following interpretation will focus solely on language-specific and sociolinguistic tendencies, while future research and developments of the study might be able to report on the influence of further variables.

The fact that modality has no significant effect on the timing of Swedish backchannels is a particularly striking result. In fact, what emerged from several previous studies in different languages (e.g., Dittmann and Llewellyn, 1968; Ferré and Renaudier, 2017; Rossi et al., 2023; Truong et al., 2011) is that, in order to be the least disruptive as possible, listeners rely heavily on the different modalities to time their feedback appropriately, with gestural items being preferred in overlap with the speech by the current speakers, and verbal ones arising more frequently after the offset of speech.

A possible insight into this result for this Swedish dataset might come from the lexical content of the backchannels. In fact, when compared to German, Swedish displays a higher frequency of very short forms, i.e., non-lexical items (*non*), followed by monosyllabic lexical items (*mlx*) (see Figure 10.3). These backchannels, among those comprising a vocal component, are the least disruptive type of form, which might explain their greater occurrence in overlap with the current turn.

The results regarding backchannels point towards the possible existence of a higher overall baseline (Stubbe, 1998; White, 1989) for backchannel rate in Swedish conversation (and, apparently, in general in North Germanic languages compared to West Germanic languages). In languages with a high baseline for feedback frequency, such as Japanese, this tendency has been connected to the speaker's culture giving a strong importance to smooth and harmonious interactions and avoidance of confrontation (e.g., Cutrone, 2005). This is a cultural trait that is also observed in Swedish culture as "partly a consequence of the postwar ideology of consensus and harmonizing in Sweden" (Ilie, 2004, p. 54), where avoiding conflict and uncomfortable situations has a particular importance in social life and decision-making processes, such as group discussions and business settings (Auer-Rizzi and Berry, 2000; Daun, 1996; Tixier, 1996). In his cross-linguistic comparison involving Spanish and Swedish, Fant (1989) suggests that turn-taking dynamics reflect language-specific communicative priorities, which, in Swedish, are to display agreement and

cooperation to maintain a positive interaction, avoiding conflict and assertive behaviors. On the other hand, the description that Byrnes (1986) gives of the German conversational style portrays an almost opposite picture. She reports that, being more focused on information transmission and facts, German conversational style “seeks, or at least should not shy away from, overt disagreement and confrontation. In fact, disagreement and confrontation are valued, and have become ritualized, in that they are deemed to further the process of establishing truth. Perhaps, in its own way, it becomes a form of social bonding for those who customarily engage in it” (Byrnes, 1986, p. 201) It is possible that the different rates of backchannels that were highlighted in our datasets in German and in Swedish could be to some extent connected to these cultural tendencies.

Another striking result that emerged from the analysis of backchannel distribution in the two languages was related to modality and timing in relation to the current speaker’s utterance. The results for German are in line with previous studies on feedback distribution, which indicate that the listener is able to carefully time their backchannels to the current speaker’s turn in order to be the least disruptive as possible for the stream of talk and, in face-to-face interaction, multimodal variation is a fundamental resource to smooth feedback giving (e.g., Dittmann and Llewellyn, 1968; Ferré and Renaudier, 2017; Rossi et al., 2023; Truong et al., 2011). In the current analysis, German backchannels comprising a vocal element (verbal and multimodal ones), i.e., the ones that would be potentially more disruptive, are significantly found to occur after the offset of speech. On the other hand, in Swedish, all backchannels occur in overlap with the speech, and the quantitative analysis indicates that the modality of the feedback is not a significant predictor for the temporal distribution within the current speaker’s turn. A potential explanation at the base of this tendency could be that the verbal component feedback is mostly a very short and minimal one, composed of either non-lexical expressions or monosyllabic lexical items, and therefore potentially minimally interrupting. An additional interpretation could again be connected to a possible cultural tendency of openly displaying agreement and understanding being more important than careful timing. As explored in Section 10.2, cross-linguistic studies have reported that listeners’ behavior of languages such as Japanese, where feedback is frequent and multimodal, or Chinese, where feedback is frequently produced at non-TRPs, might be perceived

as interrupting or inappropriate to speakers of other languages with different feedback baselines (Deng, 1998; Hayashi, 1988; Ulijn and Xiangling, 1995), while being part of the habitual interactional routines of the language.

10.8.2 Gender-specific variation

The results also highlighted significant differences related to gender-specific turn-taking dynamics between German and Swedish. In particular, while in both languages male and female speakers do exhibit distinct behaviors, especially when the dyad composition is taken into consideration, the specific dynamics of the differences between genders vary across languages. The general picture that emerges from our results is that, in German, interactional dynamics in face-to-face dialogues appear to be more in line with the traditional, or stereotypical, gender-specific conversational habits described in the literature (see Section 10.3). For instance, the results reported that German female speakers tend to receive more backchannels, especially in overlap with speech; male subjects exhibit fewer instances of overlapped backchannels, but in mixed-gender dialogues they tend to overlap female interlocutors more often, and in same-gender dyads they tend to exchange longer uninterrupted turns (i.e., they exhibit a higher frequency of keeps).

On the other hand, the patterns highlighted in Swedish conform less with previous descriptions of gendered conversational behaviors, in particular in mixed-gender and male-male dyads. In the current results, in fact, Swedish male speakers use more overlapped turn-taking instances, such as backchannels, which is something that is observed also in female-female dyads in Swedish. Moreover, the association of longer turns with male-male conversations is not evident in Swedish, and in mixed-gender dyads, female speakers seem to hold the floor more frequently. Female speakers in same-gender dyads, consistent with findings in other languages, tend to utilize more overlapped talk and exhibit fewer silent gaps between turns. Similar observations of less frequent gaps were reported for male-male dyads as well.

These results can be considered in line with the previous findings by Weirich and Simpson (2023) and Weirich et al. (2019), specifically on gender performance through speech in German and Swedish. They find phonetic differences between male and female speakers in F0 values and vowel space in both languages, but

the distinction was smaller in Swedish, validated by femininity scores indicating a larger divide between German men and women than between Swedish men and women. The smaller difference that they observe in Swedish appear to be connected to female speakers' values approximating male speakers' ones. In other words, Swedish women in their data adopted more masculine speech patterns than German women, but Swedish men did not vary towards more feminine values. As far as our data are concerned, fewer stereotypical behaviors from Swedish female speakers in mixed-gender dyads, i.e., in the dyads where gender identity becomes more salient for the speakers (e.g., Hogg, 1985), were detected. On the other hand, in the context where gender should be less salient for the participants, i.e., same-gender, both male and female speakers adopt similar conversational dynamics, which comprises more overlapped changes, and more frequent and overlapped backchannels. In other words, in mixed-gender interactions, female speakers might adopt more masculine turn-taking behaviors. In same-gender dyads, both male and female interlocutors adopt a general cooperative and involved interactional style.

A possible confounding factor in the interpretation of these results in relation to gender dynamics is the structure of the interactions analyzed in the two languages. As explained in detail in Section 10.6, the German dialogues are loosely task-based, while the Swedish data comprises of completely spontaneous conversations. Task-based or structured conversations have been found to activate more frequently gender-specific stereotypical dynamics. For instance, Leaper and Robnett (2011) report an increased use of tentative speech (hedges, tag questions, expressions of uncertainty and intensifiers) by women compared to men in task-driven dialogues. They suggest that the coordinating and organizational dynamics associated with task-based interactions might have made gender identities more salient for speakers, activating the differences in conversational behavior that they observed. Thus, it is possible that, in our German data, the presence of a loose activity assigned to the participant might have had a role in shaping, to some extent, the turn-taking dynamics observed in the results, or at least in making them more prominent. Although the German interactions did not appear structured at an impressionistic level, the presence of a conversational goal may have influenced backchannel use to some extent. However, the differences observed between German and Swedish speakers cannot be attributed solely to the nature of the conversational activity. As

Palomares (2004) suggests, gender salience is influenced by various factors, including the discussion topic, the relative importance of other identity features in a given context (e.g., student, employee, colleague), and an individual's perception of gender differences. Therefore, to determine whether the type of conversational activity limits the interpretation of the data in light of language-specific gender dynamics, additional factors must be examined to assess whether there is an actual difference in perceived gender salience between the two L1 groups. At this stage, with the available information and the current analysis, it remains unclear to what extent the presence or absence of topic-related instructions may have influenced gender-specific behaviors.

While more extensive research on fully comparable datasets is necessary to make definitive claims, the results of this study point towards the existence of gender-specific dynamics interacting with language-specific tendencies, underscoring the importance of cross-linguistic studies with a quantitative approach involving social factors. This type of investigations would allow for a more comprehensive view of gender performance through conversational behaviors and for a better understanding of the extent of its interactions with other social factors from specific cultures or communities. As Beňuš (2009) points out, turn-taking behavior is inherently socially determined, and the “choices that we make in selecting when we speak have clear implications as signals of our social roles, positions of power, or other typical sociolinguistic categories such as age, gender or race” (Beňuš, 2009, p. 20). While still relatively underinvestigated from a sociolinguistic perspective, turn-taking in conversation represents a stimulating and challenging subject to analyze in order to develop a deeper understanding of the underlying individual and social dynamics at its base.

10.9 Conclusions and outlook

Despite its limitations, the current study highlighted the existence of cross-linguistic differences in the turn-taking dynamics arising in German and Swedish data, and in the gender performance through conversational behaviors across languages. The results from the analysis of two typologically related languages, which exhibit important differences in their prosodic structures, and are socially and culturally

close, yet display several culture-specific tendencies, suggesting crucial interactions among these factors in determining the dynamics of turn-taking in conversation.

Gender performance has a key role in shaping conversational behaviors. While for the scope of the current study, gender was investigated only in interaction with the dyad composition, the native language and culture of the speakers, its performance should be investigated by also taking into consideration other social categories, such as age, or social status—all parts of the individual's social identity, which might be conveyed through speech variation in interaction, and to different extents depending on the communicative situations.

As for our analysis of turn-taking in German and Swedish, open questions leave room for future investigations. For instance, the features of verbal and multimodal backchannels, occurring mostly in overlap in Swedish, should be further analyzed from an acoustic perspective. More in detail, the prosodic features of the verbal component might reveal particularly reduced and inconspicuous forms, giving further insights into their observed distributional patterns. The acoustic configuration of speech in the current speaker's turn could also be relevant in this sense, as it could exhibit specific features that would allow the listener to provide feedback without causing excessive disruption (e.g., reduced loudness, slowed speech rate). Finally, the turn-taking dynamics observed might suggest the existence of potential differences in the use that male and female speakers make of turn-taking cues (e.g., Rossi, 2022). Given the already outlined language-specific preferences in the use of pre-final turn-taking cues between German and Swedish (Feindt et al., 2023, in preparation; Rossi et al., 2022b), as well as their gender-specific pattern of acoustic variation in speech (Weirich and Simpson, 2023; Weirich et al., 2019), a further comparison involving gender will offer further insight on both sociophonetic and cross-linguistic dynamics in interaction.

Part IV

Discussion

Chapter 11

Summary of the results

The studies described in Chapters 6, 7, 8, 9 and 10 reported the results of a series of analyses aimed at observing the variation of acoustic turn-taking cues towards TRPs and the distribution and timing of turn-taking structures, such as backchannels, speaker changes and keeps, by taking into account the gender of the speaker, the gender of the listener, the L1 of the interlocutors and, for backchannels, the availability of the visual modality in the interaction.

The first study (Chapter 6) analyzed German dialogues from the DUEL Corpus with the aim of observing the turn-taking-related variation of F0, intensity and segmental duration at different time-points approaching a TRP. The first part of the analysis observed the patterns of variation of the different acoustic cues based on the conversational action that followed the TRP (a speaker change or a turn keep). After that, the second part of the analysis tested the potential influence of the gender of the speaker and their interlocutor on the phonetic marking of conversational intentions. Specifically, the study tested the following hypothesis:

Hypothesis 1 (H1): Speakers of German use different patterns of acoustic turn-taking variation to mark their conversational intentions in the speech approaching TRPs followed by speaker changes and turn-keeps influenced by their gender and the gender of their interlocutor.

The results are consistent with previous studies on German (e.g., Niebuhr et al., 2013) by giving evidence for pre-boundary segmental lengthening accompanied by a stark decrease in intensity for speaker change cases, while for keep cases intensity remains higher, and final lengthening is less evident. The results obtained for F0 were less clearly associated with either a speaker change or a keep; however, what was evident in the results was a high degree of variability in F0 movements towards

TRPs in the speech of German speakers. As for gender-related effects, a first distributional exploration revealed a higher percentage of keeps in male-male conversations, indicating a tendency to hold the turn for longer between male speakers in the study's dataset. Moreover, the same-gender dialogues (male-male and female-female dyads) displayed a higher percentage of no-gap-no-overlap transitions, suggesting the possibility of increased effort to align with one another, between interlocutors of the same gender, to signal conversational intentions and achieve coordination using turn-taking cues.

The quantitative analysis of gender-specific variation in the use of turn-taking cues revealed different acoustic patterns between male and female speakers in the dataset. Male speakers displayed lower F0 compared to female speakers, close to their own individual baselines, and produced more rises and falls approaching TRPs. On the other hand, female speakers showed a more consistent use of intensity variation and final lengthening as turn-taking cues to distinguish between turn-holding and turn-yielding intentions. The distinction between TRPs followed by speaker change and turn keeps in terms of final values of intensity and segmental duration was observable for female speakers in general, but especially in mixed-gender dyads. For male speakers, while it was not so evident in mixed-gender dyads, the distinction was very clear in same-gender conversations. These results thus indicate that male speakers might rely more on changes in final F0 values to signal an approaching TRP; and that, in conversation with a male interlocutor, both female and male speakers put more effort into signaling their conversational intentions using intensity and duration, suggesting a greater effort to manage the floor in these situations.

H1 is therefore fully confirmed: male and female speakers of German seem to prefer different turn-taking cues. Most importantly, the results display that such a preference is manifested not only in the greater use of the acoustic feature in isolation, but specifically on its pattern of variation on the temporal dimension approaching the TRP. Moreover, the female-female, male-male and mixed-gender interactions displayed specific turn-taking dynamics as well, indicating the importance of the gender of the interlocutor as well.

One open question that arose from this study was the variability in the acoustic cues preceding backchannels. Therefore, the second study (Chapter 7) aimed at ex-

ploring the forms and the distribution of backchannels in German by considering vocal elements and the movements of the head. Lexical and non-lexical backchannels, in the forms of short verbal expressions on one hand, and non-verbal vocalizations and laughter on the other, were included in the analysis of vocal feedback. Head movements, labeled as head nods, tilts, turns, slides and protrusions, were all identified as part of listener behavior in conversation. The study aimed at exploring the influence of the form and the modality on the distribution and the timing of feedback by the listener in relation to the current speaker's turn and, more closely, in relation to the offset of speech by the current speaker. It tested the following hypothesis:

Hypothesis 2 (H2): Listeners in conversation are sensitive to the placement of backchannels within the speaking turn and, when multiple modalities are available, they tend to rely on multimodal alternation to provide feedback that is the least interrupting as possible for the speech by the current speaker.

The results show that, in line with the previous literature on backchannels, feedback with lexical content occurred turn-externally and after the offset of speech, often preceded by a silent gap. Head movements, i.e., the least disruptive type of listener behavior among those considered, occur frequently within the current turn (turn-internally) and the onset of the movement preceded the end of the turn, thus occurring frequently in overlap with speech. Laughter was observed to occur mostly turn-externally, as the loudest type of feedback, while non-lexical expressions, which are very minimal but still contain a vocal element, occurred with a similar frequency both within the current turn and after the offset of speech. Moreover, the results on the temporal alignment of the head with the vocal component in multimodal backchannels showed that the onset of the gestural part anticipated the verbal production by around 141 milliseconds.

These results provided the basis for a further analysis of multimodal feedback in face-to-face conversations, and confirmed H2, by displaying consistent patterns in the distribution and the timing of vocal and gestural feedback in relation to the current speaker's turn, in order for it to be present and frequent but causing the least amount of disruption as possible by exploiting multimodal alternation.

The third study (Chapter 8) reports a preliminary analysis of the cross-linguistic differences in the acoustic marking of TRPs in German and in Swedish, with a specific focus on the contribution of F0 to the turn-taking mechanism, and its patterns of individual variation. The following hypothesis was tested:

Hypothesis 3 (H3): German and Swedish speakers exhibit distinct patterns of variation in F0 as a turn-taking cue in the speech preceding TRPs, with German speakers displaying greater flexibility in marking conversational intentions and employing F0 for social dynamics, while the use of F0 at turn boundaries is expected to be more constrained for Swedish speakers.

The results confirmed H3 by displaying that F0 did show cross-linguistic differences in its use for turn-taking purposes between the two languages. Furthermore, the patterns of individual variation detected in the dataset seem to also suggest a differential use of F0 with a social and coordinating function between German and Swedish speakers. Specifically, in line with previous literature on the different use of pitch between intonational languages, like German, and pitch-accent languages, like Swedish, the variation of F0 towards TRPs for turn-taking purposes appears to be more restricted and limited for Swedish, where the values obtained in the analysis appear to consistently remain close to the speaker's individual F0 baseline. On the other hand, the variability observed is greater for German speakers, who appear to use F0 movements more flexibly in its interactional function. Moreover, the results from this exploration also indicated a striking similarity in the F0 values towards potential turn ends between German conversational partners, which was not detected within the Swedish data. It was therefore suggested that, together with turn-taking purposes, F0 movements might be used by German speakers as a mean to entrain with each other, in order to achieve coordination, rapport and to display understanding and cooperation between interlocutors.

The fourth study (Chapter 9) focused on investigating to what extent the gender of the listener, as well as that of the current speaker, influences the timing of multimodal backchannels in face-to-face German dialogues. Moreover, the study tests how the variation of acoustic turn-taking cues (F0, intensity, speech rate and voice quality) in the current speaker's utterances guide male and female listeners' plan-

ning and production of feedback in same-gender and mixed-gender dyads. The acoustic values were extracted from the syllable nuclei in the 1 second of speech preceding the production of a backchannel by the listener, and were observed dynamically in their patterns of variation leading up to the feedback location.

In particular, the hypothesis tested in the analysis of this study was the following:

Hypothesis 4 (H4): The distribution and the placement by the listener of multimodal backchannels within the current speaking turn in German dialogues display gender-specific patterns of variation, also based on the gender of the current speaker and the modality of the feedback.

The findings partially confirm H4. In fact, results show gender-specific behaviors in how feedback is produced by listeners, and timed across different modalities in relation to the speaker's current speech. However, the patterns of variation of the different turn-taking cues in the speech preceding feedback appear to not always have a significant relation with the gender of the listener or the gender of the interlocutor, pointing instead to a more complex picture.

As for the distributional results, several factors influence the timing of backchannels. These factors include the completeness of the syntactic/semantic context where the feedback occurs, the modality of production, as well as the genders of both the listener and the speaker. In general, it was observed that listener feedback occurs after the offset of speech when a syntactic/semantic boundary is reached, or when it involves vocal elements (such as verbal or multimodal backchannels). Additionally, backchannels overlapping with the ongoing utterance are more common when the speaker is female, whereas those occurring after the end of the speech are more prevalent when the speaker is male.

The phonetic analysis results revealed gender-specific preferences in the use of acoustic features related to backchanneling in conversation; however, the results also indicate that the acoustic variation in the speech preceding feedback is influenced by different interactions between all the considered factors. It was observed that female listeners tend to backchannel in overlap with the current speaker's turn after lower F0 values, while male listeners do so after higher F0. Gestural backchannels exhibit the highest preceding intensity. Dyad composition, in terms of gender of the two participants, and the temporal distance from the feedback location sig-

nificantly affect intensity variation as well: in mixed-gender dyads, backchannels tend to arise after a decrease in intensity, while in same-gender dyads, they follow steadier, louder intensity values. Speech rate patterns before feedback vary in relation to listener's gender, the transition type, and the distance from the location of the backchannel. Female listeners backchannel more quickly after slower speech, while male listeners do so after faster speech. Voice quality variation is linked to the completeness of the utterance, with higher CPP values associated with complete locations preceding backchannels. Therefore, while female and male speakers in this dataset appear to have specific preferences for F0 and speech rate values in order to locally place their backchannels within the current turn, other acoustic cues, such as intensity and voice quality, appear to be more connected to local decisions about the modality of the backchannel, or to the completeness status of the location.

The last study, reported in Chapter 10, investigated possible language-specific distributions of turn-taking structures, as well as language-specific gendered patterns in conversational behaviors, by comparing the German data with Swedish face-to-face spontaneous dialogues. Speaker changes, keeps and multimodal backchannels were taken into consideration, and their frequency and timing in relation to the preceding speech was observed across languages, and within different dyads, i.e., same-gender (male-male and female-female) and mixed-gender interactions. The main hypothesis that the analysis tested was:

Hypothesis 5 (H5): The distribution and the timing of speaker changes, turn keeps and multimodal backchannels display language-specific patterns in German and Swedish; the gender-specific turn-taking behavior displayed by male and female speakers in same-gender and mixed-gender dyads display language-specific dynamics as well.

The results from the quantitative analysis indicated different frequencies of backchannels between German and Swedish, with feedback constituting 27% of turn-taking structures in German, and 52% in Swedish. Moreover, while in German backchannels tend to be timed to the current speaker's turn in relation to their modality, as also reported in detail in Chapter 7 and Chapter 9, in Swedish backchannels seems to be produced mostly in overlap independently of their modality. As for the results related to gender, while in both German and Swedish male and

female speakers exhibit distinct behaviors, especially when considering the dyad composition, the dynamics of gender are language-specific. In German, female speakers tend to receive more overlapped backchannels, while male speakers show fewer instances of overlapped backchannels; in mixed-gender dialogues, males tend to overlap female interlocutors more often, whereas in same-gender dyads, they exchange longer uninterrupted turns. In Swedish conversations, male speakers display more overlapped transitions, a pattern also observed in female-female dyads. Additionally, the association of longer turns with male-male conversations is less evident in Swedish, with female speakers in mixed-gender dyads holding the floor more frequently. In same-gender dyads, females tend to engage in more overlapped talk and experience fewer silent gaps between turns, a trend similarly observed in male-male exchanges. These results on cross-linguistic differences and gender-specific preferences in turn-taking dynamics partially confirm H5.

Chapter 12

General discussion

The main objective of the current work was to explore and quantify, through empirical data and quantitative testing, the fundamental role of the sociolinguistic and communicative context in shaping the dynamics of turn-taking in conversation. More in detail, the central thesis that the studies aimed at demonstrating was that turn-taking has inherent social bases and it is rooted in the interactional situation in which the exchange takes place; therefore, the behavior of the speakers more or less adhering to, or more or less distancing themselves from, the one-speaker-at-a-time model, in which gaps and overlaps are kept to a minimum, can be partially accounted for by language-specific sociolinguistic and contextual variables that influence the realization of specific turn-taking mechanisms and dynamics.

Specifically, the present work has explored the role of the gender of the speaker and that of the listener, the modality of communication employed to provide backchannels, and the L1 of the interlocutors (German or Swedish) for the variation of turn-taking cues, and the distribution and timing of different turn-taking structures.

12.1 Gender and turn-taking dynamics

As discussed in Chapter 5, RQ1 specifically addressed the influence of gender in conversation, in order to understand to what extent the gender that is being performed by the interlocutors in that specific communicative context could be connected the observed turn-taking dynamics:

RQ1: Are the turn-taking-related variations of acoustic features and

turn-taking dynamics influenced by the gender of the speaker and the gender of the listener in conversation?

The studies reported in Chapters 6, 9 and 10 all involved the gender of both interlocutors taking part in the dialogues. The results indicate that the gender of speakers and listeners influences the use of acoustic turn-taking cues towards TRPs, as well as the frequency and timing of speaker changes, turn keeps and multimodal backchannels in the face-to-face conversations analyzed. This indicates that gender might play an important role in the way turn-taking cues are used in interaction in order to mark conversational intentions and to place backchannels within the ongoing turn.

This result has considerable implications for the investigation of both the performance of gender through phonetic detail, as well as gender-specific turn-taking behaviors. In fact, while there is an existing body of research investigating how gender is indexed in the variation of single features in speech, there seems to be a gap in the literature in analyzing how these features are used in an interactional context for turn-taking purposes. Equally, while previous research identified spontaneous conversation as one of the main social settings in which gender identity is constructed and performed, studies investigating the possibility of gender-specific strategies for turn management using acoustic means are lacking. Addressing these gaps, the present thesis suggest that there could be a fundamental connection between these two perspectives of analysis by indicating that, not only acoustic features display different patterns of variation preceding turn keeps and speaker changes, but also that their turn-taking-related variations can be gender-specific. The sociophonetic explanations of these patterns—which should be further developed in future studies, as the datasets that have been used for this dissertation are too small to make definitive claims in this regard—should at least in part differ from the ones related to the variation of individual features, since the interactional context, the turn-taking function and its outcome within the conversation should all be taken into consideration for the interpretation.

One of the main observations reported in the study in Chapter 6 is that both male and female speakers distinguish more noticeably between change and keeps using final durational and intensity values when the conversational partner is a male interlocutor. This was suggested to be indicative of a possible greater effort to main-

tain smooth turn-taking by using greater acoustic signaling. Broadly, this could potentially point towards an adaptability of cue use based on gender-specific performance expectations by the interlocutor. More specifically, if male interlocutors are expected to be less inclined to cooperate in the turn-taking mechanism, e.g., by being more prone to overlap their interlocutor or by leaving silent gaps before accepting the speaking turn, as suggested by Whiteside (1995) for English, a greater signaling effort is requested from the other party. On the other hand, male speakers putting less effort in phonetic signaling in conversation with female speakers, as do female speakers in same-gender conversations, might be indicative of higher expectations of collaboration from the other party. While these represent the possible interpretations of the results from a first exploratory study on the matter, the finding that turn-taking cues display patterns of sociophonetic variation should be further investigated in larger-scale studies. In fact, due to the absence of previous research addressing specifically the gender-specific variation of turn-taking cues, direct comparisons with prior findings are currently impossible to report. Further studies in this direction are necessary to provide a larger picture of the phenomenon and more in-depth explanations.

In this dissertation, the sociophonetic variation of turn-taking cues was further investigated in the study reported in Chapter 9 by focusing on the placement of backchannels, i.e., a conversational device that has fewer distributional restrictions than full-fledged turns, and is more directly connected to the social dimension of the interaction (Roberts et al., 2015; Stubbe, 1998, 2013). The analysis focused on the acoustic configuration of the current utterance preceding the production of a backchannel by the listener. If male and female listeners weigh turn-taking cues in gender-specific ways, then they would produce these optional responses at different acoustic locations within the current turn.

The results support this hypothesis. In fact, female listeners seem to pay more attention to produce backchannels at more predictable locations (i.e., after speech marked by low F0), thus potentially causing less disruption. Additionally, female listeners were found to backchannel more promptly after areas of slower speech rate in the current turn, while male listeners after faster speech rate. These findings confirm the results from the study in Chapter 6 on a larger dataset and on a more variable and socially-marked turn-taking structure. Specifically, a faster speech rate

might signal to male listeners an appropriate approaching opportunity to backchannel as it might be perceived as a turn-hold cue —concurrently displaying a higher effort by the current speaker in maintaining the floor by rushing through TRPs. On the other hand, female listeners backchannelling after slower speech suggest that that is perceived as a turn-hold cue, which might also be related to possible hesitations in the speech of the current speaker, and the production of feedback might function as a confirmation that the channel remains open, and highlighting cooperative effort.

The possibility of the preference for different cues across genders to place backchannels within the current turn might also be detected in the results obtained by Rivera and Ward (2007), who investigated Northern Mexican Spanish. In their study, they briefly mention that the combination of cues they identified as most predictive of backchannels —consisting of low F0 followed by a rise and a decrease in speech intensity and rate—was less accurate in predicting backchannels in conversations between two male interlocutors compared to female-female and mixed-gender dialogues. The authors only mention this tendency without going into further exploration; however, the observation that backchannels are less predictable in male speech after specific acoustic configurations would support the results obtained in this thesis. In fact, it might suggest that male and female listeners place backchannels at different acoustic locations in speech, also depending on the gender of their interlocutor. It might also indicate that, when in conversation with each other, male interlocutors might pay less attention to backchannel placement within the current turn. Both suggestions are in line with the results obtained in this dissertation.

In summary, these findings indicate that turn-taking cues display gender-specific patterns to mark conversational intentions by the speakers, and to time backchannels to the current speaker's turn for the listeners. As turn-taking cues have yet to be investigated from a sociophonetic perspective, this dissertation represents an addition to the literature by suggesting the potential influence of gender on the interactional use of acoustic features in informal conversation, and by proposing a possible methodology to expand this investigation within bigger datasets.

Gender-specific listener behavior was also observed in the timing of backchannels. In fact, female listeners were found to produce more backchannels right at the offset of speech, or in overlap with it, with other female speakers, while male

listeners produced more backchannels after a silent gap with male speakers. Conversely, in mixed-gender conversations, men produced more overlapped backchannels, while women with a gap or a no-gap-no-overlap transition. Again, as was the case for turn-taking cues, this result could suggest the adoption of a backchanneling pattern influenced by both the perceived gender of the speaker and its related expectations, with women's communication style linked to providing feedback while the turn is still ongoing, which is less common in men's conversational behavior. This tendency was previously noted by Feke (2003), who found that female listeners in her dataset adjusted to the backchanneling style of male subjects —whereas in the current study, both female and male listeners both seem to adjust to each other.

In conclusion, these results contribute to the literature on gender-specific conversational dynamics by providing an empirical and quantitative account of the extent of the influence of the gender of both interlocutors, in terms of the specific behavior of the speaker/listener and the expectations they have about listeners/speakers of a specific gender. They highlight the importance of a more sociolinguistic-oriented research approach to turn-taking cues and add to the literature on the timing and distribution of backchannels in conversations within same-gender and mixed-gender dyads.

12.2 Listener behavior across modalities

The analysis of backchannels within the current work was motivated by their nature of optional responses closely associated with the social dynamics of the specific interactional exchange, as well as by their multimodal character, as their frequency and distribution are strongly influenced by the modality in which they are produced. For these reasons, backchannels represent an ideal turn-taking device to investigate both the influence of social dimensions and the availability of multiple modalities in the interaction on the timing of listener responses in relation to the current speaker's turn. The studies reported in Chapter 7 and 9 aimed at answering RQ2:

RQ2: How does the availability of multiple modalities (i.e., the visual and the vocal modalities) in the interaction affect the timing of

backchannels in relation to the offset of speech by the current speaker and to the acoustic configuration of the current utterance?

The findings from the study reported in Chapter 7, which investigated different types of backchannels in German, indicate that multimodal alternation, i.e., the choice to produce a backchannel in the verbal modality, in the visual modality, or with both the head and the voice simultaneously, as well as the specific forms that backchannels may acquire, do significantly impact the temporal dynamics of turn-taking. Specifically, the results provide further support to the theory that listeners are very sensitive to the specific realization and timing of feedback within different contexts (c.f., Wehrle et al., 2023).

Gestural feedback has been observed to provide support to the current utterance without causing disruption, and to arise before the end of a turn, also in anticipation of a verbal backchannel arising after the offset of speech (Dittmann and Llewellyn, 1968; Ferré and Renaudier, 2017; Truong et al., 2011). These patterns were observed in the results reported in Chapter 7, in which the onset of gestural and vocal feedback had different timings in relation to the end of the current speaker's turn. Visual resources were observed to be employed in order to provide non-disruptive feedback at turn-internal locations, but also at turn-external locations, for instance as part of multimodal backchannels which, given their less reduced acoustic form, were suggested to be more emphatic and enthusiastic than verbal ones.

In the vocal modality, backchannels with different forms displayed specific timing patterns: non-lexical feedback arose both in overlap and right at the offset of speech; lexical forms, having a more complex content, also comprising multiple syllables, are preferred turn-externally and after silent gaps. In spite of being the loudest type of backchannel in the dataset, laughter was found to occur frequently both in overlap with the current turn and after the offset of speech. This result is in line with previous findings indicating that listener's laughter frequently arises during the current speaker's turn, or as a joint event involving both interlocutors (Ludusan and Wagner, 2022b; Ludusan et al., 2020; Trouvain and Truong, 2013). A specific pattern related to laughter was also observed in relation to the head movements that co-occur with it in the German dialogues analyzed. In general, confirming previous results by Dittmann and Llewellyn (1968) and Ferré and Renaudier (2017), head movements tend to be longer than lexical and non-lexical components in mul-

timodal backchannels. For laughter, a specific movement was observed to co-occur with it, i.e., protrusions, which, like in the other cases, started before the vocalization, but ended before the laughter was finished. Protrusions are not constituted by multiple phases, as was the case for other gestures (e.g., nods), and thus had a shorter duration and a different temporal alignment with the vocal feedback that they accompanied, i.e., laughter.

Overall, these results highlight the complexity of listener behavior as a system in which local decisions significantly shape temporal dynamics of the conversation. The findings contribute to the literature on the temporal distribution of various types of vocal feedback within the current turn, emphasizing the greater flexibility linked to the reduced potential for interruption of certain forms. This precise timing underscores the role of multimodal planning behind the listener's local decisions to produce a backchannel at a specific moment.

The study reported in Chapter 9 was specifically aimed at identifying the features, in the current speaker's turn, that might be used by the listener to locally plan backchannels across modalities. The results indicate that overlaps occur at syntactically/semantically incomplete locations within the current turn, possibly to encourage continuation of what is being said. Moreover, the feature of intensity of the current speech was the one that emerged as significant for the placement of backchannels in different modalities within the speaker's turn. This supports previous findings by Truong et al. (2011), who report that the intensity profiles in the speech preceding backchannels varied in relation to the modality of the responses, with gestural ones being preceded by loud speech, since they tend to occur in overlap with the current utterance. On the other hand, in this analysis, as in previous investigations (e.g., Levow and Duncan, 2012; Rivera and Ward, 2007; Truong et al., 2011), verbal backchannels were preceded by a decrease in intensity. These results reveal a close relationship between timing of backchannels across modalities and the variation of intensity within the conversational turn.

Still in relation to timing, it was observed, in the analysis in Chapter 9, that a higher intensity, or a final rising intensity, led to more closely-timed backchannels by the listener (i.e., they were produced with an overlap or with a no-gap-no-overlap transition), while after a stark decrease in intensity, backchannels were produced after gaps. It is possible that, in the latter case, where backchannels followed a turn-

yielding cue (intensity decrease) and a pause, feedback was uttered by the listener as a “perverse passive” (Ward and Tsukahara, 2000, p. 1184; cf. Jefferson, 1984) in order to refuse the floor with a minimal contribution (Cutrone, 2005). This might suggest the existence of specific loudness configurations preceding feedback with different timing and interactional purposes and, in general, point towards a primary role of intensity for the timing of backchannels in German. Specifically when considering backchannels across modalities, this acoustic feature appears to have a more decisive role for timing than F0, which was assigned a central emphasis in previous studies focused primarily on verbal feedback (e.g., Boudin et al., 2021; Gravano and Hirschberg, 2009a, 2011; Koiso et al., 1998; Levow and Duncan, 2012; Rivera and Ward, 2007; Ward and Tsukahara, 2000).

These results contribute to the research on the phonetic features in the speech preceding backchannels upon which listeners rely to time their responses appropriately across modalities. As mentioned earlier, while other cues may not be strictly tied to modality, they are associated with the gender of both the listener and the current speaker in conversation. Consequently, they could be utilized to place backchannels independently of modality, but rather based on gender-specific patterns related to turn-taking and listener behavior. This again highlights the importance of considering gender alongside other factors, such as the availability of multiple modalities, in understanding turn-taking dynamics, and emphasizes that both factors, to varying degrees, influence the structure of German face-to-face dialogues.

12.3 Language-specific turn-taking dynamics

Finally, the dissertation comprised a cross-linguistic comparison of the previously observed turn-taking dynamics to investigate the influence of the language and its cultural background in relation to general conversational habits and gender-specific language usage patterns. This comparison was carried out between German and Swedish.

RQ3: Do the same patterns of variation occur across (typologically related) languages, or do social factors and the availability of multiple modalities influence conversation in language-specific ways, connected to the language-specific conversational habits, preferences, and values?

The answer to RQ3 in this dissertation is that German and Swedish, despite being typologically related languages and sharing similarities in societal structures (cf. Weirich and Simpson, 2023), exhibit language-specific approaches to the implementation of turn-taking cues and dynamics in conversation. Moreover, the patterns of variation associated with the gender of the speaker and the gender of the listener also appear to be language-specific, indicating the existence of subtle differences, in the two groups' cultural backgrounds, that contribute to shaping the turn-taking dynamics.

Before discussing the results of the cross-linguistic comparison between German and Swedish, it is important to clarify how the term language-specificity will be used to address both acoustic variation and the cultural context of the target groups of speakers. The description "language-specific" is used to refer to phenomena that are either observed for German or for Swedish (or any other language), and represent a difference between the two. However, such language-specificity may have different sources. In the study reported in Chapter 8, that focused on the different use of F0 as a turn-taking cue between the two languages, language-specific tendencies are grounded in linguistic or, more specifically, phonological reasons. In the cross-linguistic study reported in Chapter 10, which analyzes the distribution and the temporal dynamics of different turn-taking structures, the source of the language-specific tendencies observed was suggested to be connected to extra-linguistic or, more specifically, socio-cultural factors.

The first cross-linguistic study was reported in Chapter 8 and it highlighted the different use that German and Swedish make of F0 as a turn-taking cue. The language-specific use of F0 as a turn-taking cue in German and Swedish was further investigated by Feindt et al. (in preparation), who confirmed the preference for Swedish speakers for durational features rather than F0 to mark TRPs, given the fact that the use of F0 might already be saturated as a signaling cue, given its lexical-phonological function in Swedish.

Moreover, the analysis reported in Chapter 8 also highlighted the possibility of entrainment in F0 values among the speakers in the German subset. This phenomenon was not observed among the Swedish participants. Entrainment is a conversational device (Dideriksen et al., 2023) that consists of a "cross-participant repetition of communicative behaviors" (Rasenberg et al., 2020, p. 29), such as acoustic values,

lexical structures or gestures (Brennan and Clark, 1996; Graziano et al., 2011; Levitan et al., 2011; Levitan and Hirschberg, 2011). Entrainment in conversation is employed by interlocutors in order to develop common ground and rapport and is directly connected to the social dynamics at the background of the interaction (Brennan and Clark, 1996; Gonzales et al., 2009; Lubold and Pon-Barry, 2014). In the current study, it was suggested that the observed entrainment among German conversational partners, who became more similar in their F0 values towards potential turn boundaries, while Swedish speakers did not, may be once again linked to the greater freedom in the use of pitch variation among German speakers. In fact, alongside being used to signal conversational intentions, pitch might be used by German interlocutors to establish specific social dynamics. It was not excluded, though, that Swedish speakers might rely on another, more available acoustic turn-taking cue to develop entrainment. Further analyses are needed to investigate this hypothesis.

Since the dataset used in that study was not balanced for the gender of the speaker, the gender-specific variation in the use of F0 as a turn-taking cue across languages was not tested. However, in line with the hypothesis advanced for the possible entrainment patterns observed, it could be expected that, while German speakers display patterns of gender-specific turn-taking-related variation of F0, Swedish speakers might rely on the variation of another cue to perform gender in conversation, such as, for instance, duration rather than F0. Further research is necessary to explore this hypothesis; however, previous investigations on sociophonetic variation in German and Swedish point toward its validity. In fact, Weirich et al. (2019) and Weirich and Simpson (2023) report that, while in German F0 shows gender-specific values in the speech of men and women, the difference between the F0 values of Swedish men and women was smaller. Weirich and Simpson (2023) suggested that Swedish speakers might prefer other acoustic values to index gender identity in speech rather than F0. Thus, similar results might be obtained in a larger-scale comparison of turn-taking acoustic cues across German and Swedish male and female speakers.

Finally, the analysis in Chapter 10 investigated the cross-linguistic and gender-specific conversational dynamics in German and Swedish. In this study, the datasets

used were balanced for the gender of both interlocutors, thus a comparison was possible.

The main results that emerged from the general cross-linguistic analysis show that Swedish listeners produced higher percentages of backchannels in comparison to German listeners, revealing a different backchannel frequency baseline in the two languages. Specifically, more than a half of the extracted turn-taking structures in Swedish dialogues were backchannels, while in German they represented a quarter of the data. These results are in line with previous findings related to the amount of backchannels observed between West Germanic languages and Scandinavian languages. Knudsen et al. (2020) report that, in their German dataset, they find a 16% of backchannels, with an average of 15% across Dutch, English and German. On the other hand, studies involving Swedish, such as Cerrato (2002), indicate that feedback occupies the 50% of the conversations in the dataset and, in his cross-linguistic comparison taking into account Spanish and Swedish, Fant (1989) reports a 60% of verbal backchannels for Swedish versus a 4% for Spanish for 10 minutes of dialogues. Fant (1989) connects this tendency to the different cultural backgrounds and values of the two languages, that, in conversation, are reflected in their different frequency baselines for backchannels, and in the preference, for Swedish, to display attentiveness through verbal responses, while gaze is the preferred mean in Spanish. In particular, he claims that speakers of different languages have "culturally determined communicative priorities" (Fant, 1989, p. 249), and Swedish speakers place great importance on displaying agreement and solidarity through their "relatively emphatic" (Fant, 1989, p. 260) backchanneling style, and on avoiding assertive or competing behaviors. This Swedish attitude is reflected in different spheres of social life, such as group discussions, political speech, business meetings (e.g., Auer-Rizzi and Berry, 2000; Daun, 1996; Ilie, 2004; Tixier, 1996). German conversational style, on the other hand, highly values discussion, exchange of different point of views and confrontation (Byrnes, 1986), and overt displays of agreement tend to be less frequent in conversation. These two distinct attitudes towards agreement displayed through listener behavior could explain the language-specific frequency of backchannels observed across the German and Swedish datasets analyzed for this dissertation, and confirm the influence of culture-specific values in shaping speakers' conversational behavior.

Another finding from the analysis of backchannel distribution across German and Swedish is that, in Swedish, backchannels tend to predominantly occur in overlap with the speech of the current speaker, with no apparent effect of modality on their distribution. While gestural items still occur more frequently in overlap and verbal items the least, the majority of backchannels across modalities arise while the current speaker is still talking, rather than at the offset of speech or after a silent gap. Even if the vocal component is constituted by a monosyllabic or a non-lexical item, thus minimally interrupting, the result that modality does not seem to influence the timing of feedback in Swedish is a striking finding, in that it seems to deviate from the studies outlined above highlighting the close connection between timing and multimodal alternation. This observation can be connected to culture-specific variation in the timing of feedback across modalities already evidenced in previous studies. For instance, Stubbe (1998) found that Pakehas (New Zealanders of European descent) and Maoris display preferences for different backchannels depending on the timing, with Pakehas providing a higher rate of verbal backchannels closely timed to speech, while, in comparison, Maoris displayed more frequent non-verbal feedback often after silent gaps; Lehtonen and Sajavaara (1985) report the low tolerance of Finnish speakers for simultaneous speech together with the prevalence of gestural backchannels in their interactions; Maynard (1990) found that, while Japanese and American English listeners use similar backchannels across modalities, Japanese listeners overlap more frequently with the current speaker, while American English listeners tend to provide feedback during silent gaps, highlighting the existence of language-specific preferences related to timing. Building upon these prior investigations and introducing a quantitative approach, the findings of the current thesis indicate that Swedish and German listeners within the analyzed datasets also display different backchanneling patterns across modalities: while German listeners seem to associate specific feedback modalities with specific timings within the current turn, Swedish listeners appear to do so to a lesser degree, placing more importance on the overall overt display of support and attentiveness. These results contribute further empirical evidence of how culture-specific interactional norms and social values can influence conversational dynamics, particularly regarding listener behavior in conversation. In turn, they also demonstrate how the

systematic investigation of turn-taking dynamics can be employed to study social and cultural habits.

Differences in gender dynamics were also noted between German and Swedish conversational patterns. In German dyads, turn-taking dynamics align more closely with conventional conversational habits attributed to female and male speakers as documented in existing literature (e.g., Coates, 2004; Edelsky, 1993; Plug et al., 2021; Tannen, 1990; Zimmermann and West, 1975). Specifically, in German conversations, female speakers tend to receive more backchannels by the listener, particularly during speech, while male speakers receive fewer overlapped backchannels. However, in mixed-gender dialogues, male speakers tend to overlap female interlocutors more frequently and engage in longer uninterrupted turns. In contrast, Swedish males exhibit more instances of overlapped turn-taking, similar to the behavior observed in female-female dyads. In contrast to German conversations, longer turns are not consistently associated with male-male dialogues in Swedish, and in mixed-gender dyads, female speakers often dominate the exchanges by keeping the turn for longer. Within same-gender dyads, both female and male speakers in Swedish tend to engage in more overlapped talk and exhibit fewer silent gaps between turns, a pattern observed in male-male dyads as well. These findings can be considered consistent with those reported in the comparative analysis of German and Swedish conducted by Weirich and Simpson (2023) and Weirich et al. (2019). Notably, they observe that while German women and men exhibit more conventional patterns of acoustic variation in speech, Swedish speakers show a smaller divide between genders, with female speakers adopting speech patterns more similar to those of male speakers, suggesting a less pronounced gender-based distinction in Swedish speech. The same separation from a conventional gender divide in conversational behavior was found in the current results for Swedish, even though male speakers in the Swedish dataset also appear to diverge from conventional masculine turn-taking behaviors.

In conclusion, interlocutors' L1 is observed to influence the dynamics of turn-taking in relation to the specific acoustic turn-taking cues used to mark conversational intentions in speech, and the way in which such acoustic values might be used as mean of social cohesion. Moreover, the modality and timing of listener responses, along with gender-specific conversational behaviors, are also impacted

to different extents cross-linguistically and cross-culturally. Together, these factors play a significant role in shaping turn-taking dynamics in face-to-face conversation.

12.4 Potential limitations

12.4.1 Subjects

Possible limitations to the results obtained by the studies can be identified in the subjects that were part of the datasets on which the analyses were conducted. Given the availability of speech corpora containing free conversation in both German and Swedish, the analysis was conducted using pre-existing data. Thus, the individual and social variables related to the background of the participants could be controlled only to a certain extent, based on their availability within the corpora. For instance, while the more recent research on gender urges for the rightful inclusion of non-binary categories alongside the binary male-female cisgender distinction (e.g., Gratton, 2016; Pearce, 2019) in order to develop a more comprehensive and realistic picture of the performance of gender in speech within specific societies and contexts, this dissertation only investigated male and female speakers. The availability of the subjects in the corpora used, i.e., the subjects within German sub-corpus of the DUEL Corpus (Hough et al., 2016), the Spontal Corpus of Swedish (Edlund et al., 2010), and the Lindenstraße task in the Kiel Corpus of Spoken German (Kohler et al., 2017) determined this limitation, since they self-identified as male or female. A more comprehensive description of gender-specific dynamics could thus not be provided.

Moreover, it is a central theme to the intersectional approach to sociolinguistic research that the gender of the speaker by itself is not sufficient to account for the social experience and behavior of an individual, but rather “gendered meanings are also simultaneously classed, raced, and region- and age-specific (not to mention a host of other such ‘specificities’) and [...] speakers draw on all of these underlying social connotations together” (Levon, 2015, p. 296). While the cross-linguistic analyses allowed the exploration of gender-specific patterns within two different languages in the current study, the absence of more extensive information regarding the sociolinguistic background of the speaker across the corpora limits a broader understanding of turn-taking dynamics.

Finally, the sample of the datasets used in the gender-related analyses (24 subjects for German and 12 for Swedish) poses a potential limitation for making broad claims about gender influence on conversational behavior. However, making such claims was not the objective of this dissertation. Rather, generalizations must be limited to the datasets that were employed, based on observed dynamics related to participants' gender (and/or L1), while also comparing the results with previous studies in order to hypothesize potential explanations. As mentioned in the previous Sections, the systematic inclusion of sociolinguistic variables in turn-taking studies aimed to highlight the extent to which speakers' identity features and the interactional context significantly affect turn-taking, underscoring the need to consider such factors in future, larger-scale studies. In other words, despite working with relatively small datasets, this thesis aims to present a nuanced and context-sensitive approach to turn-taking analysis. Future research will apply such approach to larger datasets, potentially allowing broader claims about gender performance and other identity variables in turn-taking dynamics.

12.4.2 Data

Another potential limitation regards the comparability of the data used for the cross-linguistic analyses reported in Chapter 8 and Chapter 10. As previously mentioned, the exploratory analysis reported in Chapter 8 did not control for the gender of the participants, which caused a limitation for a more in-depth analysis of the social dynamics that could have potentially influenced F0 as a turn-taking cue. Moreover, a further comparability issue for that specific analysis is that in the Swedish dataset speakers sat face-to-face to each other, while in the Lindenstraße task of the Kiel Corpus subjects were not able to see each other.

The issue of comparing these two corpora in the cross-linguistic analysis of turn-taking cues, already raised in Feindt et al. (in preparation), is that the availability of the visual modality might lead to a weaker signaling using acoustic cues, since speakers are able to rely on visual cues, such as gestures and eye movements (e.g., Edlund and Beskow, 2007, 2009; Mondada, 2007; Zellers et al., 2019). On the other hand, when the two speakers could not see each other, acoustic cues might be more salient, since speakers cannot simultaneously rely on the visual modality in order to coordinate with each other to time their transitions. In Feindt et al. (in prepara-

tion)’s results, they identified more cues significantly contributing to turn-taking in German compared to Swedish. Moreover, while they observed a significant contribution of segmental duration to turn-taking in Swedish, features like F0 and intensity were not significant, whereas they were in German. The difference in the corpus settings may explain the discrepancy in the number of cues used to signal conversational intentions. Nevertheless, the cross-linguistic preference for specific turn-taking cues over others remains detectable, indicating the reliability of the analysis in spite of the limitations.

While in the study reported in Chapter 10 all the subjects could see each other when talking, further comparability issues between the German DUEL corpus and Spontal Swedish corpus, as discussed in Section 10.6.1, arise from differences in the specific tasks assigned to speakers, and the selection of dialogue excerpts. While dialogues in the Swedish corpus were entirely spontaneous in terms of topic choice and interactional dynamics (Edlund et al., 2010), those in the DUEL corpus were loosely task-based: participants were given a theme for discussion so that they could start talking without spending too much time in searching for a topic (Hough et al., 2016). It is possible that the introduction of even a minimal task could influence conversational patterns, especially regarding the salience of speaker gender. In task-based dialogues, in fact, gender identity tends to be more salient, leading to more stereotypical and conventional gendered behaviors in conversation (Leaper and Robnett, 2011; Palomares, 2004; Vinciarelli et al., 2015). However, as Palomares (2004) reports, gender salience is not only modulated by one single contextual variable, such as having a task to carry out; instead, it is the result of several characteristics of the situation, as well as the participant’s identity, attitudes and perception of the interaction. For instance, other identity aspects, such as being a student or an employee (Mulac et al., 2003; Palomares, 2004), might be more relevant in specific contexts or when discussing specific topics, and might thus interact with gender performance in conversational behavior. In studies that took the type of conversational activity (structured vs. unstructured) into account, such as Leaper and Ayres (2007) and Leaper and Robnett (2011), while differences between genders were mostly detected in structured dialogues, other factors were found to carry more weight in determining conversational behaviors, such as the gender composition of the dialogue pair, the length of observation (e.g., more differences were

detected in longer exchanges), and the type of population analyzed (e.g., male and female undergraduate students presented more differences in conversational behavior between each other than older speakers, such as graduate students and working adults). Similarly, in Vinciarelli et al. (2015), while the presence of a task did elicit an increased use of non-verbal expressions associated with dominance by male subjects in phone conversations, other factors, such as gender composition, the role assigned to the participant in the task and the individual speaker's personality traits were found to play a major role in shaping the distribution and frequency of laughter, silences, backchannels, fillers and overlapping speech. Therefore, while the DUEL dialogues were spontaneous but with minimal task constraints, it is possible that the presence of a given conversational goal requiring coordination and collaboration between the two participants might have had an effect on gender-specific dynamics. However, it is not possible to quantify the extent of such effect, as several other variables (e.g., the discussed topics or the degree of familiarity between the two speakers) might have had relevance for gender performance. This is not true only for the DUEL corpus, but for the Spontal conversations as well. In fact, even if the exchanges were completely spontaneous, other aspects of the interaction, besides those taken into consideration, might have contributed to make gender identities more or less salient. A comparison of interactions in German and Swedish in which the same type of activity is being carried out (e.g., either fully spontaneous conversations or task-based interactions) will provide further insights into the potential impact of instructions given to participants in the DUEL Corpus used for the present work. Moreover, the operationalization of further identity and contextual factors is essential to determine how gender identity interacts with other variables in different situations, and how relevant they are for the dynamics of turn-taking.

As for the second limitation addressed for this analysis (Section 10.6.1), the availability of the Spontal data led the annotation and the analysis to focus on dialogue excerpts where interlocutors were in mid-conversation. On the other hand, dialogues investigated for the DUEL Corpus were available in full and annotated from their onset. Given previous findings suggesting the relevance of the time-course of the conversation for the frequency of conversational devices like backchannels or entrainment (Fusaroli et al., 2017; Levitan and Hirschberg, 2011; Weise et al., 2019), this difference posed a comparability risk, since it is possible that these dynamics

might increase over time. This would cause Swedish dialogues to potentially exhibit more of them. However, the results obtained, for example, for backchannel frequency did not differ from previous studies on Swedish or other Scandinavian languages (e.g., Cerrato, 2002; Dideriksen et al., 2023; Fusaroli et al., 2017; Knudsen et al., 2020), which might suggest a potential limited influence of the datasets' difference on observed turn-taking dynamics. For this reason, the results were interpreted in terms of language-specific and gender-specific tendencies. However, without a more comprehensive analysis of the dyads in the datasets, it is not possible to investigate more in detail how backchannel use would develop over the time course of the conversation in German and Swedish. Again, the use of more comparable data to carry out this analysis will potentially help quantifying the actual effect of the temporal dimension on the frequency of specific conversational devices in the two languages.

12.4.3 Acoustic analysis

One last potential limitation to be addressed within the current work is the use of two different extraction and visualization methods for acoustic features in Chapters 6, 8 and 9. In the two initial studies (Chapter 6 and Chapter 8), acoustic values were extracted from predetermined time-points, or test locations, within the turn leading up to a TRP. This method aimed to provide quantitatively comparable time points to investigate the time course of acoustic turn-taking cues, and to analyze the potential extent of the transition space by focusing on determined and replicable time windows. Following the design employed by Feindt et al. (2021), different test locations were explored to achieve this objective. However, this approach, while offering comparable time points, made it challenging to detect and visualize the dynamic course of acoustic features in the analysis. Moreover, since values were collected at specific locations, they were extracted independently of the corresponding segment within speech, potentially including consonants, which, even after normalization, could have skewed results (for example, segments like sibilant fricatives are characterized by very intense signals, e.g., Ohala and Solé, 2010).

To address these issues and explore further methodological approaches, the study outlined in Chapter 9 adopted different extraction and analysis parameters. Following Zellers (2017), a time window of 1 second before the TRP was taken,

and values were extracted from the nuclei of the syllables within that specific time-frame. This enabled the exclusion of potential confounding factors related to the acoustic features of specific segments, which might have caused issues in previous calculations. Additionally, it allowed for a more dynamic representation and analysis of the full acoustic patterns of variation over the time-course preceding the TRP, by accounting for the specific timing of each syllable. Consequently, the visualizations and analysis became more comprehensive, yet still enabled the identification of potential specific time locations where the acoustic pattern may have exhibited a significant change compared to previous values, indicating a potential turn-taking-relevant time-point for acoustic signaling.

12.5 Outlook

The analyses presented in this work, particularly those in Chapter 6 and Chapter 9, suggest that turn-taking cues exhibit sociophonetic variation: the dynamic use of acoustic features to mark TRPs in conversation provides insight into the speakers' interactional attitude, their expectations regarding conversational floor management, and possibly the level of cooperation anticipated from their interlocutor based on social factors. Therefore, future advancements adopting a sociophonetic approach to studying turn-taking in spontaneous interaction —currently a gap in the research —will consider the gender of the interlocutors, together with its interaction with further variables, in order to offer a more comprehensive understanding of how turn-taking is influenced by the sociolinguistic background of individuals within specific conversational contexts. The same perspective applies to broader conversational dynamics, including the distribution and timing of speaker changes, pauses, and multimodal backchannels among interlocutors. While previous research has explored the role of gender in shaping such dynamics, particularly in relation to backchannels, its interaction with additional language-specific social factors remains understudied. Future cross-linguistic studies on comparable corpora are essential for further contextualizing results within a broader framework that also takes into consideration the cultural norms of the speaker's community.

Finally, while the current thesis explored two different approaches in observing acoustic turn-taking cues approaching TRPs —namely, by extracting acoustic fea-

tures at predefined time points or over the final 1-second interval of speech —and tested their variations using linear mixed effects models, further statistical tests and data visualization methods could offer a more comprehensive picture of the transition space, as well as an improved approach to analyzing the dynamic development of acoustic contours throughout the turn. Further studies on this dataset will develop the methodology presented in the dissertation by employing more dynamic types of statistical testing, such as Generalized Additive Mixed Models (GAMMs; Hastie and Tibshirani, 1990; Wood, 2017), as done in studies by Włodarczak et al. (2023) and Feindt et al. (in preparation) to investigate turn-taking related variations of acoustic features. Allowing the modeling of dynamic contours, as well as the analysis of non-linear variation over time (Winter and Wieling, 2016), the use of GAMMs in future studies will provide additional quantitative information on the extent of the phonetic and sociophonetic variations of turn-taking cues in spontaneous conversation.

Chapter 13

Conclusions

The findings of this work shed light on the individual and combined influence of the gender of the speaker and the listener, the modality of the response and language-specific factors on several aspects of the turn-taking mechanism.

One of the main findings of this dissertation is that the gender of the speaker and of the interlocutor have a significant effect on the turn-taking-related variation of the acoustic cues approaching a potential speaker change location in German informal conversation. Backchannels were also found to arise at turn locations marked by gender-specific acoustic configurations, suggesting a varying perception of acoustic features' communicative functions, which might motivate the differences observed in backchannel placement. The modality of backchannels also plays a significant role in the distribution of listener responses and their timing in relation to speech. The results show that multimodal alternation is a key feature of listener behavior, employed to signal attentiveness and participation while reducing disruptive overlaps. These findings contribute additional empirical evidence on the interactional behavior of interlocutors in conversation, and highlight listeners' great sensitivity in achieving this dynamic through the use of the more appropriate form of feedback.

A further main finding of the current work is that the language of the interaction, its specific constraints, the related socio-cultural values and conversational priorities, are at the base of quantitative differences in regards to turn-taking dynamics. The cross-linguistic comparison of German and Swedish revealed an important influence of the language-specific affordances of acoustic cues for their contribution to turn-taking and to coordinating devices in informal conversation. Moreover, German and Swedish speakers involved in the face-to-face dialogues from the DUEL Corpus and the Spontal Corpus were found to display different preferences

in the use of backchannels, as well as the multimodal alternation for their production. Gender-specific conversational dynamics were found to be language-specific as well, with German male and female interlocutors adopting more traditional gendered patterns in relation to, e.g., overlapped speech, amount of turn-yields and floor-holds, and backchannels frequency, while the distinction between Swedish male and female interlocutors appeared to be less clear and less gender-specific.

This dissertation was the first to investigate in detail the sociophonetic variation of turn-taking cues in conversation. While additional work on more extensive datasets is needed to further explore this topic, the present work found significant associations between the use of acoustic features as means to control the conversational floor and the gender of the speaker, as well as the gender of the interlocutor involved in the dialogue, indicating that specific patterns of acoustic variation for turn-taking might carry social meaning, and might be part of the speakers' performance of social identity in spontaneous interactions. Thus, in spite of its limitations, this dissertation represents a first step towards the development of an additional approach to the study of turn-taking cues and the structure of interactions by involving sociolinguistic variables in the analysis in a more systematic and quantitative way.

An additional main contribution of the current work to the literature on natural conversation is given by the quantitative cross-linguistic analysis of turn-taking dynamics. The results obtained on the language-specific variation of listeners' multimodal behavior, as well as the gender-specific patterns observed across languages, carry fundamental significance for the investigation of spontaneous interactions and their structures. In fact, it is important to stress that the two languages considered for the current work are German and (Central) Swedish, i.e., two typologically related languages whose societies share significant structural similarities and cultural ties, but also have relatively subtle and not immediately identifiable differences. The gender-specific and language-specific patterns in the implementation of the turn-taking mechanism highlighted by this dissertation strongly point towards the ability of turn-taking dynamics to reflect fine-grained variations related to the individual speaker's socio-cultural identity and its performance within a specific context.

Altogether, these findings show that turn-taking is uniquely informative for improving our understanding of individuals' multifaceted performance of their social identities through everyday interaction.

Part V

Bibliography

Bibliography

- Agliati, A., Vescovo, A., & Anolli, L. (2005). Conversation patterns in Icelandic and Italian people: Similarities and differences in rhythm and accommodation. In L. Anolli, S. Duncan Jr., M. Magnusson, & G. Riva (Eds.), *The hidden structure of interaction: From neurons to culture patterns* (pp. 224–235). IOS Press: Amsterdam.
- Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1), 1–26.
- Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Jastrzebska, M., & Trouvain, J. (2017). Comparison of pitch range and pitch variation in Slavic and Germanic languages. *Proceedings of Speech Prosody*. <https://doi.org/10.21437/speechprosody.2014-143>
- Anolli, L. (2006). *Psicologia della comunicazione*. Il Mulino: Bologna.
- Auer-Rizzi, W., & Berry, M. (2000). Business vs. cultural frames of reference in group decision making: Interactions among Austrian, Finnish, and Swedish business students. *The Journal of Business Communication*, 37(3), 264–288.
- Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final “go-signals”. *Frontiers in Psychology*, 8, 1–10.
- Bartoń, K. (2022). *Lme4: Linear mixed-effects models using ‘eigen’ and s4* [R package version 1.47.1]. <https://CRAN.R-project.org/package=MumIn>
- Baumann, T. (2013). *Incremental spoken dialogue processing: Architecture and lower-level components* [Doctoral dissertation, Bielefeld University].
- Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6), 941.
- Beňuš, Š. (2009). Turn-taking strategies in cooperative task dialogues. *Discourse and Interaction*, 2(1), 19–31.
- Beňuš, Š. (2016). The prosody of backchannels in Slovak. *Proceedings of Speech Prosody*, 75–79.
- Beňuš, Š., Gravano, A., & Hirschberg, J. (2007). The prosody of backchannels in American English. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2007)*, 1065–1068.
- Berry, A. (1994). *Spanish and American turn-taking styles: A comparative study*. ERIC. <https://eric.ed.gov/?id=ED398747>

- Bhuta, T., Patrick, L., & Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, 18(3), 299–304.
- Birdwhistell, R. L. (1966). Some relations between American kinesics and spoken American English. *Communication and Culture*, 182–189.
- Boersma, P., & Weenink, D. (2023). Praat: doing phonetics by computer. <http://www.praat.org>
- Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52, 46–57.
- Bogetic, K. (2009). Prosodic cues in turn-taking. *Proceedings of the BAAL Annual Conference 2009, Newcastle University, UK*, 13–16.
- Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., & Blache, P. (2021). A multimodal model for predicting conversational feedbacks. *International Conference on Text, Speech, and Dialogue*, 537–549.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482–1493.
- Brescoll, V. L. (2011). Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly*, 56(4), 622–641.
- Bruce, G. (1977). *Swedish word accents in sentence perspective* [Doctoral dissertation]. Liber.
- Brusco, P., Vidal, J., Beňuš, Š., & Gravano, A. (2020). A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication*, 125, 24–40.
- Butler, J. (1993). *Bodies that matter*. Routledge: New York.
- Byrnes, H. (1986). Interactional style in German and American conversations. *Text-Interdisciplinary Journal for the Study of Discourse*, 6(2), 189–206.
- Cabarrão, V., Trancoso, I., Mata, A. I., Moniz, H., & Batista, F. (2016). Entrainment profiles: Comparison by gender, role, and feature set. *International Conference on Advances in Speech and Language Technologies for Iberian Languages*, 215–223.
- Cameron, D. (1995). Rethinking language and gender studies: Some issues for the 1990s. In S. Mills (Ed.), *Language and Gender: Interdisciplinary Perspectives* (pp. 31–44). Longman: London; New York.
- Cannon, B. C., Robinson, D. T., & Smith-Lovin, L. (2019). How do we “do gender”? Permeation as over-talking and talking over. *Socius*, 5. <https://doi.org/https://doi.org/10.1177/2378023119849347>
- Carbaugh, D. (2005). *Cultures in conversation*. Lawrence Erlbaum: Mahwah, NJ.
- Carbaugh, D., & Poutiainen, S. (2000). By way of introduction: An American and Finnish dialogue. In M. W. Lustig & J. Koester (Eds.), *Among US: Essays on Identity, Belonging and Intercultural Competence*. Longman: New York.

- Cerrato, L. (2002). Some characteristics of feedback expressions in Swedish. *Proceedings of FONETIK*, 44, 41–44.
- Cerrato, L. (2012). *Investigating communicative feedback phenomena across languages and modalities* [Doctoral dissertation, Royal Institute of Technology, Stockholm].
- Cerrato, L., & Skhiri, M. (2003). Analysis and measurement of head movements signalling feedback in face-to-face human dialogues. *Proceedings of the First Nordic Symposium on Multimodal Communication*, 9, 43–52.
- Chapman, A. J., & Foot, H. C. (1976). The social responsiveness of young children in humorous situations. In A. J. Chapman & H. C. Foot (Eds.), *Humor and laughter: Theory, research, and applications*. Wiley: New York.
- Clancy, P. M., Thompson, S. A., Suzuki, R., & Tao, H. (1996). The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26(3), 355–387.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81.
- Coates, J. (2004). *Women, men, and language: A sociolinguistic account of sex differences in language* (2nd ed.). New York: Pearson Longman.
- Coleman, R. O. (1976). A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research*, 19(1), 168–180.
- Couper-Kuhlen, E., & Selting, M. (1996). Towards an interactional perspective on prosody and a prosodic perspective on interaction. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 11–56). Cambridge University Press: Cambridge.
- Cutrone, P. (2005). A case study examining backchannels in conversations between Japanese–British dyads. *Multilingua*, (24), 237–274.
- Cutrone, P. (2011). Politeness and Face Theory: Implications for the backchannel style of Japanese L1/L2 speakers. In D. Giannoni & C. Ciarlo (Eds.), *Language studies working papers*. University of Reading: Reading.
- Cutrone, P. (2014). A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners. *Intercultural Pragmatics*, 11(1), 83–120. <https://doi.org/doi:10.1515/ip-2014-0004>
- Danner, S. G., Krivokapić, J., & Byrd, D. (2021). Co-speech movement in conversational turn-taking. *Frontiers in Communication*, 6, 779814.
- Daun, Å. (1996). *Swedish mentality*. University Park, PA: The Pennsylvania State University Press.
- de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390.
- de Ruiter, J.-P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3), 515–535.

- Deng, X. (1998). Overlapping in Australian and Chinese conversations. In J. Ingram (Ed.), *Selected Papers from the 1998 Conference of the Australian Linguistic Society*. <http://www.als.asn.au>
- Dideriksen, C., Christiansen, M. H., Dingemanse, M., Højmark-Bertelsen, M., Johansson, C., Tylén, K., & Fusaroli, R. (2023). Language-specific constraints on conversation: Evidence from Danish and Norwegian. *Cognitive Science*, 47(11), e13387.
- Dingemanse, M. (2024). Interjections at the heart of language. *Annual Review of Linguistics*, 10, 257–277.
- Dingemanse, M., Liesenfeld, A., & Woensdregt, M. (2022). Convergent cultural evolution of continuers (mhmm). *Joint Conference on Language Evolution (JCoLE)*, 160–167.
- Dittmann, A. T., & Llewellyn, L. G. (1968). Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9(1), 79–84.
- Dittmann, A. T., & Llewellyn, L. G. (1967). The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology*, 6(3), 341.
- Dixon, J. A., & Foster, D. H. (1998). Gender, social context, and backchannel responses. *The Journal of Social Psychology*, 138(1), 134–136.
- Dombrowski, E., & Niebuhr, O. (2005). Acoustic patterns and communicative functions of phrase-final rises in German: Activating and restricting contours. *Phonetica*, 62, 176–195. <https://doi.org/10.1159/000090097>
- Duncan, S. (1972). Some signals and rules for taking speaking turn in conversations. *Journal of Personality and Social Psychology*, 23, 283–292.
- Duncan, S. (1973). Toward a grammar for dyadic conversation. *Semiotica*, 9(1), 29–46.
- Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3, 161–180.
- Duncan, S., & Fiske, W. (1977). *Face-to-face Interaction. Research, Methods, and Theory*. Hillsdale, N.J.: Lawrence Erlbaum.
- Eckert, H., & Laver, J. (1994). *Menschen und ihre Stimmen. Aspekte der vokalen Kommunikation*. Beltz: Weinheim.
- Eckert, P., & McConnell-Ginet, S. (2013). *Language and Gender*. Cambridge: Cambridge University Press.
- Edelsky, C. (1993). Who's got the floor? In D. Tannen (Ed.), *Gender and conversational interaction*. Oxford: Oxford University Press.
- Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. *Proceedings of LREC*, 2992–2995.

- Edlund, J., & Beskow, J. (2007). Pushy versus meek-using avatars to influence turn-taking behavior. *Proceedings of Interspeech*. <https://doi.org/10.21437/Interspeech.2007-289>
- Edlund, J., & Beskow, J. (2009). Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. *Language and Speech*, 52(2-3), 351–367.
- Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica*, 62(2-4), 215–226.
- Engwall, O., Cumbal, R., & Majlesi, A. R. (2023). Socio-cultural perception of robot backchannels. *Frontiers in Robotics and AI*, 10, 988042.
- Erickson, F. (1986). Listening and speaking. *Languages and Linguistics: The Interdependence of Theory, Data, and Application*, 294–319.
- Ericsson, C., & Ericsson, A. M. (2001). Gender differences in vowel duration in read Swedish: Preliminary results. *Proceedings of FONETIK*, 49, 34–37.
- Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, 142, 27–41.
- Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 1, 22–30.
- Fant, L. (1989). Cultural mismatch in conversation: Spanish and Scandinavian communicative behaviour in negotiation settings. *HERMES-Journal of Language and Communication in Business*, (3), 247–265.
- Feindt, K., Rossi, M., Esfandiari-Baiat, G., Ekström, A. G., & Zellers, M. (2023). Cues to next-speaker projection in conversational Swedish: Evidence from reaction times. *Proceedings of Interspeech*, 1040–1044. <https://doi.org/10.21437/Interspeech.2023-778>
- Feindt, K., Rossi, M., & Zellers, M. (2021). The time course of pitch variation towards possible places of speaker transition in German and Swedish. *Proceedings of the 1st International Conference on Tone and Intonation (TAI)*, 112–116. <https://doi.org/10.21437/TAI.2021-23>
- Feindt, K., Rossi, M., & Zellers, M. (in preparation). Pre-final prosodic cues to turn-taking in German and Swedish conversation.
- Feke, M. (2003). Effects of native-language and sex on back-channel behavior. *Selected Proceedings of the First Workshop on Spanish Sociolinguistics*, 96–106.
- Felleggy, A. M. (1995). Patterns and functions of minimal response. *American Speech*, 70(2), 186–199.
- Ferrand, C. T., & Bloom, R. L. (1996). Gender differences in children's intonational patterns. *Journal of Voice*, 10(3), 284–291.
- Ferré, G., & Renaudier, S. (2017). Unimodal and bimodal backchannels in conversational English. *Proceedings of SEMDIAL 2017*, 27–37.
- Féry, C. (2017). *Intonation and prosodic structure*. Cambridge: Cambridge University Press.

- Ford, C., Fox, B., & Thompson, S. (1996). Practices in the construction of turns: The TCU revisited. *Pragmatics*, 6(3), 427–454.
- Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), 1–27. <https://doi.org/10.18637/jss.v008.i15>
- Fox, J., & Weisberg, S. (2019). *An R companion to Applied Regression* (3rd). SAGE Publishing: Thousand Oaks, CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html>
- Fox, J., Weisberg, S., Price, B., Friendly, M., & Hong, J. (2022). *Effects: Effect displays for linear, generalized linear, and other models* [R package version 4.2-2]. <https://CRAN.R-project.org/package=effects>
- Fraile Muñoz, R., & Godino Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14, 42–54.
- Fries, C. C. (1952). *The structure of English: An introduction to the construction of English sentences*. SAGE Publishing: Thousand Oaks, CA.
- Funk, R., & Simpson, A. P. (2023). The acoustic and perceptual correlates of gender in children's voices. *Journal of Speech, Language, and Hearing Research*, 66(9), 3346–3363.
- Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. *The 39th Annual Conference of the Cognitive Science Society (CogSci)*, 2055–2060.
- Gårding, E. (1989). Intonation in Swedish. *Lund University Department of Linguistics Working Papers*, 35, 63–88.
- Gardner, R. (1997). The listener and minimal responses in conversational interaction. *Prospect*, 12(2), 12–32.
- Gardner, R. (2001). *When listeners talk*. Amsterdam: John Benjamins Publishing Company.
- Gardner, R. (2004). Conversation analysis. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics* (pp. 262–284). Oxford, UK: Blackwell Publishing.
- Gardner, R., & Mushin, I. (2015). Expanded transition spaces: The case of Garrwa. *Frontiers in Psychology*, 6(251), 1–14.
- Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6(751), 1–12.
- Gelfer, M. P., & Young, S. R. (1997). Comparisons of intensity measures and their stability in male and female speakers. *Journal of Voice*, 11(2), 178–186.
- Goffman, E. (1976). Gender display. In E. Goffman (Ed.), *Gender advertisements: Communication and culture*. Palgrave: London.
- Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1), 3–19.

- Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9, 205–217. <https://api.semanticscholar.org/CorpusID:145355164>
- Gratton, C. (2016). Resisting the gender binary: The use of (ing) in the construction of non-binary transgender identities. *University of Pennsylvania Working Papers in Linguistics*, 22(2). <https://repository.upenn.edu/pwpl/vol22/iss2/7>
- Gravano, A., & Hirschberg, J. (2009a). Backchannel-inviting cues in task-oriented dialogue. *10th Annual Conference of the International Speech Communication Association*, 1019–1022.
- Gravano, A., & Hirschberg, J. (2009b). Turn-yielding cues in task-oriented dialogue. *Proceedings of SIGDIAL*, 253–261.
- Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3), 601–634.
- Graziano, M., Kendon, A., & Cristilli, C. (2011). Parallel gesturing in adult-child conversations. In G. Stam & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 89–101). John Benjamins.
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279.
- Guaitella, I., Santi, S., Lagrue, B., & Cavé, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. *Language and Speech*, 52(2-3), 207–222.
- Gussenhoven, C., & Bruce, G. (1999). Word prosody and intonation. In G. Bossong & B. Comrie (Eds.), *Empirical approaches to language typology* (pp. 233–272). Mouton de Gruyter.
- Habalet, B. (2019). The impact of the listener's gender on the frequency and diversity of the English audible backchannels. *Human Sciences Journal*, 59–68.
- Hadar, U., Steiner, T. J., Grant, E., & Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2), 35–46.
- Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). The timing of shifts of head postures during conversations. *Human Movement Science*, 3, 237–245.
- Hamilton, A. F. d. C., & Holler, J. (2023). Face2face: Advancing the science of social interaction. *Philosophical Transactions of the Royal Society B*, 378, 20210470. <https://doi.org/10.1098/rstb.2021.0470>
- Hannah, A., & Murachver, T. (2007). Gender preferential responses to speech. *Journal of Language and Social Psychology*, 26(3), 274–290. <https://doi.org/10.1177/0261927X06303457>
- Hardy, T. L., Rieger, J. M., Wells, K., & Boliek, C. A. (2020). Acoustic predictors of gender attribution, masculinity–femininity, and vocal naturalness ratings amongst transgender and cisgender speakers. *Journal of Voice*, 34(2), 300–e11.

- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall: New York.
- Hayashi, R. (1988). Simultaneous talk—from the perspective of floor management of English and Japanese speakers. *World Englishes*, 7(3), 269–288.
- Heinz, B. (2003). Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics*, 35(7), 1113–1142.
- Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *Journal of the Acoustical Society of America*, 130(1), 508–513.
- Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555–568.
- Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. *Proceedings of Interspeech*, 3054–3057.
- Heldner, M., Hjalmarsson, A., & Edlund, J. (2013). Backchannel relevance spaces. *Proceedings of Nordic Prosody*, 137–146.
- Heldner, M., Włodarczak, M., Beňuš, Š., & Gravano, A. (2019). Voice quality as a turn-taking cue. *Proceedings of Interspeech*, 4165–4169.
- Helweg-Larsen, M., Cunningham, S. J., Carrico, A., & Pergram, A. M. (2004). To nod or not to nod: An observational study of nonverbal communication and status in female and male college students. *Psychology of Women Quarterly*, 28, 358–361.
- Henton, C. (1995). Cross-language variation in the vowels of female and male speakers. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 1995)*, 4, 420–423.
- Henton, C. G., & Bladon, R. A. (1985). Breathiness in normal female speech: Inefficiency versus desirability. *Language & Communication*, 221–227.
- Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of Social Interaction: Studies in Conversation Analysis* (pp. 299–345). Cambridge University Press: Cambridge.
- Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, 39(2), 311–321.
- Hirst, D. J., & de Looze, C. (2021). Measuring speech. Fundamental frequency and pitch. In R.-A. Knight & J. Setter (Eds.), *Cambridge Handbook of Phonetics* (pp. 336–361, Vol. 1). Cambridge University Press: Cambridge.
- Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1), 23–35.
- Hjalmarsson, A., & Laskowski, K. (2011). Measuring final lengthening for speaker-change prediction. *Proceedings of Interspeech*, 2065–2068.

- Hogg, M. A. (1985). Masculine and feminine speech in dyads and groups: A study of speech style and gender salience. *Journal of Language and Social Psychology*, 4(2), 99–112.
- Holmes, J. (1995). *Women, men and language*. Longman: London.
- Hömke, P., Holler, J., & Levinson, S. C. (2018). Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, 50(1), 54–70. <https://doi.org/10.1080/08351813.2017.1262143>
- Hough, J., Tian, Y., de Ruiter, L., Betz, S., Kousidis, S., Schlangen, D., & Ginzburg, J. (2016). DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. *Proceedings of LREC*, 1784–1788.
- House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue. *Speech Communication*, 46, 268–283.
- House, D. (2004). Final rises and Swedish question intonation. *Proceedings of FONETIK*, 56–59.
- Howes, C., & Eshghi, A. (2021). Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *Journal of Logic, Language and Information*, 30(2), 331–362.
- Ilie, C. (2004). Insulting as (un)parliamentary practice in the British and Swedish parliaments. *Cross-cultural Perspectives on Parliamentary Discourse*, 26(2), 45–68.
- Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, 2, 255. <https://doi.org/10.3389/fpsyg.2011.00255>
- Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2), 101–144.
- Ishimoto, Y., Teraoka, T., & Enomoto, M. (2017). End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous Japanese speech. *Proceedings of Interspeech*. <https://doi.org/10.21437/Interspeech.2017-837>
- Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens “yeah” and “mm hm”. *Paper in Linguistics*, 17(2), 197–216. <https://doi.org/10.1080/08351818409389201>
- Jiahong Yuan, C. C., Mark Liberman. (2007). Towards an integrated understanding of speech overlaps in conversation. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2007)*, 1337–1340.
- Johnson, K. (2003). *Acoustic and auditory phonetics* (2nd Edition). Blackwell Publishing: Malden, MA.
- Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 638–646.

- Keitel, A., Prinz, W., Friederici, A. D., Von Hofsten, C., & Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *Journal of Experimental Child Psychology*, 116(2), 264–277.
- Kendall, T., Pharaoh, N., Stuart-Smith, J., & Vaughn, C. (2023). Advancements of phonetics in the 21st Century: Theoretical issues in sociophonetics. *Journal of Phonetics*, 98, 101226.
- Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26(1), 22–63.
- Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics*, 23(3), 247–279.
- Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B*, 378(1875), 20210473.
- Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326–347. <https://doi.org/10.1016/j.csl.2017.01.005>
- Kjellmer, G. (2009). Where do we backchannel? On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics*, 14, 81–112.
- Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2), 820–857.
- Knight, D. (2011). *Multimodality and active listenership: A corpus study*. Continuum Publishing: London.
- Knudsen, B., Creemers, A., & Meyer, A. S. (2020). Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, 11, 593671.
- Kohler, K. J., Peters, B., & Scheffers, M. (2017). *The Kiel Corpus of Spoken German—read and spontaneous speech. New edition, revised and enlarged*. Institut für Phonetik und digitale Sprachverarbeitung, Christian-Albrechts-Universität.
- Kohler, K. J. (1983). Prosodic boundary signals in German. *Phonetica*, 40, 89–134.
- Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: Duration, F0 movement, and F0 level. *Language and Speech*, 29, 115–139.
- Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41, 295–321.
- Koreman, J. (2006). Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *The Journal of the Acoustical Society of America*, 119(1), 582–596. <https://doi.org/10.1121/1.2133436>

- Köser, S. (2014). Phrasen-finale Phonationsänderungen und ihre Rolle beim turn taking. *Prosodie und Phonetik in der Interaktion*, 20–45.
- Kousidis, S., Pfeiffer, T., & Schlangen, D. (2013). Mint. tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. *Proceedings of Interspeech*, 2649–2653. <https://doi.org/10.21437/Interspeech.2013-609>
- Kraaz, M., & Bernaisch, T. (2022). Backchannels and the pragmatics of South Asian Englishes. *World Englishes*, 41(2), 224–243.
- Krause-Ono, M. (2004). *Change in backchanneling behaviour: The influence from L2 to L1 on the use of backchannel cues* [Doctoral dissertation, Muroran Institute of Technology].
- Krepsz, V., Horváth, V., Hátori, Á., Gyarmathy, D., & Dér, C. I. (2022). Backchannel responses in Hungarian conversations: A corpus-based study on the effect of the partner's age and gender. *Linguistica Silesiana*, 113–140.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/JSS.V082.I13>
- Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2, 205–254. <https://api.semanticscholar.org/CorpusID:4688836>
- Ladefoged, P., & Johnson, K. (2006). *A Course in Phonetics*. Thomson Wadsworth: Boston, MA.
- Lakoff, R. (1973). Language and woman's place. *Language in Society*, 2(1), 45–80.
- Lakoff, R. (1975). *Language and woman's place*. Harper & Row: New York.
- Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press: Cambridge.
- Leaper, C., & Ayres, M. M. (2007). A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology Review*, 11(4), 328–363. <https://doi.org/10.1177/1088868307302221>
- Leaper, C., & Robnett, R. D. (2011). Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. *Psychology of Women Quarterly*, 35(1), 129–142. <https://doi.org/10.1177/0361684310392728>
- Lehtonen, J., & Sajavaara, K. (1985). The silent Finn. In D. Tannen & M. Saville-Troike (Eds.), *Perspectives on Silence* (pp. 193–201). Ablex: Norwood, New Jersey.
- Lenth, R. V. (2022). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.8.2]. <https://CRAN.R-project.org/package=emmeans>
- Levinson, S. C. (1986). *Pragmatics*. Cambridge University Press: Cambridge.

- Levinson, S. C. (2013). Action formation and ascription. In J. Sidnell & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 103–130). Wiley-Blackwell: Oxford. <https://doi.org/10.1002/9781118325001.CH6>
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731.
- Levitan, R., Gravano, A., & Hirschberg, J. B. (2011). Entrainment in speech preceding backchannels. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 113–117.
- Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of Interspeech*, 3081–3084. <https://doi.org/10.21437/Interspeech.2011-771>
- Levon, E. (2015). Integrating intersectionality in language, gender, and sexuality research. *Language and Linguistics Compass*, 9(7), 295–308. <https://doi.org/10.1111/lnc3.12147>
- Levow, G.-A., & Duncan, S. (2012). Contrasting cues to verbal and non-verbal backchannels in multi-lingual dyadic rapport. *Proceedings of Interspeech*, 835–838.
- Liesenfeld, A., & Dingemanse, M. (2022). Bottom-up discovery of structure and variation in response tokens (backchannels) across diverse languages. *Proceedings of Interspeech*, 1126–1130. <https://doi.org/10.21437/Interspeech.2022-11288>
- Local, J., & Walker, G. (2012). How phonetic features project more talk. *Journal of the International Phonetic Association*, 42, 255–280.
- Local, J., Kelly, J., & Wells, W. H. G. (1986). Towards a phonology for conversation: Turn-taking in Tyneside English. *Journal of Linguistics*, 22, 411–437.
- Local, J., & Walker, G. (2005). Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica*, 62(2-4), 120–130.
- Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. *MLA '14: Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 5–12.
- Lüdecke, D. (2022). *Sjplot: Data visualization for statistics in social science* [R package version 2.8.12]. <https://CRAN.R-project.org/package=sjPlot>
- Ludusan, B., & Dupoux, E. (2015). A multilingual study on intensity as a cue for marking prosodic boundaries. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2015)*.
- Ludusan, B., & Schuppler, B. (2022). To laugh or not to laugh? The use of laughter to mark discourse structure. *Proceedings of SIGDIAL*, 76–82.
- Ludusan, B., & Wagner, P. (2022a). ha-HA-hha? Intensity and voice quality characteristics of laughter. *Proceedings of Speech Prosody*, 560–564.

- Ludusan, B., & Wagner, P. (2022b). Laughter entrainment in dyadic interactions: Temporal distribution and form. *Speech Communication*, 136, 42–52.
- Ludusan, B., Wesemann, M., & Wagner, P. (2020). A distributional analysis of laughter across turns and utterances. *Proceedings of the Workshop on Laughter and Other Non-verbal Vocalisations*, 28–31.
- Malisz, Z., Włodarczyk, M., Buschmeier, H., Skubisz, J., Kopp, S., & Wagner, P. (2016). The ALICO Corpus: Analysing the Active Listener. *Language Resources and Evaluation*, 50(2), 411–442. <https://doi.org/10.1007/s10579-015-9342-3>
- Maltz, D. N., & Borker, R. A. (1982). A cultural approach to male-female miscommunication. In J. J. Gumperz (Ed.), *Language and Social Identity* (pp. 195–216). Cambridge University Press: Cambridge.
- Marche, T. A., & Peterson, C. (1993). On the gender differential use of listener responsiveness. *Sex Roles*, 29, 795–816.
- Markel, N. (1975). Coverbal behavior associated with conversation turn. In A. Kendon, R. M. Harris, & M. R. Key (Eds.), *Organization of behavior in face-to-face interaction*. Mouton: The Hague.
- Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *Journal of the Acoustical Society of America*, 126(5), 2619–2634. <https://doi.org/10.1121/1.3224706>
- Maynard, S. K. (1990). Conversation management in contrast: Listener response in Japanese and American English. *Journal of Pragmatics*, 14(3), 397–412.
- McCarthy, M. (2003). Talking back: Small interactional response tokens in everyday conversation. *Research on Language and Social interaction*, 36(1), 33–63.
- McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7), 855–878.
- Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317(5834), 82–82.
- Mendoza, E., Valencia, N., Muñoz, J., & Trujillo, H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (ltas). *Journal of Voice*, 10(1), 59–66.
- Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: A comparison of English and German. *Journal of the Acoustical Society of America*, 131(3), 2249–2260.
- Menz, F., & Al-Roubaie, A. (2008). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *Journal of the Acoustical Society of America*, 19(5), 645–666.
- Meteor, M. W., Taylor, A. A., MacIntyre, R., & Iyer, R. (1995). *Disfluency annotation stylebook for the Switch-Board Corpus* (tech. rep.). University of Pennsylvania.

- Mizutani, N. (1982). The listener's response in Japanese conversation. *Sociolinguistics*, 13(1), 33–38.
- Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2), 194–225.
- Mulac, A., Bradac, J. J., & Palomares, N. A. (2003). A general process model of the Gender linked language effect: Antecedents for and consequences of language used by men and women. *Annual Conference of the International Communication Association*.
- Mulac, A., Erlandson, K. T., Farrar, W. J., Hallett, J. S., Molloy, J. L., & Prescott, M. E. (1998). "Uh-huh. What's that all about?": Differing interpretations of conversational backchannels and questions as sources of miscommunication across gender boundaries. *Communication Research*, 25(6), 641–668. <https://doi.org/10.1177/009365098025006004>
- Mullany, L. (2010). Gender and interpersonal pragmatics. In M. Locher & S. L. Graham (Eds.), *The Handbook of Interpersonal Pragmatics* (pp. 225–250). Mouton de Gruyter: The Hague.
- Müller, F. E. (1996). Affiliating and disaffiliating with continuers: Prosodic aspects of reciprocity. In M. Couper-Kuhlen & E. Selting (Eds.), *Prosody in Conversation: Interactional studies* (pp. 131–176). Cambridge University Press: Cambridge.
- Murphy, B. (2011). Gender Identities and Discourse. In G. Andersen & K. Aijmer (Eds.), *Pragmatics of Society* (pp. 53–77). Mouton de Gruyter: The Hague.
- Murton, O., Hillman, R., & Mehta, D. (2020). Cepstral peak prominence values for clinical voice evaluation. *American Journal of Speech-Language Pathology*, 29(3), 1596–1607.
- Mushin, I., & Gardner, R. (2009). Silence is talk: Conversational silence in Australian Aboriginal talk-in-interaction. *Journal of Pragmatics*, 41(10), 2033–2052.
- Myers, G. E., & Myers, M. T. (1976). *The dynamics of human communication: A laboratory approach*. McGraw-Hill Book Company: London.
- Nakane, I. (2012). Silence. In C. B. Paulston, S. F. Kiesling, & E. S. Rangel (Eds.), *The Handbook of Intercultural Discourse and Communication*. Wiley-Blackwell: Hoboken, NJ.
- Niebuhr, O., Görs, K., & Graupe, E. (2013). Speech reduction, intensity, and F0 shape are cues to turn-taking. In *Proceedings of SIGDIAL* (pp. 261–269).
- Norwine, A., & Murphy, O. J. (1938). Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, 17(2), 281–291.
- Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11(8), 1017.
- Ochs, E. (1992). Indexing gender. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon*. Cambridge University Press: Cambridge.

- Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1), 139–152.
- Ohala, J., & Solé, M.-J. (2010). Turbulence and phonology. In S. Fuchs, M. Toda, & M. Zygis (Eds.), *Turbulent sounds: An interdisciplinary guide* (pp. 37–102). De Gruyter Mouton: The Hague. <https://doi.org/10.1515/9783110226584.37>
- O’Keeffe, A., & Adolphs, S. (2008). Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. In K. P. Schneider & Barron (Eds.), *Variational Pragmatics*. John Benjamins: Amsterdam.
- Oreström, B. (1983). *Turn-taking in English conversation*. Gleerup: Lund.
- Ozkan, D., & Morency, L.-P. (2013). Prediction of visual backchannels in the absence of visual context using mutual influence. *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings 13*, 189–202.
- Paggio, P., & Navarretta, C. (2011). Head movements, facial expressions and feedback in Danish first encounters interactions: A culture-specific analysis. In C. Stephanidis (Ed.), *Universal Access in Human-Computer Interaction. Users Diversity - 6th International Conference* (pp. 583–590, Vol. 6766). Springer: Berlin. https://doi.org/10.1007/978-3-642-21663-3%5C_63
- Palomares, N. A. (2004). Gender schematicity, gender identity salience, and gender-linked language use. *Human Communication Research*, 30(4), 556–588.
- Pammi, S., & Schröder, M. (2009). A corpus based analysis of backchannel vocalizations. *Proceedings of Interdisciplinary Workshop on Laughter and other Interactional Vocalisations in Speech*, 2–3.
- Pearce, J. (2019). Identity, socialization and environment in transgender speakers: Sociophonetic variation in creak and /s/. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2019)*, 3290–3294.
- Pépiot, E., & Arnold, A. (2021). Cross-gender differences in English/French bilingual speakers: A multiparametric study. *Perceptual and Motor Skills*, 128(1), 153–177. <https://doi.org/10.1177/0031512520973514>
- Pépiot, E. (2014). Male and female speech: A study of mean F0, F0 range, phonation type and speech rate in Parisian French and American English speakers. *Proceedings of Speech Prosody*, 305–309.
- Peters, B. (2006). *Form und Funktion prosodischer Grenzen im Gespräch* [Doctoral dissertation]. Christian-Albrechts-Universität zu Kiel.
- Pilkington, J. (1998). ‘Don’t try and make out that I’m nice!’ The different strategies women and men use when gossiping. In J. Coates (Ed.), *Language and Gender: A reader* (pp. 254–269). Blackwell: Oxford.
- Plug, I., Stommel, W., Lucassen, P. L., olde Hartman, T. C., van Dulmen, S., & Das, E. (2021). Do women and men use language differently in spoken face-to-face interaction? A scoping review. *Review of Communication Research*, 9, 43–79.

- Plug, L., Lennon, R., & Smith, R. (2022a). Measured and perceived speech tempo: Comparing canonical and surface articulation rates. *Journal of Phonetics*, 95, 101193. <https://doi.org/https://doi.org/10.1016/j.wocn.2022.101193>
- Plug, L., Lennon, R., & Smith, R. (2022b). Schwa deletion and perceived tempo in English. *Proceedings of Speech Prosody 2022*, 470–474. <https://doi.org/10.21437/SpeechProsody.2022-96>
- Poppe, R., Truong, K. P., & Heylen, D. (2011). Backchannels: Quantity, type and timing matters. *Proceedings of the Intelligent Virtual Agents International Conference*, 228–239.
- Poppe, R., Truong, K. P., & Heylen, D. (2013). Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous Agents and Multi-agent Systems*, 27(2), 235–253.
- Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. *Proceedings of the Intelligent Virtual Agents International Conference*, 146–158.
- Provine, R. R. (1993). Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology*, 95(4), 291–298.
- R Core Team. (2022). *R A language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, 44(e12911). <https://doi.org/10.1111/cogs.12911>
- Reichel, U. D., Beňuš, S., & Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication*, 100, 46–57.
- Reid, S. A., Keerie, N., & Palomares, N. A. (2003). Language, gender salience, and social influence. *Journal of Language and Social Psychology*, 22(2), 210–233.
- Rietveld, A. C. M., & Gussenhoven, C. (1987). Perceived speech rate and intonation. *Journal of Phonetics*, 15, 273–285.
- Rivera, A. G., & Ward, N. G. (2007). *Three Prosodic Features That Cue Back-Channel Feedback in Northern Mexican Spanish* (tech. rep. No. 106). Departmental Technical Reports (CS).
- Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, 6, 509.
- Rohrer, P. L., Tütüncübası, U., Vilà-Giménez, I., Florit-Pons, J., Esteve-Gibert, N., Ren-Mitchell, A., Shattuck-Hufnagel, S., & Prieto, P. (2023). The MultiModal MultiDimensional (M3D) labeling system.
- Rohrer, P. L., Vila-Giménez, I., Florit-Pons, J., Gurrado, G., Esteve Gibert, N., Ren, A., Shattuck-Hufnagel, S., & Prieto, P. (2020). The MultiModal MultiDimen-

- sional (M3D) labelling scheme for the annotation of audiovisual corpora. *Proceedings of GESPIN*.
- Rossi, M., Feindt, K., & Zellers, M. (2022a). Perception of F0 movements towards potential turn boundaries in German and Swedish conversation: Background and methods for an eye-tracking study. *Proceedings of FONETIK*.
- Rossi, M. (2022). Gender influence on phonetic turn-taking cues at potential transition locations in German. In R. Orrico & L. Schettino (Eds.), *The position of the speaker in interaction: Attitudes, intentions, and emotions in verbal communication* (pp. 127–147, Vol. 9). Officinaventuno: Milan. <https://doi.org/10.17469/O2109AISV000007>
- Rossi, M., Feindt, K., & Zellers, M. (2022b). Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish. *Proceedings of Speech Prosody 2022*, 185–189. <https://doi.org/10.21437/SpeechProsody.2022-38>
- Rossi, M., Feindt, K., & Zellers, M. (2022c). *Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish* [Poster presented at Speech Prosody 2022, Lisbon, Portugal]. <https://doi.org/10.13140/RG.2.2.23889.51046>
- Rossi, M., Schröer, M., Ludusan, B., & Zellers, M. (2023). A multimodal account of listener feedback in face-to-face interactions. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2023)*, 4121–4125.
- Runge, T. E., Frey, D., Gollwitzer, P. M., Helmreich, R. L., & Spence, J. T. (1981). Masculine (instrumental) and feminine (expressive) traits: A comparison between students in the United States and West Germany. *Journal of Cross-Cultural Psychology*, 12(2), 142–162.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, 50(4), 696–735.
- Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of ‘uh huh’ and other things that come between sentences. *Analyzing Discourse: Text and talk*, 71, 71–93.
- Schegloff, E. A. (1987). Analyzing single episodes of interaction: An exercise in conversation analysis. *Social Psychology Quarterly*, 101–114.
- Schegloff, E. A. (2020). Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In *Roots of Human Sociality* (pp. 70–96). Routledge: Abingdon, Oxfordshire, UK.
- Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 1999)*, 607–610.
- Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press: Cambridge.

- Schnur, T. T., Costa, A., & Caramazza, A. (2006). Planning at the phonological level during sentence production. *Journal of Psycholinguistic Research*, 35, 189–213.
- Scollon, R., & Scollon, S. B. K. (1981). *Narrative, literacy, and face in interethnic communication*. Ablex: Norwood, NJ.
- Scott, S. K., Lavan, N., Chen, S., & McGettigan, C. (2014). The social life of laughter. *Trends in Cognitive Sciences*, 18(12), 618–620.
- Selting, M. (1992). Prosody in conversational questions. *Journal of Pragmatics*, 17, 315–345.
- Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructive units and turns in conversation. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 6(3), 371–388.
- Shelley, L., & Gonzalez, F. (2013). Backchanneling: Function of backchanneling and L1 effects on back channeling in L2. *Linguistic Portfolios*, 2(1), 9.
- Simpson, A. P. (1998). *Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung* (Vol. 33). Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK).
- Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2), 621–640.
- Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech and Language*, 67, 101178.
- Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65, 50–66.
- Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. <https://doi.org/10.3115/1609067.1609150>
- Speer, S. A., & Stokoe, E. (2011). An introduction to conversation and gender. In S. A. Speer & E. Stokoe (Eds.), *Conversation and gender* (pp. 1–27). Cambridge University Press: Cambridge.
- Steensig, J. (2011). Turn-taking in conversation. In *Pragmatics of Society* (pp. 499–532). De Gruyter Mouton: Berlin.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.
- Stokoe, E., & Smithson, J. (2001). Making gender relevant: Conversation analysis and gender categories in interaction. *Discourse and Society*, 12, 217–244.
- Strömbergsson, S., Hjalmarsson, A., Edlund, J., & House, D. (2013). Timing responses to questions in dialogue. *Proceedings of Interspeech*, 2584–2588.

- Stubbe, M. (1998). Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3), 257–289.
- Stubbe, M. (2013). Active listening in conversation: Gender and the use of verbal feedback. In S. Yamazaki & R. Sigley (Eds.), *Linguistic Insights - Studies in Language and Communication: Approaching language variation through corpora: A festschrift in honor of Toshio Saito* (pp. 365–416, Vol. 167). Peter Lang: Bern.
- Susana Finger, L., Aparecida Cielo, C., & Schwarz, K. (2009). Acoustic vocal measures in women without voice complaints and with normal larynxes. *Brazilian Journal of Otorhinolaryngology*, 75(3), 432–440. [https://doi.org/10.1016/S1808-8694\(15\)30663-7](https://doi.org/10.1016/S1808-8694(15)30663-7)
- Swann, J. (2009). Doing gender against the odds: A sociolinguistic analysis of educational discourse. In E. Eppler & P. Pichler (Eds.), *Gender and Spoken Interaction* (pp. 18–41). Palgrave Macmillan.
- Tannen, D. (1990). *You just don't understand: Women and men in conversation*. Morrow: New York.
- Tannen, D. (1984). *Conversational style: Analysing talk among friends*. Ablex: Norwood, New Jersey.
- Tannen, D. (1994). *Gender and discourse*. Oxford University Press: Oxford.
- Tannen, D. (1998). The relativity of linguistic strategies: Rethinking power and solidarity in gender and dominance. In D. Cameron (Ed.), *The Feminist critique of language: A reader*. Routledge: London.
- Tannen, D. (2012). Turn-taking and intercultural discourse and communication. In C. B. Paulston, S. F. Kiesling, & E. S. Rangel (Eds.), *The Handbook of Intercultural Discourse and Communication*. Wiley-Blackwell: Hoboken, NJ.
- Tao, H., & Thompson, S. A. (1991). English backchannels in Mandarin conversations: A case study of superstratum pragmatic 'interference'. *Journal of Pragmatics*, 16(3), 209–223.
- Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis – Jitter, shimmer and HNR parameters. *Procedia Technology*, 9, 1112–1122. <https://doi.org/10.1016/j.protcy.2013.12.124>
- Tiede, M., Mooshammer, C., & Goldstein, L. (2019). Noggin nodding: Head movement correlates with increased effort in accelerating speech production tasks. *Frontiers in Psychology*, 10.
- Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4), 1699–1707. <https://doi.org/10.1121/1.397959>
- Tixier, M. (1996). Cultural adjustments required by expatriate managers working in the Nordic countries. *International Journal of Manpower*, 17(6/7), 19–42.
- Tottie, G. (1991). Conversational style in British and American English: The case of backchannels. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman: New York.

- Traunmüller, H., & Eriksson, A. (1993). *The frequency range of the voice fundamental in the speech of male and female adults*. <https://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Asu%3Adiva-10290>
- Trouvain, J., & Truong, K. (2012). Comparing non-verbal vocalisations in conversational speech corpora. *Proceedings of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 36–39.
- Trouvain, J., & Truong, K. P. (2013). Exploring sequences of speech and laughter activity using visualisations of conversations. *Workshop on Affective Social Speech Signals 2013*.
- Truong, K. P., Poppe, R., de Kok, I., & Heylen, D. (2011). A multimodal analysis of vocal and visual backchannels in spontaneous dialogues. *Proceedings of Interspeech*, 23–25.
- Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35(4), 445–472.
- Turnbull, W. (2003). *Language in action: Psychological models of conversation*. Psychology Press: London.
- Uchida, A. (1992). When “difference” is “dominance”: A critique of the “anti-power-based” cultural approach to sex differences. *Language in Society*, 21(4), 547–568.
- Ueno, J. (2004). Gender differences in Japanese conversation. *Intercultural Communication Studies*, 13, 85–100.
- Ulijn, J. M., & Xiangling, L. (1995). Is interrupting impolite? Some temporal aspects of turn-taking in Chinese-Western and other intercultural business encounters. *Text & Talk*, 15(4), 589–628.
- Valentine, T. (1986). Language and power: Cross-sex communicative strategies in Hindi and Indian English. *Economic and Political Weekly*, 75–87.
- Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, 38(3), 253–265.
- Van Borsel, J., Janssens, J., & De Bodt, M. (2009). Breathiness as a feminine voice characteristic: A perceptual approach. *Journal of Voice*, 23(3), 291–294.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth) [ISBN 0-387-95457-0]. Springer: New York. <https://www.stats.ox.ac.uk/pub/MASS4/>
- Vettin, J., & Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28, 93–115.
- Vinciarelli, A., Chatziioannou, P., & Esposito, A. (2015). When the words are not everything: The use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, 2.
- Waara, E., & Shaw, P. (2006). Male and female witnesses’ speech in Swedish criminal trials. *Journal of Language and Communication in Business*, (36), 129–156. <https://doi.org/10.7146/HJLCB.V19I36.25842>

- Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232. <https://doi.org/10.1016/j.specom.2013.09.008>
- Ward, N., & Tsukahara, W. (2000). Prosodic features which cue backchannel responses in English and Japanese. *Journal of Pragmatics*, 32(8), 1177–1207.
- Ward, N. (1999). Low-pitch regions as dialogue signals? Evidence from dialog-act and lexical correlates in natural conversation. *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*.
- Ward, N. G. (2019). *Prosodic patterns in English conversation*. Cambridge University Press: Cambridge.
- Wehrle, S., Vogeley, K., & Grice, M. (2023). Backchannels in conversations between autistic adults are less frequent and less diverse prosodically and lexically. *Language and Cognition*, 1–26.
- Wei, Z., & Lee, S. H. (2019). Prosodic turn-yielding cues in English spontaneous speech by Chinese EFL learners. *Eoneohak*, (85), 227–242.
- Weirich, M., & Simpson, A. P. (2018). Gender identity is indexed and perceived in speech. *PLOS ONE*, 13(12).
- Weirich, M., & Simpson, A. P. (2023). Changes in parent’s phonetic parameters during the first year of a child’s life. Comparative study of German and Swedish. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2023)*, 3517–3521.
- Weirich, M., Simpson, A. P., Öjbro, J., & Ericsson Nordgren, C. (2019). The phonetics of gender in Swedish and German. *Proceedings of FONETIK*, 49–53.
- Weirich, M., & Zahner-Ritter, K. (2023). Indexing femininity in vowel acoustics: A comparison between speakers in eastern and western parts of Germany. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2023)*, 3562–3566.
- Weise, A., & Levitan, R. (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, 297–302.
- Weise, A., Levitan, S. I., Hirschberg, J., & Levitan, R. (2019). Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, 115, 78–87.
- West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender and Society*, 1(2), 125–151.
- White, S. (1989). Backchannels across cultures: A study of Americans and Japanese. *Language in Society*, 18(1), 59–76.
- Whiteside, S. P. (1995). Temporal-based speaker sex differences in read speech: A sociophonetic approach. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 1995)*, 516–519.

- Whiteside, S. P. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association*, 26(1), 23–40.
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wieland, M. (1991). Turn-taking structure as a source of misunderstanding in French-American cross-cultural conversation. *Pragmatics and Language Learning*, 2(1), 101–118.
- Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, 1(1), 7–18. <https://doi.org/10.1093/jole/lzv003>
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. *Proceedings of LREC*.
- Włodarczak, M., & Heldner, M. (2022). Contribution of voice quality to prediction of turn-taking events. *Proceedings of Speech Prosody*, 485–489.
- Włodarczak, M., Heldner, M., Bruggeman, A., & Wagner, P. (2023). Voice quality dynamics of turn-taking events in Swedish and German. *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2023)*, 3477–3481.
- Wong, D., & Peters, P. (2007). A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification. *International Journal of Corpus Linguistics*, 12(4), 479–510.
- Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC press.
- Yngve, V. H. (1970). On getting a word in edgewise. *CLS-70*, 567–578.
- Zellers, M. (2013). Pitch and lengthening as cues to turn transition in Swedish. *Proceedings of Interspeech*, 248–252.
- Zellers, M. (2014). Duration and pitch in perception of turn transition by Swedish and English listeners. *Proceedings of FONETIK*, 41–46.
- Zellers, M. (2017). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. *Language and Speech*, 60(3), 454–478.
- Zellers, M. (2021). An overview of forms, functions, and configurations of backchannels in Ruruuli/Lunyala. *Journal of Pragmatics*, 175, 38–52.
- Zellers, M., Gorisch, J., House, D., & Peters, B. (2019). Timing properties of hand gestures and their lexical counterparts at turn transition places. *Proceedings of FONETIK*, 119–124.
- Zellers, M., & Schweitzer, A. (2017). An investigation of pitch matching across adjacent turns in a corpus of spontaneous German. *Proceedings of Interspeech*. <https://doi.org/10.21437/Interspeech.2017-811>
- Zimmermann, D. H., & West, C. (1975). Sex roles, interruptions and silences in conversation. In B. Thorne & N. Henley (Eds.), *Language and sex: Difference and dominance* (pp. 105–129). Newbury House: Rowley.

Part VI

Appendix

Appendix 1

Full model results from the analysis reported in Chapter 6. No post-hoc tests were conducted in this analysis.

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmer-ModLmerTest'] Formula: *lmer*(*F0* ~ *distance* * *gender* * *listener* + (1 | *speaker*))
Data: change

REML criterion at convergence: 8732.3

Scaled residuals:

Min 1Q Median 3Q Max

-3.0853 -0.5693 -0.0326 0.4627 4.5108

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 10.11 3.180

Residual 20.74 4.554

Number of obs: 3537, groups: speaker, 12

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	11.0322	1.6712	9.2173	6.601	8.88e-05 ***
F0_pointsP2	-2.2745	0.6473	1465.0749	-3.514	0.000455 ***
F0_pointsP1	-2.5282	0.6473	1465.0749	-3.906	9.82e-05 ***
genderM	-4.9311	2.3359	8.8010	-2.111	0.064651 .
listenerother	1.1346	2.8237	8.3554	0.402	0.697904
F0_pointsP2:genderM	2.4714	0.8134	1465.0749	3.038	0.002420 **
F0_pointsP1:genderM	4.4082	0.8134	1465.0749	5.420	6.98e-08 ***
F0_pointsP2:listenerother	1.9198	0.8134	1465.0749	2.360	0.018391 *
F0_pointsP1:listenerother	4.5724	0.8134	1465.0749	5.621	2.26e-08 ***
genderM:listenerother	-2.2018	4.0219	8.5934	-0.547	0.597998
F0_pointsP2:genderM:listenerother	-1.5617	1.2932	1465.0749	-1.208	0.227398
F0_pointsP1:genderM:listenerother	-7.1274	1.2932	1465.0749	-5.511	4.20e-08 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

```
(Intr) F0_pP2 F0_pP1 gendrM lstnrt F0_pP2:M F0_pP1:M F0_pP2: F0_pP1:
gendrM: F0_P2:M:
F0_pointsP2 -0.194
F0_pointsP1 -0.194 0.500
genderM -0.715 0.139 0.139
listenerthr -0.592 0.115 0.115 0.423
F0_pntsP2:M 0.154 -0.796 -0.398 -0.174 -0.091
F0_pntsP1:M 0.154 -0.398 -0.796 -0.174 -0.091 0.500
F0_pntsP2:l 0.154 -0.796 -0.398 -0.110 -0.144 0.633 0.317
F0_pntsP1:l 0.154 -0.398 -0.796 -0.110 -0.144 0.317 0.633 0.500
gendrM:lstnr 0.416 -0.080 -0.080 -0.581 -0.702 0.101 0.101 0.101 0.101
F0_pntP2:M: -0.097 0.501 0.250 0.110 0.091 -0.629 -0.314 -0.629 -0.314
-0.161
F0_pntP1:M: -0.097 0.250 0.501 0.110 0.091 -0.314 -0.629 -0.314 -0.629
-0.161 0.500
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmer-ModLmerTest'] Formula: *lmer(F0 ~ distance * gender * listener + (1 | speaker))*
Data: keep

REML criterion at convergence: 16776.5

Scaled residuals:

Min 1Q Median 3Q Max

-3.8938 -0.3637 0.0221 0.5013 3.7820

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 6.35 2.520

Residual 18.30 4.277

Number of obs: 2916, groups: speaker, 12

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	9.1696	1.2859	7.3667	7.131	0.000148 ***
F0_pointsP2	1.3409	0.3408	2896.9797	3.934	8.55e-05 ***
F0_pointsP1	0.8831	0.3408	2896.9797	2.591	0.009613 **
genderM	-3.6594	1.8152	7.3140	-2.016	0.081880 .
listenerother	2.9748	2.2156	7.2154	1.343	0.220092
F0_pointsP2:genderM	-0.8469	0.4724	2896.9797	-1.793	0.073108 .
F0_pointsP1:genderM	-1.4471	0.4724	2896.9797	-3.063	0.002209 **

```

F0_pointsP2:listenerother -0.3076 0.5112 2896.9797 -0.602 0.547383
F0_pointsP1:listenerother -2.3349 0.5112 2896.9797 -4.567 5.15e-06 ***
genderM:listenerother -2.5707 3.6319 7.3266 -0.708 0.500957
F0_pointsP2:genderM:listenerother -3.6392 0.9743 2896.9797 -3.735
0.000191 ***
F0_pointsP1:genderM:listenerother 1.3532 0.9743 2896.9797 1.389 0.164965
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Correlation of Fixed Effects:

```

(Intr) F0_pP2 F0_pP1 gendrM lstnrt F0_pP2:M F0_pP1:M F0_pP2: F0_pP1:
gendrM: F0_P2:M:
F0_pointsP2 -0.133
F0_pointsP1 -0.133 0.500
genderM -0.708 0.094 0.094
listenerthr -0.580 0.077 0.077 0.411
F0_pntsP2:M 0.096 -0.721 -0.361 -0.130 -0.055
F0_pntsP1:M 0.096 -0.361 -0.721 -0.130 -0.055 0.500
F0_pntsP2:l 0.088 -0.667 -0.333 -0.063 -0.115 0.481 0.240
F0_pntsP1:l 0.088 -0.333 -0.667 -0.063 -0.115 0.240 0.481 0.500
gendrM:lstnr 0.354 -0.047 -0.047 -0.500 -0.610 0.065 0.065 0.070 0.070
F0_pntP2:M: -0.046 0.350 0.175 0.063 0.061 -0.485 -0.242 -0.525 -0.262
-0.134
F0_pntP1:M: -0.046 0.175 0.350 0.063 0.061 -0.242 -0.485 -0.262 -0.525
-0.134 0.500

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [*'lmer-ModLmerTest'*] Formula: *lmer(intensity ~ distance * gender * listener + (1 | speaker))* Data: change

REML criterion at convergence: 24436.7

Scaled residuals:

```

Min 1Q Median 3Q Max
-3.1364 -0.6345 0.1177 0.6927 2.5126

```

Random effects:

```

Groups Name Variance Std.Dev.
speaker (Intercept) 3.874 1.968
Residual 58.473 7.647
Number of obs: 3537, groups: speaker, 12

```

Fixed effects:

```

Estimate Std. Error df t value Pr(>|t|)

```

```
(Intercept) 10.1865 1.0628 9.7709 9.585 2.78e-06 ***
IntensityI2 0.4834 0.5630 3517.0583 0.859 0.390567
IntensityI1 -11.9678 0.5630 3517.0583 -21.259 < 2e-16 ***
genderM 3.7963 1.5193 10.2027 2.499 0.031085 *
listenerother -0.8750 1.8088 9.1111 -0.484 0.639985
IntensityI2:genderM -3.2728 0.8580 3517.0583 -3.815 0.000139 ***
IntensityI1:genderM -6.6951 0.8580 3517.0583 -7.804 7.88e-15 ***
IntensityI2:listenerother 0.1572 0.8503 3517.0583 0.185 0.853326
IntensityI1:listenerother -6.7422 0.8503 3517.0583 -7.929 2.94e-15 ***
genderM:listenerother 1.0235 2.5763 9.3741 0.397 0.700063
IntensityI2:genderM:listenerother -1.8130 1.2741 3517.0583 -1.423
0.154837
IntensityI1:genderM:listenerother 8.8455 1.2741 3517.0583 6.942 4.57e-12
***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correlation of Fixed Effects:

```
(Intr) IntnI2 IntnI1 gendrM lstnrt InI2:M InI1:M IntI2: IntI1: gndrM:
II2:M:
IntensityI2 -0.265
IntensityI1 -0.265 0.500
genderM -0.700 0.185 0.185
listenerthr -0.588 0.156 0.156 0.411
IntnstyI2:M 0.174 -0.656 -0.328 -0.282 -0.102
IntnstyI1:M 0.174 -0.328 -0.656 -0.282 -0.102 0.500
IntnstyI2:l 0.175 -0.662 -0.331 -0.123 -0.235 0.434 0.217
IntnstyI1:l 0.175 -0.331 -0.662 -0.123 -0.235 0.217 0.434 0.500
gndrM:lstnr 0.413 -0.109 -0.109 -0.590 -0.702 0.167 0.167 0.165 0.165
IntnstI2:M: -0.117 0.442 0.221 0.190 0.157 -0.673 -0.337 -0.667 -0.334
-0.247
IntnstI1:M: -0.117 0.221 0.442 0.190 0.157 -0.337 -0.673 -0.334 -0.667
-0.247 0.500
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [*lmer-ModLmerTest*] **Formula:** *lmer(intensity ~ distance * gender * listener + (1 | speaker))* **Data:** keep

REML criterion at convergence: 20789.6

Scaled residuals:

Min 1Q Median 3Q Max

-2.4891 -0.6946 0.0181 0.7225 3.0017

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 16.95 4.117

Residual 72.93 8.540

Number of obs: 2916, groups: speaker, 12

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	10.3673	2.1214	7.5234	4.887	0.00145 **
IntensityI2	0.3900	0.6805	2896.9487	0.573	0.56661
IntensityI1	-7.5689	0.6805	2896.9487	-11.123	< 2e-16 ***
genderM	4.8643	2.9921	7.4452	1.626	0.14547
listenerother	-2.4356	3.6464	7.2980	-0.668	0.52473
IntensityI2:genderM	-2.6697	0.9432	2896.9487	-2.831	0.00468 **
IntensityI1:genderM	-7.3551	0.9432	2896.9487	-7.798	8.69e-15 ***
IntensityI2:listenerother	-2.7057	1.0207	2896.9487	-2.651	0.00807 **
IntensityI1:listenerother	-6.9076	1.0207	2896.9487	-6.767	1.58e-11 ***
genderM:listenerother	-0.4589	5.9880	7.4643	-0.077	0.94093
IntensityI2:genderM:listenerother	3.8497	1.9452	2896.9487	1.979	0.04790
					*
IntensityI1:genderM:listenerother	8.0387	1.9452	2896.9487	4.133	3.69e-05 ***

					--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	IntnI2	IntnI1	gndrM	lstnrt	InI2:M	InI1:M	IntI2:	IntI1:	gndrM:
IntensityI2	-0.160									
IntensityI1	-0.160	0.500								
genderM	-0.709	0.114	0.114							
listenerthr	-0.582	0.093	0.093	0.412						
IntnstyI2:M	0.116	-0.721	-0.361	-0.158	-0.067					
IntnstyI1:M	0.116	-0.361	-0.721	-0.158	-0.067	0.500				
IntnstyI2:l	0.107	-0.667	-0.333	-0.076	-0.140	0.481	0.240			
IntnstyI1:l	0.107	-0.333	-0.667	-0.076	-0.140	0.240	0.481	0.500		
gndrM:lstnr	0.354	-0.057	-0.057	-0.500	-0.609	0.079	0.079	0.085	0.085	
IntnstI2:M:	-0.056	0.350	0.175	0.076	0.073	-0.485	-0.242	-0.525	-0.262	-0.162
IntnstI1:M:	-0.056	0.175	0.350	0.076	0.073	-0.242	-0.485	-0.262	-0.525	

-0.162 0.500

**Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmer-
ModLmerTest'] Formula: *lmer(duration ~ distance * gender * listener + (1 |
speaker))* Data: change**

REML criterion at convergence: -11645.4

Scaled residuals:

Min 1Q Median 3Q Max

-3.2905 -0.5589 -0.0347 0.3733 6.9745

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 0.0001434 0.01197

Residual 0.0020964 0.04579

Number of obs: 3537, groups: speaker, 12

Fixed effects:

Estimate Std. Error df t value Pr(>|t|)

(Intercept) 7.724e-02 6.451e-03 9.551e+00 11.974 4.57e-07 ***

seg_durationS2 1.488e-02 3.371e-03 3.517e+03 4.414 1.05e-05 ***

seg_durationS1 3.707e-02 3.371e-03 3.517e+03 10.998 < 2e-16 ***

genderM 1.271e-02 9.219e-03 9.962e+00 1.378 0.19830

listenerother 6.645e-03 1.098e-02 8.923e+00 0.605 0.56027

seg_durationS2:genderM 1.480e-02 5.137e-03 3.517e+03 2.881 0.00399 **

seg_durationS1:genderM 4.744e-02 5.137e-03 3.517e+03 9.235 < 2e-16 ***

seg_durationS2:listenerother 4.184e-03 5.091e-03 3.517e+03 0.822

0.41120

seg_durationS1:listenerother 3.020e-02 5.091e-03 3.517e+03 5.932

3.29e-09 ***

genderM:listenerother -1.568e-02 1.564e-02 9.173e+00 -1.003 0.34177

seg_durationS2:genderM:listenerother -1.979e-02 7.629e-03 3.517e+03

-2.594 0.00953 **

seg_durationS1:genderM:listenerother -5.805e-02 7.629e-03 3.517e+03

-7.609 3.53e-14 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr) sg_dS2 sg_dS1 gendrM lstnrt sg_S2:M sg_S1:M sg_S2: sg_S1:

gndrM: s_S2:M:

seg_durtnS2 -0.261

```

seg_durtnS1 -0.261 0.500
genderM -0.700 0.183 0.183
listenerthr -0.587 0.153 0.153 0.411
sg_drtnS2:M 0.171 -0.656 -0.328 -0.279 -0.101
sg_drtnS1:M 0.171 -0.328 -0.656 -0.279 -0.101 0.500
sg_drtnS2:l 0.173 -0.662 -0.331 -0.121 -0.232 0.434 0.217
sg_drtnS1:l 0.173 -0.331 -0.662 -0.121 -0.232 0.217 0.434 0.500
gndrM:lstrn 0.412 -0.108 -0.108 -0.589 -0.702 0.164 0.164 0.163 0.163
sg_drtS2:M: -0.115 0.442 0.221 0.188 0.155 -0.673 -0.337 -0.667 -0.334
-0.244
sg_drtS1:M: -0.115 0.221 0.442 0.188 0.155 -0.337 -0.673 -0.334 -0.667
-0.244 0.500

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [*'lmer-ModLmerTest'*] Formula: *lmer(duration ~ distance * gender * listener + (1 | speaker))* Data: keep

REML criterion at convergence: -10692.9

Scaled residuals:

Min 1Q Median 3Q Max

-2.4578 -0.5574 -0.0930 0.4043 7.1164

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 0.0003852 0.01963

Residual 0.0014274 0.03778

Number of obs: 2916, groups: speaker, 12

Fixed effects:

Estimate Std. Error df t value Pr(>|t|)

(Intercept) 7.394e-02 1.007e-02 7.546e+00 7.341 0.000108 ***

seg_durationS2 7.429e-03 3.010e-03 2.897e+03 2.468 0.013661 *

seg_durationS1 1.286e-02 3.010e-03 2.897e+03 4.271 2.01e-05 ***

genderM 8.735e-03 1.421e-02 7.477e+00 0.615 0.557050

listenerother 1.755e-02 1.733e-02 7.349e+00 1.013 0.343466

seg_durationS2:genderM 2.835e-03 4.173e-03 2.897e+03 0.679 0.496981

seg_durationS1:genderM 1.135e-02 4.173e-03 2.897e+03 2.721 0.006549 **

seg_durationS2:listenerother 4.000e-03 4.516e-03 2.897e+03 0.886
0.375803

seg_durationS1:listenerother 1.715e-02 4.516e-03 2.897e+03 3.798
0.000149 ***

genderM:listenerother -2.451e-02 2.844e-02 7.494e+00 -0.862 0.415566

```
seg_durationS2:genderM:listenerother 4.308e-03 8.605e-03 2.897e+03
0.501 0.616661
seg_durationS1:genderM:listenerother -1.422e-02 8.605e-03 2.897e+03
-1.652 0.098642 .
```

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

```
(Intr) sg_dS2 sg_dS1 gendrM lstnrt sg_S2:M sg_S1:M sg_S2: sg_S1:
gendrM: s_S2:M:
seg_durtnS2 -0.149
seg_durtnS1 -0.149 0.500
genderM -0.709 0.106 0.106
listenerthr -0.581 0.087 0.087 0.412
sg_drtnS2:M 0.108 -0.721 -0.361 -0.147 -0.063
sg_drtnS1:M 0.108 -0.361 -0.721 -0.147 -0.063 0.500
sg_drtnS2:l 0.100 -0.667 -0.333 -0.071 -0.130 0.481 0.240
sg_drtnS1:l 0.100 -0.333 -0.667 -0.071 -0.130 0.240 0.481 0.500
gendrM:lstnr 0.354 -0.053 -0.053 -0.500 -0.609 0.073 0.073 0.079 0.079
sg_drtS2:M: -0.052 0.350 0.175 0.071 0.068 -0.485 -0.242 -0.525 -0.262
-0.151
sg_drtS1:M: -0.052 0.175 0.350 0.071 0.068 -0.242 -0.485 -0.262 -0.525
-0.151 0.500
```

Appendix 2

Full model results from the analysis reported in Chapter 7.

Model and post-hoc test head onset time:

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`lmer-ModLmerTest`] **Formula:** $head_onset \sim item + (1 \mid speaker)$ **Data:** DUEL

REML criterion at convergence: 241.7

Scaled residuals:

Min 1Q Median 3Q Max

-6.4558 -0.2972 0.0895 0.3815 5.6143

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 0.01883 0.1372

Residual 0.15647 0.3956

Number of obs: 222, groups: speaker, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	-0.12647	0.07077	96.41346	-1.787	0.0771
itemLex+Head	0.06211	0.07487	215.08229	0.829	0.4078
itemNonLex+Head	0.01362	0.09026	218.66124	0.151	0.8802

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr) itmL+H

itemLex+Hed -0.769

itemNnLx+Hd -0.615 0.599

`emmeans::emmeans(model.onset, specs = "item")`

item	emmean	SE	df	lower.CL	upper.CL
------	--------	----	----	----------	----------

Laugh+Head	-0.1265	0.0714	89.7	-0.268	0.0153
------------	---------	--------	------	--------	--------

Lex+Head	-0.0644	0.0504	26.0	-0.168	0.0393
----------	---------	--------	------	--------	--------

NonLex+Head -0.1128 0.0740 79.2 -0.260 0.0344

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

pairs(emm55)

contrast estimate SE df t.ratio p.value

(Laugh+Head) - (Lex+Head) -0.0621 0.0758 215 -0.819 0.6914

(Laugh+Head) - (NonLex+Head) -0.0136 0.0911 219 -0.150 0.9878

(Lex+Head) - (NonLex+Head) 0.0485 0.0764 209 0.635 0.8013

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Model and post-hoc test head offset time:

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmer-ModLmerTest'] Formula: *head_offset ~ item + (1 | speaker)* Data: DUEL

REML criterion at convergence: 366.8

Scaled residuals:

Min 1Q Median 3Q Max

-4.3171 -0.2785 -0.1404 0.3674 4.5737

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 0.009289 0.09638

Residual 0.288530 0.53715

Number of obs: 222, groups: speaker, 24

Fixed effects:

Estimate Std. Error df t value Pr(>|t|)

(Intercept) -0.35436 0.08733 113.08631 -4.058 9.14e-05 ***

itemLex+Head 0.46910 0.09796 193.96997 4.789 3.32e-06 ***

itemNonLex+Head 0.51101 0.11905 207.71367 4.293 2.71e-05 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr) itmL+H

itemLex+Hed -0.835

itemNnLx+Hd -0.676 0.611

```
emmeans::emmeans(model.offset, specs = "item")
```

```
item emmean SE df lower.CL upper.CL
Laugh+Head -0.354 0.0884 114 -0.52958 -0.179
Lex+Head 0.115 0.0566 22 -0.00262 0.232
NonLex+Head 0.157 0.0910 71 -0.02477 0.338
```

```
Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95
```

```
pairs(emm56)
contrast estimate SE df t.ratio p.value
(Laugh+Head) - (Lex+Head) -0.4691 0.0999 194 -4.696 <.0001
(Laugh+Head) - (NonLex+Head) -0.5110 0.1211 208 -4.219 0.0001
(Lex+Head) - (NonLex+Head) -0.0419 0.1004 167 -0.418 0.9084
```

```
Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 3 estimates
```

Model and post-hoc test intensity:

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmer-ModLmerTest'`] **Formula:** *intmax_n feedback* + (1|*speaker*) + (1|*verbal_bc*) **Data:** DUEL

```
REML criterion at convergence: 3333
```

```
Scaled residuals:
Min 1Q Median 3Q Max
-3.9190 -0.5318 0.0819 0.5796 2.5021
```

```
Random effects:
Groups Name Variance Std.Dev.
verbal_bc (Intercept) 4.721 2.173
speaker (Intercept) 12.959 3.600
Residual 34.167 5.845
Number of obs: 512, groups: verbal_bc, 124; speaker, 24
```

```
Fixed effects:
              Estimate Std. Error df t value Pr(>|t|)
(Intercept)  3.1139  0.9433  38.6514  3.301  0.00208 **
feedbacknon-lexical -2.4693  1.3374  38.3470 -1.846  0.07256 .
feedbacklaughter  3.8514  1.1342  44.7404  3.396  0.00144 **
```

```
--
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr) fdbck-

fdbcknn-lxc -0.272

fedbcklghtr -0.327 0.244

emmeans::emmeans(m1, specs = "feedback")

feedback emmean SE df lower.CL upper.CL

lexical 3.114 0.952 40.7 1.19 5.04

non-lexical 0.645 1.439 51.3 -2.24 3.53

laughter 6.965 1.243 53.3 4.47 9.46

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

pairs(emm1)

contrast estimate SE df t.ratio p.value

lexical - (non-lexical) 2.47 1.37 41.2 1.798 0.1828

lexical - laughter -3.85 1.17 48.0 -3.290 0.0053

(non-lexical) - laughter -6.32 1.58 34.8 -4.011 0.0009

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Model and post-hoc CPP:

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmer-ModLmerTest'] Formula: $CPP \sim feedback + (1 | speaker) + (1|verbal_bc)$ Data: DUEL

REML criterion at convergence: 2405.8

Scaled residuals:

Min 1Q Median 3Q Max

-3.2278 -0.5684 -0.0010 0.6010 3.3596

Random effects:

Groups Name Variance Std.Dev.

verbal_bc (Intercept) 2.924 1.710

speaker (Intercept) 1.333 1.155

Residual 5.059 2.249

Number of obs: 512, groups: verbal_bc, 124; speaker, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	10.4201	0.3742	73.1996	27.849	< 2e-16 ***
feedbacknon-lexical	1.3758	0.7277	86.6174	1.891	0.062 .
feedbacklaughter	-3.2907	0.5954	107.5619	-5.527	2.29e-07 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	fdbck-
fdbcknn-lxc	-0.307	
fedbcklgtr	-0.379	0.200

	emmeans::emmeans(m2, specs = "feedback")
feedback	emmean SE df lower.CL upper.CL
lexical	10.42 0.375 68.2 9.67 11.17
non-lexical	11.80 0.713 86.6 10.38 13.21
laughter	7.13 0.574 104.5 5.99 8.27

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

	contrast	estimate	SE	df	t.ratio	p.value
lexical - (non-lexical)	-1.38	0.732	79.9	-1.879	0.1514	
lexical - laughter	3.29	0.599	99.6	5.492	<.0001	
(non-lexical) - laughter	4.67	0.848	78.9	5.500	<.0001	

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 3 estimates

Model and post-hoc test for presence of the head on intensity:

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmer-ModLmerTest'] Formula: $intmax_n \sim item + (1 | speaker)$ Data: DUEL

REML criterion at convergence: 3321.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.6804	-0.5126	0.0785	0.5740	2.4024

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 13.06 3.614

Residual 36.07 6.006

Number of obs: 512, groups: speaker, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	5.0055	0.9552	42.8452	5.240	4.63e-06 ***
itemLaugh+Head	0.9126	1.1765	488.0073	0.776	0.438285
itemLex	-3.4756	0.7851	500.3467	-4.427	1.18e-05 ***
itemLex+Head	-1.0537	0.8721	504.5868	-1.208	0.227528
itemNonLex	-4.7251	1.2003	492.8239	-3.937	9.45e-05 ***
itemNonLex+Head	-3.7889	1.1373	497.5158	-3.331	0.000928 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr)	itmLg+H	itemLx	itmLx+H	itmNnL
itemLagh+Hd	-0.291			
itemLex	-0.478	0.358		
itemLex+Hed	-0.451	0.339	0.575	
itemNonLex	-0.298	0.216	0.372	0.353
itemNnLx+Hd	-0.308	0.252	0.364	0.389

```
emmeans::emmeans(m3, specs = "item")
item emmean SE df lower.CL upper.CL
Laugh 5.01 0.956 44.5 3.079 6.93
Laugh+Head 5.92 1.282 130.6 3.381 8.46
Lex 1.53 0.902 35.9 -0.299 3.36
Lex+Head 3.95 0.962 45.1 2.015 5.89
NonLex 0.28 1.294 132.4 -2.278 2.84
NonLex+Head 1.22 1.241 112.6 -1.242 3.68
```

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

```
pairs(emm3)
contrast estimate SE df t.ratio p.value
Laugh - (Laugh+Head) -0.913 1.177 489 -0.775 0.9716
Laugh - Lex 3.476 0.787 501 4.418 0.0002
Laugh - (Lex+Head) 1.054 0.875 505 1.204 0.8347
Laugh - NonLex 4.725 1.202 493 3.933 0.0013
```

Laugh - (NonLex+Head) 3.789 1.139 498 3.326 0.0121
 (Laugh+Head) - Lex 4.388 1.159 493 3.787 0.0024
 (Laugh+Head) - (Lex+Head) 1.966 1.205 496 1.631 0.5781
 (Laugh+Head) - NonLex 5.638 1.490 494 3.784 0.0024
 (Laugh+Head) - (NonLex+Head) 4.702 1.417 494 3.318 0.0124
 Lex - (Lex+Head) -2.422 0.769 498 -3.149 0.0214
 Lex - NonLex 1.250 1.166 495 1.072 0.8924
 Lex - (NonLex+Head) 0.313 1.125 501 0.279 0.9998
 (Lex+Head) - NonLex 3.671 1.211 496 3.033 0.0305
 (Lex+Head) - (NonLex+Head) 2.735 1.135 499 2.410 0.1545
 NonLex - (NonLex+Head) -0.936 1.421 492 -0.659 0.9862

Degrees-of-freedom method: kenward-roger

P value adjustment: tukey method for comparing a family of 6 estimates

Model and post-hoc test for presence of the head on CPP:

Linear mixed model fit by REML. t-tests use Satterthwaite's method [`'lmer-ModLmerTest'`] **Formula:** $CPP \sim item + (1 | speaker)$ **Data:** DUEL

REML criterion at convergence: 2448

Scaled residuals:

Min 1Q Median 3Q Max

-2.95529 -0.61295 -0.01387 0.69082 2.99956

Random effects:

Groups Name Variance Std.Dev.

speaker (Intercept) 1.345 1.16

Residual 6.556 2.56

Number of obs: 512, groups: speaker, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	6.2808	0.3491	62.9146	17.992	<2e-16 ***
itemLaugh+Head	0.4375	0.5007	492.7253	0.874	0.383
itemLex	4.8864	0.3325	505.1519	14.698	<2e-16 ***
itemLex+Head	5.2734	0.3681	505.0838	14.326	<2e-16 ***
itemNonLex	5.6422	0.5100	498.5972	11.064	<2e-16 ***
itemNonLex+Head	6.9104	0.4822	503.5742	14.330	<2e-16 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

```
(Intr) itmLg+H itemLx itmLx+H itmNnL
itemLagh+Hd -0.338
itemLex -0.553 0.358
itemLex+Hed -0.522 0.340 0.572
itemNonLex -0.346 0.218 0.372 0.353
itemNnLx+Hd -0.359 0.252 0.366 0.388 0.261
```

```
emmeans::emmeans(m4, specs = "item")
item emmean SE df lower.CL upper.CL
Laugh 6.28 0.350 57.8 5.58 6.98
Laugh+Head 6.72 0.505 201.5 5.72 7.71
Lex 11.17 0.323 43.8 10.52 11.82
Lex+Head 11.55 0.352 57.9 10.85 12.26
NonLex 11.92 0.510 201.2 10.92 12.93
NonLex+Head 13.19 0.485 165.1 12.23 14.15
```

```
Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95
```

```
pairs(emm4)
contrast estimate SE df t.ratio p.value
Laugh - (Laugh+Head) -0.438 0.501 491 -0.873 0.9528
Laugh - Lex -4.886 0.333 505 -14.652 <.0001
Laugh - (Lex+Head) -5.273 0.370 505 -14.257 <.0001
Laugh - NonLex -5.642 0.511 498 -11.046 <.0001
Laugh - (NonLex+Head) -6.910 0.484 503 -14.291 <.0001
(Laugh+Head) - Lex -4.449 0.493 497 -9.031 <.0001
(Laugh+Head) - (Lex+Head) -4.836 0.512 501 -9.445 <.0001
(Laugh+Head) - NonLex -5.205 0.633 499 -8.220 <.0001
(Laugh+Head) - (NonLex+Head) -6.473 0.602 499 -10.748 <.0001
Lex - (Lex+Head) -0.387 0.326 503 -1.185 0.8438
Lex - NonLex -0.756 0.495 500 -1.526 0.6476
Lex - (NonLex+Head) -2.024 0.477 506 -4.245 0.0004
(Lex+Head) - NonLex -0.369 0.514 501 -0.717 0.9798
(Lex+Head) - (NonLex+Head) -1.637 0.481 504 -3.401 0.0094
NonLex - (NonLex+Head) -1.268 0.604 497 -2.098 0.2899
```

```
Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 6 estimates
```

Appendix 3

Full model results from the analysis reported in Chapter 9.

Multinomial model testing feedback timing: Formula: *multinom(transitiontype ~ completeness + modality + listener + (listener : speaker) + (listener : modality), data = Backchannels, family = multinomial)*

z test of coefficients:

	Estimate	Std. Error	z	value	Pr(> z)
Nogap:(Intercept)	0.088861	0.555784	0.1599	0.8729726	
Nogap:completenessIncomplete	0.742787	0.389635	1.9064	0.0566027	.
Nogap:modalityMulti	-0.395261	0.638232	-0.6193	0.5357147	
Nogap:modalityVerb	-0.358657	0.661409	-0.5423	0.5876378	
Nogap:listenerMale	-0.328831	0.775370	-0.4241	0.6714959	
Nogap:listenerFemale:interlocutorSame	0.131473	0.418544	0.3141	0.7534290	
Nogap:listenerMale:interlocutorSame	-0.434981	0.426096	-1.0209	0.3073242	
Nogap:modalityMulti:listenerMale	0.093345	0.872377	0.1070	0.9147886	
Nogap:modalityVerb:listenerMale	0.017591	0.907107	0.0194	0.9845276	
Over:(Intercept)	-0.157264	0.552130	-0.2848	0.7757738	
Over:completenessIncomplete	3.336594	0.373847	8.9250	< 2.2e-16	***
Over:modalityMulti	-2.493189	0.669162	-3.7258	0.0001947	***
Over:modalityVerb	-1.669654	0.729189	-2.2897	0.0220364	*
Over:listenerMale	0.072524	0.735056	0.0987	0.9214047	
Over:listenerFemale:interlocutorSame	0.432236	0.521273	0.8292	0.4069946	
Over:listenerMale:inter.Same	-1.085140	0.456662	-2.3762	0.0174900	*
Over:modalityMulti:listenerMale	1.618891	0.882671	1.8341	0.0666418	.
Over:modalityVerb:listenerMale	0.820469	0.948209	0.8653	0.3868839	

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmer-ModLmerTest'] **Formula:** *lmer(f0mean ~ listener + speaker + modality + completeness + transition_type + syllable_number + (1 | speaker_id) + listener : speaker : modality + listener : modality : transition_type*

Data: Backchannels

REML criterion at convergence: 5487.3

Scaled residuals:

Min 1Q Median 3Q Max
-5.2466 -0.5427 -0.0770 0.5268 4.0304

Random effects:

Groups Name Variance Std.Dev.

speaker_id (Intercept) 3.280 1.811

Residual 7.267 2.696

Number of obs: 1131, groups: speaker_id, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	-5.78262	1.64950	42.62688	-3.506	0.001086 **
listenerMale	3.95731	1.28541	34.82672	3.079	0.004040 **
speakerOther	0.67192	1.13896	22.66432	0.590	0.561065
modalityGest	-0.36720	0.72123	1099.56785	-0.509	0.610758
modalityMulti	1.48208	0.62168	1089.77842	2.384	0.017296 *
completenessIncomplete	0.49512	0.23263	1096.57289	2.128	0.033528 *
transition_typeGap	-0.87902	0.52472	1091.85231	-1.675	0.094180 .
transition_typeOver	-0.89510	0.56022	1095.11601	-1.598	0.110384
syllable_number	0.16347	0.05881	1091.15151	2.780	0.005537 **
listenerFemale:speakerSame:modalityVerb	4.54413	1.86097	37.37585	2.442	0.019463 *
listenerMale:speakerSame:modalityVerb	0.67850	0.61406	1101.81613	1.105	0.269427
listenerFemale:speakerOther:modalityVerb	3.06108	0.94076	1101.34692	3.254	0.001173 **
listenerFemale:speakerSame:modalityGest	6.64617	1.98260	47.81772	3.352	0.001574 **
listenerMale:speakerSame:modalityGest	-0.08483	0.66883	1104.88809	-0.127	0.899097
listenerFemale:speakerOther:modalityGest	3.73403	1.04412	1099.36110	3.576	0.000364 ***
listenerFemale:speakerSame:modalityMulti	3.83879	1.67361	24.83133	2.294	0.030547 *
listenerFemale:modalityVerb:transition_typeGap	0.98655	0.67428	1094.41720	1.463	0.143722
listenerMale:modalityVerb:transition_typeGap	1.55073	0.75233	1098.77839	2.061	0.039516 *
listenerFemale:modalityGest:transition_typeGap	0.08860	0.92762	1088.86421	0.096	0.923926
listenerMale:modalityGest:transition_typeGap	2.42632	0.92670	1100.40294	2.618	0.008960 **
listenerFemale:modalityMulti:transition_typeGap	0.46503	0.78591	1095.86805	0.592	0.551065

```

0.592 0.554165
listenerFemale:modalityVerb:transition_typeOver 0.33732 0.68956 1094.33761
0.489 0.624810
listenerMale:modalityVerb:transition_typeOver 1.33619 0.72747 1090.69054
1.837 0.066516 .
listenerFemale:modalityGest:transition_typeOver -0.98922 0.81518
1093.50348 -1.213 0.225204
listenerMale:modalityGest:transition_typeOver 3.36227 0.84298 1103.70160
3.989 7.09e-05 ***
listenerFemale:modalityMulti:transition_typeOver -0.36779 0.87652
1094.15822 -0.420 0.674860
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [*lmer-ModLmerTest*] Formula: *intmean* ~ *listener* + *speaker* + *modality* + *completeness* + *transition_type* + (1 | *speaker_id*) + *listener* : *speaker* : *transition_type* + *listener* : *modality*

Data: Backchannels

REML criterion at convergence: 6318.8

Scaled residuals:

Min 1Q Median 3Q Max

-2.9343 -0.6633 -0.0124 0.6593 3.1410

Random effects:

Groups Name Variance Std.Dev.

speaker_id (Intercept) 0.2053 0.4531

Residual 15.7659 3.9706

Number of obs: 1131, groups: speaker_id, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	-2.21015	1.07419	152.44791	-2.058	0.04134 *
listenerMale	-0.83867	0.71820	68.78173	-1.168	0.24694
speakerOther	0.42239	0.57608	43.94382	0.733	0.46732
modalityGest	0.04473	0.49156	237.62867	0.091	0.92757
modalityMulti	-1.27651	0.45795	778.79493	-2.787	0.00544 **
completenessIncomplete	0.37360	0.32763	905.10177	1.140	0.25446
transition_typeGap	0.64815	0.63601	1045.73383	1.019	0.30840
transition_typeOver	1.65953	0.57056	1090.76560	2.909	0.00370 **
listenerMale:modalityGest	1.43267	0.65241	300.81628	2.196	0.02886 *

```
listenerMale:modalityMulti 1.61581 0.61679 784.90906 2.620 0.00897 **
listenerFemale:speakerSame:transition_typeNogap 1.16504 1.16029 117.70133
1.004 0.31739
listenerMale:speakerSame:transition_typeNogap 1.40503 0.91372 1074.19156
1.538 0.12442
listenerFemale:speakerOther:transition_typeNogap 1.86662 0.83237
1088.01614 2.243 0.02513 *
listenerFemale:speakerSame:transition_typeGap 1.78706 1.15294 109.45595
1.550 0.12403
listenerMale:speakerSame:transition_typeGap 0.81627 0.79497 746.93179
1.027 0.30485
listenerFemale:speakerOther:transition_typeGap 1.48010 0.83164 1027.45250
1.780 0.07541 .
listenerFemale:speakerSame:transition_typeOver 1.36559 0.86914 42.98962
1.571 0.12347
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [*lmer-ModLmerTest*] Formula: *durmean ~ listener + modality + completeness + transition_type + (1 | speaker_id) + listener : completeness : transition_type*
Data: Backchannels

REML criterion at convergence: -2274.7

Scaled residuals:

Min 1Q Median 3Q Max

-2.0703 -0.7054 -0.1701 0.5130 4.9085

Random effects:

Groups Name Variance Std.Dev.

speaker_id (Intercept) 0.005168 0.07189

Residual 0.006788 0.08239

Number of obs: 1131, groups: speaker_id, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	-9.868e-02	3.326e-02	1.133e+02	-2.967	0.00367 **
listenerMale	-2.788e-02	3.135e-02	2.433e+01	-0.889	0.38246
modalityGest	-7.171e-03	7.425e-03	1.112e+03	-0.966	0.33435
modalityMulti	-3.093e-03	6.678e-03	1.108e+03	-0.463	0.64332
completenessIncomplete	6.925e-03	1.306e-02	1.104e+03	0.530	0.59608
transition_typeGap	7.700e-02	2.883e-02	1.103e+03	2.671	0.00768 **

```

transition_typeOver 6.559e-03 1.792e-02 1.100e+03 0.366 0.71447
listenerFemale:completenessComplete:transition_typeNogap -3.269e-04
2.613e-02 1.104e+03 -0.013 0.99002
listenerMale:completenessComplete:transition_typeNogap -2.748e-03
2.344e-02 1.103e+03 -0.117 0.90671
listenerFemale:completenessIncomplete:transition_typeNogap 4.477e-02
2.399e-02 1.101e+03 1.866 0.06228 .
listenerFemale:completenessComplete:transition_typeGap -6.821e-02
2.969e-02 1.102e+03 -2.298 0.02176 *
listenerMale:completenessComplete:transition_typeGap -5.784e-02 2.647e-02
1.098e+03 -2.185 0.02907 *
listenerFemale:completenessIncomplete:transition_typeGap -8.936e-02
2.808e-02 1.101e+03 -3.183 0.00150 **
listenerFemale:completenessComplete:transition_typeOver 1.141e-02
2.077e-02 1.101e+03 0.549 0.58309
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmer-ModLmerTest'] Formula: *cppmean* ~ *listener* + *speaker* + *modality* + *completeness* + *listener : speaker* + (1 | *speaker_id*)

Data: Backchannels

REML criterion at convergence: 6359.9

Scaled residuals:

Min 1Q Median 3Q Max

-2.9525 -0.6533 0.1058 0.7483 2.6107

Random effects:

Groups Name Variance Std.Dev.

speaker_id (Intercept) 0.08079 0.2842

Residual 16.16170 4.0202

Number of obs: 1131, groups: speaker_id, 24

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	0.31401	0.31374	16.88351	1.001	0.3310
listenerMale	-0.84372	0.40201	16.26003	-2.099	0.0518 .
speakerOther	-0.52557	0.40612	13.76347	-1.294	0.2169
modalityGest	-0.02529	0.31877	251.97018	-0.079	0.9368
modalityMulti	-0.42950	0.30206	673.87295	-1.422	0.1555
completenessIncomplete	0.52659	0.25618	824.37123	2.056	0.0401 *

listenerMale:speakerOther 0.86936 0.55369 14.74307 1.570 0.1376

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr) lstnrM spkrOt mdltyG mdltyM cmpltI

listenerMal -0.600

speakerOthr -0.576 0.501

modalityGst -0.135 -0.135 -0.212

modalityMlt -0.273 -0.085 -0.109 0.417

cmpltInssInc -0.349 0.074 0.074 -0.291 -0.011

lstnrMl:sp0 0.443 -0.728 -0.726 0.124 0.029 -0.066

Appendix 4

Full model results from the analysis reported in Chapter 10.

Multinomial model testing feedback timing in German: Formula:

multinon(transitiontype ~ completeness + modality + dyad, data = DUEL, family = multinomial)

z test of coefficients:

	Estimate	Std. Error	z	value	Pr(> z)
Nogap:(Intercept)	-0.27131	0.46568	-0.5826	0.560158	
Nogap:completenessIncomplete	0.75428	0.38887	1.9397	0.052418	.
Nogap:modalitymultimodal	-0.26997	0.45405	-0.5946	0.552124	
Nogap:modalityverbal	-0.29446	0.43648	-0.6746	0.499912	
Nogap:dyadFemale_Same	0.41023	0.39400	1.0412	0.297794	
Nogap:dyadMale_Other	0.28255	0.41393	0.6826	0.494856	
Nogap:dyadMale_Same	-0.43987	0.42393	-1.0376	0.299451	
Over:(Intercept)	0.32515	0.43907	0.7405	0.458967	
Over:completenessIncomplete	3.19845	0.36351	8.7989	< 2.2e-16	***
Over:modalitymultimodal	-1.44227	0.45265	-3.1863	0.001441	**
Over:modalityverbal	-1.25069	0.42744	-2.9260	0.003433	**
Over:dyadFemale_Same	-0.33539	0.43372	-0.7733	0.439357	
Over:dyadMale_Other	-0.81424	0.45739	-1.7802	0.075042	.
Over:dyadMale_Same	-1.08387	0.45334	-2.3908	0.016809	*

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multinomial model testing feedback timing in Swedish: Formula:

multinon(transitiontype ~ completeness + modality + dyad, data = Spontal, family = multinomial)

z test of coefficients:

	Estimate	Std. Error	z	value	Pr(> z)
Nogap:(Intercept)	-0.43982	0.46561	-0.9446	0.344856	
Nogap:completenessIncomplete	0.77838	0.42579	1.8281	0.067538	.
Nogap:modalitymultimodal	-0.10604	0.53354	-0.1987	0.842463	
Nogap:modalityverbal	-0.51082	0.46452	-1.0997	0.271477	
Nogap:dyadFemale_Same	0.23373	0.73541	0.3178	0.750622	

```

Nogap:dyadMale_Other 0.41230 0.42584 0.9682 0.332945
Nogap:dyadMale_Same 1.13332 0.55363 2.0471 0.040652 *
Over:(Intercept) -4.75717 1.18048 -4.0299 5.581e-05 ***
Over:completenessIncomplete 6.60156 1.08633 6.0770 1.225e-09 ***
Over:modalitymultimodal 0.46727 0.63337 0.7378 0.460658
Over:modalityverbal -0.61413 0.51650 -1.1890 0.234431
Over:dyadFemale_Same 1.24569 0.71174 1.7502 0.080085 .
Over:dyadMale_Other 0.47517 0.52773 0.9004 0.367902
Over:dyadMale_Same 2.26147 0.77364 2.9232 0.003465 **
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```