Individual variability in production and comprehension of prosodically disambiguated structural ambiguities

Doctoral dissertation

in partial fulfillment of the requirements for the academic degree of

Doctor of Philosophy in Cognitive Science

submitted to the

Faculty of Human Sciences at the University of Potsdam

by

Clara Huttenlauch

Date of defense: Potsdam, 08.11.2023

Supervision: Prof. Dr. Isabell Wartenburger, Dr. Sandra Hanne
1st reviewer: Prof. Dr. Isabell Wartenburger
2nd reviewer: PD Dr. Gerrit Kentner

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-61926
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-619262

An den Zusammenhalt

To cohesion

Danksagung

Mein Dank gilt meinen beiden Betreuerinnen, Isabell Wartenburger und Sandra Hanne, ohne die es dieses Projekt nicht gegeben hätte. Vielen Dank für eure Unterstützung all die Jahre und Höhen und Tiefen hindurch, für eure Geduld und die Bereitschaft, euch auf unkonventionelle Wege einzulassen. Danke, Carola, für deine aufmunternden, unterstützenden Worte, dein offenes Ohr, die gemeinsame Arbeit und das miteinander und voneinander Lernen. Ein ganz lieber Dank gilt allen im Projekt B01, mit denen ich zusammen lernen und gemeinsam forschen durfte. Danke an den SFB1287 für die Möglichkeit, die Arbeit auf Konferenzen vorstellen zu dürfen und die Menschen, die ich durch den SFB treffen durfte. Danke Michaela für deine scheinbar unerschöpfliche Geduld, die Funktion von Abläufen und Antragswegen zu erklären. Danke für die technische Unterstützung an Jan Ries und die Unterstützung bei der statistischen Auswertung an Reinhold Kliegl und Daniel Schad, sowie die Möglichkeit, bei der SMLP mehrere Jahre dabei sein zu dürfen. Danke an die vier Menschen, die uns ihre Stimme und ihr Bild für die Kontexte zur Verfügung gestellt haben. Ohne Antonia, Marc und Anne wäre Golm für mich nicht das gewesen, was es war. Danke für den Sonnenschein, mit dem aus dunklen Regenwolken Regenbogen entstehen konnten.

Die Arbeit wäre in ihrer jetzigen Form nicht entstanden ohne den Beistand und die Unterstützung derer, die mit großer Bereitschaft unbezahlte Fürsorgearbeit übernommen und dafür gesorgt haben, dass ich immer wieder neu gestärkt zur Arbeit zurückkehren konnte: sei es durch Verköstigung einer spontan entstehenden WG-internen Mensa, körperlichen Ausgleich zu Wasser und zu Land, mit und ohne Musik, das Auffangen, wenn der Boden wegzubrechen drohte, Gespräche über ganz anderes als Linguistik, die Bereitschaft, über meine Arbeit zu sprechen, gemeinsame Motivationstreffen über jitsi während wir entfernt voneinander an unseren Schreibtischen saßen, ... euer immer für mich da sein.

Ich danke allen, die an mich geglaubt haben, auch als ich noch nicht glaubte, dass es jemals so weit kommen würde. Allen, die meine Begeisterung für die (Sprach-)Wissenschaft geteilt, gefüttert und wachgehalten haben. Danke an alle, die mich auf einzelnen Metern oder der ganzen Strecke begleitet und unterstützt haben und allen, die mich zum Endspurt ermutigt und dabei unterstützt haben. Ohne euch würde es diesen Text jetzt nicht geben. Danke an alle meine Freund:innen innerhalb und außerhalb der Linguistik, die die letzten Jahre gemeinsam mit mir gegangen sind.

Mit liebem Gruß an die eigenen Zweifel. *¡Toma, impostora! Take this, impostor!*

List of abbreviations

ACC	accusative case
brack, bracket	coordinate condition with internal grouping
CHILD	context with a child interlocutor
CI	confidence interval
CLES	Common Language Effect Size
CPS	Closure Positive Shift
coordinate	name sequence of three names coordinated with ‘and’
CR	confidence rating
DAT	dative case
ELDERLY	context with an elderly adult interlocutor
ERP	Event-Related Potentials
f0, F0	fundamental frequency
FEM	grammatical femininum
g	gate
GAMM	Generalised Additive Mixed Models
GEN	genitive case
GLMM	Generalised Linear Mixed Models
H	local f0 maximum (high)
HL	Hearing Loss
Hz	Hertz
L	local f0 minimum (low)
L1	first language
LMM	Linear Mixed Models
MASC	grammatical masculinum
NEU	grammatical neutrum
nobrack, no bracket	coordinate condition without internal grouping
NOISE	context with a young adult in background white noise
NON-NATIVE	context with a non-native German-speaking young adult interlocutor
NP	nominal phrase/noun phrase
NOM	nominative case
OVS	object-verb-subject word order
PP	prepositional phrase
RPT	Rapid Prosody Transcription
st	semitones
SVO	subject-verb-object word order
ToBI	Tone and Break Indices
V	vowel
VP	verbal phrase
YOUNG	context with a young adult interlocutor

The fundamental frequency measured to describe the difference between the two grouping conditions of coordinates: without internal grouping (*Moni and Lilli and Manu*) and with internal grouping (*[Moni and Lilli] and Manu*, bracket indicates grouping) was described with the following terms: *f0-range* and *f0-movement* (the latter as a more general term in Background and Discussion), *rise* and *f0-range* (study I), *F0 range* (study II), and *F0-range* (study III). All terms refer to the same calculation described in section 2.2.4 on page 17.

Summary

Strings of words can correspond to more than one interpretation or underlying structure, which makes them ambiguous. Prosody can be used to resolve this structural ambiguity. Besides that, prosody has many other functions, for instance, it also transmits communicative aspects including individual features of the speaker and the situation in which the communication takes place.

In this thesis, we investigated the use of prosodic cues in the domains of fundamental frequency and duration to disambiguate between two interpretations of ambiguous structures when speakers addressed different interlocutors. The thesis includes three production studies and one comprehension study. Prosodic disambiguation was studied with a focus on German name sequences of three names (*coordinates*) in two conditions: without (*Name1 and Name2 and Name3*) and with (*[Name1 and Name2] and Name3*) internal grouping of the first two names in two production studies (studies I and II) and one comprehension study (study III). The study of coordinates was complemented with production data of locally ambiguous sentences with a case-ambiguous first noun phrase (study IV).

To evoke prosodic adaptations to different conversational contexts we elicited productions with a within-subject manipulation of context in a referential communication task. Context had five levels and involved interlocutors in three age groups (child, young adult, elderly adult) with German as L1 in the absence of background white noise, the young adult with background white noise, and a young adult without German as L1. The interlocutors were audio-visually present on a screen.

The thesis addresses three aims. The first aim is to improve our understanding of the form of prosodic grouping studied in the distribution of the three prosodic cues, f0-movement, final lengthening, and pause, involved in the ambiguity resolution in the case of coordinates addressing two sub-points. With the first sub-point we aim to replicate the involvement of the three prosodic cues, f0-range, final lengthening, and pause, in the ambiguity resolution of coordinates and to extend them to older adult speakers. With the second sub-point we aim to deepen the insights of the distribution of prosodic cues within the utterance addressing the question whether the cues are globally or locally used in production and comprehension. The data were discussed in terms of the Proximity/Similarity model by Kentner & Féry (2013). This model makes predictions for structures with internal grouping compared to the baseline without internal grouping. For elements inside a group, the model predicts smaller prosodic cue values (weakening of the prosodic boundary within a group on Name1, *proximity*), while for elements across groups, larger prosodic cue values are predicted (strengthening of the prosodic boundary at the group edge on Name2, *anti-proximity*).

Our results and previous findings on the production and comprehension of coordinates without and with internal grouping show that the group-internal name (Name1) carries a lot of information about the grouping structure already. The marking of prosodic grouping is not restricted to a specific location but appears as a more widespread phenomenon. This has implications for the analysis of production and comprehension data. Therefore, material for comprehension studies should be carefully selected accordingly. The cues build up over time of a coordinate and speakers and listeners differ in the amount of information they produce and use to reliably mark and predict the upcoming structure. For some speakers, some listeners are able to decode these early cues effectively and use them to predict the

upcoming structure. Prosodic grouping is a global phenomenon involving prosodic cues at distal (non-local) positions. Even these prosodic cues at distal positions are an important source of information and can be sufficient for predicting the upcoming structure.

The second aim is to deepen our knowledge of the relationship between prosody and syntax by investigating whether the close link between prosody and syntax is maintained in different conversational contexts or whether the aforementioned disambiguating prosodic cues are modified when speakers address different interlocutors with possibly different needs. If disambiguating prosody is ‘automatically’ present independent of the context (or the situation), we interpret this as a rather direct link between prosody and syntax (situational independence of disambiguating prosody). However, if disambiguating prosody is less automatically connected to the structural properties of the utterance, but used in a more controlled way by the speaker to support the interlocutor’s parsing of an ambiguous utterance, then disambiguating prosody appears as rather situationally dependent. To study this aim, productions of coordinates in different conversational contexts were analysed.

In our data, speakers only slightly modified the prosodic cues marking the disambiguation when addressing interlocutors differing in age or L1, and in the absence/presence of background white noise. Listeners were unable to identify to which interlocutor the sequence had been produced. We interpret this intra-individual consistency in the production of disambiguating prosodic cues as support for a strong link between prosody and syntax. The findings support models in favour of situational independence of disambiguating prosody. The internal structure of coordinates was disambiguated irrespective of the type of addressee and the absence/presence of background white noise. Prosodic disambiguation is interpreted as part of the production process rather than dependent on the situation.

The third aim is to discuss possible generalisations of the findings on prosodic grouping in three sub-points. In the first sub-point, we discuss structured variability and how it supports a phonological category of grouping. In the second sub-point, we discuss whether a relative character of the strength of prosodic cues in grouping conflicts with reliable decoding of early cues. In the third sub-point, we come back to the starting point looking for prosodic disambiguation in another syntactically ambiguous structure, namely data on locally ambiguous sentences. In study IV, we did not find a clear prosodic pattern to resolve the local ambiguity at the group level. Two distinct f0-contours for the two word order conditions present in the data, could not be clearly discriminated by naïve listeners.

Our findings on coordinates support the existence of a phonological category of prosodic grouping that allows for individual variability at the phonetic realisation. Prosodic grouping was consistently marked group-internally on the first name and at the group edge irrespective of the age of speaker (young adults and older adults) using f0-range, final lengthening, and pause. Prosodic grouping appears as a global phenomenon building up along the utterance. For future research it would be interesting to study, whether group-internal weakening of boundaries observed in prosodic grouping is also observable in grouping outside speech. One promising tool to approach this question could be the notion of expectation discussed by Huron (2006). The chunking of complex processes into smaller parts facilitates processing (cf. Frazier et al. 2006, Jackendorff 2009). We conclude that grouping is a common phenomenon and not restricted to prosody.

Contents

List of Tables viii

List of Figures ix

1 Introduction 1

2 Background 4

2.1 Prosody 4

2.2 Prosody and structural ambiguity 6

2.2.1 Ambiguous structures in general 6

2.2.2 Coordinates: About the ambiguity in three elements in a row 8

2.2.3 Case syncretism: About the role of case-ambiguous noun phrases 12

2.2.4 Acoustic correlates of prosodic grouping 14

2.2.5 On the form of prosodic grouping 21

2.3 Prosody and (in)variability 26

2.3.1 Group-level variability 27

2.3.2 Individual-level variability 29

2.3.3 Situational (in)dependence of prosodic disambiguation 34

3 Aims of this thesis 36

4 Experimental Work 38

4.1 Methodological considerations: Production 38

4.1.1 Contexts & referential communication 38

4.1.2 Parallel experimental procedure 43

4.2 Methodological considerations: Comprehension 45

4.2.1 Perception check 45

4.2.2 Gating paradigm 47

5 Study I 49

5.1 Introduction 49

5.2 Methods 59

5.3 Results 69

5.4 Discussion 76

5.5 Conclusion 81

6 Study II **89**
6.1 Introduction 90
6.2 Methods and material 97
6.3 Results 102
6.4 Discussion 112
6.5 Conclusion 115

7 Study III **117**
7.1 Introduction 117
7.2 Aims and hypotheses of the current study 121
7.3 Methods and procedures 123
7.4 Results 133
7.5 Discussion 141
7.6 Conclusion 145

8 Study IV **146**
8.1 Introduction 146
8.2 Methods and procedure 150
8.3 Analyses and results 155
8.4 Discussion 166
8.5 Conclusion 169

9 General Discussion **171**
9.1 Summary of major results 172
9.2 Aim 1: The form of prosodic grouping of coordinates 176
9.3 Aim 2: Situational independence of disambiguating prosody 181
9.4 Aim 3: Generalisations of grouping and beyond coordinates 185

10 Conclusion **192**

References **193**

List of Tables

1 Durations of videos and trigger questions of the five contexts. 43
2 Information on the five contexts. 62
3 Number of productions entering statistical analyses across contexts and conditions in the final data set. 66
4 Criteria for the three possible types of distinction of a cue between the two conditions used for the exploratory analysis. 68
5 Matrix of possible combinations of the cues rise, final lengthening, and pause. 69
6 Estimates of the model for rise on Name1. 70
7 Estimates of the model for final lengthening on Name1. 70
8 Estimates of the model for rise on Name2. 71
9 Estimates of the model for final lengthening on Name2. 72
10 Estimates of the model for pause after Name2. 74
11 Fictional names, ages, origins, and further information of the interlocutors present in the five contexts. 98
12 Distribution of productions entering statistical analyses across age groups, conditions, and contexts in the final data set. 101
13 Descriptive statistics of absolute durational measurements by age group and statistical group comparison. 103
14 Selected model estimates and 95% confidence intervals of the fixed effects for F0 range on Name1 including main effect of condition and main effect and interactions of age group. 106
15 Selected model estimates and 95% confidence intervals of the fixed effects for final lengthening on Name1 including main effect of condition and main effect and interactions of age group. 107
16 Selected model estimates and 95% confidence intervals of the fixed effects for F0 range on Name2 including main effect of condition and main effect and interactions of age group. 109
17 Selected model estimates and 95% confidence intervals of the fixed effects for final lengthening on Name2 including main effect of condition and main effect and interactions of age group. 111
18 Selected model estimates and 95% confidence intervals of the fixed effects for pause after Name2 including main effects and interactions of age group. . . . 112
19 Means and *SDs* of accuracy and confidence ratings by gate. 133
20 Number and percent of participants with accuracy above chance by gate. . . 134
21 Fixed effects of the model on accuracy by gate. 134
22 Fixed effects of the model including *subgroup* nested under *gate* as a predictor of *accuracy*. 138
23 Fictional names, ages, origins, and further information of the interlocutors present in the two contexts. 153
24 Schematic representation of the five situational contexts. 172
25 Distribution of the four studies of the thesis with regard to different levels at which variability was analysed in production and comprehension. 172
26 Visualisation of the syntagmatic and paradigmatic level in sequences of names. 188

List of Figures

1 Spectrograms and smoothed f0-contours for coordinated name sequence with and without internal grouping of the first two names. 24
2 Fictional and true demographics of YOUNG and NOISE interlocutor 41
3 Fictional and true demographics of CHILD interlocutor 41
4 Fictional and true demographics of ELDERLY interlocutor 41
5 Fictional and true demographics of NON-NATIVE interlocutor 42
6 Experimental procedure and timing of one trial for the elicitation of coordinates and locally ambiguous sentences. 44
7 Pictograms used in comprehension studies on coordinates for the condition without internal grouping and for the condition with internal grouping. . . . 46
8 Experimental procedure and timing of trials in the gating paradigm with gate1 and the beginning of gate2 of a coordinate stimulus. 48
9 Pictures and fictional names, ages, origins, and further information of the interlocutors present in the five contexts. 61
10 Experimental setting and timing of two trials. 63
11 Pictograms used in the perception check. Picture A depicts the condition without internal grouping, picture B the condition with internal grouping. . 64
12 Time-normalized f0-contours of coordinates in two conditions produced in five contexts by a subset of 13 female speakers 72
13 Mean values and 95% confidence intervals for rise, final lengthening, and pause on Name2 for each context and condition. 73
14 Speaker variability of cue combinations across contexts on Name2, showing the patterns of cue combinations used by individual speakers in the contexts YOUNG, CHILD, ELDERLY, NON-NATIVE, and NOISE. 75
15 Waveform and spectrogram with F0 contour of the coordinated name sequence *MOni und LIlli und MAnu* produced without internal grouping and with internal grouping by a young female speaker. 91
16 Experimental setting and timing of two trials. 99
17 Pictograms used in the perception check depicting the condition without grouping and with grouping. 100
18 Distribution of raw values of F0 range and final lengthening on Name1 divided by context, condition, and age group. 104
19 Distribution of raw values of F0 range, final lengthening, and pause on Name2 divided by context, condition, and age group. 104
20 Model predictions for F0 range on Name1 divided by age group, condition, and context. 106
21 Model predictions for final lengthening on Name1 divided by age group, condition, and context. 108
22 Model predictions for F0 range on Name2 divided by age group, condition, and context. 109
23 Model predictions for final lengthening on Name2 divided by age group, condition, and context. 110
24 Example of the segmental material in each of the seven gates. 124

25 Oscillogram/spectrogram with F0 contour for the bracket and the no bracket condition. 125
26 Distribution of the three prosodic cues F0-range, final lengthening, and pause on Name1 and Name2 by condition separated for speakers. 127
27 Time normalized smoothed F0-contours for bracket and no bracket condition separated for speakers. 128
28 Pictograms indicating button press in the experiment 129
29 Scatterplots of response patterns of participants assigned to the identification and to the waiting pattern subgroup. 136
30 Boxplots of accuracy by gate and subgroup. 137
31 Boxplots of accuracy by gate and speakers. 139
32 Bar plots of CRs by gate. 140
33 Example of black-and-white line-drawings used as visual stimuli for SVO and OVS word order of a locally ambiguous sentence. 151
34 f0-contours of SVO and OVS with predicted difference in ambiguous part. . . 157
35 f0-contour of SVO and OVS with predicted difference in unambiguous part. . 160
36 f0-contours of SVO and OVS with no predicted difference 161
37 Durations of silent intervals preceding the verb and NP2. 163
38 Selected time-normalised f0-contours of an individual speaker 165
39 Example of black-and-white line-drawings used as visual stimuli for SVO and OVS word order of a globally ambiguous sentence. 168
40 Distribution of prosodic cues on Name2 in productions excluded after the perception check. 178

1 Introduction

No two single utterances are equal to another. Even if produced by the same speaker, they exhibit some variation such as slight phonetic differences in segment durations, vowel quality, or intensity. Between speakers, the variations are even larger: a short segment in the speech of one person can be equal in duration to a long segment in the speech of another person. In the same sense, an utterance-final rise in fundamental frequency (f0) can be phonetically identical in one person's question and another person's statement (Xie et al. 2021) resulting in an ambiguous signal. There is no clear one-to-one mapping between form and meaning. Despite this variability in the speech signal, humans are able to communicate. Nevertheless, variability has long been considered noise. Nowadays, we know not only that variability exists in a structured way, but also that it even supports language processing, including first language acquisition (Rost & McMurray 2009, 2010, Seidl et al. 2014, Höhle et al. 2020, 2021, Bulgarelli & Bergelson 2022). Knowledge and understanding of variability and its limits are further indispensable to form a theory about the complex system of language and communication, for example in distinguishing between neurotypical and deviant language use. Structured variability is also informative to discover phonological categories. These are a few of the reasons why current research stopped circumventing variability and turned towards studying it.

Prosody is one aspect of speech where (individual) variability can be observed. Prosody is always present in non-written language and transmits linguistic information (e. g., distinction between statement and question, as mentioned previously) and paralinguistic information at the same time. If someone calls another person by their name, we are able to retrieve more than just the segmental information of the sounds of the name from the call: pragmatic information (e. g., the purpose of the call: the speaker wants to greet, to reprimand, to request) as well as paralinguistic and situational information regarding the emotional state of the speaker (e. g., happy, annoyed, bored), the relationship between speaker and addressee (e. g., formal, informal, close, distant), and more general information about the speaker (e. g., has a rather low or high voice; has a cold). The prosodic signal is therefore quite complex and signals meaning at many different levels of linguistic and paralinguistic specification. The existence of individual prosody can be seen as a logical consequence or a surprising finding; in the words of Cole (2015):

> Given that prosody has the capacity to convey information about the grammatical and discourse context that is critical to the intended meaning of the utterances, and given that listeners are sensitive to prosodic cues in comprehending speech, it is remarkable that speakers are not more consistent in the expression of prosody. Of course, the fact that prosody performs many functions may itself be a reason for individual differences [...]. (Cole 2015: 18)

A string of sounds such as [a͡ɪskriːm] can correspond to several meanings, and thus be ambiguous: *I scream*, *ice cream*[1], and *ice, cream* as in *I would like to have ice, cream, and a coffee.* In written language, strings of words (in the absence of punctuation) can be ambiguous in their structure. Some of such structural ambiguities can be resolved in spoken language by prosodic means, mainly from the domains of duration and f0. The word sequence *ice cream* can either be produced as belonging together (grouping, describing one concept[2]) or the sequence can be produced describing two individual concepts[3]. This distinction is achieved by modification of prosodic cues such as f0-movement, segmental lengthening, and pause insertion. These disambiguating prosodic cues open space for individual variability in cue use and cue combinations. Further, individual differences in the production of prosody can be triggered by various sources outside the acoustic composition of the signal such as addressing different interlocutors. Several studies report that speakers vary prosodically when addressing different age groups or when facing background noise (cf. DePaulo & Coleman 1986, van Summers et al. 1988, Kemper et al. 1995, Thimm et al. 1998, Biersack et al. 2005, Smith 2007, Zollinger & Brumm 2011, Smiljanic & Gilbert 2017, Piazza et al. 2021). The question arises whether prosodic cues to resolve structural ambiguities are modified in conversational situations involving different types of interlocutors and/or background noise.

This thesis investigates individual variability in disambiguating prosody of structural ambiguities in German. The main focus is on production and comprehension of sequences of three coordinated names with and without internal grouping of the first two names (cf. (1), referred to as *coordinates*), complemented with production data of sentences with either subject-verb-object (*SVO*) or object-verb-subject (*OVS*) word order.

(1) (a) Caro and Toni and Jana. without internal grouping
    (b) (Caro and Toni) and Jana. with internal grouping

The thesis comprises three production studies and one comprehension study and investigates whether and how speakers and listeners used prosodic cues to disambiguate between two conditions (without and with internal grouping, word order). The analyses focused on prosodic cues in the domains of duration and f0. The work contributes to a deeper understanding of the form of prosodic grouping in the disambiguation of coordinates. The data of production indicate a global distribution of disambiguating prosodic cues within the structure as opposed to a local marking of the group edge that are recoverable for predicting the upcoming structure in comprehension. We further investigate the relationship between prosody and syntax by exploring whether it is maintained in different conversational situations or whether the disambiguating prosodic cues are modified when different interlocutors are addressed. To this end, productions were elicited in different conversational situations.

The thesis is structured as follows: Chapter 2 introduces prosody as a means to resolve structural ambiguities and as a channel of variability. Besides ambiguous structures in gen-

[1] also written as *ice-cream*

[2] ice cream: a sweet frozen flavoured food typically made of milk

[3] ice: e. g., frozen water; cream: e. g., fatty part of milk

eral, the two structures investigated here are introduced together with acoustic correlates and the form of prosodic grouping. Further, we present the levels at which variability is investigated and discuss the situational (in)dependence of prosodic disambiguation. Chapter 3 presents the aims of the thesis and Chapter 4 summarises the major results of the four studies and discusses methodological considerations regarding production and comprehension. Chapters 5 to 7 contain the studies I to III that investigate production and comprehension of coordinates. In study I (Chapter 5) we investigated the production of prosodic boundaries used to disambiguate the syntactic structure of coordinates in young adult speakers. This study was intended to replicate the findings of the syntax-prosody model by Kentner & Féry (2013) regarding the prosodic marking of internal grouping. The study elicited productions of coordinates in different contexts (i. e., interlocutors differing in age and L1 and in the absence/presence of background white noise), analysing whether the contexts trigger variability in prosodic grouping at the inter- and intra-group level. Further, study I addressed the question whether disambiguating prosody is produced dependent or independent of the context discussing the nature of the relation between prosody and syntax. Study II (Chapter 6) builds on and extends the results on prosodic boundary production of young adult speakers in study I and compared them to productions of older speakers. By using the same design, stimuli, and elicitation procedure, the study generates valuable and controlled data that make it possible to deepen our understanding of age-related aspects of prosody production and to broaden our insights into the variability of the use of prosodic cues used for disambiguation. Study III (Chapter 7) complements the results on prosodic grouping with the comprehension of coordinates by young adult speakers. Stimuli consisted of recordings of coordinates that were cut into seven parts (gates) and presented gate by gate with increasing duration of the utterance and increasing amount of prosodic information (gating paradigm). We tested whether listeners can decide about the internal grouping of a coordinate structure by already exploiting prosodic information on Name1. Data were collected in a two-alternative forced choice decision task, considering accuracy and confidence of responses. Chapter 8 contains study IV, which expands the investigation of individual variability in prosodic cue production in German by the second structure: locally ambiguous sentences. We elicited productions of SVO and OVS verb-second main clauses in young adult speakers of German. Sentences began with a case-ambiguous NP1 and were string-identical up to the post-verbal case-unambiguous NP2 (2).

(2) (a) Das Kamel sieht nun den Tiger. (SVO)
the.NOM/ACC camel sees now the.ACC tiger
(b) Das Kamel sieht nun der Tiger. (OVS)
the.NOM/ACC camel sees now the.NOM tiger

Chapter 9 and 10 discuss the form of prosodic grouping and possible generalisations, the relationship between prosody and syntax, prosodic (in)variability, and conclude the thesis.

2 Background

2.1 Prosody

"It is impossible to speak without using prosody" (Peppé 2009: 258) – prosody is inherent to non-written language production, both speaking and signing (Cutler et al. 1997, Pfau & Quer 2010, Herrmann 2016)[4]. At the same time, it is difficult to speak and especially to write *about* prosody. There is no clear correspondence between prosodic features and written form. Besides length, pitch, and loudness, prosody further includes aspects related to voice quality. The acoustic correlates of these perceptual aspects are duration, fundamental frequency (f0), intensity, and spectral quality (correlate of voice quality), respectively (Grice & Baumann 2007). A review of them and their articulatory features follows later on in the text. The interweaving of the components of prosody was vividly expressed by Cutler & Isard (1980):

> Prosody is the sauce of the sentence - it adds to, enhances or subtly changes the flavour of the original. And like a good sauce, the realization of a sentence's prosodic structure is a blend of different ingredients none of which can be separately identified in the final product. (Cutler & Isard 1980: 245)

Turk (2009) cites these "multiple physical attributes" as reasons for the difficulty in understanding the mechanism of prosodic production as they "are simultaneously used for prosodic (and other) purposes" (Turk 2009: 319). Regarding neurotypical as opposed to atypical prosody, she speaks about a "conceptual challenge of defining prosody in a meaningful way in normal speech" (Turk 2009: 318). This conceptual challenge is reflected in the several ways researchers describe prosody and possibly explains Peppé's (2009) critique of a "lack of agreement on the terminology and scope of the topic" (Peppé 2009: 258).

Definitions of prosody differ with regard to the viewpoint: defining prosody by its function or by its (phonetic) form (Wagner & Watson 2010). Following Cole's review, "prosody conveys information about the linguistic context of an utterance at every level of linguistic organisation, from the word up to the discourse context" (Cole 2015: 1). Grice & Baumann (2007) attribute the following communicative functions to prosody: (i) lexical and morphological marking (e. g., in tone and pitch accent languages)[5], (ii) disambiguation of syntactic structure, (iii) marking of information structure (e. g., distinction between background vs. focus and given vs. new), (iv) disambiguation between speech acts (e. g., distinction between statements and questions), and (v) transmission of paralinguistic information (e. g., indicating emotional state, affect, and attitude, such as surprise, politeness, and boredom, among others). Further, Grice & Baumann (2007) name *highlighting* and *phrasing* as two main tasks of prosody (Grice & Baumann 2007: 26). Highlighting refers to the "marking of

[4] This thesis focuses exclusively on prosody of spoken language.

[5] The authors note that lexical and morphological marking involve f0 and other prosodic cues, but are not part of intonation in the strict sense (Grice & Baumann 2007).

prominence relations" between words and phrasing to the "division of speech into chunks" (Grice & Baumann 2007: 26). The grammatical aspect of prosody was described as "spoken equivalent of written punctuation" by Peppé (2009), which illustrates why it probably becomes "somewhat invisible" in written language (Peppé 2009: 260). A functional similarity between a comma in written language and a prosody break in spoken language, both leading to early disambiguation of a local ambiguity in Dutch, was found by Kerkhofs et al. (2008).

It is beyond the limits of this thesis to give a detailed definition of prosody (for comprehensive reviews in this context see Cutler et al. 1997, Wagner & Watson 2010, Cole 2015). In the present work, the term *prosodic* is used to refer to acoustic correlates in the domains of duration and fundamental frequency. The analysed measurements include duration of segments and pause/silent intervals and descriptions of the f0-movement especially the f0-range of a change in pitch.

Before turning towards the functions of prosody, we briefly review the aspects of speech that contribute to prosody following Grice & Baumann (2007). The temporal aspects of prosody include all kinds of durational measurements of speech gestures that can be extracted from the speech stream: speech or articulation rate, duration of individual segments, syllables, and if present of silent intervals/pauses. Duration is usually acoustically measured in milliseconds (ms) and corresponds to the perceptual concept of length (Grice & Baumann 2007). The tonal aspect (intonation in its narrow definition, cf. Grice & Baumann 2007: 1) refers to the perceptual concept of pitch, which means "the auditory sensation of tonal height" (Gussenhoven 2004: 1). The articulatory source are quasi-periodic vibrations of the vocal folds that are acoustically measured in Hertz (Hz) as the frequency of the vocal folds' vibrations, referred to as fundamental frequency (f0) (Grice & Baumann 2007). Pitch is perceived as high or low, rising or falling (Grice & Baumann 2007). The faster the vocal folds vibrate, the higher is the perceived pitch. Aperiodic pulses of the vocal folds and the shape of the glottis result in different phonation types that are referred to as voice quality, including modal, creaky, and breathy (Ladefoged 2003). Prosody further includes intensity and vowel quality. The acoustic concept of intensity is measured in decibel (dB) and represents the articulatory effort, that is, the subglottal air pressure, which is perceived as loudness. Vowel quality encompasses several modes of articulatory precision. The vocal tract configuration influences vowel quality, which we perceive on a scale from full to reduced, acoustically measured as spectral quality in formant values (Grice & Baumann 2007).

In summary, prosody is an indispensable component of language. It is relevant in production and comprehension. Prosody is expressed through various linguistic aspects and fulfils a variety of functions. The work in this thesis focuses on tonal and temporal aspects of prosody, on the syntactic function of prosody (i. e., resolution of syntactic ambiguities), and possible prosodic modifications in the speech directed at different interlocutors.

2.2 Prosody and structural ambiguity

Prosody as a means to disambiguate In this section, we consider the structuring function of prosody and, in particular, its use to structure sentence elements in a way that resolves structural ambiguities. We start with ambiguous structures in general (2.2.1) followed by the two subtypes of ambiguous structures investigated in this dissertation. In short, in the first structure (used in studies I, II, and III), the ambiguity arises from different possibilities of internal grouping in sequences with three equally weighted elements ("A and B and C", 2.2.2). In the second structure (used in study IV), the ambiguity arises from form syncretism between nominative and accusative case in the German determiner system leading to ambiguity between agent and patient role (locally ambiguous sentences, 2.2.3). We present the three main acoustic correlates involved in the resolution of the grouping ambiguity: f0-range, final lengthening, and pause (2.2.4) as well as the form of prosodic grouping (2.3.3).

2.2.1 Ambiguous structures in general

Strings of words can form utterances that are lexically identical (i. e., same sequence of word form) but correspond to more than one meaning. In the Lexikon der Sprachwissenschaft (Encyclopaedia of Linguistics), Bußmann (2008) distinguishes between four types of ambiguities: (3) lexical, (4) syntactical, (5) scopal, and (6) relational.

(3) Lass uns an der Bank treffen.
reading (a): 'Let's meet at the bank.'
reading (b): 'Let's meet at the bench.'

(4) Flying airplanes can be dangerous.
reading (a): It is dangerous to fly airplanes.
reading (b): Airplanes that are flying are dangerous.

(5) All books were written by one author.
reading (a): One author wrote all the books.
reading (b): Each book has an author.

(6) Jana's letter.
reading (a): Jana received the letter.
reading (b)[6]: Jana wrote the letter.

Considering two alternative meanings, Lehiste (1973b) speaks of "sentence pairs" that are "syntactically ambiguous, but lexically identical" (Lehiste 1973b: 1231). A lot of ambiguities pass unnoted in communication as they either get resolved by prior context (i. e., the

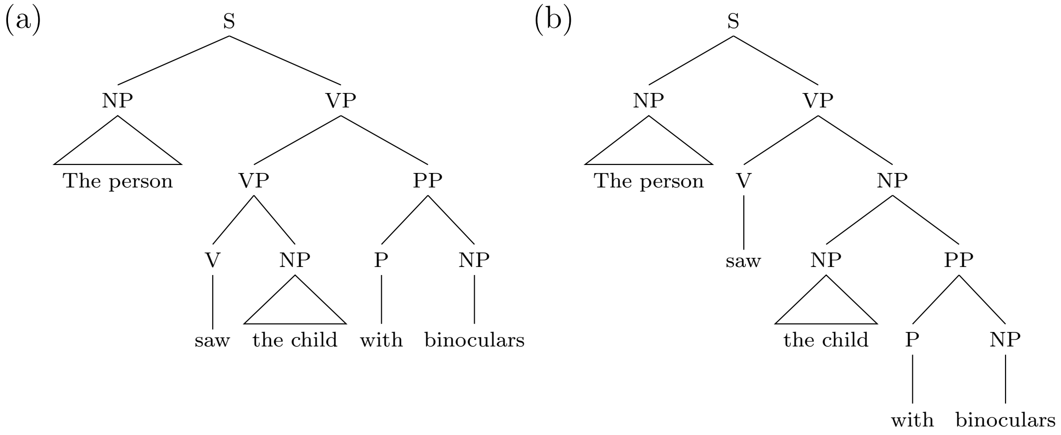
[6] The relational ambiguity is not restricted to two readings, further relations between Jana and the letter are possible.

ambiguity never arises) or get resolved in the further context of the utterance through linguistic or extra-linguistic context (Bußmann 2008), by world knowledge, or shared context by the communication partners (Price et al. 1991). Another possibility to resolve ambiguities are prosodic means. In this thesis, we investigated two types of syntactical ambiguity[7], also referred to as structural ambiguity and the prosodic means to disambiguate them. The following description is thus restricted to syntactical/structural ambiguities.

There are different ways to characterise (ambiguous) structures. Sentences can be categorised as either (i) unambiguous (no structural ambiguity), (ii) locally or temporally ambiguous (structural ambiguity in the first part of the sentence that gets resolved in the course of the sentence, also referred to as garden path sentences), or (iii) globally ambiguous (structural ambiguity that continues until the end of the sentence). Further, sentences can be characterised by the relation between meaning interpretations and underlying syntactic bracketing: different interpretations can be reflected in different surface bracketings versus identical surface bracketings (more details in Lehiste et al. 1976).

(7) The person saw the child with binoculars.
reading (a): The person used binoculars to see the child.
reading (b): The person saw the child that had binoculars.

(8)



In examining various ambiguous structures, previous research revealed that not all ambiguous sentences can be disambiguated equally well by prosodic means. Successful disambiguation was measured as "correctly" interpreted by naïve listeners. Prosodic disambiguation turned out possible for sentence pairs "in which a meaning difference was associated with a difference in the surface phrase structure" (Lehiste et al. 1976). One way of visualising different possible syntactic analyses is by syntactic trees with different branching structures. An example for such a structurally ambiguous sentence for which the two readings can be

[7] For syntactical ambiguity, sentences do not need to be word-identical. Price et al. (1991) used in their stimuli sentence pairs that share the same string of phones and are associated with two contrasting syntactic structures such as *Dave will never know why he's enraged, will he?* vs. *Dave will never know why he's enraged Willy.* (Price et al. 1991: 2968).

represented with different tree structures is *The person saw the child with binoculars* (example (7) with the corresponding tree structures in (8)). In reading (a), the person uses binoculars to see the child, the prepositional phrase (PP) *with binoculars* modifies the verb *saw*. This reading would be marked by a prosodic boundary after *the person*, thus rather early in the sentence (also referred to as early boundary). In reading (b), the child has binoculars and is seen by the person. The PP modifies the noun phrase (NP) *the child*, which would be prosodically marked with a prosodic boundary after *the child*. In the tree structure (8a), the PP *with binoculars* is a sister of the VP, whereas in (8b), it is a sister of the NP *the child*. The PP in (8a) is attached higher in the tree as in (8b). Referring to this, the reading in (a) is called *high attachment* and the reading in (b) *low attachment*. To sum up, in cases where syntax, the order of words, is undetermined between two or more possible analyses prosody can influence the decision in favour of one of the different tree structures (Bögel 2015: 76).

This thesis focuses on two types of structural ambiguities in German: coordinated name sequences differing in their internal grouping of elements (cf. examples (11) – (13)) and locally ambiguous sentences (cf. examples (18) and (19)). Both structures will be introduced separately in the following.

2.2.2 Coordinates: About the ambiguity in three elements in a row

One of the structures studied in this thesis consists of a sequence of three names combined by the coordinating conjunction *and* (in German *und*): "A and B and C". Translated into a mathematical equation, this structure corresponds to an addition: $A + B + C$. Following the associative law, it does not play a role in which order we sum the elements (i. e., whether we calculate $2 + 3 + 4$ or $3 + 4 + 2$), the sum is always 9, unaltered by reordering[8]. In any case, the plus sign operates on two elements by summing them. If, for instance, the first plus sign is replaced with a multiplication sign, for example, the order in which addition and multiplication are performed affects the result and the result is either 10 or 14 ($2 * 3 + 4 = 10$, while $2 * (3 + 4) = 14$). The order of operations is no longer interchangeable. Again, the two operations are each carried out between two items and one operation after the other (applied to the first example: first the multiplication of 2 with 3, second the addition of 6 with 4). Thus, there is an internal grouping of elements and a temporal ordering of the operations: two items or elements are grouped together (around the operator). Regarding the order, the priority rule determines in math that multiplication precedes addition. If the order is to be reversed, the addition needs to be written in parentheses: $2 * (3 + 4)$, thus, the coordination elements are grouped together and prioritised in order. Going back to linguistics, the structure with two different operations can be compared to a sequence containing a disjunction and a conjunction: "A or B and C". Any calculation with two different operators

[8] $a + (b + c) = (a + b) + c$, this also applies to multiplication: $a(bc) = (ab)c$.

works. Kentner & Féry (2013) compared the same sequence to the "arithmetic procedure" $3 - 2 + 1$ that either resolves as 0 or as 2 (Kentner & Féry 2013: 277).

In linguistics, we lack a rule of priority. This results in an ambiguity regarding the internal grouping of the elements in a sequence of more than two elements in which the associative law does not apply (Wagner 2005: 92). For example, the sentence "Steve or Sam and Bob will come" (Lehiste 1973a,b, Lehiste et al. 1976) says that either one person (Steve) or two persons will come. Regarding the two persons, the structure leaves open, which two persons will come: whether Steve and Bob come or Sam and Bob come. In non-written communication, the ambiguity regarding the group of two can be prosodically resolved by marking the corresponding grouping (as in the math example, the prosodic grouping is indicated by parentheses). To get the reading that either one or two persons are coming, Sam and Bob have to be grouped prosodically (9).

(9) Steve or (Sam and Bob) will come.
reading (a): Steve will come.
reading (b): Sam and Bob will come.

For the reading that in any case two persons will come, Steve and Sam have to be grouped prosodically (10).

(10) (Steve or Sam) and Bob will come.
reading (a): Steve and Bob will come.
reading (b): Sam and Bob will come.

These example show that the internal grouping of three elements connected with two different connectors leads to a difference in meaning.

It might appear less obvious that in language, contrary to math, a coordinated sequence such as "A and B and C", with only one connector, is also more complex than it seems at the surface. This becomes clearer when we embed the utterance in a context. In "Caro and Toni and Jana will come", there are not different outcomes in terms of number of persons that will come (always all three of them). Nevertheless, different internal groupings can transmit information about relationships within the group of people and, thus, result in different meanings that are transported: (i) two of them are siblings and the third is a close friend or (ii) two of the persons form a couple and the third one is a child or a friend. In another context, if "Caro and Toni and Jana." is the short answer to the question "Who will plant a tree?", different internal groupings will lead to different numbers of planted trees: one, two, or three.

(11) (Caro and Toni and Jana).

(12) (a) (Caro and Toni) and Jana.
(b) Caro (and Toni and Jana).

(13) Caro and Toni and Jana.

There can be one planted tree, if all three persons plant one together (11). There can be two planted trees, if Caro and Toni plant one and Jana plants a second one or if Caro plants one and Toni and Jana plant a second one (12). If all three persons individually plant a tree, there will be three trees planted (13).

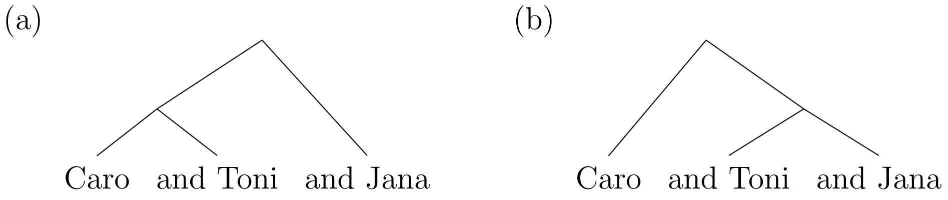
Slightly different sequences are lists of three nouns that either refer to three elements (14) or to two elements with a compound noun in the first position (15) (Peppé et al. 2000, Zhang 2012). In written form, the two versions are distinguished by punctuation and are, thus, different from the previous examples, but in the spoken form, the prosodic disambiguation works in the same way as in the previous examples.

(14) Ice, cream, and fruits.

(15) Ice cream and fruits.

The present dissertation compares sequences with internal grouping (12) to sequences without internal grouping (13). In a tree structure, the two examples in (12) can be displayed as in (16). The two trees differ in their branching direction, which means that the position of the forking branch (node) differs with respect to head node. The example (16a) corresponds to the internal grouping in (12a) and is called a left-branching structure since the forking branch (Caro and Toni) is at the left part of the tree and is then grouped with Jana at the right. Conversely, the example (16b) corresponds to the internal grouping in (12b) and is called a right-branching structure, as the forking branch with the internal group is positioned at the right.

(16)

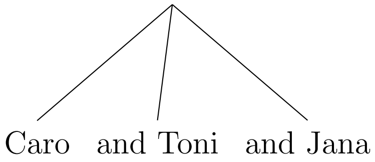


Usually, structures are assumed with binary branching nodes. In a strictly binary branching structure, the name sequence without internal grouping in (13) would be displayed with the tree in (16b). This would mean that the number of phonological and semantic interpretations (3: (12a), (12b), and (13)) would not correspond to the number of associated syntactic representations (2). A solution is to capture the sequence without internal grouping in (13) with the tree in (17)[9]. More details on tree structures of coordinates and arguments in favour

[9] The question may arise how the structure corresponding to the example in (11) looks like. So far, we were concerned with finding corresponding syntactic representations to possible internal groupings within a coordinate sequence of three conjuncts. The example in (11) is a further reading but does not contain another internal grouping. Moreover, as it is not further relevant for the work presented here, we will not deepen this aspect.

of trees with unbounded branching can be found in Wagner (2005). The studies presented here focus on left-branching structures.

(17)



The syntactically ambiguous structure introduced in this chapter has been referred to by different terms in the course of research, including *coordinated names* (Kentner & Féry 2013), *coordinations*, *coordination structures* or *coordination groupings* (Féry & Kentner 2010, Bögel 2015), *coordinate structures* (Wagner 2005, 2010), *coordinated structures* (Wellmann et al. 2012), *bracketed lists* (Petrone et al. 2017), or *conjunctions* (Wagner 2005). In this dissertation, we will use the terms *coordinated name sequences*, in short *coordinates*, that were also used in the related publications (Huttenlauch et al. 2021, 2023, Hansen et al. 2022). The use of prosody to mark the underlying branching structure in coordinates has been described with terms such as *prosodic grouping* (Cutler & Isard 1980, Wagner & Watson 2010, Kentner & Féry 2013, Bögel 2015, Wellmann et al. 2023), *chunking* (Peppé et al. 2000), and *prosodic phrasing* (Wagner & Watson 2010, Wagner 2010, Wellmann et al. 2012, Zhang 2012, Cole 2015, Holzgrefe-Lang 2017). The term prosodic phrasing is not restricted to coordinate structures, but refers in general to the grouping of words in continuous speech into prosodic phrases (cf. Cole 2015, see Grice & Baumann 2007 for a list of terms). The prosodic phenomenon itself that leads to grouping and prosodic phrasing is broadly referred to as *prosodic boundary* (Wagner 2005, Wellmann et al. 2012, Zhang 2012, Kentner & Féry 2013, Cole 2015, Holzgrefe-Lang et al. 2016, Holzgrefe-Lang 2017) or *juncture* (Ip & Cutler 2022). The function of such a boundary is to "signal the relative independence of the upcoming words to the immediately preceding words" (Cole 2015: 5). We will use the terms *prosodic grouping/phrasing* and *prosodic boundary.* The main acoustic correlates of prosodic grouping are described in the chapter after next.

Coordinated sequences are by no means restricted to names or a maximum of three elements. Besides three-name sequences, Kentner & Féry (2013) investigated four-name sequences with different combinations of internal grouping. Coordinates also appear outside the speech materials for linguistic research: The children of Bullerbü, for example, are introduced by Lisa, one of them, as

> Lasse und Bosse und ich und Ole und Britta und Inga (Lindgren 1988: 8).

Preceding to this sequence of coordinated names, Lisa introduces Lasse and Bosse as her brothers, Ole as a child living in one of the neighbouring farm houses, and Britta and Inga as sisters living in a third farm house. In the larger context, the sequence of names is not unstructured. Using parentheses to mark the family relationships (internal groupings) would lead to the following structure: (Lasse und Bosse und ich) (und Ole) (und Britta und Inga).

2.2.3 Case syncretism: About the role of case-ambiguous noun phrases

The other structure investigated in this thesis is an ambiguity that arises from transitive verbs in combination with two noun phrases (NPs) that leave undetermined who does the action to whom. In the processing of a sentence, thematic roles (e. g., agent or patient) are assigned and mapped onto syntactic functions (e. g., subject or object). In a case when both noun phrases are equally likely to be assigned the agent (= subject) and the patient (= object) role, the sentence is globally ambiguous. If the role-ambiguity only exists for the first NP (NP1) in the sentence, the sentence is called locally or temporally ambiguous. The latter one is in the focus of the investigation of the thesis. This type of ambiguity is more language-dependent than the coordinate structures, as it depends on how a language marks grammatical function, thematic roles and how flexible the word order is.

German[10], the language investigated in the present work, has a rich morphological case system for marking grammatical function and allows for a relatively free word order. Despite the rich case marking system, the surface form of NPs can be ambiguous. For instance, for NPs involving feminine and neuter nouns, respectively, the surface form is identical in nominative and accusative case. Case is marked on the determiner: *die* for feminine NPs and *das* for neuter NPs both in nominative as well as in accusative case, respectively. Such NPs are, thus, considered case-ambiguous. Regarding the flexible word order: In addition to subject-verb-object (SVO) sentences, the non-canonical word order of object-verb-subject (OVS) is also possible. Thus, if the determiner *die* or *das* is part of an NP at the beginning of a sentence, the syntactic function of that NP as well as its thematic role remains open: it is ambiguous between subject and direct object as well as between agent and patient. Therefore, the word order configuration could potentially be both, SVO or OVS. If the ambiguity gets resolved at later points in the sentence (e. g., by a case-marked post-verbal NP or by verb inflection), the sentence is called temporarily or locally ambiguous (see (18) and (19)). Besides morphological case markers, prosody, verb semantics, and (visual) context can resolve or influence such thematic role-assignment ambiguities.

(18)	Das	Kamel	sieht	nun	den	Tiger.
	the.NOM/ACC	camel	sees	now	the.ACC	tiger

‘The camel now sees the tiger.’

[10] A modified version of this paragraph is published in Huttenlauch et al. (2022) (study IV of this dissertation).

(19) Das Kamel sieht nun der Tiger.
the.NOM/ACC camel sees now the.NOM tiger
'The tiger now sees the camel.'

If both NPs are case-ambiguous with regard to nominative and accusative case, no thematic roles can be assigned on the basis of case marking and the ambiguity remains until the end of the sentence. Those sentences are called globally ambiguous sentences.

Besides case-syncretism between nominative and accusative case in the German feminine and neuter determiner, the feminine determiner *die* shares the same surface form in other two cases: both genitive and dative case have the surface form *der.* An example for a globally ambiguous sentence is given in (20) (taken from Bögel 2015: 83).

(20) Alle waren überrascht, dass der Partner der
Everyone was surprised that the.MASC.NOM partner the.FEM.GEN/DAT
Freundin zuhörte.
friend.FEM listened to.
'Everyone was surprised that [the partner listened to the friend/ the friend's partner listened].'

The sentence in (20) contains an ambiguity regarding the theme that surprised everyone: that the friend's partner listens or that the partner listens to the friend. The ambiguity remains until the end of the sentence.

Studies on German structural ambiguities are not restricted to transitive verbs, but also include sentences with ditransitive verbs (Gollrad et al. 2010, Häussler & Bader 2012). Ditransitive verbs request two objects (a direct and an indirect one), which can correspond to the thematic roles of theme and recipient requiring accusative and dative case, respectively. The theme is manipulated by the action of the agent and the recipient receives the outcome of the action. Local ambiguities can arise in subordinate clauses with three NPs with case ambiguous determiners and a final verb (cf. (21) in Gollrad et al. 2010). One sentence ends in a ditransitive verb, the other in a transitive verb.

(21) Neulich hat der Mann der Nachbarin
Recently did the.MASC.NOM man the.FEM.GEN/DAT neighbour.FEM
ein Haus geschenkt/gesehen,
a.NEU.NOM/ACC house give/see.PAST PARTICIPLE
Recently, the man the neighbour a house gave/saw,
'Recently the man [gave the neighbour/of the neighbour saw] a house, that'

In both sentences in (21) the man is the agent of the action. In one version, the neighbour is the recipient of a ditransitive action and receives a house as a gift. In the other version, the neighbour modifies the man, who sees a house (transitive verb). This type of locally ambiguous sentence is also referred to as garden path sentence.

In the present work, we focus on locally ambiguous sentences as in (18) and (19).

2.2.4 Acoustic correlates of prosodic grouping

We now turn from ambiguous structures towards their prosodic disambiguation. Prosodic grouping[11] in production and comprehension is established by acoustic correlates. Early studies investigating the production of prosodic disambiguation of different syntactic groupings found a primary role of duration (cf. Lehiste 1973a, Lehiste et al. 1976). The predominant locations to look for prosodic grouping were the group edges, more specifically the right edge. In section 2.2.5 we will see evidence for a more global view of prosodic grouping. Studies nowadays agree on the joint use of duration and fundamental frequency (f0): Across languages, there are three main acoustic correlates of prosodic boundaries: f0-movement, final lengthening, and pause insertion (Wagner & Watson 2010, Kentner & Féry 2013, Cole 2015, Petrone et al. 2017). Additional acoustic boundary cues in some languages include voice quality and articulatory strengthening (Cole 2015) as well as intensity (Wagner & Watson 2010: 907). An example for the articulatory strengthening is (domain) initial strengthening at the beginning of a constituent (cf. Napoleão de Souza 2023 for Spanish and Portuguese).

In describing prosodic boundaries, the observation of production and comprehension are intertwined. The recognition of prosodic boundaries is difficult (Grice & Baumann 2007: 4). Prosodic boundaries have no clear counterpart outside prosody, and they do not always coincide with syntactic boundaries (Grice & Baumann 2007, Cole 2015). In order to explore the acoustic correlates in their production, the perceptual side is needed to identify their location. Vice versa, to investigate the influences of individual acoustic cues on comprehension, production is required.

Before describing the three main cues, which this thesis focuses on (i. e., f0-movement, final lengthening, and pause) separately, we will make some general notes on f0. Over the time course of an utterance, the fundamental frequency constantly drops, which is called declination (cf. the dashed line in the bottom panel of Figure 1 on page 24). This means that successive peaks (high points) are lower than the preceding ones. The same applies to low values, they also drop relative to the preceding ones. The opposite phenomenon is called f0-reset: an f0-peak is higher than the immediately preceding one. Thus, the phenomenon of declination is interrupted and f0 is reset to a higher value from which declination starts again.

There are many ways to describe an f0-pattern. The analysis starts therefore with the decision for which way to go (cf. researchers degrees of freedom, Roettger 2019) and is accompanied by further decisions along the way (Grice & Baumann 2007: 6). F0-patterns

[11] Terms such as *prosodic phrasing* or *prosodic phrase boundary* are more broadly used than prosodic grouping. Since the coordinate structures studied here are no full sentences, we prefer the term *group* rather than *phrase* to refer to the internal chunks. Accordingly, we use *prosodic boundary* rather than *prosodic phrase boundary*. This is not to say that phrase would not apply, as the names can be considered noun phrases (NP).

are time-varying data that can be transferred for analysis into discrete values or events (e. g., individual values, categorical types of contours) or they “can be analysed holistically as continuous trajectories” (Roettger 2019: 7). The challenge is to describe the f0 curve as such and in relation to the text to which it is produced. Decisions include the time domain in which the contour is analysed, the theoretical framework in which the contour is viewed, the number of points with which the contour is described and many more. There is a variety of values that can be picked individually or in relation to each other within the f0-pattern: turning points, minimum and maximum values within a certain domain, mean, standard deviation, range between two points, points in relation to segmental landmarks (alignment), shape of a pattern and so on. Often the description of f0/pitch involves a combination of perceptual categorisation and acoustical measurements.

F0 can be measured in each voiced segment (in Hertz, Hz). The significance of a single Hertz value is limited in that it can only be classified as high or low in comparison to other values. 250 Hz can be high when produced by a person with a mean frequency lower than 250 Hz and low when the person has a higher mean frequency. Ways to characterise the magnitude of a change in f0 include measuring the range of an f0-movement (difference between minimum and maximum values) or the slope (range divided by the time between start and end of the movement). A common unit for f0-range is semitones (st) calculated with the formula

$$f0_{range}(st) = 12 * log_2(\frac{f0_{max}(Hz)}{f0_{min}(Hz)})$$

which gives the relative difference between two Hertz values independent of the absolute pitch height. In comprehension, the same absolute difference of for example 50 Hz is perceived as much smaller between 400 and 450 Hz (2.3 st) compared to between 100 and 150 Hz (7.0 st). Since speakers differ in their pitch range, relative measures in semitones allow pitch-independent comparisons between speakers. For a single Hertz value, an equivalent in semitones can be calculated in several ways: relative to (i) the minimum of the speakers’ range, (ii) the speakers’ mean f0 value, (iii) a speaker independent value such as 1 or 100 Hz (cf. comments in Bögel 2015: 65 and Hazan et al. 2016).

There are different phonological models to describe intonation (cf. Grice & Baumann 2007 for the description of two of them: the British School and the autosegmental-metrical model). It would go beyond the scope of this work to give an extensive overview of different models. We will briefly introduce the Tone and Break Indices (ToBI) annotation system, which is based on one of them, the autosegmental-metrical model of intonation (Ladd 2008 and references therein). ToBI was originally established for American English (Silverman et al. 1992) and is now widely adapted to other languages (for the German adaption: GToBI Grice & Baumann 2002, Grice et al. 2005). It is a phonological system that describes the pitch contour in terms of two events, pitch accents and boundary tones, in addition to the break index, marking the perceived strength of a boundary (Grice & Baumann 2007).

Pitch accents mirror the f0-movement around lexically stressed syllables (Grice & Baumann 2007). Boundary tones are f0-contours that "are distinct from f0 contours that express pitch accents primarily in *not* being prominence-lending, which is to say that they extend over one or more final syllables regardless of the status of those syllables as prominent" (highlighted as in original Cole 2015: 6). These f0-movements are therefore not associated with stressed syllables, but function as structure-forming features[12]. Pitch accents and boundary tones are described in terms of low (L) and high (H) values and the movement (or interpolation) between these values: Changes from low to high result in a rise, changes from high to low result in a fall, and movement without much vertical change results in level pitch or a plateau. Pitch accents and boundary tones are composed of combinations of Ls and Hs and language-specific pitch accents and boundary tones are inventoried and related to pragmatic functions. ToBI uses categorical events and also allows take into account gradient values by adding diacritica to the labels (e. g., ! for downstep, ˆ for upstep). The ToBI annotation is widely used in intonation research[13]. In the analyses carried out in the coordinates, we do not use the categorical ToBI labels to describe the pitch contour, but instead use continuous measures of f0-range. This allows us a better comparison with final lengthening and pause duration, which are also on a continuous scale. Nevertheless, the analysis is inspired by the ToBI labeling system.

Another way to describe intonation contours is by comparing complete contours with each other. This allows to capture more fine-grained information of the continuous f0-trajectories, such as the shape of the contour. A statistical way to model continuous data is by fitting Generalised Additive Mixed Models (GAMMs, Wood 2017, Baayen et al. 2017, Sóskuthy 2017, Wieling 2018). GAMMs allow to model time-varying data with non-linear patterns and have been successfully used in previous analyses of f0-contours (Chuang et al. 2020, Zahner et al. 2020, Sóskuthy 2021). Instead of choosing individual points, a larger part of the f0-contour is entered into the analyses. In the time domain of analysis, f0 is measured at equidistant times and normalised in this way. The data is still discrete, but with a small sampling rate, closer to continuous data. Considering the shape of a contour allows to distinguish between convex and concave rises that have been shown to differ between pragmatic conditions (e. g., in productions in Neapolitan Italian, convex rises were correlated to narrow focus questions and concave rises to partial topic statements, Cangemi 2009; in French, increased concavity of the rising f0-movement yielded more question responses than continuation responses in a comprehension study by Dorokhova & D'Imperio 2019). Further, the shape of an f0 movement in American English was shown to matter for perception of peak height (plateaus are psychoacoustically better discriminated than peaks) and also for

[12] If assigned to phrase-final syllables, the f0-movement is called edge tone.

[13] Thoughts on the relationship between intonation (as a categorical aspect) and f0 (as a continuous aspect) especially in the context of the investigation of variability and thinking intonation beyond the AM model are given in Arvaniti (2019).

memory (larger advantage with higher memory load) (Kimball & Cole 2016).

The three main cues for prosodic grouping, f0-movement, final lengthening, and pause, are not restricted to prosodic marking of groups, but used more in general to mark prosodic boundaries. In a corpus with German spontaneous speech data containing a variety of prosodic boundaries, 93.1% of the phrase boundaries were marked by at least of these three cues (Peters et al. 2005: 159). In the following, f0-movement, final lengthening, and pause are described separately with respect to their involvement in the marking of prosodic grouping.

F0-movement The tonal marking of prosodic groups often involves "an abrupt change in pitch across unaccented syllables", which can go either upwards or downwards (Grice & Baumann 2007: 4f.). Tonal marking is observed in the f0-movement on (i) group-final elements (e. g., described as categorical boundary tones or continuous measures of f0-change in the contour) and on (ii) the element following the group (e. g., f0-reset).

For items in a list, Grice & Baumann (2007) report that

> all but the last phrase end at a relatively high pitch [...] or with a high level pitch. The high pitch indicates that there is still at least one more item to come. After it the pitch is reset (i. e., there is a jump down), marking the beginning of the next phrase. (Grice & Baumann 2007: 5)[14]

A high f0 value at the end of a non-final group (endpoint of a final rise or plateau) is also reported for German coordinates as the most frequent cue besides a low f0 value (endpoint of a fall) and overall speaker-specific preferences (Petrone et al. 2017: 77). The coordinates consisted of disyllabic trochaic names and the f0 contour was acoustically characterised by the difference in f0, measuring the f0 minimum on the stressed syllable (first syllable of the name) and the f0 maximum on the final syllable of each name.

On lists of either two or three elements (cf. examples (15) and (14)) produced in American English and Mandarin Chinese, Zhang (2012) reported language dependent f0 cues: Speakers of American English used f0-slope to differentiate between list conditions, while speakers of Mandarin Chinese used f0-reset.

Besides languages that differ with regard to which f0 feature they use distinctively, studies differ in how they quantify and name the f0-movement involved in the prosodic boundaries (even when studying the same language). Some speak about the *pitch contour* and *high tones* (Kentner & Féry 2013), others use *pitch change* (Wellmann et al. 2012, Holzgrefe et al. 2013, Holzgrefe-Lang et al. 2016) or directly specify the direction of the contour (e. g., *pitch rise*, Wellmann et al. 2012, van Ommen et al. 2020), and again others use *f0* instead of *pitch*, referring to the acoustic instead of the perceptual correlate, or use *edge tone*. As far as acoustic characterisation is concerned, different measured values are used in the studies.

[14] Note that reset is used here in the other direction than in the previous definition of f0-reset, namely as a downward jump back into the region of declination.

For German coordinates, Petrone et al. (2017) run statistical analyses on the f0 value in Hz at the end of each name, while Wellmann et al. (2012) measured the pitch change in Hz as the difference between the f0 maximum on the final vowel and the f0 value in the center of the first segment of the name. For French coordinates, van Ommen et al. (2020)[15] calculated the pitch rise (the difference between the f0 maximum in the final vowel and the f0 minimum in the prefinal vowel) in semitones (van Ommen et al. 2020: 4).

In the studies in this dissertation, the f0-movement is characterised in the following ways: for coordinates (studies I, II, and III), the difference between the f0 minimum in the first syllable and the f0 maximum in the second syllable (mostly situated on the final vowel) was calculated in semitones[16], for locally ambiguous sentences (study IV), the complete f0 contours were analysed using GAMMs.

Final lengthening describes "an increase in segmental duration at the right edge of different types of prosodic domains above the word level" (Paschen et al. 2022: 1). This phenomenon is also referred to as *pre-boundary lengthening* (Zhang 2012, Schubö & Zerbian 2023) and *(phrase-)final* and *boundary-related lengthening* (Turk & Shattuck-Hufnagel 2007). The amount of final lengthening is assessed by comparing the duration of segments preceding a prosodic boundary or pause to the duration of the same segments without a following boundary or to segments in non-final positions, depending on the structure of the available speech data. Final lengthening is distinct from accentual lengthening induced by accent (Grice & Baumann 2007).

Final lengthening is widespread in the worlds' languages and no counterexample to its universality is known so far according to a recent comprehensive review by Paschen et al. (2022). Paschen et al. (2022) themselves observed final lengthening in final, phonemically non-long vowels in natural speech of 25 languages from 19 linguistic families across all six geographical macro-areas, supporting the view of final lengthening as a "common process across languages" (Paschen et al. 2022: 13). Languages differ with regard to the number of segments affected by lengthening, referred to as domain or scope, and the degree of final lengthening (Turk & Shattuck-Hufnagel 2007, Paschen et al. 2022). According to Paschen et al. (2022), final lengthening can be affected but not overridden by stress placement and interacts with phonological vowel length. For some languages, final lengthening is reported to be gradient: strongest in segments directly adjacent to the boundary and decreasing with increasing distance from the boundary (Paschen et al. 2022; for American English: Cole 2015; for German: Kohler 1983, Schubö & Zerbian 2023). The gradient increase in final

[15] The coordinates contained disyllabic trochaic names that were segmentally similar to the names used in the studies by Wellmann et al. (2012) and Petrone et al. (2017), and also to the material used in the studies presented here.

[16] In study I this measure is referred to as *rise* or *f0-range*, in study II and III it is called *F0 range* and *F0-range*, respectively, taking into account that a rising contour is not the only way how f0 was produced.

lengthening is not necessarily linearly, as some regions in a word can be less lengthened than preceding and following regions as reported for syllables between the rime carrying the main-stress and the final rime in American English by Turk & Shattuck-Hufnagel (2007). A recent study on final lengthening at boundaries with different strengths in German prose reported a U-shape pattern for the relation between degree of final lengthening and the strength of a boundary (Kentner et al. 2023). In the data by Kentner et al. (2023), final lengthening increased up to the level of a "short comma phrase" determined by only one or two preceding words (Kentner et al. 2023: 5) and then decreased with further increase in boundary strength.

Previous research on German used different measures of final lengthening, including the domain of the syllable rime (Peters et al. 2005 and Petrone et al. 2017, for the latter, the rime consisted of a final vowel in open syllables and thus a single segment), the duration of a complete name in name sequences (Féry & Kentner 2010, Kentner & Féry 2013). Data by Schubö & Zerbian (2023) suggest that in German final lengthening starts on the main stress vowel, and increases by segments from left to right. This is in line with earlier work on German laboratory speech, reporting larger lengthening in the syllable directly preceding the boundary than in the stressed syllable (Kohler 1983).

In the studies on coordinates (studies I, II, and III) in this dissertation, final lengthening is used as a relative measure, calculated as the duration of the final vowel in a name (rime) relative to the duration of the whole name.

Pause The term pause refers to a break in the speech signal at the right edge of a prosodic group. Pauses can go along with breaths that may be audible in the speech signal. The term is widely used in a general manner, sometimes distinguishing between filled and unfilled or silent pauses, even though studies showed that silent pauses are not really silent either (Trouvain et al. 2016) and more fine-grained distinctions would be appropriate. Pauses are an important indication of prosodic boundaries. In a study on boundary markers, boundaries were annotated in a semi-automatic manner using the presence of pauses as indicators to locate boundaries in the speech stream (Peters et al. 2005). It is therefore interesting that, at least for German spontaneous speech according to Peters et al. (2005), pauses are not the most frequent cues, as pauses and breathing are reported as cues involved in the marking of 38.3% of the boundaries.

In contrast to final lengthening and f0-movement, pause is not superimposed on existing speech material. The pause cue has a categorical nature as it is either present or absent (cf. Peters 2006: 60f.). Nevertheless, when present, its duration correlates positively "with the syntactic complexity [...] of the upcoming utterance" (Petrone et al. 2017: 72). The speaking rate influences pause duration in production (with shorter pauses in faster speaking rate, Šturm & Volín 2023) and the threshold for perceiving pauses in perception (Cole 2015). It is therefore not surprising that studies differ in whether and what minimum duration they specify for silent breaks.

In the studies in this dissertation pause is characterised in the following ways: for coordinates (studies I, II, and III), the durations of gaps in the spectrogram of at least 20 ms following Name1 and Name2 were measured and calculated relative to the duration of the whole utterance in percent. The minimum duration of 20 ms was defined following the production study on similar name sequences by Petrone et al. (2017). For locally ambiguous sentences (study IV), the absolute duration of silent intervals (corresponding either to pauses or closures of stops) preceding the verb and the second noun phrase were measured in seconds.

Cue combinations Acoustic correlates are referred to as *prosodic cues*[17] in this thesis. Although they can be described individually, they interact in the construction of prosodic phrases. In the dataset with German spontaneous speech analysed by Peters et al. (2005), 23.3% of the boundaries were signalled by only one of the cues (composed of 10.3% f0-reset, 9.4% final lengthening, 2.3% separating contours, and 1.3% pause/breathing, Peters et al. 2005: 157f.). For 6.9% of the cases a boundary was perceived despite the absence of any of the cues investigated (Peters et al. 2005: 158f.)[18]. The remaining boundaries (about two third) were marked by at least two of the cues (Peters et al. 2005: 159). The data suggest that prosodic boundaries are complexly composed of cue combinations. Given this complexity in the phonetic composition of prosodic boundaries, the question arises whether individuals are free in their choice of how to mark prosodic boundaries. Early studies not only identified the phonetic mechanisms of prosodic grouping, but also observed individual differences between speakers in the phonetic implementation of disambiguation (Lehiste 1973b, Cutler & Isard 1980).

With regard to the comprehension of prosodic boundaries, cues are weighted differently and the weighting differs cross-linguistically. *Weighting* means that individual cues contribute to different degrees to the perception of a boundary. According to Grice & Baumann (2007) pauses are the "most obvious indicators of boundaries" and their duration correlates positively with boundary strength, although boundaries can be perceived despite the absence of pauses too (Grice & Baumann 2007: 4). Data of 33 participants on boundary perception in English support this, as post-pause duration was used by all listeners as a cue to boundary marking (Roy et al. 2017: 28f.). Most listeners used two predictors as cues.

[17] We use the term *cue* for both, production and comprehension.

[18] Half of these cases showed f0-patterns that were not classified into the separating contours by the authors and further cues not comprised in the automatic annotation including glottalisation, changes in speech rate or speech register (Peters et al. 2005: 179). This is an example of how the researcher's decisions about the inclusion/exclusion of variables in the analysis influence the conclusions that can be drawn (cf. Roettger 2019). Further it underlines the complexity of the phonetic composition of prosodic boundaries (cf. Peters et al. 2005: 160f.).

The perception of prosodic boundaries can be assessed using behavioural and electrophysiological measures. In event-related potentials (ERPs), the processing of a prosodic boundary is reflected in a positive peak that coincides with the closure of a prosodic boundary (i. e., the end of a group) called closure positive shift (CPS) (Steinhauer et al. 1999, Holzgrefe et al. 2013). Studies on the comprehension of coordinates without and with internal grouping of the first two names in German-speaking adults showed that f0 (instantiated as pitch change on Name2) and final lengthening as single cues are not sufficient for the perception of a prosodic boundary but that they are sufficient in combination (Holzgrefe-Lang et al. 2016). The study elicited ERPs and behavioural data (prosodic judgement task) and the results of both correspond. In the stimuli, the acoustic boundary cues were locally added to natural productions without internal grouping increasing the duration of the final vowel in Name2 and implementing a rising f0 contour on Name2. The materials contained two versions of f0 manipulation, one with an implemented f0 rise on Name2 and the second with additional flattening of the f0 contour on Name1, as this contour is closer to natural productions with internal grouping (cf. weakened group-internal prosodic cues discussed in the next section). Stimuli with both f0 manipulations (and implemented final lengthening) elicited a stronger CPS and larger mean proportions of boundary responses compared to stimuli with local f0 boundary cue only on Name2. Similar to adults, 8-month-old German-learning infants do not need a pause cue to perceive a prosodic boundary. A pitch change and final lengthening are sufficient in combination but not as single cues (Wellmann et al. 2012). For 6-month-olds, however, final lengthening is only sufficient in combination with a pause but not with a pitch change (Wellmann et al. 2023), which indicates that the weighting of boundary cues develops in the first year of life. A cross-linguistic study with French- and German-learning infants showed that the development of boundary perception is language-specific (van Ommen et al. 2020). In contrast to German-learning 8-month-olds, French-learning infants did not perceive a boundary cued by only two cues. A natural boundary, however, was perceived by French-learning infants at 8 and already at 6 months of age (van Ommen et al. 2020).

2.2.5 On the form of prosodic grouping

A large body of research concentrated mainly on the right edge of a group of elements, investigating the prosodic nature of the prosodic phrase boundary at the location of the syntactic ambiguity (for production: Lehiste 1973b, Price et al. 1991, Cutler et al. 1997, Peters et al. 2005, Turk & Shattuck-Hufnagel 2007, among others and for comprehension: Weber et al. 2006, Wellmann et al. 2012, Henry et al. 2017, van Ommen et al. 2020, Wellmann et al. 2023, among others). However, as Cutler et al. (1997) mention in their review, the internal structure can be viewed from two perspectives, “as prosodic signaling of syntactic breaks, or as prosodic signaling of grouping” (Cutler et al. 1997: 169). Cutler et al. (1997) stress, that these two views are not independent, but nevertheless correspond to different questions that researchers should be explicit about (Cutler et al. 1997: 169). Cutler et al. (1997) use the term *cohesion* for the opposite of a break. Kentner & Féry (2013) use the term *proximity* and speak about *sisters*. Internal grouping, thus, is not only a matter of divergence, separating a group from the neighbouring elements by a boundary, but also about cohesion or proximity between the elements within the group.

In the following we will briefly review some relevant models on prosodic phrasing in coordinates, which have been proposed in the past[19].

With respect to the question how coordinates are prosodically phrased in English, Taglicht (1998) formulated the "Coordination Constraint". It specifies that the same hierarchical level of intonational boundaries must be applied to all elements at the same syntactic level. Watson & Gibson (2004) argue in their "Left hand side/Right hand side Boundary hypothesis (LRB)" that the likelihood of the presence of a prosodic boundary depends on the size of the preceding and following constituents because of processing demands: For larger constituents, the speaker needs more refractory time to recover from the preceding constituent and more time to plan the upcoming constituent, respectively. Wagner (2005, 2010) demonstrates that, in coordinate structures, the relative strength of the boundary reflects the level of embedding at the syntactic level and thereby confirms the close match between prosody and syntax in coordinate structures. This observation is summarised in the principle of the "Scopally Determined Boundary Rank" (SBR) stating that "if Boundary Rank at a given level of embedding is n, the rank of the boundaries between constituents of the next higher level is n+1" (Wagner 2005). According to Wagner (2010) more deeply embedded constituents "are separated from each other by weaker boundaries than constituents that are less deeply embedded" and "constituents separated by relatively weaker boundaries are perceived as grouping together" (Wagner 2010: 186). Wagner uses the term of "relational prosodification" (Wagner 2005: 82) as the strength of a boundary is in this relational theory always *relative* to some other boundary. He proposes that the syntax-to-prosody mapping assigns relative boundary ranks that can then be realised by prosodic means with a certain flexibility (Wagner 2005: 155). In his opinion, the "relational theory is incompatible with the idea that particular syntactic categories map to particular prosodic categories" (Wagner 2005: 155) and he therefore provides no specific phonological labels.

In a similar vein as this relative theory, Féry & Kentner (2010) and Kentner & Féry (2013) developed a model on German data of coordinates that aims to account for both, processing demands, depending on the constituent size or complexity, and demands of the syntactic structure, depending on the depth of syntactic embedding. Their so called Proximity/Similarity model assumes two principles that "interact to shape the prosody of syntactic structures" (Kentner & Féry 2013: 283) such as coordinated name sequences. Proximity is related to the syntactic constituent structure and states that "adjacent ele-

[19] The following text is a modified version of the corresponding introduction of study I published as Huttenlauch et al. (2021).

ments which are syntactically grouped together into one constituent should be realised in close proximity" (Kentner & Féry 2013: 282). Similarity is related to the depth of syntactic embedding and refers to the idea that "constituents at the same level of embedding should be realised in a similar way, that is, they should be similar in pitch and duration, irrespective of their inherent complexity"; this principle is comparable to the models of Taglicht (1998) and Wagner (2005, 2010), which assume that elements at the same syntactic level are prosodically matched.

The principle of proximity predicts a weakening of the prosodic cues at an element x if the neighbouring element to the right is part of the same group as x. The strength of the cues in a coordinate with internal grouping is compared to a coordinate without internal grouping (considered the baseline form). In the baseline form, all names are expected to be separated by boundaries of the same strength. With boundary cue *weakening*, Kentner & Féry (2013) refer to less final lengthening, a lower f0 peak, a smaller f0-range, and a shorter pause duration. Reversely, anti-proximity predicts a strengthening of a prosodic boundary if the right-adjacent element of x does not form a group with x. A strengthened boundary at the right edge of the grouped element is expressed by higher f0 values and longer durations.

(22) (a) Caro and Toni and Jana.
(b) [Caro and Toni] and Jana.

Applied to the two coordinate structures in (22), in the condition (b) with internal grouping, the first name (i. e., Caro) is produced with weaker prosodic cues while the second name (i. e., Toni) is produced with stronger prosodic cues compared to the condition (a) without internal grouping, respectively.

The Similarity principle predicts that a simplex element at the same level of embedding as a complex constituent is, for instance, lengthened to adjust its duration to the length of a complex constituent. This would be relevant for groupings in which the first name is followed by a complex sequence on the right (right-branching structure: Caro and [Toni and Jana]). As the materials studied here do not contain such structures, the Similarity principle will not be discussed any further. On the final element of the coordinates (cf. Jana in (22)) in either condition, Kentner & Féry (2013) observed neutralisation in duration and f0-movement (cf. the similarity of the shapes of the solid (upper panel) and dashed line (bottom panel) on the name *Manu* in Figure 1).

Figure 1 shows spectrogram and smoothed f0-contours for the name sequence *Moni und Lilli und Manu* with internal grouping of the first two names in the upper panel and without grouping in the lower panel. The dashed line in the bottom panel shows the f0-contour of the baseline condition with declination from the f0 peak on *Moni* to the f0 peak on *Lilli*, the typical list intonation, which Wagner (2005, 2010) calls "prosodically flat" in the context of coordinates. In the condition with internal grouping, peak height increased from Name1 to Name2 with a larger f0-range on Name2 compared to Name1 (upstep, cf. the solid line in

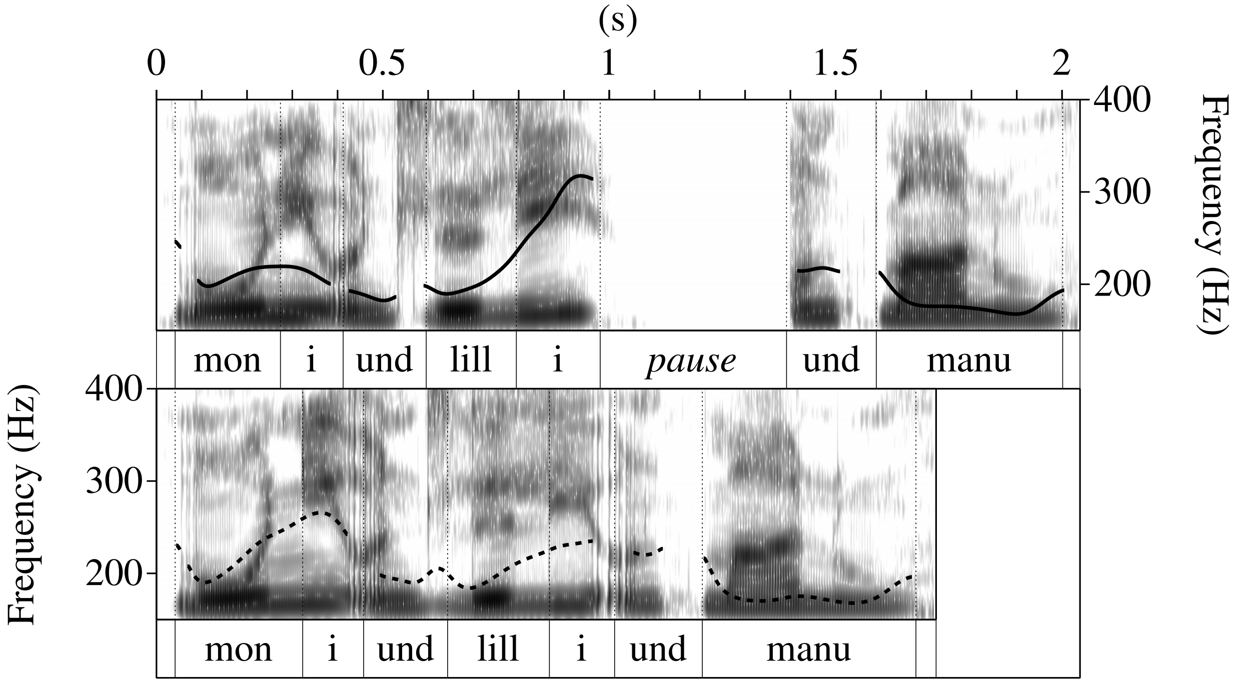


Figure 1: Spectrograms and smoothed f0-contours (solid line in upper panel, dashed line in bottom panel) for coordinated name sequence *Moni und Lilli und Manu* with internal grouping of the first two names (upper panel) and without internal grouping (bottom panel). Both examples were produced by the same young adult, as female identifying speaker in study I. The tiers below the spectrograms illustrate the segmentation of the individual words, the final vowels of the first two names, and a pause (in the upper panel).

the upper panel of Figure 1 has a smaller maximal value on the final *i* of *Moni* than the maximal value on the final *i* of *Lilli*, Name2). Besides a comparison on the syntagmatic level, between Name1 and Name2 in the same structure, we can compare the two structures on the paradigmatic level (i.ė., comparing between the condition with internal grouping and the baseline condition without internal grouping). The solid line in the upper panel of Figure 1 shows a smaller f0-range and a lower peak on *Moni* than the dashed line in the bottom panel (principle of proximity or downstep). Proximity also includes less final lengthening in the condition with compared to the condition without internal grouping, which is not clearly visible in Figure 1. The principle of anti-proximity is visible on the name *Lilli* in both panels of Figure 1 with the solid line showing a larger f0-range compared to the dashed line, along with more final lengthening and the insertion of a pause following *Lilli* (cf. the longer final vowel *i* in *Lilli* and the pause in the upper panel of Figure 1).

Petrone et al. (2017) expanded the exploration of three-name sequences with grouping of either the first two or the last two names with a control condition. In the control condition, a simple name was followed by a complex name (word condition: *Lola [oder Mona Urlena]* compared to *Lola [oder Mona und Lena]*). The authors adopted a more phonological view on the coordinate structures. The word condition was a control to the right-branching condition as different prosodic structures were hypothesised: a phonological phrase break between *Mona* and *Lena* in the right-branching condition, but not within *Mona Urlena*. Their results, however, do not show acoustic differences between the right-branching structure and the word condition on *Mona*, which is, along other reasons, interpreted as lacking evidence for a phonological phrase break (Petrone et al. 2017: 86). The same applies to the group-internal name in the left-branching condition (*Lola*). The absence of acoustic cues to a phonological phrase boundary is interpreted as being in accordance with the principle of proximity by Kentner & Féry (2013). To account for the deviation of the observed prosodic realisation from the expected prosodic structure, Petrone et al. (2017) propose an extended formulation of proximity, referring to task-specific effects: “To emphasise the presence of a constituent x (e. g., in rendering brackets in the stimuli of an experiment around the constituent), a speaker may render x as a phonological phrase, even if the syntax-prosody mapping would otherwise assign the word in x to two (or more) phonological phrases” (Petrone et al. 2017: 87).

In summary, so far we have seen that prosody can be used to resolve structural ambiguities. If a string of words has more than one meaning (interpretation) and if these different interpretations correspond to different underlying structures (e. g., visualised in different syntactic trees) prosody can be used to disambiguate between the different meanings at the surface (e. g., by prosodic boundaries at the position of syntactic ambiguity). Prosodic grouping involves mainly a combination of three prosodic cues: f0-movement, final lengthening, and pause. These cues are not only present at the right edge of the group (position of syntactic ambiguity) but also preceding the edge marking the proximity or cohesion of the elements within the group. Prosody is by no means restricted to the resolution of structural ambiguities. Besides many further functions regarding the organisation from the word up to the discourse level, prosody comes along with a lot of variability.

2.3 Prosody and (in)variability

Prosody as a channel of variability Language is variable and “speakers should not be assumed to be homogeneous even though they speak the same language” (Ouyang & Kaiser 2015: 153). This is no new information (for prosody: Lehiste 1973b), as a quarter of a century ago Cutler et al. (1997) wrote in their review on prosody that “there has also in general been regrettably little attention paid to characterizing the acoustic dimensions making up the prosodic information” (Cutler et al. 1997: 170) and Peppé et al. (2000) wrote that the characterisation of variability in the use of prosodic features is a “neglected topic” (Peppé et al. 2000: 309). In the meantime, variability is no longer neglected or only mentioned in passing, but is a main focus of research and provides topics for entire volumes[20]. The authors of one of those books described the interest in inter-individual variation in speech as increasing (Fuchs et al. 2015: 7). More and more studies systematically investigate variability. This includes viewing variability between individuals as a source of information instead of considering it “noise in the data, which could be eliminated” (Fuchs et al. 2015: 7f.). Still, “variability in the prosodic signal across talkers and contexts” complicates the understanding of the mechanism of communication (Xie et al. 2021: 1).

Variability in general occurs between and within languages, between and within groups of language users, and between and within individuals. Further, variability exists at different levels of language (e. g., lexical, syntactical, prosodic). By variability we mean a range of possible linguistic forms or behaviours available for the transmission of a specific concept to an individual language user, a group of language users, or a language. Variability can be looked at from different angles: Focusing on the sources and focusing on the outcomes (e. g., dialectal vs. lexical; phonological vs. prosodic).

In this thesis, the focus is on variability in the prosodic domain. With regard to sources of prosodic variability, Peppé et al. (2000) list six types labelled A to F: (A) dialectal variation, (B) variation across groups within a single dialectal community, (C) individual differences between speakers of the same speech community, (D) phonologically conditioned variation within an individual speaker, (E) contextually conditioned variation within an individual speaker, (F) random variation (Peppé et al. 2000: 309ff.). Type B includes variability due to differences in socio-economic class, gender, age, sexual orientation, educational level, and hearing status, among others. In this thesis, this type of variability was investigated between two groups of speakers differing in age, referred to as *inter-group* or *between-group* variability. Type C includes variability due to qualitative differences in articulation (e. g., place of articulation) between individuals. With respect to prosody, this includes differences in patterns as well as degree of cues. This type of variability is henceforth referred to as *inter-individual* variability. Type D includes variability dependent on the phonological nature of

[20] cf. Arvaniti (2019) on a critical view with respect to approaches to intonation that either make variability “the main focus of modelling or a problem to be solved” (Arvaniti 2019: 3)

the speech material (including syllable structure, duration, and vowel quality) that can be avoided by controlling "lexical content and segmental phonological structure" (Peppé et al. 2000: 311). Type E includes variability due to the "semantic, pragmatic, and interactional context in which the informant is required to produce an utterance" (Peppé et al. 2000: 311). This broad definition also includes different registers (differences in formality and politeness, among others) and comprises the variability henceforth referred to as *intra-individual* and *intra-group* variability.

The studies presented in this thesis consider variability at the group and at the individual level, both within and between each level: intra-group, inter-group, inter-individual, and intra-individual. Each level will be individually described in the next sections focusing on previous research in the field of prosody, although not always on disambiguating prosody.

2.3.1 Group-level variability

The investigation of linguistic behaviour between and within groups of speakers of the same dialect allows insights into the inner mechanisms of language use and communication. Group comparisons at the inter-group level are relevant in sociolinguistic research (e. g., for the identification of drivers of changes in language use) and in clinical research (e. g., defining ranges of neurotypical and disturbed behaviour), to just name a few.

Intra-group level: Intra-individual manipulation Variability at the intra-group level is not included as an individual type of variability by Peppé et al. (2000). It can be considered as closely related to contextually conditioned variation at the intra-individual level (type E of Peppé et al. 2000), provided the factors are manipulated intra-individually. If each subject provides data for each level of context manipulation and the responses have some consistency, intra-individually manipulated situational contexts can also become visible at the group level. The individual is seen as an individual *in* a speech community. However, intra-group and intra-individual effects occur not necessarily in line with each other, as will be addressed later on in the text. A closer description of the intra-individual manipulation used in this thesis that includes different types of interlocutors and the absence/presence of background white noise is given in the section on the *intra-individual level*.

In the domain of intonation, intra-group variability with intra-individual manipulation was studied for instance with focus on the influence of different communicative contexts on the intonation contours of vocatives (address forms) in Central Catalan (Borràs-Comes et al. 2015). The contexts involved two sociopragmatic features and two situational factors with two levels each. This resulted in the following factors: Social distance (at work vs. at home), power (superior vs. inferior), physical distance (close vs. distant), and insistence (first vs. second call). Vocatives were elicited by means of a discourse completion task (Blum-Kulka et al. 1989), speakers were asked to call their addressee by the name to then express a request. In general, mainly three different vocative contours were produced. Results of the production data show that distinct f0-contours were used for different levels of insistence, social and physical distance. A subsequent comprehension task showed consistent results. Additionally, in production, the communicative contexts affected pitch range and syllable duration of the stressed and post-stressed syllables (Borràs-Comes et al. 2015: 78).

Inter-group level The inter-group level corresponds to the type B in the list of sources of prosodic variability by Peppé et al. (2000), including differences due to socio-economic class, gender, age, sexual orientation, educational level, and hearing status, among others. Peppé et al. (2000), for instance, measured whether social and individual factors affect prosodic performance in production and comprehension. They conclude "that differences in age, sex, and age related hearing acuity are not, on their own at least, factors which affect prosodic performance" (Peppé et al. 2000: 318). These results contrast with findings by other studies that report on tonal and durational differences between younger and older speakers, especially increasing f0 ranges and durations as age increases.

In the dissertation, age is investigated as inter-group factor, comparing productions of young and older adult speakers (study II). The terms *young* or *younger speakers* are used to refer to the age range between 18 and 30 years and *older speakers* for ages above 60 years.

The evidence of age-related changes in the tonal and durational domain entails the question whether these interact with the modulation of prosodic cues. Studies on English speakers found an unaffected ability to modulate prosody to convey linguistic meaning in older speakers (Scukanec et al. 1996, Tauber et al. 2010, Barnes 2013). Prosodic cues were even produced in a more extreme way for disambiguation by older compared to younger speakers. A comparison by Scukanec et al. (1996) of maximal f0 values within the vowel of elicited monosyllabic words in either contrastive or non-contrastive stress position in younger and older female English speakers revealed higher f0 values in words with contrastive stress and lower maximal f0 values in words in non-contrast positions for older compared to young speakers. The difference between conditions was thus larger in older compared to younger speakers. Similarly, older speakers used f0 to a greater extent than young adults when producing lexical stress to differentiate noun-verb pairs with strong-weak and weak-strong stress patterns (Barnes 2013: 43). With respect to duration, Tauber et al. (2010) found longer intonational boundaries (defined as pause duration plus duration of the critical word at the boundary) in productions of older compared to younger speakers. The authors tested for age differences in the realisation of disambiguating prosody in English ambiguous sentences (e.g., *The lake froze over a month ago*). The percentage of sentences, which were successfully disambiguated via prosody was 66% for older speakers (above chance, $p < .05$) and 59% for the younger age group (not significantly above chance, $p > .06$) (Tauber et al. 2010). As far as we know, no study has yet investigated age differences in the use of prosody to resolve ambiguities in coordinate structures.

2.3.2 Individual-level variability

Variability between or within speakers is interesting for two reasons. First, if all speakers reliably use disambiguating prosody to distinguish between conditions of ambiguous structures, this would indicate a close link between syntax and prosody (e. g., Nespor & Vogel 1986) and situational independence (e. g., Speer et al. 2011). If, in addition to disambiguation, we find variability between speakers in the way they realise prosodic boundaries, this would be further evidence that the link between syntax and phonology is relational rather than categorical (e. g., Clifton et al. 2002, Wagner 2005). Second, if speakers do not reliably distinguish between different syntactic structures or vary within a speaker in different contextual settings, this would argue for situationally dependent models and against a close prosody-syntax link.

Inter-individual level Inter-individual refers in this context to variability between individuals in the production and comprehension of linguistic categories (type C by Peppé et al. 2000). In production, this means that the phonetic realisation of a (phonological) category varies from speaker to speaker. Other terms are *between-subject variability* and *talker variability*. This is despite the fact that the phenomenon is found in both production and comprehension. Each language user has their individual way of producing and perceiving language. We are able to recognise each other by the way we express ourselves. Applied to comprehension, it means that individual listeners react differently to a stimulus. Since prosody is composed of several cues (see section on prosody and ambiguous structures), there is a wide scope for variability: Each cue can be individually varied in degree and cues can be combined individually.

The studies in this dissertation investigate inter-individual differences between young adults in the production of single and combined prosodic cues to resolve structural ambiguities in coordinates (study I) and locally ambiguous sentences (study IV) and in the comprehension of coordinates (study III). In the following we will review previous work on inter-individual differences in general and more specifically in the production of coordinated name sequences and prosodic cue combinations.

Inter-individual variability "can occur qualitatively and quantitatively, both on a general level and in specific cases" (Ouyang & Kaiser 2015: 153). On the one hand, speakers differ in their ranges of absolute values in a given acoustic dimension and on the other hand, they can use different strategies and combinations of cues to signal a linguistic contrast (Ouyang & Kaiser 2015: 153f.). More concretely, in terms of f0-contours, qualitative differences correspond to different shapes of f0-contours, while quantitative differences correspond to different ranges of f0 values. Both are reported for speakers of American English producing sentences in different focus and givenness conditions (Ouyang & Kaiser 2015: 166). Individual speakers produced different numbers of f0 peaks and valleys that additionally differed in location and relative height. Furthermore, speakers varied inter-individually in the proportion of their f0-range employed to mark different meanings (Ouyang & Kaiser 2015: 166ff.). Overall, speakers showed intra-individual consistency in the produced f0 shapes between conditions. Another example for a qualitative difference in strategy can be found in the already mentioned production study with two versions of the sentence "Steve or Sam and Bob will come", one with grouping of the first two names (version A: "(Steve or Sam) and Bob will come", parentheses mark the grouping), the other with grouping of the last two names (version B: "Steve or (Sam and Bob) will come") in English (Lehiste 1973b). The productions of two speakers contained a durational difference between the two versions that was correctly distinguished by listeners in 94.4% of the cases (Lehiste 1973b: 1231). However, the two speakers differed in the strategy that led to the durational difference: One speaker inserted a pause (in version A after "Sam", in version B after "Steve"), while the other lengthened the segments of the connector ("and" in version A, "or" in version B) (Lehiste 1973b: 1231). Cutler & Isard (1980) reported a similar observation where one speaker manipulated duration, while the other manipulated pitch to distinguish between two conditions. They propose the idea of a trade-off for a similar observation which they view as a

> justification for an abstract level of prosodic groupings, where different speakers would have in common the intention of marking off a syntactic unit by assigning it a grouping of its own, and would then diverge as to the way in which the presence of this grouping would be signalled. (Cutler & Isard 1980: 260)

An example for a quantitative difference between individuals is provided by a production study on German. Speakers read a text at self-paced slow, normal, and fast rates. One speakers' normal rate measured in syllables per second was similar to the other two speaker's slow rates. Additionally, all three speakers differed in the relative increase or decrease from normal to fast or slow rate, respectively (Trouvain & Grice 1999). Thus, even though all speakers modulated tempo on the same dimension (i. e., syllables per second), they differed in absolute values and degree of difference. In a similar vein, Xie et al. (2021) point out that inter-individual variability can go so far that "one person's production of a statement and another person's production of a question can be phonetically identical" (Xie et al. 2021: 1). Their statement refers to American English, in which statements and questions are generally distinguished by the sentence-final intonation contour: falling for statements and rising for questions. Taking into account the phenomenon of uptalk (a finally rising f0-contour in statements) and a smaller f0-range in the rise of children's productions, the identical f0-contour can possibly be produced on both sentence types (Xie et al. 2021: 1,19).

Effects on the group level are not always visible on the individual level and vice versa. What seems clear at the group level can appear blurred at the individual level (Niebuhr et al. 2011). Individual differences were reported for a well established tonal difference between a high pitch accent (H*) and a high falling pitch accent (H+L*) in Standard Northern German.

On the group level (35 speakers), these two f0-contours showed an earlier peak alignment in the H+L* compared to the H* pitch accent and a rather symmetrical peak shape for H* compared to a left-tailed peak shape for H+L*. On the individual level, however, alignment values formed a continuum with productions of five speakers at one end using only alignment and no shape difference (called "aligners") and five other speakers at the other end using no alignment but only shape difference (called "shapers") between the two accent types (Niebuhr et al. 2011: 121f.). The authors reported similar findings for two Italian varieties and advocate taking the behaviour of the individual into account when studying prosodic cues (Niebuhr et al. 2011: 123).

A rather contrary observation was reported by Xie et al. (2021) for the distinction between statement and question prosody in American English. Both categories overlapped on the group level in duration and utterance-final f0 "primarily caused by variability between talkers, rather than variability within talkers". Models considering talker information performed better in the categorisation paralleling findings that listeners compute prosodic cues in relation to specific speakers (Xie et al. 2021: 11).

The issue of inter-individual variability concerning the prosodic disambiguation of structural ambiguities in German has so far been explored only scarcely (exceptions being the work by Petrone et al. 2017 on coordinates with different internal groupings and the mentioning of inter-individual variability in the use of f0-contours on OVS sentences by Weber et al. 2006.). Petrone et al. (2017) reported that speakers largely differed in how they used f0 to mark the right group edge in the grouping condition: Only two out of 12 speakers consistently used the same f0-contour, namely a rise. Another six participants produced predominantly a rise in addition to a high plateau. Another three speakers varied between rise, high plateau, and final fall to different degrees and one speaker produced either rises or falls. Inter-individual variability in cue combinations was observed on speakers of British English producing similar stimuli: a list of three nouns that either formed a list of two items (a compound and a simple noun: *cream-buns and cheese*) or three items (three simple nouns: *cream, buns, and cheese*) (Peppé et al. 2000: 320). The results show inter-individual differences in the degree and use of f0 pattern, segmental lengthening, and absence/presence of a pause, as well as in their combination. Speakers had in common that the majority of them used more than one cue. Lengthening and pause appeared in the data as more reliable than pitch movement and pitch reset to distinguish between 2-item and 3-item lists (Peppé et al. 2000: 323ff.).

As for comprehension, Cangemi et al. (2015) reported listener-specific variability in the decoding of three linguistic focus structures in German (broad, narrow, and contrastive focus) realised by five speakers and differing with regard to combinations and degrees of prosodic cues. Listeners had to match the target sentences to one of the three focus conditions. Results show inter-listener differences in the percentage of correct responses ranging from 56% to 75% (Cangemi et al. 2015: 136f.). For English boundary comprehension, Roy et al.

(2017) reported inter-individual differences with regard to the number of predictors listeners used in combination as cues to boundary marking (Roy et al. 2017: 28f.).

Intra-individual level Variability on the intra-individual level, also called *within-talker* or *within-individual* level, refers to "contextually conditioned [variability] within an individual speaker" (type E, Peppé et al. 2000: 309). In terms of prosody, this means whether individuals vary their prosodic realisations depending on the context. Context here refers to the semantic, pragmatic, and interactional level of speech production. A more detailed description is provided by Pescuma et al. (2023), who use "situational-functional context" to refer to "the extra-linguistic situation in which language is produced and processed" including "time and place of the communication, the number and identity of participants (their age, gender, ethnicity, status, education, and social role, among others)" but also the "intra-textual linguistic context, such as surrounding sentences" (Pescuma et al. 2023: 2)[21].

Studies differ in how implicitly or explicitly information about the type of interlocutor and the relationship of the participant to the interlocutor(s) is presented and in what way the participants are presented with different situations. Differences range from written or orally presented descriptions of situations and interlocutors (e. g., in a discourse completion task, cf. Blum-Kulka et al. 1989 and proposals for modifications by Vanrell et al. 2018), to visually presented situations and interlocutors (e. g., showing pictures or pre-recorded videos on a screen) to interactions with physically present interlocutors in the situation. Regarding information about the interlocutor or speaker, respectively for production or comprehension studies, possibilities include descriptions that contain evaluative descriptions of the persons (cf. as if talking to a hearing-impaired person), rather neutral instructions (cf. "as if you were telling someone a story that you wanted them to understand", Allbritton et al. 1996: 716), explicit information triggering prejudices about a group of people (cf. social information and linguistic stereotypes based on the names of two Berlin districts: Kreuzberg vs. Zehlendorf, Jannedy & Weirich 2014: 91), and visual cues (cf. attire and hair style indicating differences in formality, Pescuma et al. 2023: 18).

Intra-individual variability in production was addressed in two recent studies on German. Both of them reported preliminary results in phonetic changes in spontaneous speech directed at different interlocutors on the group level. The first study investigates the effect of situations differing in perceived formality (in terms of topic of conversation) and social constellation between speaker and addressee on fine phonetic details in German. The addressee is in all situations performed by the same person wearing different clothes representing a boss, a professor, a fellow student, and a neighbour and appeared in pre-recorded videos.

[21] This intra-individual variability arising from situational-functional contexts is referred to with the term *register* in their collaborative research center in Berlin, Germany https://sfb1412.hu-berlin.de/, Lüdeling et al. 2022, Pescuma et al. 2023. Register is defined as "conventionalised and recurrent linguistic patterns of (individuals in) a speech community" (Pescuma et al. 2023: 2).

The speech recorded in the different contexts shows more variability in f0 and vowel dispersion in the formal compared to the informal situation (Pescuma et al. 2023: 18f.). The second study investigates accommodations of German L1 speakers addressing an English L1 speaker with mid or high proficiency in German in comparison to addressing a German native speaker (baseline). Speech addressing the non-native interlocutor was slowed down and contained less filled pauses as revealed by preliminary results (Pescuma et al. 2023: 22).

In the studies conducted within this dissertation, the term *context* is used in a more specific way. In the design of the production studies, context refers to five different conversational situations involving four interlocutors and one condition with white background noise. Consequently, with respect to intra-group and intra-individual variability we are mainly interested in the variability induced by different types of interlocutors and the presence of background noise. To this end, the rest of this section reviews[22] previous research that has focused on differences in prosodic realisations when children, elderly adults, or non-native speakers are being addressed in comparison to young adult native speakers. Most studies take the speech addressed to an adult native speaker of the language under investigation as a baseline for comparisons. For easier reading, we will refrain from mentioning this adult baseline in the following. For example, for attachment disambiguation, Kempe et al. (2010) reported lengthened vowels when English-speaking adults addressed 2–4-year old real or imaginary children and, in addition, found longer pause durations. Other studies investigated intra-individual variability in prosodic information per se (i. e., not focusing on disambiguating prosody): Biersack et al. (2005) reported an increased pitch range and higher f0-maxima as well as longer durations due to the lengthening of vowels in semi-spontaneous speech addressed to a two-year old imaginary child in English. DePaulo & Coleman (1986) also reported longer pauses in spontaneous English speech addressing a 6-year old child. With respect to prosodic cues in speech addressing a non-native interlocutor, results are inconclusive: while one study involving English speakers found no differences (DePaulo & Coleman 1986), another one found a lowered speech rate due to lengthened pauses (Biersack et al. 2005), and Smith (2007) reported an increased f0-range and segmental modifications leading to a more emphatic style in French. Regarding prosodic cues when addressing elderly interlocutors in English, Kemper et al. (1995) reported a slower speech rate due to prolonged vowels and more frequent pauses in spontaneous speech of a map task with a physically present interlocutor. Although expected, they did not find exaggerated pitch ranges. For German, Thimm et al. (1998) also reported more pauses as well as more variation in intonation in spoken explanations of an alarm clock when a positively stereotyped elderly person was addressed as opposed to a young adult.

Besides different interlocutors, our design contains a conversational setting with background white noise. Speech production in noisy environments leads to increased f0-values and f0-range, increased signal amplitude, increased word or segment durations, and spec-

[22] The vast majority of this section is literally taken from study I published as Huttenlauch et al. (2021).

tral changes such as smaller spectral slope (van Summers et al. 1988, Junqua 1993, 1996, Jessen et al. 2003, Davis et al. 2006, Garnier et al. 2006, Varadarajan & Hansen 2006, Lu & Cooke 2008, Folk & Schiel 2011, Zollinger & Brumm 2011, Landgraf et al. 2017). These noise-dependent changes are summarised under the term Lombard speech, tracing back to Étienne Lombard who first described the noise-dependent increase in speech amplitude for French (Lombard, 1911; as cited in Zollinger & Brumm, 2011). Lombard speech is also described as a source of inter- and intra-speaker variability (Stanton et al. 1988, Junqua 1993, Jessen et al. 2003). For a review on the neural mechanisms of the Lombard effect in humans and animals see Luo et al. (2018).

The intra-individual level in comprehension, as the inter-individual level, was addressed in the study by Cangemi et al. (2015) on the encoding and decoding of three linguistic focus structures (broad, narrow, and contrastive focus). Listeners categorised productions by five speakers into the three focus conditions. Each speaker differed with regard to combinations and degrees of prosodic cues used. The results show that listeners' percentage of correct responses differs intra-individually between the productions of the five speakers (Cangemi et al. 2015: 138f.). At the same time, speakers' productions vary in how well they are perceived by individual listeners. This suggests "the existence of interacting speaker- and listener-specific strategies" (Cangemi et al. 2015: 141).

2.3.3 Situational (in)dependence of prosodic disambiguation

This final section of the introduction brings together the previously introduced aspects of prosody (i) as a means to disambiguate and (ii) as a channel of situational variability, more specifically the (in)variability induced by the conversational situation (the type of interlocutor and the absence/presence of background white noise, context). It addresses the question whether disambiguating prosody and the contextual situation are dependent in any sense. More precisely, which function disambiguating prosody fulfils in the situation in which it is being produced: is prosody produced mainly for the interlocutors or for the speakers themselves[23]. This question goes in line with the question in how far the prosodic realisation of an utterance is dependent or independent from the actual situation in which it is being produced. If there is a rather direct link between syntax and prosody, disambiguating prosody should be "automatically" present in any case – independent of the situation. However, if prosody is less automatically connected to the structural properties of the utterance, but used in a more controlled way by the speaker to support the interlocutor's parsing of an ambiguous utterance, then the use of prosody may vary more depending on the situation and/or properties of the interlocutor. The latter assumption can be subsumed under models of "situational dependence" and the former under models of "situational independence".

Situationally dependent models, on the one hand, assume that prosodic realisations de-

[23] The following paragraphs are literally taken from study I published as Huttenlauch et al. (2021).

pend on the actual communicative situation. Prosodic cues are only necessary, and therefore expected, if the speaker is aware of the ambiguity and the possible misunderstanding of the interlocutor and if the context does not provide other, non-prosodic disambiguating cues. Those other cues can be linguistic or non-linguistic. Models assuming situational dependence of prosodic realisations predict that speakers use prosody differently when addressing interlocutors with different needs or, more generally, that speakers use prosody differently in different communicative or contextual situations. Situational dependence supports the view that prosody is realised for the interlocutor – to help them derive the intended meaning. In that sense, the speech planning mechanism would be required to foreshadow for any stage of the upcoming speech whether it is in fact ambiguous and lacks disambiguating cues of any kind in order to evaluate the necessity for disambiguation (Speer et al. 2011: 87f.). Furthermore, in a strict interpretation of context dependence of prosodic cues, their occurrence would then be more likely in situations, which do not provide any disambiguating information and, thus, they should appear rather inconsistent and infrequent (Speer et al. 2011: 36f.). This inconsistency, however, would render them unreliable for comprehension (see e. g., Kraljic & Brennan 2005: 196).

Situationally independent models, on the other hand, assume that prosodic realisations are largely independent from actual interlocutors or the communicative/contextual situation. Under such accounts prosodic cues are produced automatically and their realisation is affected by grammatical factors such as phrase structure, information status, or phonological length (Kraljic & Brennan 2005, Speer et al. 2011: 37). In this view, prosody is not primarily realised for the interlocutor, but more automatically "for" the speaker. Since prosodic cues are interpreted as depending on linguistic factors, their occurrence should be rather common and frequent, which would make them reliable for comprehension (Kraljic & Brennan 2005, Speer et al. 2011).

Overall, there is evidence supporting the situationally dependent (Allbritton et al. 1996 and Snedeker & Trueswell 2003) and the situationally independent account (Schafer et al. 2000, Kraljic & Brennan 2005, Wagner 2005, and Speer et al. 2011) on the disambiguating function of prosody. Detailed descriptions of the exemplar studies are given in the corresponding section of study I and are not repeated here to limit repetitions. Differential findings might be related to task differences (e. g., instruction, presence of an interlocutor, degree of interaction between speaker and interlocutor, potential for misunderstandings, awareness of the ambiguities) or the complexity or length of the to-be-produced structures (e. g., longer utterances in Speer et al. 2011 than in Snedeker & Trueswell 2003). For a detailed discussion on these differences see Kraljic & Brennan (2005), Snedeker & Trueswell (2003), and Speer et al. (2011). The latter also discuss the option of an intermediate position between situational dependence and independence (Speer et al. 2011: 37f.).

3 Aims of this thesis

Prosody is highly relevant for the resolution of structural ambiguities and for transmitting communicative aspects including individual features of the speaker and the situation in which the communication takes place. To evoke prosodic adaptations to different conversational contexts (*context*) we elicited productions with a within-subject manipulation of context in a referential communication task (studies I, II, and IV). Context had five levels and involved interlocutors in three age groups (child, young adult, elderly adult) with German as L1 in the absence of background white noise, the young adult with background white noise, and a young adult without German as L1. The interlocutors were audio-visually present on a screen. We considered (individual) prosodic variability at different levels: (i) between-group variability in the productions of young and older adult speakers (study II) and, within the age group of young adults, (ii) inter-individual and (iii) intra-individual variability (studies I, III, and IV). Prosodic disambiguation was studied with a focus on German name sequences of three names (*coordinates*) in two conditions: without (Name1 and Name2 and Name3) and with ([Name1 and Name2] and Name3) internal grouping of the first two names in two production studies (studies I and II) and one comprehension study (study III). The study of coordinates was complemented with production data of locally ambiguous sentences with a case-ambiguous first noun phrase (NP1, study IV). The thesis focuses on the following three aims.

The **first aim** is to improve our understanding of the form of prosodic grouping studied in the distribution of the three prosodic cues, f0-movement, final lengthening, and pause, involved in the ambiguity resolution in the case of coordinates in German. This aim addresses two sub-points. With the **first sub-point** we aim to replicate the involvement of the three prosodic cues, f0-range, final lengthening, and pause, in the ambiguity resolution of coordinates and to extend them to older adult speakers. With the **second sub-point** we aim to deepen the insights of the distribution of prosodic cues within the utterance addressing the question whether the cues are globally or locally used in production and comprehension.

The data were discussed in terms of the Proximity/Similarity model by Kentner & Féry (2013). This model makes predictions for structures with internal grouping compared to the baseline without internal grouping. For elements inside a group, the model predicts smaller prosodic cue values (weakening of the prosodic boundary on Name1, *proximity*), while for elements across groups, larger prosodic cue values are predicted (strengthening of the prosodic boundary on Name2, *anti-proximity*). Data of two production studies (studies I and II) and one comprehension study (study III) are included in the analysis of the first aim.

The **second aim** is to deepen our knowledge of the relationship between prosody and syntax by investigating whether the close link between prosody and syntax is maintained in different conversational contexts or whether the aforementioned disambiguating prosodic cues are modified when speakers address different interlocutors with possibly different needs.

If disambiguating prosody is ‘automatically’ present independent of the context (or the situation), we interpret this as a rather direct link between prosody and syntax (situational independence of disambiguating prosody). However, if disambiguating prosody is less automatically connected to the structural properties of the utterance, but used in a more controlled way by the speaker to support the interlocutor’s parsing of an ambiguous utterance, then disambiguating prosody appears as rather situationally dependent. To study this aim, productions of coordinates addressed to different interlocutors were analysed.

The **third aim** is to discuss possible generalisations of the findings on prosodic grouping. The aim is divided into three sub-points. In the **first sub-point**, we discuss structured variability and how it supports a phonological category of grouping. In the **second sub-point**, we discuss whether a relative character of the strength of prosodic cues in grouping conflicts with reliable decoding of early cues. In the **third sub-point**, we come back to the starting point looking for prosodic disambiguation in another syntactically ambiguous structure, namely data on locally ambiguous sentences (study IV).

4 Experimental Work

In the following section, we address methodological considerations of production and comprehension regarding the investigation of variability. With respect to production, we present thoughts regarding the controlled elicitation of variability in production data and the procedure of the production studies. With respect to comprehension, we reflect about a researchers' access to the intention of the speakers via a so called perception check and about a researchers' access to the time domain of comprehension of disambiguation with a gating paradigm.

4.1 Methodological considerations: Production

Production was targeted in the studies I, II, and IV. When eliciting speech productions, researchers try to control speakers' productions with regard to the aspects under study by manipulating the items accordingly. To do so, researchers create items in a way that triggers speakers to produce speech samples that contain the features they want to study. When researching natural speech production, it is often impossible to explain to speakers in detail what the researchers are interested in. On the one hand, in exploratory research, fine-grained details of certain aspects can be still unknown. On the other hand, researchers may be interested in speaker-specific and explicitly non-normative productions. In both cases, it is not in the spirit of independent research to recite the items to the speakers, which are therefore only presented in written form or triggered in another non-verbal way. The study of prosody bares the additional complication that for some prosodic phenomena written equivalents are lacking. At this point, we do not want to distract from the fact that scripted speech collected in the laboratory is in its naturalness incomparable to real spontaneous speech. One can argue, that "this is compensated for by the control over linguistic and contextual variables" offered by formal test procedures (Peppé et al. 2000: 332). Since the linguistic phenomena studied in this thesis do not occur very frequently in spontaneous speech, we have opted for semi-spontaneous speech, which makes it possible to collect enough comparable repetitions for statistical analysis. Another challenge is the controlled study of a research topic that is variable in its very name: variability. In the next section, we address the challenge of controlled elicitation of variability.

4.1.1 Contexts & referential communication

It is a challenge to capture variability in prosodic productions in a controlled way, as there are many causes of variability. Here, we focus on variability within the same language, between age groups and between as well as within individuals.

Inter-individual variability can be studied in data containing productions of the same speech material produced by different speakers in a comparable setting. Apart from speech samples that are comparable between individual speakers, no special design is required. However, for the study of intra-individual variability, a source that triggers individual variability is needed. In the context of the present studies, we aimed to elicit variability in response to different conversation partners (interlocutors) and the absence or presence of background white noise.

The creation of a methodological design requires various decisions including: (i) type and number of interlocutors, (ii) virtual or physical presence of interlocutor(s), (iii) degree of control on the setting in which interlocutor and speaker interact, and, along with it, (iv) type of speech to be recorded. It is clear that the above-mentioned aspects are mutually dependent, as the decision on the type of interlocutor can determine whether the interlocutor can be physically present or not. With respect to the presence of the interlocutor(s), both options have their advantages and disadvantages. An advantage of a physically present interlocutor is the larger naturalness of the communicative situation. For aspects of naturalness, spontaneous speech data are desirable. However, spontaneous speech is highly variable and each person present in the recording situation, including the experimenter, possibly contributes to further sources of variability, simply by day-to-day changes in mood, clothing, and the rendering of stimuli if they are presented orally[24]. Furthermore, if the data recording extends over a longer period of time (e. g., months or years), it becomes more difficult to avoid changes in the performance of a physically present interlocutor due to habituation and natural changes such as ageing, especially in the case of a child interlocutor. Not to mention the challenge of coordinating a larger number of interlocutors for each recording session in terms of time and availability. These aspects accumulate to the disadvantages of a physically present interlocutor. In contrast, a recording situation with a virtual interlocutor permits to expose each participant to the same scene and to record in different locations (cf. advantages listed by Enzinna & Tilsen 2019). Further advantages are, that the data are easier to control, all speakers hear the same speech sample (Enzinna & Tilsen 2019: 30). The recording situation remains artificial, however, the speech is recorded in an interaction-like setting. The speech responses are triggered by audio input compared to written stimuli, which makes them closer to a natural conversation. With respect to the setting in which interlocutor and speaker interact, a referential communication task provides the possibility of a controlled and semi-interactive setting. It allows to control the verbal response in terms of structural complexity, lexical, and segmental content (Leinonen & Letts 1997: 54). The term referential communication refers “to communicative acts [...] in which some kind of information is exchanged between two speakers” (Yule 1997: 1). The need to communicate is created by a kind of barrier that prevents one participant from seeing materials that might be available to the other, so that the latter has to provide knowledge to the first participant (Leinonen & Letts 1997: 54). A barrier can be a screen between the speakers or different

[24] The influence of the experimenter on participants’ behaviour is still little explored. The influence of the experimenter’s identity (race and accent) on children’s waiting behaviour was studied by Pierre et al. 2023.

rooms that make it necessary to communicate via telephone or another type of device. The materials include pictures, maps, or game boards of which there are two similar versions with small differences. Interactive/cooperative communication is encouraged by tasks such as finding differences between the pictures (e. g., Diapix in Van Engen et al. 2010, Hazan & Baker 2011), describing directions on a map (e. g., map task in Anderson et al. 1991), or helping the partner to complete their board (e. g., in Enzinna & Tilsen 2019, Hwang et al. 2015). The tasks can be combined with more or less restrictive instructions with respect to the formulations or wordings to be used by interlocutor and speaker. This allows researchers to influence the type of the elicited speech data.

We will now turn towards the methodological design used in the production studies I, II, and IV. With respect to the previously discussed decisions, we chose (i) four female interlocutors differing in age and first language in situations without and with background white noise that were (ii) virtually present in pre-recorded videos, (iii) in a semi-interactive referential communication task, (iv) triggering the production of two different types of responses from the speakers by requesting missing information that was accessible to the speakers. Our aim was to explore the effects of the type of interlocutor and the absence/presence of background noise on the variability between and within speakers' productions of disambiguating prosody. The speakers' task was to produce the study-specific items in response to the interlocutors' trigger question: in studies I and II coordinated name sequences and in study IV locally ambiguous sentences[25]. Our design contained five different contextual settings, that will henceforth be referred to as *contexts*. The contexts involve four different female interlocutors: a young adult (YOUNG), a child (CHILD), an elderly adult (ELDERLY), a young non-native adult speaker of German (NON-NATIVE), and the young adult in a noisy environment (with background white noise, NOISE). In each context, the speakers enroled in a referential communication task with the current interlocutor. The coordinates were elicited in all five contexts, while the locally ambiguous sentences were elicited in the YOUNG and ELDERLY context only.

The interlocutors presented themselves audio-visually to the speakers in two video clips in each context. Presenting the interlocutors in pre-recorded videos limited influencing factors and made us independent of the availability of our interlocutors. Furthermore, it allowed us to collect data with the same interlocutors over several months in 2018 and 2019 and to keep the differences in recording settings between speakers minimal. One video contained a personal introduction of the interlocutor and the other the instructions for the task. To

[25] The structure of the items is presented in more detail in the respective studies. For a better understanding of the design, both types of items will be described here only briefly. In the studies on coordinates, the items consisted of a sequence of three names coordinated with *und* ('and') displayed on the screen. The internal grouping was indicated by parentheses around the grouped names. The absence of parentheses displayed the condition without internal grouping. In the study on locally ambiguous sentences, the items consisted of written transitive verb-second main clauses and corresponding black-and-white line drawings depicting agent, patient, and the action.

minimise the influence of non-person-specific factors, the interlocutors in the videos wore similar monochrome clothing and all sat in front of a light neutral background (cf. Figures 2, 3, 4, and 5; *note*: faces were not pixelated in the experiment). During video recording, the interlocutors were asked to look into the camera and speak with few gestures and little movement.

Figure 2: YOUNG and NOISE interlocutor.
Fictional persona Hannah (24 years old), originally from Eberswalde (Brandenburg area). Hannah moved to Potsdam to study biology, lives there in a shared flat and likes the parks in Potsdam. White noise was presented auditorily in the NOISE context.
(True) demographic data: 21 years old, first language: German, origin and current residence in the Berlin-Brandenburg area.

Figure 3: CHILD interlocutor.
Fictional persona Carlotta (6 years old). Carlotta was born and raised in Potsdam. Carlotta goes to school and her parents pick her up from there. She likes horse riding and is good at swimming.
(True) demographic data: 7 years old, first language: German. Current residence in the Berlin-Brandenburg area since the age of 4.

Figure 4: ELDERLY interlocutor.
Fictional persona Maria Korbmacher (82 years old). Lives currently in Potsdam, for two years in an old-age home with her husband. Maria Korbmacher is a retired school teacher and starts to forget things from time to time.
(True) demographic data: 89 years old, first language: German, origin and current residence in the Berlin-Brandenburg area.

In the introduction video, each interlocutor introduced her character in a few sentences and told about her fictional demographic background, including information about her name, age, origin, occupation, place of residence, and some interests (see texts next to the Figures 2, 3, 4, and 5; the exact wording of their presentations is given in Appendix A of study I.).

Figure 5: NON-NATIVE interlocutor.
Fictional persona Zsófi (26 years old). Zsófi is an exchange student living in Potsdam that started to learn German one year ago. Zsófi lives in a shared flat and enjoys doing sports. (True) demographic data: 32 years old, first language: Hungarian, origin: Hungary, current residence in the Berlin-Brandenburg area, in Germany for < 3 years.

The (true) demographic data of the person behind the fictional interlocutor are also given in the captions next to the Figures 2, 3, 4, and 5.

In the instruction video, the interlocutor explained the upcoming task. The text of the instruction differed between the studies on coordinates and the study on locally ambiguous sentences. For the elicitation of coordinates, the interlocutor instructed the speaker to utter the name sequences in a way that would allow the interlocutor "to understand as rapidly and accurately as possible who is coming together". The wording of the instruction for the task was nearly the same for all contexts but the adult interlocutors addressed the speakers in the formal way using German *Sie* (you), while the child used the informal *Du* (you), which reflects prescriptive German pronoun use. The complete wording of the instructions is given in Appendix A of study I. For the elicitation of locally ambiguous structures, the speakers were asked to utter the sentence written below the highlighted one of two pictures in a way that would allow the interlocutor "to understand as rapidly and accurately as possible" what they see. As only adult interlocutors were present in study IV, speakers were addressed by formal *Sie* (you) throughout. The exact wordings of the instructions are given in Appendix A of study IV. The durations of the instruction videos are given in Table 1. For the locally ambiguous sentences, all five contexts were recorded, although only two contexts were used in study IV (the durations for the unused contexts are given in smaller font size).

For the noise context (NOISE), the young interlocutor was exposed to the same white noise that was later played to the speakers in the recording session. She heard the noise via in-ear headphones, which were invisible in the video clip. Instead of presenting herself again, she reminded the speakers of who she was and that they should do the task with her again. Furthermore, she commented on the noise in the background and repeated the instruction for the task, adding to the usual wording that she was going to be the interlocutor *again*.

To create the referential communication, each trial began with a question produced by the interlocutor of the current context (henceforth referred to as *trigger question*). This triggered the production of the study-specific item and reminded the speaker of their current interlocutor. For eliciting the coordinates, the trigger question was *Wer kommt?* ('Who is coming?') with a mean duration of 0.94 seconds (*SD*: 0.028 seconds, for individual values see Table 1). The production of locally ambiguous sentences was triggered by the question *Was sehen Sie?* ('What do you see?') with a mean duration of 1.03 seconds (*SD*: 0.073 seconds, for individual values see Table 1). For the two contexts used in study IV (YOUNG and the ELDERLY interlocutor), the mean duration of the trigger question is 0.98 seconds (*SD*: 0.074 seconds).

Table 1: Durations of videos and trigger questions of the five contexts (in seconds). Numbers in small font size mark videos that are not included in the presented studies.

	YOUNG	CHILD	ELDERLY	NON-NATIVE	NOISE
introduction video	28	18	41	30	17
instruction video					
coordinates	21	19	35	24	22
locally ambiguous sentences	22	23	32	28	21
trigger question					
who is coming?	0.892	0.956	0.961	0.932	0.953
what do you see?	0.931	1.134	1.036	1.034	1.010

In summary, the aim was to study variability in speakers' prosodic realisations in a controlled way. Variability was elicited in a referential communication task with different interlocutors in the absence or presence of background white noise (contexts). The interlocutors were pre-recorded so that all speakers received the same audio-visual input, which enables comparability of the data collected. Each trial began with a question from the interlocutor that triggered the speaker to respond producing the item shown on a screen.

4.1.2 Parallel experimental procedure

Closely related to the decision for an experimental design that enables comparability between participants is the question of the experimental procedure. A uniform experimental procedure makes it possible to compare the results of different studies with each other. We used the same procedure to collect data in the production studies independent of the type of stimuli (coordinates in studies I and II and locally ambiguous sentences in study IV). The speakers were asked to utter the stimuli to the interlocutors in a kind of question-answer dyads. The interlocutors asked questions (trigger question) and the speakers answered with the given item. There was no further turn by the interlocutor (i. e., no queries).

The studies took place in a sound-attenuated recording booth in the acoustics laboratory at the University of Potsdam. Speakers were seated in front of a screen wearing a headset with over-ear headphones and a microphone. Speakers were presented with a referential communication task including contexts with different interlocutors and the absence/presence of noise (contexts, further details in the previous section). Contexts were presented in blocks, always starting with the YOUNG context, which served as a baseline in the analysis. Each block started with a presentation of the interlocutor and the instruction for the task, both presented audio-visually on the screen spoken by the interlocutor. In the test phase, the video of the interlocutor was replaced by the presentation of the target items. The interlocutor addressed the speakers auditorily with the trigger question.

In the studies on coordinates, the test phase consisted of five blocks, corresponding to the five experimental contexts (YOUNG, CHILD, ELDERLY, NON-NATIVE, NOISE) in which speakers were asked to produce the coordinated name sequences in the two conditions: without or with internal grouping. Each item was presented in each context once (study I) or twice (for older speakers in study II), hence, speakers produced each item five or ten times, respectively. The YOUNG context, as the baseline context, was always presented first, the other four contexts were presented in randomised order. In each context, items were pseudo-randomised using different lists (a more detailed description of the randomisation is given in study I). Each block started with the two video clips and during the test phase, for each trial, speakers saw a fixation cross on the screen while they heard the trigger question *Wer kommt?* ('Who is coming?'). After 1000 ms, the fixation cross was replaced by the visual presentation of a name sequence in one of the two conditions, which stayed on the screen for 5000 ms. The task was to produce the name sequence in a way that would allow the interlocutor to know who was coming together: all three persons together or the first two persons together with the third person alone. The sound recording started together with the presentation of the name sequence and continued for 1000 ms after the names disappeared, see Figure 6 left panel.

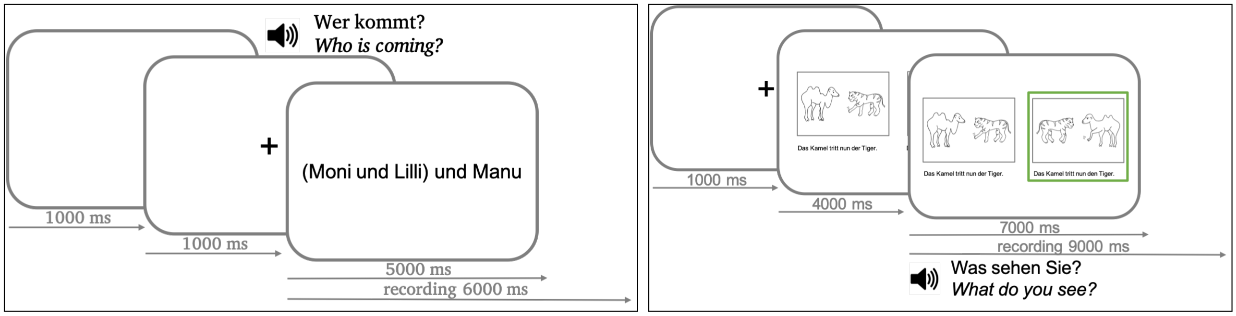


Figure 6: Experimental procedure and timing of one trial for the elicitation of coordinates (left panel) and locally ambiguous sentences (right panel).

In the study on locally ambiguous sentences, the test phase consisted of two blocks, the young interlocutor (baseline) presented first and the elderly interlocutor in a second block. Parallel to the procedure on coordinates, each block started with the two video clips of the current interlocutor. Then, each trial started with a fixation cross on the screen that was replaced after 1000 ms by the target and the foil picture with the corresponding sentences written below the pictures. After a preview time of 4000 ms, the target picture was highlighted with a green frame along with the auditory presentation of the question *Was sehen Sie?* (‘What do you see?’), see Figure 6 right panel. The task was to produce the target sentence in a way that would allow the interlocutor to identify the target picture “as rapidly and accurately as possible”; the speakers were told that the interlocutor would see the same pictures (without the sentences) and had to find the matching picture. Speakers produced each item twice, once addressing the young and once the elderly interlocutor.

4.2 Methodological considerations: Comprehension

Similar to inter-individual variability in production, inter-individual variability in comprehension can be studied in data containing responses to the same speech stimuli provided by different listeners in a comparable setting. For the study of intra-individual variability, potential sources to trigger individual variability are needed. Potential sources are any kind of describable variability in the stimuli including (i) stimuli produced by different speakers, versions of stimuli differing in (ii) degree or (iii) combinations of prosodic cues. In the comprehension study (study III), we analysed in an exploratory way, whether the coordinate productions of different speakers lead to different responses in comprehension. Binary responses were collected in a gating paradigm.

As mentioned earlier, in the study of prosodic boundaries, production and comprehension are intertwined. Researchers try their best to control speakers’ productions with regard to their research interest. Speech productions are variable in many dimensions. Some of this variability is random and some systematic. In the case of the coordinates (studies I and II), the stimuli were intended to elicit two different conditions: (i) three names without internal grouping and (ii) three names with the first two names grouped together. The written stimuli contained parentheses around the names to be grouped to indicate the intended grouping. In the analysis, we were interested in the prosodic realisation of the two conditions (ungrouped vs. grouped). More specifically, we focused the analysis on the three prosodic cues: f0-range, final lengthening, and pause. Further, the speech stimuli were elicited addressing varying interlocutors. Measurable differences between productions addressed to different interlocutors would gain in meaning if perceivable by listeners. In order to check whether naïve listeners are able to perceive the conditions and addressed interlocutors, we conducted a perception check. Listeners judgements were used to “pre-process” the productions before running acoustic analyses.

4.2.1 Perception check

The perception check was conducted on the production data of studies I and II, to separate productions where the grouping structure was perceived in the intended way by naïve listeners from those where it was not. This separation allowed us to only include the first ones into the analysis on the use of f0-range, final lengthening, and pause between coordinates without and with internal grouping (cf. Peppé et al. 2000: 322 for a similar selection of productions for analysis). *Intended* refers to the conditions without or with parentheses, respectively, presented to the speakers on the screen in the production study. As researchers we do not have access to the speakers' intentions while speakers conduct the study. For this reason, we lack knowledge about whether a stimulus coded as grouped was produced intended as grouped by a particular speaker. We are aware of the intrusion into the dataset that we make by excluding individual productions and that, by this, our researchers' degrees of freedom influence the result (Roettger 2019). For a different research question, the exclusion of data points may be less legitimate.

After each data collection (studies I and II), all recordings were auditorily presented to naïve listeners (who had not taken part in the production experiments). Answers were elicited in a two-alternative forced-choice decision task. Listeners were asked to indicate for each production the perceived condition. To this end they were given two pictograms with three persons each, one pictogram per condition (cf. Figure 7, picture A without and picture B with internal grouping).

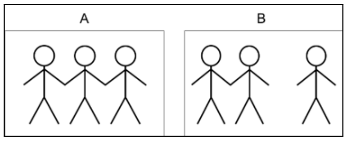


Figure 7: Pictograms used in comprehension studies on coordinates, panel A for the condition without internal grouping, panel B for the condition with internal grouping.

The production data of the young and older age groups were rated separately. The recordings were distributed across different lists with 147 to 267 items. Each listener judged one list and each list was judged by seven or eight listeners. Each list contained the complete productions of several speakers plus a subset of productions from various speakers. The subset was judged by all listeners of the corresponding perception check and constituted a semi-random sample of all productions within age-group, containing at least one production of each speaker and of each context.

The perception check was run in two versions: (i) in presence with a paper-and-pen version and several listeners in the same room and (ii) as a web-based study. The main task was in both versions the same: to listen to each production and to choose the matching one of two pictograms given in Figure 7 (i. e., to identify the condition). The first version contained the additional task to indicate the most probable addressee the name sequence was uttered to (young adult, child, elderly, non-native, in noise; i. e., to identify the context). In the in-person version there was no practice phase, while the web-based version started with four practice items (as there was no possibility to ask questions). Independent of testing mode, the perception check started with the presentation of the subset, followed by the remaining productions of individual speakers, presented in blocks. Each of the lists contained some productions twice, those, which were part of the subset and the following productions. In the case of repetitions, only the first judgement was considered.

In the analysis of the perception check, the exclusion threshold for individual productions was set to a hit-ratio of 2 *SD* below the mean ratio, as suggested by standard assumptions on the exclusion of data points (e.g., Howell et al. 1998). First, for each listener, we counted the number of congruent rates (i. e., correct identifications of the intended grouping/condition and context, referred to as *hit-rate*). If, for a given listener, the hit-rate was 2 *SD* below the mean hit-rate of all listeners, all ratings of this listener were excluded. Second, for each individual production, we calculated the ratio of the hit-rate to the number of total rates.

$$\textit{hit-ratio} = \frac{\sum \textit{hit-rate}}{\sum \textit{total rates}}$$

We used the ratio instead of the absolute hit-rate since individual productions were rated by a varying number of listeners. We then calculated the mean ratio of all productions as well as the standard deviation. Only the ratings of condition influenced the exclusion of individual productions: productions for which the ratio of the hit-rate was more than 2 *SD* below the overall mean ratio of hit-rates were excluded from further analyses.

4.2.2 Gating paradigm

The speech stream is processed online. In addition to investigating whether listeners perceive prosodic disambiguation between two conditions in auditory stimuli in the intended way, the question of how early in the stimulus listeners can reliably make this distinction is another topic to investigate. In the perception check on coordinates, listeners were asked to choose between the two conditions after hearing a complete three-name sequence. However, the two conditions differ in the structure of the first two names. Therefore, the question arises, whether listeners can distinguish between the two conditions, even when they hear the initial part of the sequence. And further, what minimal number of segments of the three-name sequence is sufficient for listeners to reliably distinguish between the two conditions? To investigate this, a speech stimulus can be played repeatedly with its duration increasing with each iteration. At the beginning it is a short part, in the next iteration the duration is increased. After each iteration, the listeners are asked to guess the stimulus presented and indicate how confident they are about their guess. Grosjean (1980) introduced such an experimental paradigm that allows to study the online processing of speech stimuli as “gating paradigm”. This paradigm “is particularly useful for assessing how much acoustic/phonetic information is needed for a word to be identified correctly” (Grosjean et al. 1994: 597). In their studies, Grosjean (1980) and Grosjean et al. (1994) used the gating paradigm to investigate spoken word recognition, but Grosjean (1980) also mentions its potential use “to study such issues as the effect of prosodic cues and perceptual strategies on sentence processing” (Grosjean 1980: 267). The varying duration of the stimulus is referred to as “presentation time” or “gate” (Grosjean 1980: 267). In the area of sentence processing, the gating design has been used, among other, to investigate the prediction of sentence length

(Grosjean & Hirt 1996) and the distinction between questions and statements (Petrone & Niebuhr 2014).

We used the gating paradigm in study III to investigate at what time in the coordinate structure listeners are reliably able to predict the structure of the presented coordinate sequence. The gates were extended syllable by syllable, starting with one syllable at gate 1. Gate 2 contained two syllables (the complete first name), gate 3 three syllables and so on until gate 7, that contained the complete sequence. Figure 8 shows the timing of the first two gates of a stimulus. All seven gates belonging to a coordinate sequence were presented one after the other with increasing duration.

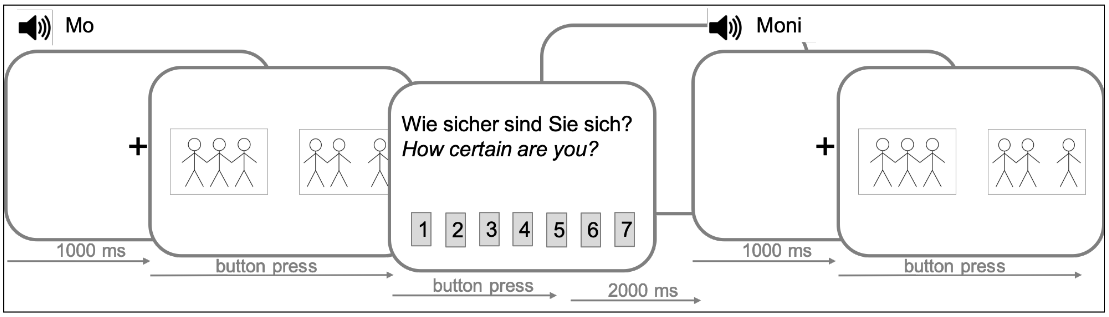


Figure 8: Experimental procedure and timing of trials in the gating paradigm with gate1 and the beginning of gate2 of a coordinate stimulus.

The study took place in the same setting as the production studies in a sound-attenuated booth in the acoustics laboratory at the University of Potsdam. Listeners participated in individual sessions. They were seated in front of a wide screen wearing a headset with over-ear headphones. Answers were elicited in a two-alternative forced-choice decision task and given via button press. The two alternatives were exemplified in two pictograms showing three persons without or with internal grouping (cf. Figures 7 and 8). After their decision for one of the two conditions, listeners were asked about their confidence on a 7-point scale. The test phase was preceded by a practice phase with two gated utterances, one for each condition.

5 Study I

Production of prosodic cues in coordinate name sequences addressing varying interlocutors[26]

Abstract

Prosodic boundaries can be used to disambiguate the syntactic structure of coordinated name sequences (*coordinates*). To answer the question whether disambiguating prosody is produced in a situationally dependent or independent manner and to contribute to our understanding of the nature of the prosody-syntax link, we systematically explored variability in the prosody of boundary productions of coordinates evoked by different contextual settings in a referential communication task. Our analysis focused on prosodic boundaries produced to distinguish sequences with different syntactic structures (i.e., with or without internal grouping of the constituents). In German, these prosodic boundaries are indicated by three major prosodic cues: f0-range, final lengthening, and pause. In line with the Proximity/Anti-Proximity principle of the syntax-prosody model by Kentner & Féry (2013), speakers clearly use all three cues for constituent grouping and prosodically mark groups within and at their right boundary, indicating that prosodic phrasing is not a local phenomenon. Intra-individually, we found a rather stable prosodic pattern across contexts. However, inter-individually speakers differed from each other with respect to the prosodic cue combinations that they (consistently) used to mark the boundaries. Overall, our data speak in favour of a close link between syntax and prosody and for situational independence of disambiguating prosody.

5.1 Introduction

Syntactic ambiguities, like the internal grouping of sequences, see example (23), are a common phenomenon in many languages. In spoken language, such ambiguities can be resolved by prosodic phrasing, phonetically indicated by modified prosodic cues. If the answer to the question *Who will bring a spare bike for the trip?* were (23), the lexical string alone would not clearly indicate whether there will be one, or two, or three bikes. This is because the phrase has three possible readings depending on the grouping of the coordinated names: one bike could be brought by all three persons together, or two of them could bring one bike together and another person brings a second bike, or each of them could bring their own

[26] An adapted version of this chapter has been published as Huttenlauch, Clara, Carola de Beer, Sandra Hanne & Isabell Wartenburger. 2021. Production of prosodic cues in coordinate name sequences addressing varying interlocutors. *Laboratory Phonology* 12(1). 1–31. doi:10.5334/labphon.221. This study is written in American English.

bike, respectively.

(23) Caro and Lea and Jana.

The syntactic grouping of the names in (23), however, can be disambiguated by prosodic cues which lead to the perception of a boundary that will be referred to as prosodic boundary (Frazier et al. 2006, Holzgrefe-Lang 2017, Kentner & Féry 2013, Wagner 2005), marking the intended syntactic grouping. As such, there is a close link between syntax and prosody from the perspective of the listener/interlocutor and the speaker as well. At the phonetic level, in German, the language studied here, three main cues in two domains are used for prosodic boundary marking in spoken language production: in the tonal domain, pitch change, mostly realized as a rise in fundamental frequency (f0) and in the durational domain, lengthening of the syllable or segment immediately preceding the boundary (final lengthening) and pause at the boundary (for German: Gollrad et al. 2010, Kentner & Féry 2013, Peters et al. 2005, Petrone et al. 2017). The pitch change is operationalized as fundamental frequency, abbreviated to f0 and used interchangeably with pitch in this paper, even though pitch refers to the perceptual correlate and f0 to the acoustic measure. Pitch changes have been shown to be relevant in coordinates already in the seminal works of Ladd (1986) for English and van den Berg et al. (1992) for Dutch.

These examples illustrate that syntactic structure and prosody are closely related to each other. However, it is still a matter of debate, how this link is represented in the linguistic system: Whether the phonology-syntax mapping follows a fixed, categorical, phonological hierarchy in which certain syntactic categories are mapped to certain phonological units, such as the phonological phrase or the intonational phrase with particular phonetic characteristics (e. g., Nespor & Vogel 1986), or whether this mapping is more flexible and characterized by rather relative or gradient phonetic correlates (e. g., Wagner 2005). Moreover, it is being discussed which function disambiguating prosody actually fulfils in the situation in which it is being produced, that is, whether it is produced mainly for the sake of the interlocutors or for the speakers themselves (e. g., Speer et al. 2011). The latter case would point towards situational independence of prosodic realizations whereas the first scenario would indicate that prosody production is situationally dependent.

To address the question whether disambiguating prosody is produced in a situationally dependent or independent manner and to contribute to our understanding of the nature of the prosody-syntax link, we will compare prosodic realizations in varying situations between and within individuals. Specifically, we study inter- and intra-individual variability in spoken productions of name sequences in German coordinated with *und* (Engl. *and*), hereafter referred to as *coordinates.* We will focus on two conditions of these coordinates: one without internal grouping referred to as *nobrack* (see 24) and another condition with internal grouping, in which the first two names are grouped together in one sequence and the third name is a separate sequence, referred to as *brack* (see 25). For easier reading, brackets around the grouped names will indicate the structure. Regarding the question of the number of bikes, in (24) there would be one spare bike while in (25) there would be two spare bikes.

(24) without internal grouping (nobrack): [Moni und Lilli und Manu]

(25) with internal grouping (brack): [Moni und Lilli] und Manu

In the following we will briefly introduce previous findings on the functional role of disambiguating prosody (5.1.1, on page 51) and on individual variability in prosody production (5.1.2, on page 53). Then we summarize theories on the prosodic phrasing in coordinates (5.1.3, on page 55).

Function of disambiguating prosody: For the speaker or for the interlocutor

The function of (disambiguating) prosody concerns, in short, the question whether prosody is produced mainly for the interlocutors or for the speakers themselves. This goes in line with the question in how far the prosodic realization of an utterance is dependent or independent from the actual situation in which it is being produced. If there is a rather direct link between syntax and prosody, disambiguating prosody should be "automatically" present in any case–independent of the situation. However, if prosody is less automatically connected to the structural properties of the utterance, but used in a more controlled way by the speaker to support the interlocutor's parsing of an ambiguous utterance, then the use of prosody may vary more depending on the situation and/or properties of the interlocutor. The latter assumption can be subsumed under models of "situational dependence" and the former under models of "situational independence".

Situationally dependent models, on the one hand, assume that prosodic realizations depend on the actual communicative situation. Prosodic cues are only necessary, and therefore expected, if the speaker is aware of the ambiguity and the possible misunderstanding of the interlocutor and if the context does not provide other, non-prosodic disambiguating cues. Those other cues can be linguistic or non-linguistic. Models assuming situational dependence of prosodic realizations predict that speakers use prosody differently when addressing interlocutors with different needs or, more generally, that speakers use prosody differently in different communicative or contextual situations. Situational dependence supports the view that prosody is realized for the interlocutor–to help them derive the intended meaning. In that sense, the speech planning mechanism would be required to foreshadow for any stage of the upcoming speech whether it is in fact ambiguous and lacks disambiguating cues of any kind in order to evaluate the necessity for disambiguation (Speer et al. 2011: 87f.). Furthermore, in a strict interpretation of context dependence of prosodic cues, their occurrence would then be more likely in situations which do not provide any disambiguating information and, thus, they should appear rather inconsistent and infrequent (Speer et al. 2011: 36f.).

This inconsistency, however, would render them unreliable for perception (see e. g., Kraljic & Brennan 2005: 196).

Situationally independent models, on the other hand, assume that prosodic realizations are largely independent from actual interlocutors or the communicative/contextual situation. Under such accounts prosodic cues are produced automatically and their realization is affected by grammatical factors such as phrase structure, information status, or phonological length (Kraljic & Brennan 2005, Speer et al. 2011: 37). In this view, prosody is not primarily realized for the interlocutor, but more automatically "for" the speaker. Since prosodic cues are interpreted as depending on linguistic factors, their occurrence should be rather common and frequent, which would make them reliable for perception (Kraljic & Brennan 2005, Speer et al. 2011). In the following we introduce some exemplar studies which support either the situationally dependent or the situationally independent account.

Allbritton et al. (1996) addressed the issue of situational in/dependence by testing whether untrained, naïve speakers (vs. trained speakers) would spontaneously use prosody to resolve syntactic ambiguities in various kinds of sentence types. The speakers were instructed to read aloud "as if you were telling someone a story that you wanted them to understand" (Allbritton et al. 1996: 716). It turned out that most naïve and trained speakers did not prosodically disambiguate most of the sentences. Only if the instruction made them aware of the ambiguity and asked them explicitly to produce two different versions, trained speakers used prosody for disambiguation. This can be interpreted as a finding supporting situational dependence. The authors concluded that either the role of prosodic cues for conveying the underlying syntactic structure is limited or laboratory recordings cannot be generalized to real-world settings (Allbritton et al. 1996: 732). Applying a more real-world setting, namely a game-like interactive referential communication task, Snedeker & Trueswell (2003) confirmed the hypothesis that the relation between syntax and prosody is mediated by the context. In their study, naïve participants produced clear prosodic groupings of attachment ambiguities ("Tap the frog with the flower") only in situations in which the context did not provide sufficient information to situationally disambiguate the two possible meanings. The authors concluded that "speakers produce [prosodic cues] primarily when they appear to be necessary for clear communication" (Snedeker & Trueswell 2003: 128). Based on these two example studies, one could conclude that the use of prosody for disambiguation depends on the awareness of the speaker about the ambiguities and/or on whether the actual context made both readings plausible. Prosody is thus mainly used for the interlocutor (cf. audience design, Bell 1984).

In contrast, others found evidence in favour of situational independence: Using a cooperative interactive game-board task, similar to Snedeker & Trueswell (2003), Schafer et al. (2000), and Speer et al. (2011) found no evidence for a dependency of prosodic cues on situational disambiguation or discourse factors using global attachment ambiguities and temporal closure ambiguities. The speakers produced prosodic cues independent of the communicative situation (even in only locally ambiguous sentences), but still with some flexibility or variability in the choice of cues (Speer et al. 2011). This was also confirmed in another interactive game-like study design by Kraljic & Brennan (2005), who also found overall limited effects of the context. In their interactive setting involving attachment ambiguities ("Put the dog in the basket on the star"), speakers produced clear prosodic cues for disambiguation, irrespective of the needs of their interlocutors (i.e., regardless of whether the contextual setting provided disambiguating information or not) and irrespective of whether an interlocutor was present or not. With respect to the function of prosody, they conclude that prosodic marking emerges from the level of planning and articulation, that is, prosody is not produced dependent on the situation but rather automatically and situationally independent. Similarly, for coordinate name sequences, Wagner (2005) found that prosodic boundaries are produced independent of the context and independent of the need of the interlocutors to comprehend, which implicates that prosody is mainly used "for" the speakers themselves, in an automatic manner.

In sum, there is evidence supporting either of the two accounts on the function of prosody. The differential findings might be related to task differences (e. g., instruction, presence of an interlocutor, degree of interaction between speaker and interlocutor, potential for misunderstandings, awareness of the ambiguities) or the complexity or length of the to-be-produced structures (e. g., longer utterances in Speer et al. 2011 than in Snedeker & Trueswell 2003). For a detailed discussion on these differences see Kraljic & Brennan (2005), Snedeker & Trueswell (2003), and Speer et al. (2011). For the option of intermediate positions between situational dependence and independence see also (Speer et al. 2011: 37f.).

Individual variability in prosody production

We now turn from the group level to the individual level and to the question of whether individuals vary in their prosodic realizations. Variability between or within speakers is interesting for two reasons. First, if all speakers reliably use disambiguating prosody to distinguish between coordinates without and with grouping (example (24) vs. (25)), this would indicate a close link between syntax and prosody (e. g., Nespor & Vogel 1986) and situational independence (e. g., Speer et al. 2011). If, on top of the disambiguation, we would also find variability across speakers in how they realize prosodic boundaries, this would add evidence that the link between syntax and phonology is relational rather than categorical (e. g., Clifton et al. 2002, Wagner 2005). Second, if speakers do not reliably disambiguate the different syntactic structures (example (24) vs. (25)) or show within-speaker variability in different contextual settings, this would speak in favour of situationally dependent models–and against a close prosody-syntax link. So far, the issue of inter-individual variability concerning prosodic boundary production in coordinates has been explored only scarcely (one exception, for German, being the work by Petrone et al. 2017, or, for English, the findings by Allbritton et al. 1996, and Lehiste 1973b), and variability induced by different situational contexts has, to the best of our knowledge, not been studied yet.

Regarding *variability between speakers (i.e., inter-individual variability)*, Petrone et al. (2017) found that their speakers differed in how the prosodic boundary was realized in coordinates with internal grouping (i.e., they found multiple types of prosodic boundaries): Only two out of 12 speakers consistently used the same f0-contour, namely a rise. Although production of a rise was also the predominant contour in six further participants, these additionally employed a high plateau. Another three speakers varied between rise, high plateau, and final fall to different degrees and one speaker produced either rises or falls. Using similar three-name sequences, Lehiste (1973b) reported that two (English) speakers differed in how they used durational cues for disambiguation (insertion of a pause vs. lengthening of the coordinating element).

With respect to *variability within speakers (i.e., intra-individual variability)* induced by contextual settings, specifically concerning the type of interlocutor, previous research has focused on differences in prosodic realizations when children, elderly adults, or non-native speakers are being addressed in comparison to young adult native speakers. Most studies take the speech addressed to an adult native speaker of the language under investigation as a baseline for comparisons. For easier reading, we will refrain from mentioning this adult baseline in the following. For example, for attachment disambiguation, Kempe et al. (2010) reported lengthened vowels when English-speaking adults addressed 2–4-year old real or imaginary children and, in addition, found longer pause durations. Other studies investigated intra-individual variability in prosodic information per se (i.e., not focusing on disambiguating prosody): Biersack et al. (2005) reported an increased pitch range and higher f0-maxima as well as longer durations due to the lengthening of vowels in semi-spontaneous speech addressed to a two-year old imaginary child in English. DePaulo & Coleman (1986) also reported longer pauses in spontaneous English speech addressing a 6-year old child. When it comes to prosodic cues in speech addressing a non-native interlocutor, results are inconclusive: while one study involving English speakers found no differences (DePaulo & Coleman 1986), another one found a lowered speech rate due to lengthened pauses (Biersack et al. 2005), and Smith (2007) reported an increased f0-range and segmental modifications leading to a more emphatic style in French. Regarding prosodic cues when addressing elderly interlocutors in English, Kemper et al. (1995) reported a slower speech rate due to prolonged vowels and more frequent pauses in spontaneous speech of a map task with a physically present interlocutor. Although expected, they did not find exaggerated pitch ranges. For German, Thimm et al. (1998) also reported more pauses as well as more variation in intonation in spoken explanations of an alarm clock when a positively stereotyped elderly person was addressed as opposed to a young adult.

As an alternative to the experimental manipulation of type of the (imaginary or real) interlocutor, some studies varied the contextual setting via the presence or absence of noise.

Speech production in noisy environments leads to increased f0-values and f0-range, increased signal amplitude, increased word or segment durations, and spectral changes such as smaller spectral slope (Davis et al. 2006, Folk & Schiel 2011, Garnier et al. 2006, Jessen et al. 2003, Junqua 1993, 1996, Landgraf et al. 2017, Lu & Cooke 2008, van Summers et al. 1988, Varadarajan & Hansen 2006, Zollinger & Brumm 2011). These noise-dependent changes are summarized under the term Lombard speech, tracing back to Étienne Lombard who first described the noise-dependent increase in speech amplitude for French (Lombard 1911; as cited in Zollinger & Brumm 2011). Lombard speech is also described as a source of inter- and intra-speaker variability (Jessen et al. 2003, Junqua 1993, Stanton et al. 1988). For a recent review on the neural mechanisms of the Lombard effect in humans and animals see (Luo et al. 2018).

Prosody of coordinates (in German)

As our study specifically investigates the prosody of coordinates, we will briefly review some relevant models on prosodic phrasing in coordinates which have been proposed in the past. We will focus on the Proximity/Similarity model (Kentner & Féry 2013) since it has been tested with German speakers in similar structures as we use in the current study.

With respect to the question how coordinates are prosodically phrased in English, Taglicht (1998) formulated the "Coordination Constraint". It specifies that the same hierarchical level of intonational boundaries must be applied to all elements at the same syntactic level. Watson & Gibson (2004) argue in their "Left hand side/Right hand side Boundary hypothesis" that the likelihood of the presence of a prosodic boundary depends on the size of the preceding and following constituents because of processing demands: For larger constituents, the speaker needs more refractory time to recover from the preceding constituent and more time to plan the upcoming constituent, respectively. Wagner (2005, 2010) demonstrates that, in coordinate structures, the relative strength of the boundary reflects the level of embedding at the syntactic level and thereby confirms the close match between prosody and syntax in coordinate structures. According to (Wagner 2010: 186) more deeply embedded constituents "are separated from each other by weaker boundaries than constituents that are less deeply embedded" and "constituents separated by relatively weaker boundaries are perceived as grouping together". In a similar vein, Kentner & Féry (2013) developed a model that aims to account for both, processing demands, depending on the constituent size or complexity, and demands of the syntactic structure, depending on the depth of syntactic embedding. Their so called Proximity/Similarity model assumes two principles that "interact to shape the prosody of syntactic structures" (Kentner & Féry 2013: 283) such as coordinated name sequences. Proximity is related to the syntactic constituent structure and states that "adjacent elements which are syntactically grouped together into one constituent should be realized in close proximity" (Kentner & Féry 2013: 282). Similarity is related to the depth of syntactic embedding and refers to the idea that "constituents at the same level of embedding should be realized in a similar way, that is, they should be similar in pitch and duration, irrespective of their inherent complexity"; this principle is comparable to the models of Taglicht (1998) and Wagner (2005, 2010) which assume that elements at the same syntactic level are prosodically matched.

For the structures used in the present study ((24) and (25)), the Proximity/Similarity model makes the following predictions: In coordinates with internal grouping, the principle of Proximity predicts a weakening of the prosodic cues–compared to a coordinate without internal grouping–at an element x if the neighbouring element to the right is part of the same group as x (cf. the first name, i.e., Moni, in (25)). Reversely, Anti-Proximity predicts a strengthening of a prosodic boundary if the right-adjacent element of x does not form a group with x (cf. the second name, i.e., Lilli, in (25)). The Similarity principle predicts that a simplex element at the same level of embedding as a complex constituent is, for instance, lengthened to adjust its duration to the length of a complex constituent. This would be relevant for groupings in which the first name is followed by a complex sequence on the right (cf. Moni und [Lilli und Manu]). As the current study will focus on coordinates with an internal grouping of the first two names (as in (25)), the Similarity principle will not be discussed any further. In coordinates without internal grouping (considered the baseline form, see (24)), all names are expected to be separated by boundaries of the same strength.

With boundary cue weakening, Kentner & Féry (2013) refer to the use of lower pitch and shorter durations on the first grouped element compared to a non-grouped baseline, while a strengthened boundary at the right edge of the grouped element is expressed by a higher boundary tone and longer durations. On the final element of the coordinates (cf. Manu in (24) and (25)), Kentner & Féry (2013) observed neutralization in duration and f0-movement. The findings of increased duration of the word preceding the boundary and a possible pause along with higher pitch at the prosodic boundary have been confirmed in a further study on elicited coordinate productions in German (Petrone et al. 2017) and are in line with results on prosodic marking of syntactic boundaries in spontaneous German speech (Peters et al. 2005). What is still unclear with respect to the Proximity/Similarity account is whether its assumptions also hold situationally independent. Until now, variations in prosodic phrasing of coordinates within speakers across different situations/interlocutors have not been explored sufficiently. In addition, it is unclear to which extend there is variability across speakers and, specifically, whether the speakers differ in how they use and combine the different prosodic cues to mark the prosodic boundaries.

Therefore, in our study, we use coordinate name sequences ((24) and (25)) to replicate the findings of Kentner & Féry (2013) under different contextual settings, that is, in different situations. At the same time, due to the focus on (24) and (25), our data will not be sufficient to adjudicate among the models briefly introduced above in this section. Thus, the current study will not directly contribute to the question as to whether–or to which extent–processing demands or the level of syntactic embedding drive prosodic realizations. Instead, the main focus of our study is on inter- and intra-individual variability and its limits in prosodic boundary production, the relation of the different prosodic boundary cues to one another, and on the situational in/dependence of prosodic phrasing.

In summary, the functional role of disambiguating prosody or its situational in/dependence has been studied by means of the presence or absence of contextual effects on prosodic realizations–but remains largely inconclusive. The fact that participants are aware of an ambiguity, the task setting (reading-out loud vs. interactive setting with real vs. imagined interlocutors), the type of ambiguity (e. g., attachment ambiguities vs. pragmatic ambiguities), the length of the to-be-produced utterance, the type of interlocutor (e. g., child vs. adult), and other contextual factors, such as absence/presence of noise, seem to influence if and how individuals use prosody to disambiguate syntactic structures. We are going to address the question of situational in/dependence by comparing prosodic realizations in varying situations within individuals. In addition, we will explore differences between speakers as they will give us further insights into the nature of the prosody-syntax link.

Aims and hypotheses

In this study we systematically explore inter- and intra-individual variability in the production of prosodic boundaries to get insights into the prosody-syntax relation and the function of prosody. According to situationally dependent models of prosodic phrasing (Allbritton et al. 1996, Snedeker & Trueswell 2003, cf. audience design hypothesis, Bell 1984) we would predict that, if speakers use prosody to disambiguate different syntactic structures at all (e. g., because they are aware of the ambiguity and/or because an interlocutor is present in the communicative situation), they vary considerably in their prosodic productions between interlocutors with different needs. Contrary, according to situationally independent models of prosodic phrasing (Kraljic & Brennan 2005, Schafer et al. 2000, Speer et al. 2011), we would predict that speakers use prosody to disambiguate different syntactic structures in any event, because they are doing it "automatically" during speech planning stages. The prosodic realizations should hence be rather clear between conditions without and with internal grouping (example (24) vs. (25)), and consistent across different interlocutors–although some variability between speakers is also expected (Speer et al. 2011: 88ff.).

We argue that the issues of within-speaker situational in/dependence and of between-speaker in/variability are related to the underlying nature of the prosody-syntax link: If there is a fixed relationship (and dependency) between prosody and syntax, we would predict that speakers "automatically" produce prosodic boundaries in a rather fixed or stable manner to disambiguate the syntactic structure, irrespective of the situation they are confronted with (i.e., situationally independent). If, at the same time, the relationship between syntax and phonology is relational or gradient (e. g., Wagner 2005), we would additionally predict some variability between speakers with respect to the phonetic correlates they employ to disambiguate the syntactic structure.

Our study thus explores the effect of the type of the interlocutor and presence/absence of noise on variability between and within speakers' prosodic boundary realizations in a controlled, semi-interactive setting. Specifically, the speakers are asked to utter coordinates with vs. without internal grouping (such as (25) and (24)). The five different contextual settings will henceforth be referred to as contexts. The contexts involve four different female interlocutors: a young adult (YOUNG), a child (CHILD), an elderly adult (ELDERLY), and a young non-native adult speaker of German (NON-NATIVE) and a noisy environment (the young adult with white background noise, NOISE).

Speakers are completely aware of the intended syntactic grouping of the coordinates and are asked to utter the name sequences in such a way that the different virtual interlocutors can resolve them. We will focus on the prosodic cues f0-range, final lengthening, and pause at/after the first and the second name, as these are known to be modulated to indicate prosodic boundaries. The results will be discussed referring to the Proximity/Similarity model of syntax-prosody mapping (Kentner & Féry 2013). We will describe the interplay and combined use of the prosodic cues of prosodic boundaries and how these are affected by inter- and intra-individual variability as these will contribute to our understanding of the prosody-syntax relation and the functional role of disambiguating prosody.
Our research questions are as follows:

(Q1) Prosodic disambiguation of coordinates: Can the findings of previous studies concerning differences in the use of f0-range, final lengthening, and pause on the first and on the second name in coordinates without internal grouping, such as (24), and with internal grouping of the first two names, such as (25), be replicated?

(Q2) General context-dependent prosodic variability: To what extent do these prosodic boundary cues vary in the five different contexts?

(Q3) Inter-speaker variability: Do different speakers show different patterns in their combined use of the three prosodic cues within contexts?

(Q4) Intra-speaker variability: Do speakers show different patterns in their combined use of the three prosodic cues between contexts?

Regarding Q1, based on the literature outlined in 5.1 (on page 55), we expect speakers to mark the difference between coordinates with (25) and without (24) internal grouping in line with the Proximity/Similarity model by Kentner & Féry (2013). More specifically, we expect a prosodic boundary realized by an increase of final lengthening and an increased f0-range at the right edge of the group (i.e., on the second name), as well as the insertion of a pause after the grouping in (25) compared to (24). On the first name, we expect a decrease in final lengthening and a smaller f0-range in (25) compared to (24).

With respect to Q2, we confront speakers with five different contexts (YOUNG (baseline), CHILD, ELDERLY, NON-NATIVE, and NOISE) to disentangle the question of situational in/dependence of prosodic variability. If speakers vary their productions between contexts, we expect those variations to be in line with the literature mentioned in 5.1.2 (on page 53): we expect speakers to mark the difference between conditions with and without internal grouping in a more pronounced way in the non-baseline contexts. If the findings of prosodic cue modifications for contextual situations in different sentence types are transferable to coordinates and if the modifications found for English and French speakers also hold for German, we expect an increase in segmental and pause durations as well as increased f0-ranges for CHILD and ELDERLY. Due to inconsistent findings in previous studies regarding a non-native interlocutor, we explore this context at a rather exploratory level and refrain from a specific hypothesis. Regarding the presence of noise (NOISE), the literature predicts an increase in f0 and segment durations. Note that the (NOISE condition is the only condition, which directly affects the speaker, because the virtual interlocutor AND the speaker are confronted with the noise.

With Q3 and Q4, we expect to further disentangle the nature of the prosody-syntax link. Regarding Q3, between-speaker in/variability in the combined use, that is, in the interplay of the prosodic cues, will inform us about the type of link between syntax and phonology (i.e., fixed and categorical or relative and allowing for some flexibility). Regarding Q4, within-speaker in/variability will give further insights into situational in/dependence of prosodic cues on the individual level. These two research questions will be addressed in an exploratory manner.

Overall, marked differences between contexts would speak in favour of situationally dependent models and their absence for models of situational independence (given speakers would prosodically disambiguate the conditions with and without grouping). If speakers show inter-individual variability in how they employ and combine prosodic cues at the surface to mark prosodic boundaries, this would speak in favour of models of relative boundary strength (e. g., Wagner 2005).

5.2 Methods

Participants

16 monolingual German native speakers (sex: 13 female, 2 male, 1 other; age range: 19–34 years, mean: 25.75 years, *SD*: 4.6) took part in the study. They were recruited at the University of Potsdam and were reimbursed or received course credits. Written informed consent was obtained from all participants prior to the study. They were naïve to the purpose of the study. The procedure for this study was approved by the Ethics Committee of the University of Potsdam (approval number 72/2016). Participants (henceforth *speakers*) reported normal or corrected to normal vision. Normal hearing was also confirmed by a

hearing screening using an audiometer (Hortmann DA 324 series).

Stimuli

Items Stimuli were taken from (Holzgrefe-Lang et al. 2016) Holzgrefe-Lang et al. (2016) and consisted of six sequences of three German names coordinated by *und* (Engl. *and*). Each name sequence appeared in two conditions: without internal grouping (26) or grouping the first two names together (27), resulting in 12 items overall. Grouping was visually indicated by parentheses around the grouped names (see (27)).

(26) Name1 and Name2 and Name3

(27) (Name1 and Name2) and Name3

A total of nine different names was used. Six of these occurred as Name1 and as Name2 and ended in the high frontal vowel /i/ (Moni, Lilli, Leni, Nelli, Mimmi, or Manni) in order to decrease glottalization. The remaining three names (Manu, Nina, or Lola) ended either in /u/ or in /a/ and occurred only in the position of Name3. The names were controlled for number of syllables (disyllabic), syllable structure (trochaic), and sonority of the segments (only sonorant material was used to allow for better pitch tracking). Two corpora (*Google Ngram Viewer*, https://books.google.com/ngrams retrieved on 06.08.2020 and dlexDB Heister et al. 2011) confirmed that the name combinations we used (e. g., “Moni und Lilli”) were all non-frequent (no hits).

Contexts The five contexts (YOUNG, CHILD, ELDERLY, NON-NATIVE, NOISE, see Figure 9 and Table 2) were evoked by videos, giving the speakers a visual-auditory impression of their interlocutors. The noise for the NOISE context was created in Praat (Boersma & Weenink 2017) using the formula *randomGauss(0,0.7)*. For each context, the corresponding interlocutor appeared in two video clips (introduction and instruction) and produced a trigger question (see below).

In the introduction video, each interlocutor presented her character in a few sentences, talking about her fictional demographic background, including information on her name, age, origin, occupation, place of living, and some interests (Figure 9, for the exact wording of their presentation see Appendix A on page 83). (True) demographic data of the interlocutors are given in Table 2. Note, however, that these data were unknown to the speakers of the production study.

In the instruction video, the interlocutor instructed the speakers to utter the name sequences in a way that would allow the interlocutor “to understand as rapidly and accurately as possible who is coming together”. The wording of the instruction for the task was nearly the same for all contexts but the adult interlocutors addressed the speakers in the formal

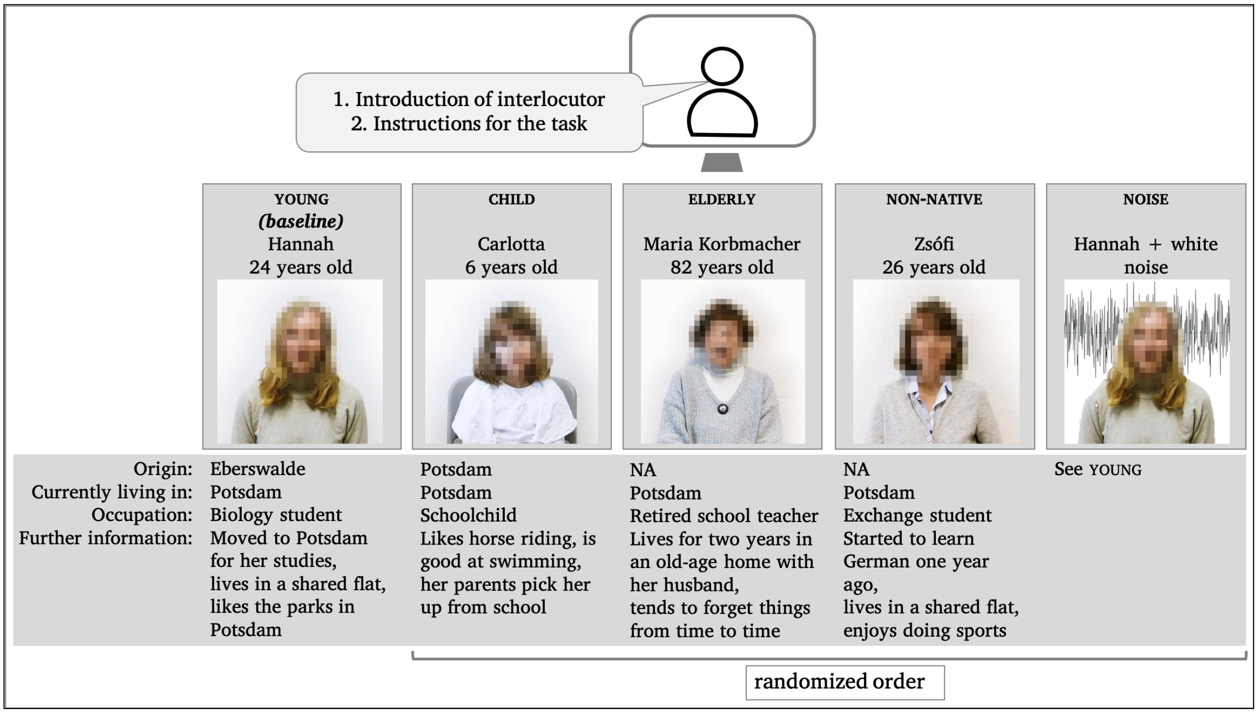

1. Introduction of interlocutor
2. Instructions for the task
YOUNG
(baseline)
Hannah
24 years old
CHILD
Carlotta
6 years old
ELDERLY
Maria Korbmacher
82 years old
NON-NATIVE
Zsófi
26 years old
NOISE
Hannah + white noise
Origin: Eberswalde | Potsdam | NA | NA | See YOUNG
Currently living in: Potsdam | Potsdam | Potsdam | Potsdam
Occupation: Biology student | Schoolchild | Retired school teacher | Exchange student
Further information: Moved to Potsdam for her studies, lives in a shared flat, likes the parks in Potsdam | Likes horse riding, is good at swimming, her parents pick her up from school | Lives for two years in an old-age home with her husband, tends to forget things from time to time | Started to learn German one year ago, lives in a shared flat, enjoys doing sports
randomized order

Figure 9: Pictures and fictional names, ages, origins, and further information of the interlocutors present in the five contexts. *Note*: faces were not pixelated in the experiment; noise was presented auditorily.

way using German *Sie* (you), while the child used the informal “Du” (you), which reflects prescriptive German pronoun use.

For the noise context (NOISE), the young interlocutor was exposed to the same white noise that was later played to the speakers in the recording session. She heard the noise via in-ear headphones which were invisible in the video clip. Instead of presenting herself again, she reminded the speakers of who she was and that they should do the task with her again. Furthermore, she commented on the noise in the background and repeated the instruction for the task, adding to the usual wording that she was going to be the interlocutor *again*.

In order to reduce the influence of non-person specific factors to a minimum, the interlocutors in the videos wore similar unicoloured clothes and were all seated in front of a light neutral background (Figure 9). They were asked to look into the camera and to talk with few gestures and little moving. The introduction and instruction videos had comparable durations (cf. Table 2). In order to trigger the production of the name sequences and to remind the speaker of their interlocutor, the speakers were played the question *Wer kommt?* (‘who is coming?’) produced by the respective interlocutor of each context. The trigger questions had a mean duration of 0.94 seconds (*SD*: 0.028 sec, see Table 2) and preceded each trial (see 5.2 *Procedure*).

Table 2: Information on the five contexts.

	YOUNG	CHILD	ELDERLY	NON-NATIVE	NOISE
(True) demographic data of the person behind the character of the fictional interlocutor (unbeknownst to the speakers)					
age (in years)	21	7	89	32	See YOUNG
mother tongue	German	German	German	Hungarian	
origin	Berlin-Brandenburg area	moved to Berlin-Brandenburg area at the age of 4	Berlin-Brandenburg area	Hungary	
currently living in	Berlin-Brandenburg area	Berlin-Brandenburg area	Berlin-Brandenburg area	Berlin-Brandenburg area (in Germany for < 3 years)	
Technical details of videos: durations in seconds					
introduction video	28	18	41	30	17
instruction video	21	19	35	24	22
trigger question *wer kommt?* *('who is coming')*	0.892	0.956	0.961	0.932	0.953

Procedure

Before the start of the recording session, the white noise was played to the speakers for one second to familiarize them with the sound to be played in the NOISE condition in order to prevent surprisal or scare effects during the experiment. The experiment then started with a practice phase (four items which were not used in the actual experiment) followed by the test phase. The test phase consisted of five blocks, corresponding to the five experimental contexts (YOUNG, CHILD, ELDERLY, NON-NATIVE, NOISE, see above) in which speakers were asked to produce the coordinated name sequences in the two conditions, that is, with or without internal grouping. Each of the 12 items was presented in each context, hence, speakers produced each item five times. The textscyoung context, as the baseline context, was always presented first, the other four contexts were presented in randomized order (cf. Figure 9). In each context, items were pseudo-randomized using different lists. No more than two items of the same condition followed one another. In addition, Name1 and Name2 were never repeated in two subsequent trials.

Each block started with the two video clips and during the test phase, for each trial, speakers saw a fixation cross on the screen while they heard the trigger question *Wer kommt?* (‘who is coming?’) via headphones. After 1000 ms, the fixation cross was replaced by the visual presentation of a name sequence (i.e., the item) in one of the two conditions which stayed on the screen for 5000 ms. The sound recording started together with the presentation of the name sequence and continued for 1000 ms after the names disappeared, see Figure 10.

Recordings took place in a sound-attenuated booth in the acoustics laboratory of the University of Potsdam via an Alesis io12 interface. Speakers wore a headset HSC 271 (AKG Acoustics by Harman, www.akg.com) with over-ear headphones and a condenser microphone and were seated in front of a wide screen monitor with 1920 x 1200 resolution and saw the stimuli in Arial font of size 50. The experiment was run from a Dell laptop standing outside of the recording booth using the software Presentation (Neurobehavioural Systems, https://www.neurobs.com/; Version 20.1)

After the recording session, speakers completed some questionnaires which will not be analysed.

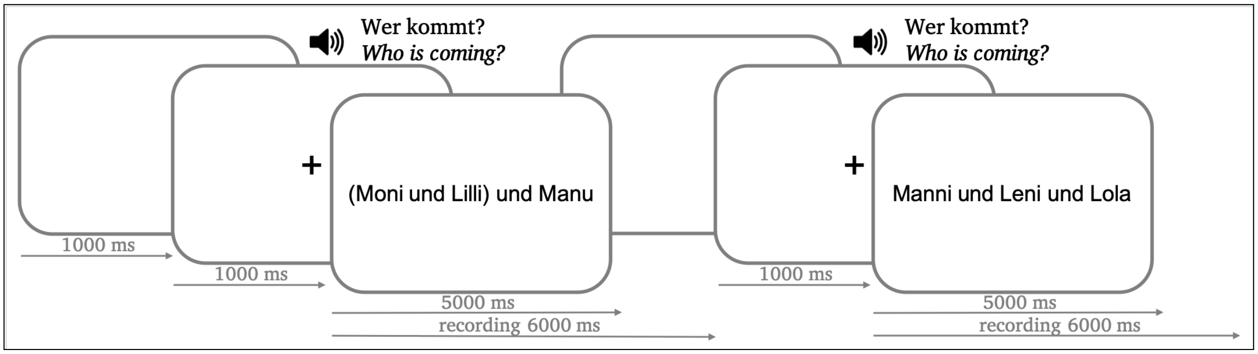


Figure 10: Experimental setting and timing of two trials.

Perception check

Procedure In the production study described above, we recorded a total of 960 individual productions: 6 name sequences ∗ 2 conditions ∗ 5 contexts ∗ 16 speakers. In order to verify whether the *intended* internal structure (i.e., the *grouping of constituents*) is congruent with the structure perceived by other naïve listeners, we ran a perception check of all 960 productions. Note that *intended* refers to the conditions with or without parentheses presented to the speakers on the screen in the production study. We lack information about the intention of the speakers at the time of production.

The perception check encompassed 32 listeners, who had not taken part in the production experiment (sex: 22 female, 10 male; age range: 18–41 years, mean: 24.25 years, *SD*: 5.8). They were recruited at the University of Potsdam and were reimbursed or received course credits. Another 10 listeners took part, but had to be excluded from the analysis due to technical problems (n = 9) or German as a non-native language (n = 1).

Each listener judged a set of 267 out of the 960 productions, which consisted of the total 60 productions of 4 speakers (4 speakers ∗ 60 recordings = 240) plus a subset of 27 productions from various speakers. The subset of 27 productions was judged by all listeners and constituted a semi-random sample of all productions, containing at least one production of each speaker and of each context. Furthermore, the subset included three productions which to the first author seemed to mismatch between intended and perceived grouping. The perception check started with the presentation of the subset, followed by the remaining 240 productions of four speakers, each presented in a block. The 960 productions of the 16 speakers were judged in four testing lists (4 lists ∗ 240 recordings = 960). Each list, and therefore the productions of each speaker, was judged by eight listeners (8 listeners ∗ 4 lists = 32 listeners in total). Each of the four lists contained some productions twice, those which were part of the subset and the following 240 productions. In the case of repetitions, only the first judgement was considered.

The perception check was run in sessions with several listeners at the same time. Two pictograms with three persons each were used to depict the two conditions (Figure 11, picture A without and picture B with internal grouping). The task was twofold: to listen to each production and (1) to choose the matching pictogram (i.e., to identify the condition) and (2) to indicate the most probable addressee the name sequence was uttered to (young adult, child, elderly, non-native, in noise; i.e., to identify the context).

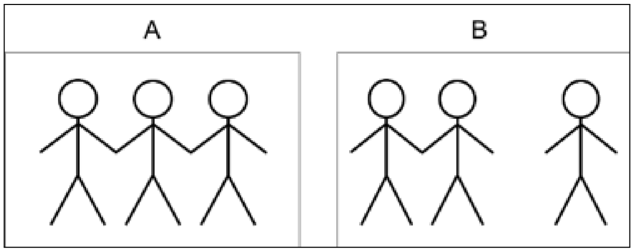


Figure 11: Pictograms used in the perception check. Picture A depicts the condition without internal grouping, picture B the condition with internal grouping.

Analysis and results First, for each listener, we counted the number of congruent rates (i.e., correct identifications of the intended grouping/condition and context, referred to as *hit-rate*). Following standard assumptions on the exclusion of data points (e. g., Howell et al. 1998), if, for a given listener, the hit-rate was 2 *SD* below the mean hit-rate of all listeners, all ratings of this listener were excluded. This was the case for one out of the 32 listeners, thus ratings from altogether 31 listeners, rendering 8067 ratings overall entered the perception check analyses (the 8067 ratings result from 27 productions in the subset ∗ (rated by) 32 listeners + 933 remaining productions ∗ (rated by) 8 listeners (= 8328) – 261 ratings of the excluded listener).

Second, for each individual production (n = 960), we calculated the ratio of the hit-rate to the number of total rates. We used the ratio instead of the absolute hit-rate since individual productions were rated by a varying number of listeners (in the subset: 31 listeners, for the rest of the productions: 8 listeners, or 7 in the case of the excluded listener). We then calculated the mean ratio of all productions as well as the standard deviation.

In what follows, we will report the ratings of condition and context separately. Only the ratings of condition influenced the exclusion of individual productions: productions for which the ratio of the hit-rate was more than 2 *SD* below the overall mean ratio of hit-rates were excluded for further analyses.

With respect to the rating of condition, the mean ratio was 0.936 (*SD*: 0.1545), and we thus used an accuracy cut-off level of 0.627. Applying this criterion, 38 productions were excluded since their ratios fell more than 2 *SD* below the mean (3 productions of the subset and 35 of the remaining productions). Nevertheless, the majority of all productions was perceived with the intended grouping (689 out of 960).

A closer look at the excluded items reveals that productions with internal grouping were twice as often not perceived as intended (25 with internal grouping, 13 without internal grouping). Looking at the context of the excluded items, we observed that most incongruent rates involved productions produced in the NOISE context (n = 15, with n = 11 in condition brack), followed by productions produced in the YOUNG context (n = 10, with n = 7 in condition brack), the CHILD context (n = 7, with n = 2 in condition brack), the NON-NATIVE context (n = 5, with n = 4 in condition brack), and the ELDERLY context (n = 1 in condition brack).

Regarding the rating of the probable interlocutor (i.e., the listeners had to select the context in which the coordinates were most probably produced), the hit-rates are overall much lower than for condition. For only 21 out of the 960 productions (2%), all listeners correctly identified the context. A closer look at these 21 productions reveals that 17 of them were productions in the NOISE context and the other four in the YOUNG context. An extended analysis revealed that, overall, in 12% of the productions (119 out of 960 productions), at least 75% of the listeners perceived the context in the intended way. These 119 productions are distributed across the five contexts as follows: 65 stem from the NOISE context, 44 from the YOUNG context, 8 from the CHILD context, 1 from the ELDERLY, and 1 from the NON-NATIVE context.

Segmentation and measurements

In addition to the 38 productions excluded based on the perception check, production data of one speaker was excluded completely, because this speaker did not comply with the task specified in the instructions for the experiment. Visual inspection revealed that the speaker–consciously or not–misinterpreted the whole experimental setting: the productions include quirky, inconsistent prosodic behaviour in the use of the prosodic cues we are interested in. Thus, a total of 96 productions (10% of the overall data) were excluded, consisting of the 38 items following the perception check and 58 productions of the excluded speaker (note that two of their productions are included in the 38 excluded items of the perception check). The remaining data comprise 864 productions from 15 speakers (sex: 13 female, 1 male, 1 other; age range: 19–34 years, mean: 25.47 years, *SD*: 4.6). Table 3 provides the distribution of the number of remaining productions across contexts and conditions.

Table 3: Number of productions entering statistical analyses across contexts and conditions in the final data set.

	YOUNG	CHILD	ELDERLY	NON-NATIVE	NOISE
brack	87	85	90	90	87
nobrack	83	88	89	86	79

For further analyses, segment boundaries and pauses were manually labelled following the criteria in Turk et al. (2006) using Praat (Boersma & Weenink 2017). In unclear cases, the boundary between the last vowel of Name1 and Name2, and the following *und*, respectively, was set to the mid of the F2 transition. The end of the utterance was set to the point where the intensity profile fell below 50 dB. The f0-minima on the first syllable and the f0-maxima on the second syllable of both, Name1 and Name2, were annotated. For phonetic analyses, we extracted three acoustic measures regarding duration and f0 each on Name1 and Name2: *rise, final lengthening*, and *pause*. The variable rise captures the range between the f0-minimum and the f0-maximum on NameX calculated in semitones (st; formula used for calculation: $12 * log_2(f0_{max}/f0_{min})$). Rise and f0-range will be used interchangeably in this paper when referring to the f0-measurements taken in the study. The second variable captures the final lengthening on each name (in %) and was calculated by dividing the duration of the final vowel in NameX by the duration of NameX. The variable pause captures the relative duration (in %) of the possible pause following NameX and was calculated by dividing the duration of the pause after NameX by the duration of the whole utterance. Relative values for durational measurements were chosen in order to normalize for differences in speech rate.

Another method to transcribe prosodic boundaries based on f0 would be GToBI (Grice et al. 2005, Grice & Baumann 2002), the German adaption of the ToBI system based on autosegmental-metrical theory of intonation (Ladd 2008 and references therein) and originally established for American English (Silverman et al. 1992). Since our main focus is on the combined realization of several acoustic cues, we opted for an analysis that can be applied to tonal and durational cues equally.

Statistical analysis of prosodic disambiguation and general context variability

For each dependent variable (rise, final lengthening, pause) on Name1 and Name2 we ran separate linear mixed-effects regression models using the function *lmer* from the R (R Development Core Team 2018) packages *lme4* (Bates et al. 2015b) and lmerTest (Kuznetsova et al. 2017). Context was entered as an independent variable and four contrasts were coded comparing each of the contexts CHILD, ELDERLY, ELDERLY, NON-NATIVE, and NOISE against YOUNG (baseline) using the general inverse (Schad et al. 2018). The model, thus, estimates the difference in the dependent variables between addressing the child compared to the young adult (CHILD VS. YOUNG), addressing the elderly compared to the young adult (ELDERLY VS. YOUNG), addressing the non-native speaker of German compared to the young (native German-speaking) adult (NON-NATIVE VS. YOUNG), and addressing the young adult in the presence of noise compared to a non-noisy environment (NOISE VS. YOUNG). For final lengthening and rise, condition was coded with a sum contrast, with the condition brack coded as 1 and the condition nobrack as -1. Pause was modelled for the condition brack only, due to the absence of a pause after Name2 in most nobrack productions (i.e., values of zero in the dataset).

For model fitting, we always started with a maximal model including the interaction of context and condition as fixed-effects terms (except for the pause measure), as well as a random-effects structure with all possible variance components and correlation parameters associated with the four within-subject contrasts (CHILD VS. YOUNG, ELDERLY VS. YOUNG, NON-NATIVE VS. YOUNG, NOISE VS. YOUNG). Following the approach outlined in Bates et al. (2015a), in order to avoid overfitting of the random effects structure, we fitted the corresponding zero correlation parameter model using the *double-bar* ('‖') syntax. The complexity of the random-effects structure was then reduced in a step-wise manner, dropping those components with a proportion of variance close to zero in a random-effects Principal Component Analysis (using the *rePCA* function in the *RePsychLing* package Baayen et al. 2015). We assessed improvements in model fit of the maximal model and the zero correlation parameter models using the log-likelihood ratio test and comparisons of the Akaike Information Criterion. For the zero correlation parameter model with the best fit, we returned to a model that included correlations of random effects (i.e., the *single-bar* syntax). In cases in which the reduction of variance components in the zero correlation parameter model did not lead to a better fit than the fit of the maximal model, we kept the maximal model. If, however, the maximal model did not converge (which happened for the pause measure) or if the maximal model had a high degree of correlations in the fixed effects, and the degree of correlations was less pronounced in the zero correlation parameter model, we kept the suppression of the random effects' correlations (i.e., we did not return to the *single-bar* syntax).

Exploratory analysis of inter- and intra-speaker variability

Following the statistical analyses for rise, final lengthening, and pause, we further explored the interplay of the three cues in combination. Specifically, we were interested in observable patterns in this interplay that differ between speakers within a given context (inter-speaker variability in cue combinations, here we will focus on the context YOUNG) or within speakers between all contexts (intra-speaker variability in cue combinations). Since pause after Name1 was not used by all speakers (see below) and since Name2 is the critical element before the syntactic boundary we decided to do the exploratory analysis of cue combinations on Name2.

We developed a classification system which was applied to each individual cue within each speaker and context, resulting in two parameters as indicators for how effectively each cue distinguishes between the brack and nobrack condition. In order to determine the degree of distinction between conditions, we estimated for each cue within speaker and context the statistical probability of the respective cue distinguishing between the two conditions using a Mann-Whitney U-Test (Mann & Whitney 1947). The two parameters were (1) the *p*-value of the Mann-Whitney U-Test computed in Matlab (MATLAB) and (2) the common language effect size (CLES, McGraw & Wong 1992). The CLES returns a value between 0 and 1, and indicates the probability that a random pair of data points belongs to two independent groups. Thus, a value of 1 for the CLES of our comparisons refers to a case in which this cue clearly separates the two conditions (brack and nobrack) from each other. For our analysis, we differentiated between three types of distinction (Table 4): (i) *clear distinction* (abbreviated to C) for cases in which the Mann-Whitney U-Test returns a *p*-value $< .05$ and the CLES $= 1$, (ii) *partial distinction* (abbreviated to P) for comparisons with a Mann-Whitney U-Test resulting in a *p*-value $< .05$ and a CLES <1), and (iii) *no distinction* (abbreviated to N) for cases in which the Mann-Whitney U-Test returns a *p*-value $< .05$, meaning that this cue does not separate the two conditions.

Table 4: Criteria for the three possible types of distinction of a cue between the two conditions used for the exploratory analysis.

	Estimated probability of the Mann-Whitney U-Test	Common language effect size (CLES)
Clear distinction (C)	$p < .05$	CLES $= 1$
Partial distinction (P)	$p < .05$	CLES < 1
No distinction (N)	$p > .05$	–

In order to explore possible patterns in the use of cues, the individual types of distinction of each of the three cues were combined for each speaker and context. The three cues and their three distinction types combine to 27 possible patterns shown in Table 5. The cues are always given in the following order: rise, final lengthening (abbreviated to “lengthening” to ease the reading), and pause, altogether shortened to RLP and given as subscript at the end of the pattern label. For example, if all three cues are clearly used to distinguish between the two conditions, this pattern is characterized as CCC_{RLP}, as given in the upper leftmost cell of Table 5. If rise and pause are clearly used and final lengthening is partially used to distinguish between the two conditions, the pattern is called CPC_{RLP}, as given in the third cell from the left in the upper row of Table 5. The raw data of the three cues on Name2 is given in two-dimensional space in the appendixes B–D, plotting the magnitude of two of the three cues, each, separated for speaker and context. In order to cover all possible combinations, we visualized three comparisons: (a) pause on the x-axis to rise on the y-axis (Appendix B on page 86), (b) final lengthening on the x-axis to pause on the y-axis (Appendix C on page 87), and (c) final lengthening on the x-axis to rise on the y-axis (Appendix D on page 88).

Table 5: Matrix of possible cue combinations (patterns) of the cues rise (R), final lengthening (L), and pause (P) (in this order). Differentiation between three types of distinction: clear distinction (C), partial distinction (P), no distinction (N).

Rise **L**engthening **P**ause:	CCC	CCP	CPC	CPP	CCN	CNC	CNN	CPN	CNP
Rise **L**engthening **P**ause:	PPP	PPC	PCP	PCC	PPN	PNP	PNN	PCN	PNC
Rise **L**engthening **P**ause:	NNN	NNP	NPN	NPP	NNC	NCN	NCC	NCP	NPC

5.3 Results

Statistical analyses of prosodic disambiguation and general context-dependent variability of individual prosodic cues on Name1

Rise on Name1 For rise, the estimates for the fixed-effects were extracted from the maximal model (Table 6). Regarding prosodic disambiguation of coordinates, we found a main effect of condition. On average, speakers produced an f0-range on Name1 which was 3 st smaller in the brack compared to the nobrack condition. Regarding general context variability, there was a marginally significant interaction between condition and the CHILD and the ELDERLY contexts, indicating that speakers showed a tendency to decrease their f0-range in the brack condition even more when speaking to a child and to an elderly adult.

Final lengthening on Name1 For final lengthening, the estimates for the fixed-effects were extracted from the maximal model (Table 7). Regarding prosodic disambiguation of coordinates, there was a main effect of condition, indicating that the duration of the final segment of Name1 was shorter in the brack compared to the nobrack condition. With respect to general context variability, we found a marginally significant interaction between condition and the NON-NATIVE context, indicating that the difference between nobrack and brack was larger when speakers addressed the non-native adult, since the duration of the final segment tended to be even shorter in the brack condition.

Pause on Name1 Since six out of 15 speakers did not produce a pause after Name1 neither in the brack nor in the nobrack condition and there were only three speakers who

Table 6: Estimates of the model for rise on Name1 (i.e., f0-range on Name1). Statistically significant effects are marked in bold ($p < .05$).

Predictor	*Estimate*	*SE*	*t-value*	*p-value*
condition_brack	-1.53	0.27	-5.76	< **.001**
CHILD VS. YOUNG	0.30	0.21	1.43	.168
ELDERLY VS. YOUNG	0.41	0.33	1.24	.233
NON-NATIVE VS. YOUNG	0.12	0.32	0.39	.704
NOISE VS. YOUNG	0.68	0.38	1.80	.091
condition_brack: CHILD VS. YOUNG	-0.45	0.23	-1.98	.064
condition_brack: ELDERLY VS. YOUNG	-0.34	0.18	-1.94	.064
condition_brack: NON-NATIVE VS. YOUNG	-0.33	0.21	-1.53	.142
condition_brack: NOISE VS. YOUNG	0.09	0.22	0.43	.675

Table 7: Estimates of the model for final lengthening on Name1. Statistically significant effects are marked in bold ($p < .05$).

Predictor	*Estimate*	*SE*	*t-value*	*p-value*
condition_brack	-2.87	0.49	-5.85	< **.001**
CHILD VS. YOUNG	-0.33	0.70	-0.47	.644
ELDERLY VS. YOUNG	1.41	0.93	1.51	.152
NON-NATIVE VS. YOUNG	0.64	0.62	1.04	.305
NOISE VS. YOUNG	-0.06	0.92	-0.75	.085
condition_brack: CHILD VS. YOUNG	-1.02	0.58	-1.75	.085
condition_brack: ELDERLY VS. YOUNG	-0.99	0.57	-1.75	.084
condition_brack: NON-NATIVE VS. YOUNG	-1.23	0.62	-1.96	.058
condition_brack: NOISE VS. YOUNG	-0.72	0.57	-1.25	.213

produced a pause in each of the contexts, we did not run any statistical analyses of pause duration after Name1.

Statistical analyses of prosodic disambiguation and general context-dependent variability of individual prosodic cues on Name2

Rise on Name2 For rise, the estimates for the fixed-effects were extracted from the maximal model (Table 8). We found a main effect of condition and a main effect of the contexts CHILD and ELDERLY. Regarding prosodic disambiguation, speakers, overall, produced a larger f0-range in the brack than in the nobrack condition. Regarding contexts, when addressing the child as well as the elderly person, speakers increased the f0-range of the rise compared to addressing the young adult (cf. Figure 12 and top panel of Figure 13). For a subset of 13 female speakers this increased f0-range can be seen in Figure 12, where the green and blue (CHILD and ELDERLY context, respectively) dashed and solid lines start below the black lines (YOUNG context) at the beginning of Name2 and rise to a level above the black lines towards the f0-peak of Name2. Note, Figure 12 cannot be compared directly to the results of the statistical model, since values in Hertz are plotted in the Figure, while the model is calculated on semitones. A similar, though statistically non-significant tendency is observable for the other two contexts, addressing the non-native speaker and in noise. The model revealed no statistically significant interactions between contexts and condition.

Table 8: Estimates of the model for rise on Name2 (i. e., f0-range on Name2). Statistically significant effects are marked in bold ($p < .05$).

Predictor	*Estimate*	*SE*	*t-value*	*p-value*
condition_brack	2.86	0.27	10.77	< **.001**
CHILD vs. YOUNG	1.10	0.31	3.56	**.003**
ELDERLY vs. YOUNG	0.94	0.36	2.59	**.021**
NON-NATIVE vs. YOUNG	0.59	0.31	1.89	.078
NOISE vs. YOUNG	0.53	0.30	1.76	.097
condition_brack: CHILD vs. YOUNG	0.03	0.20	0.16	.874
condition_brack: ELDERLY vs. YOUNG	0.09	0.22	0.40	.694
condition_brack: NON-NATIVE vs. YOUNG	0.04	0.24	0.18	.856
condition_brack: NOISE vs. YOUNG	-0.09	0.26	-0.33	.746

Final lengthening on Name2 For final lengthening, the estimates for the fixed-effects were extracted from a model that included principal components but not the correlation parameters in the random-effects structure (Table 9). Regarding prosodic disambiguation, the data show a main effect of condition, indicating that speakers marked the brack condition with increased final lengthening compared to the nobrack condition. Regarding general context variability, there was an interaction between condition and the ELDERLY context with a negative estimate and an interaction between condition and the NOISE context with a positive estimate (cf. mid panel Figure 13). This indicates that speakers increased the final lengthening when addressing the elderly as opposed to the young interlocutor in the nobrack condition. In the noise context, however, they increased final lengthening in the brack condition, but not in the nobrack condition.

Pause on Name2 For pause, the model was run on a subset containing the brack condition only; the nobrack condition was excluded, due to the large number of zero values. The estimates of the fixed-effects were extracted from the zero correlation parameter model including all variance components (Table 10). Regarding general context variability, there

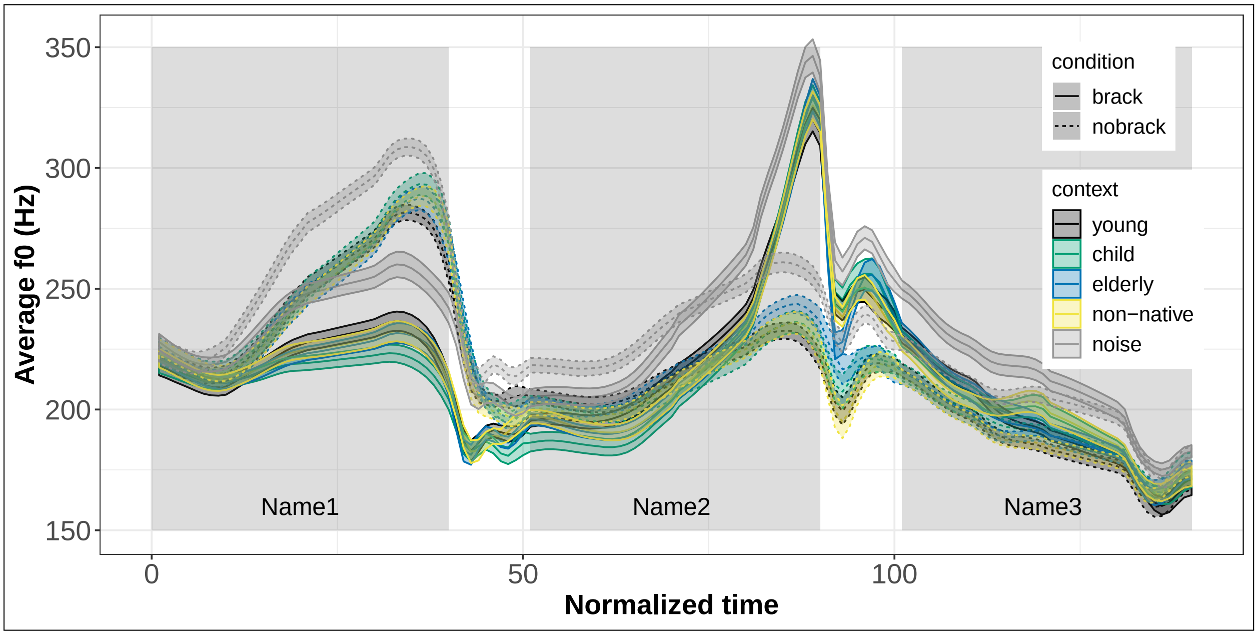


Figure 12: Time-normalized f0-contours (in Hz) of coordinates in brack (solid lines) and nobrack (dashed lines) conditions produced in five contexts (cf. colours) by a subset of 13 female speakers.

Table 9: Estimates of the model for final lengthening on Name2. Statistically significant effects are marked in bold ($p < .05$).

Predictor	*Estimate*	*SE*	*t-value*	*p-value*
condition_brack	4.96	0.58	8.59	< **.001**
CHILD VS. YOUNG	-0.39	0.58	-0.67	.511
ELDERLY VS. YOUNG	1.34	0.62	2.17	**.039**
NON-NATIVE VS. YOUNG	0.61	0.57	1.07	.293
NOISE VS. YOUNG	1.45	0.71	2.04	.056
condition_brack: CHILD VS. YOUNG	-0.56	0.51	-1.09	.274
condition_brack: ELDERLY VS. YOUNG	-1.24	0.51	-2.43	**.015**
condition_brack: NON-NATIVE VS. YOUNG	-0.65	0.51	-1.27	.205
condition_brack: NOISE VS. YOUNG	1.43	0.52	2.75	**.006**

was a main effect of the ELDERLY context indicating that speakers produced a longer pause addressing the elderly compared to the young adult interlocutor (cf. bottom panel Figure 13). Additionally, speakers showed a tendency to reduce pause duration in the noisy environment (NOISE), though this was not statistically significant.

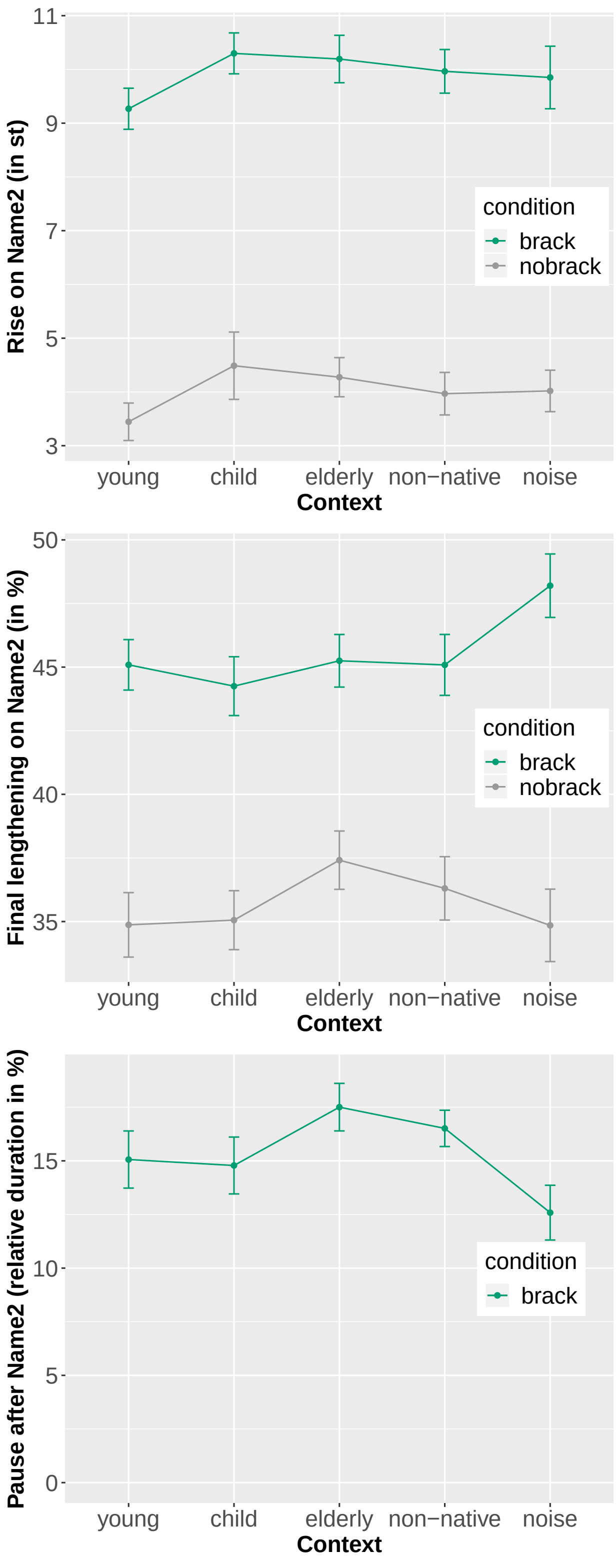


Figure 13: Mean values and 95% confidence intervals for rise (top panel), final lengthening (mid panel), and pause (bottom panel) on Name2 for each context and condition (green = condition brack, grey = condition nobrack).

Table 10: Estimates of the model for pause after Name2. Statistically significant effects are marked in bold ($p < .05$).

Predictor	*Estimate*	*SE*	*t-value*	*p-value*
CHILD VS. YOUNG	0.02	0.96	0.02	.984
ELDERLY VS. YOUNG	2.67	1.08	2.47	**.025**
NON-NATIVE VS. YOUNG	1.45	0.97	1.5	.153
NOISE VS. YOUNG	-2.56	1.36	-1.87	.08

Exploratory analyses of inter- and intra-speaker variability of cue combinations on Name2

The cue combinations for each speaker (cf., y-axis) and context (cf., x-axis) are plotted in Figure 14. For each speaker and context, the cell is divided into three rows, with the distinction type of rise given in the uppermost row of the cell, final lengthening in the middle row, and pause in the bottom row. The three types of distinction are represented by shading: full colour for clear distinction, light shade for partial distinction, and the lightest shade for no distinction.

Regarding *inter-speaker variability*, we focused on whether there are different patterns of cue combinations between speakers within the young context only, represented in the left-most column of the plot in Figure 14. In general, five different patterns are observable: CCC_{RLP} (i.e., all three cues in full green), CNC_{RLP}, CPC_{RLP}, CCP_{RLP}, and PPC_{RLP}, however they differ in number of occurrences. Seven speakers out of 15 (2, 3, 4, 7, 13, 15, and 16) produced the pattern CCC_{RLP}, indicating that all three cues were clearly used to distinguish between the two conditions. Further four speakers (1, 6, 9, and 11) produced the pattern CNC_{RLP}, which means that they clearly distinguished between brack and nobrack using rise and pause, but not using final lengthening. The other three patterns were produced by either two or one speakers. Overall, in four of the five patterns brack and nobrack were clearly distinguished by at least two of the cues. While both rise and pause were used clearly distinctively by 14 out of 15 speakers, only eight speakers used final lengthening in a clearly distinctive way. Notably, the pattern with no distinction in all three cues was never observed in the young context.

Regarding *intra-speaker variability*, we focused on whether speakers vary the patterns of cue combinations when addressing different interlocutors or speaking in noise. For that purpose, we examined the patterns of cue combinations within speaker across contexts (i.e., the three rows in each cell for each speaker across columns in Figure 14). For speaker 2, the pattern is identical across all five contexts, thus showing stability in the use of the prosodic cues for distinguishing between the brack and nobrack condition across different contexts. Most other speakers show two or three different patterns across contexts (cf., speakers 6,

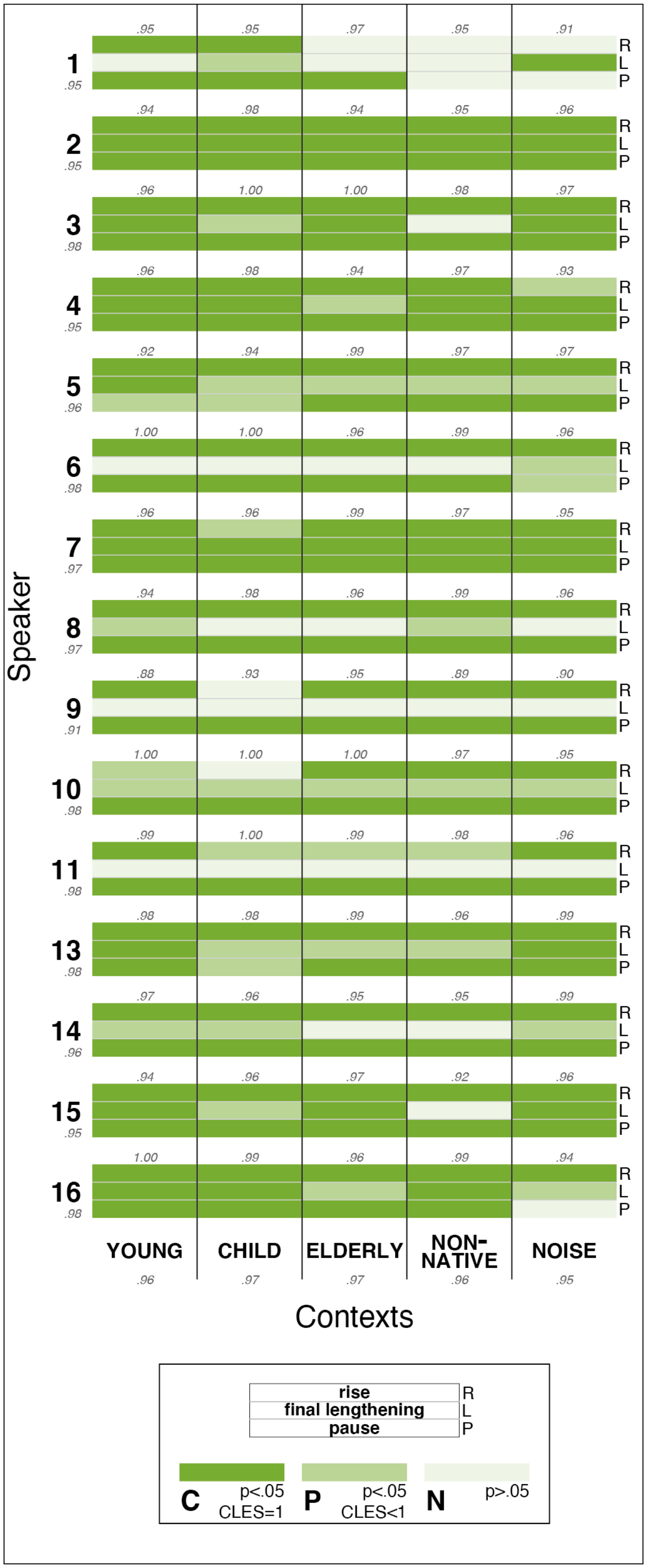


Figure 14: Speaker variability of cue combinations across contexts on Name2, showing the patterns of cue combinations (shades of green) used by individual speakers (y-axis) in the contexts YOUNG, CHILD, ELDERLY, NON-NATIVE, and NOISE (x-axis). The shades of green indicate the type of distinction: full = clear distinction (C), light = partial distinction (P), lightest = no distinction (N). For each speaker, the three rows indicate the different cues (R: rise, L: final lengthening, P: pause). The small numbers in italics indicate the mean ratios of the hit-rates for condition in the perception check (i.e., ratio of correct identifications of the intended grouping to all rates; numbers to the left: average per speaker; lowest line: average per context; above cells: average per speaker per context). For example, speaker 16 clearly distinguishes between the brack and nobrack condition, using all three cues in the YOUNG, CHILD and NON-NATIVE context, but in the context NOISE the speaker uses final lengthening only partially and pause not at all to distinguish between the two conditions. In the YOUNG context, 100% of the rates in the perception check were congruent rates, while in the NOISE context 94% were congruent rates.

7, 8, 9, 11, 14 and 3, 4, 5, 10, 13, 15, 16, respectively). Overall, there is more variability across contexts in the use of final lengthening than in rise and pause, visualized by more varying shading in the middle row of the speaker-specific cells. There is one speaker who in one context shows no distinction between brack and nobrack in any of the three cues (cf., NNN_{RLP} speaker 1, context non-native), for all other speakers and contexts at least one cue is clearly distinctive. In addition, we plotted the mean ratio of the hit-rates for condition in the perception check in Figure 14 for the respective speakers and contexts. This allows to get an impression of the relation of the produced types of distinction in the three prosodic cues to how well the prosodic boundaries (i.e., conditions) have been perceived by naïve listeners.

5.4 Discussion

In the current study, we aimed to gain insights into the situational dependence or independence of disambiguating prosody and to learn more about the nature of the prosody-syntax relation. To this end, we explored the production of prosodic boundaries used to disambiguate coordinated sequences of three names (coordinates) between two conditions: without (nobrack) and with (brack) internal grouping of the first two names. We focussed on the variability induced by speakers and contextual settings, such as interlocutors differing in age and mother tongue, as well as the absence/presence of noise (contexts). Besides the distinction between the two conditions (prosodic disambiguation of coordinates, research question Q1), we were interested in the type and size of cues produced at the prosodic boundaries and whether and how speakers varied in producing them depending on the context. Coordinate productions were elicited by means of a referential communication task with five contexts: addressing a young adult (YOUNG), a child (CHILD), an elderly adult (ELDERLY), a young non-native adult (NON-NATIVE), and the young adult in a noisy environment (NOISE). Variability was addressed on three levels: across speakers between contexts (general context-dependent prosodic variability, research question Q2), between speakers within contexts (inter-speaker variability of cue combinations, research question Q3), as well as within speakers between contexts (intra-speaker variability of cue combinations, research question Q4).

Prosodic disambiguation of coordinates (research question Q1)

Our findings replicate previous studies, showing that the internal grouping of coordinates in German is marked by a prosodic boundary consisting of three prosodic cues from the tonal and durational domain: f0-range, final lengthening, and pause. As expected, speakers used prosodic cues on Name1 as well as on Name2 to clearly distinguish between the two conditions. A perception check with naïve listeners showed that the distinction between the conditions was perceptually recoverable: 96% of the productions were correctly recognized as the intended grouping.

The results of the production study are in line with the Proximity and Anti-Proximity principles that form part of the Proximity/Similarity model introduced by Kentner & Féry (2013) and along with this, they are in line with the literature (Taglicht 1998, Wagner 2005, 2010, Watson & Gibson 2004). Thus, our hypothesis (Q1) was confirmed: In the condition with internal grouping compared to the condition without grouping, we found a statistically significant decrease in final lengthening and f0-range on Name1 along with an increase in final lengthening and f0-range on Name2 as well as the insertion of a pause after Name2. In terms of Proximity, durational and tonal cues of Name1 were decreased, indicating that the neighbouring element to the right (i.e., Name2) forms part of the same group. In terms of Anti-Proximity, the prosodic boundary after Name2 was strengthened, indicating that the neighbouring element to the right (i.e., Name3) does not form a group with Name2. This finding also underlines the assumption that prosodic phrasing is not a local phenomenon with changes of prosodic cues occurring only at the prosodic boundary (cf. in our case Name2) but rather depends on globally distributed prosodic changes (cf. in our case Name1 and Name2) (e. g., Clifton et al. 2002, Frazier et al. 2006, Wagner 2005, 2010).

We further found that speakers use the pause cue in a slightly different way than f0-range and final lengthening in marking the difference between conditions. Following Name2, a pause was mostly absent in the condition without internal grouping, while it was present in the condition with grouping. The pause, thus, appears rather as a categorical than a continuous variable. Since we were interested in differences in pause duration between contexts, however, we kept pause as a continuous variable for our analyses.

Overall, the syntactic structure (with or without internal grouping) was clearly disambiguated by means of prosody. This can be interpreted as evidence in favour of a close link between syntax and prosody.

General context-dependent prosodic variability (research question Q2)

The current study is, to the best of our knowledge, the first to systematically investigate prosodic variability in production of coordinates across speakers between various contexts to explore the situational in/dependence of disambiguating prosody and to find out whether the principles of Proximity/Anti-Proximity also hold across situations.

At the group level, we found some variability driven by the different contexts. Nevertheless, variability was rather small and not as distinct as expected. In the following, the contexts CHILD, ELDERLY, NON-NATIVE, and NOISE will be discussed individually in comparison to the baseline context (YOUNG).

In the context CHILD, when addressing the child as opposed to the young adult, speakers changed their productions in the tonal domain: they increased the f0-range on Name2 independent of condition. This can be interpreted as an adaptation to the interlocutor, but without affecting the ease of disambiguation between conditions. The increased f0-range when addressing a child is partly in line with semi-spontaneous speech data from English speakers (Biersack et al. 2005), who additionally showed lengthened vowels. These differences might be due to differences in age of the interlocutor. For a child addressee of the same age as in our study, DePaulo & Coleman (1986) reported longer pauses; a finding that was not evident in our data. A possible explanation for the absence of statistically significant effects in the durational prosodic cues (i.e., final lengthening and pause) in our study might be related to differences in speech style as well as in language-specific factors. Our data were highly restricted with respect to the wording, whereas the data of DePaulo & Coleman (1986) consisted of spontaneous speech and the data of Biersack et al. (2005) of semi-spontaneous speech, both in English.

In the context ELDERLY, when addressing the elderly adult compared to the young adult, speakers modified their speech in the tonal as well as in the durational domain. On Name2, speakers produced an overall larger f0-range in the ELDERLY context along with a longer pause (in the condition with internal grouping). In contrast, final lengthening on Name2 was not used to make the conditions more distinct in the ELDERLY context: unexpectedly, speakers increased the lengthening in the condition without grouping compared to coordinates addressed to the young adult. Yet, with the increased pause duration, the smaller difference in final lengthening between the conditions was probably levelled out. The findings of increased pause durations and increased f0-ranges, thus, partly confirm our hypotheses and are comparable to observations on other structures in English and German (Kemper et al. 1995, Thimm et al. 1998). Those studies found slower speech due to prolonged vowels and more pauses as well as increased variation in intonation, among other speech adaptations. Regarding the increased number of pauses in the reported studies, again, it needs to be mentioned that the respective data stem from spontaneous speech which probably allows for more pause insertion than the relatively restricted stimuli used in our study. Nevertheless, we suggest that the increased pause durations in our data can be interpreted as comparable speech adaptations. In previous research on speech directed at elderly persons, Kemper et al. (1998) distinguished two sets of parameters that speakers modify in order to adapt to the needs of their elder interlocutor: semantic and discourse information on the one hand, and fluency, prosody, and grammatical complexity on the other. Kemper et al. (1998: 53) discuss that the latter set of parameters does not “appear to benefit” perception, but to the contrary, decreases self-esteem on the side of the interlocutor. This type of speech is referred to as patronizing communication (Kemper et al. 1998, Ryan et al. 1995, Thimm et al. 1998, Torrey et al. 2005) and includes the changes in prosodic cues found in our data.

In the context NON-NATIVE, in response to the non-native interlocutor, the data show no clear effects. This contrasts with reports in the literature, in which non-native speakers were addressed with increased f0-ranges and a more emphatic style compared to native speakers (Smith 2007).

Finally, in the context NOISE, the interlocutor was the same young adult as in the baseline context. For adaptation to the noise, speakers increased final lengthening on Name2 in the condition with grouping while at the same time, they decreased the relative duration of the following pause. The increase in final lengthening is in line with our hypotheses and findings in the literature, although we would have expected an additional increase in the f0-range. A possible explanation for the unexpected decrease in pause duration is that a silent pause is a less effective cue in a noisy environment than in a quiet one. Instead of a silent pause, speakers lengthened the final segment to mark the boundary. Furthermore, speakers might have tried to fill the noise with their own voice, in order to distract themselves from the noise. Varadarajan & Hansen (2006) interpreted this result as "a sense of urgency on the part of the speaker [. . .] due to persistent exposure of the environmental noise" (Varadarajan & Hansen 2006: 938).

With respect to our research question Q2, we can conclude that we found only some small differences in the three prosodic boundary cues produced in coordinates elicited in different contexts. In addition, the small differences between the contexts could hardly be discriminated on the perceptual side as shown by the weak performance in the perception check regarding the assignment of productions to the differential contexts: listeners were not able to reliably identify to whom the utterance was addressed.

With regard to the question of situational in/dependence of prosodic disambiguation, the finding of clear production of a prosodic boundary to disambiguate the conditions with/without grouping (Q1) together with the only small contextual adaptations (Q2) in our data, speaks in favour of situational independence. In the context of our study, the prosodic distinction between coordinates with or without internal grouping might have been considered to be more "relevant" than a prosodic adaptation to possibly different needs of the interlocutors.

In the following we will discuss two limitations of our study, before turning to research questions Q3 and Q4:

First, another explanation for the fact that the context effects in our production data were smaller than expected might be based on the somewhat artificial design of the study: the interlocutors were auditorily present before the recording of each stimulus, however, there was no feedback of their perceptual performance. A request for repetition or a misunderstanding may have triggered further accommodations in the speech addressed to the interlocutors. As mentioned above, accommodation to possible needs of an interlocutor can also be interpreted as patronizing by the interlocutor, as Kemper et al. (1998) reported for the speech used by young adults when addressing elderly adults. In our study, speakers either may have perceived no need to adapt any further to their interlocutors or they might have been sensible and avoided an over-exaggerated speech style since no feedback was given. This can apply especially for the elderly adult and the young non-native speaker, as they are both adults. Future studies, nevertheless, might want to include feedback of the interlocutors in order to increase the necessity of speakers to adapt to their interlocutors and to make the interaction more natural.

Second, we focused on three particular prosodic boundary cues and, therefore, cannot disregard the possibility that speakers may have produced additional prosodic cues to adapt to their interlocutors. This could, for instance, apply to the NOISE context: the context NOISE was best identified in the perception check (17 out of the 21 productions that were correctly identified by all listeners had been produced in the NOISE context and a total of 65 productions in NOISE was correctly identified by at least 75% of all listeners). This suggests that speakers used additional (prosodic) cues to adapt to the noise. Other studies looking at speech in noise reported, for instance, increased intensity in the presence of noise, as well as spectral changes (e. g., Davis et al. 2006, Junqua 1996, Landgraf et al. 2017, Lu & Cooke 2008, van Summers et al. 1988). This could be seen as further evidence that disambiguating prosody is not primarily produced for the interlocutor but automatically produced "for" the speaker during planning and articulation: When speakers are confronted with noise, this might affect the cognitive resources used for the planning and articulation and hence get reflected in their prosodic output. Future studies are needed to test this hypothesis.

In the final two sections, we discuss the results of the exploratory analysis regarding which cue combinations are used by individual speakers to mark the prosodic boundary in the grouped name sequences.

Inter-speaker variability of cue combinations (research question Q3)

With regard to inter-speaker variability of prosodic cues and cue combinations (i.e., the interplay of prosodic cues) in the YOUNG context, the data show that the majority of the speakers (14 out of 15) employed at least two cues distinctively to mark the prosodic boundary in the condition with grouping on Name2. Furthermore, for 13 speakers these two clearly distinctive cues were rise and pause. To put it simply: the vast majority of speakers clearly used pause and rise on Name2 to distinguish between conditions. In comparison to rise and pause, final lengthening was used more variably in the YOUNG context: some speakers produced it clearly distinctively, others with partial or no distinction. A post-hoc exploratory visual inspection of the data points that were excluded after the perception check further showed that the "clear distinction"-pattern in either of the three prosodic cues was beneficial for perception: often the perception of the non-intended condition went along with one of the three prosodic cues falling within the range of the values of the perceived condition. In other words, if for instance a grouped item was perceived as having no internal grouping, the value of one of the three prosodic cues was more similar to other items without grouping of that speaker than to grouped items.

Overall, most speakers combined at least two cues to clearly disambiguate the conditions, but still, there is some variability between speakers. This speaks in favour of a close relation of syntax and prosody that nonetheless allows for some flexibility in how prosodic boundaries are phonetically realized at the surface (Wagner 2005: 155). Despite this variability between speakers at the phonetic level, the boundaries are easily and reliably detected by the listeners, as shown by the perception check.

Intra-speaker variability of cue combinations (research question Q4)

This discussion concerns the question whether individual speakers mark the boundaries on Name2 differently in the five contexts. Mirroring the group analysis (see 5.4 on page 77), almost half of the speakers (7 out of 15) were stable across contexts also with regard to the relation between cues, as they used one or two patterns only. A closer look at these speakers revealed that the patterns they used mostly contained alternations in one cue only and were, consequently, quite similar to each other. Again, final lengthening emerges as the cue used least distinctively of the three cues investigated, while rise and pause in most cases clearly distinguish between the two conditions–also across contexts. In conclusion, in terms of cue patterns used, the differences between contexts were quite small and individual speakers rather stuck to their individual “prosodic strategy” of marking the boundaries in the condition with grouping independent of their interlocutor.

Overall, we can summarize that individual speakers showed a limited set of cue patterns with only slight shifts in cue distribution between contexts. Hence, also the analysis of individual speakers in varying contexts is in favour of a relatively limited range of variability or rather stable intra-individual “prosodic strategies” to disambiguate coordinates with vs. without internal grouping. This adds to the notion of situational independence of disambiguating prosody that is produced automatically by the speakers in a rather invariant manner.

5.5 Conclusion

In conclusion, speakers in our production study used prosodic boundaries to reliably mark constituent grouping in sequences of three coordinated names. At the phonetic level, speakers mainly used f0-range and pause for prosodic disambiguation, while final lengthening was used more flexibly. Across contexts, speakers behaved in accordance to the Proximity/Anti-Proximity principle of the syntax-prosody model by Kentner & Féry (2013): when the first two names were grouped together, the durational and tonal cues of the first name were weakened, while the boundary on the second name was strengthened. We found only limited contextual effects within speakers, but inter-speaker variability in how the prosodic boundaries were phonetically realized. The data hence indicate a close link between syntax and prosody that is employed independently of the actual communicative situation with some flexibility at the surface.

Additional files

An Open Science Framework project page (https://osf.io/rnxej/) has been created to store the data and code. Further additional files for this article can be found in Appendices A–D.

Acknowledgements

We would like to thank J. Ries for his help with setting-up the experiment and for the support with the exploratory analysis as well as L. Junack, A. Hofmann, and E. Weiß for their help with data acquisition and pre-processing, and D. Schad and R. Kliegl for their advice on data analysis. Another thank goes to the four interlocutors for lending us their voices and faces. We also thank B. Höhle and two anonymous reviewers as well as the associate editor for providing valuable comments to an earlier version of this manuscript.

Funding information

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 317 633 480 – SFB 1287, Project B01.

Competing interests

The authors have no competing interest to declare.

Appendix A: Wording of introduction and instruction in the five contexts

German original	Translation into English
YOUNG Introduction Hallo, mein Name ist Hannah. Ich bin 24 Jahre alt und studiere in Potsdam Biologie. Ich bin geboren und aufgewachsen in Eberswalde und bin vor vier Jahren hier zum Studium nach Potsdam gekommen. Zurzeit wohne ich in einer WG. Besonders gut an Potsdam gefällt(s) mir, dass hier so viel Parks sind und dementsprechend alles so schön grün ist.	YOUNG Introduction Hello, my name is Hannah. I am 24 years old and I study biology in Potsdam. I was born and raised in Eberswalde and moved to Potsdam for my studies four years ago. Currently, I am living in a shared flat. In Potsdam I like especially the many parks and that everything is therefore green.
YOUNG Instruction Ich bin jetzt Ihre Gesprächspartnerin. Auf dem Bildschirm sehen Sie gleich drei Namen mit Klammern. Ich sehe die Namen ohne Klammern. Sagen Sie mir die Namen bitte so, dass ich so genau und schnell wie möglich verstehe, wer gemeinsam kommt. Ich frage Sie gleich immer "wer kommt?" und Sie sagen mir die Namen.	YOUNG Instruction I am your interlocutor now. On the screen you will see three names with brackets. I see the names without brackets. Please say the names (to me), in such a way that I understand as accurately and rapidly as possible, who is coming together. I will ask you "who is coming?" and you will say the names (to me).
CHILD Introduction Hallo, ich bin Carlotta und bin sechs Jahre alt und komme aus Potsdam. Ich gehe gerne in die Schule, meine Mama oder mein Papa holen mich ab. Und dann mach ich meine Hausaufgaben. Ich reite sehr gerne und kann schon gut schwimmen.	CHILD Introduction Hello, I am Carlotta and I am six years old and I am from Potsdam. I like going to school, my mum or my dad pick me up. And then I do my homework. I like horse riding and I am good at swimming.

child Instruction
Jetzt sprichst du mit mir. Auf dem Bildschirm siehst du gleich drei Namen. Sag mir bitte die Namen so, dass ich immer gut verstehe, wer zusammen kommt. Ich frage immer “wer kommt?” und dann sagst du mir die Namen.

child Instruction
Now you are going to talk to me. On the screen you will see three names. Please say the names in such a way that I can understand well who is coming together. I will ask you “who is coming?” and then you say the names (to me).

elderly Introduction
Hallo, ich bin Frau Korbmacher, Maria. Und 82 Jahre bin ich. Früher, da hab ich als Lehrerin gearbeitet und jetzt bin ich schon länger in Rente. Seit zwei Jahren, wohne ich mit meinem Mann in einem Seniorenheim in Potsdam. Zu Hause war uns Vieles schon sehr anstrengend. Und im Alter, da wird man auch ein bisschen schusselig und ich vergesse häufiger mal was.

elderly Introduction
Hello, I am Mrs. Korbmacher, Maria. And I am 82 years old. In the past, I worked as a school teacher, but now I retired a while ago. For two years, I live in an old-age home in Potsdam with my husband. At home many things got demanding. And with increasing age, one tends to become scatty and I forget things from time to time.

elderly Instruction
So, ich bin jetzt Ihre Gesprächspartnerin. Und auf dem Bildschirm sehen Sie Namen, die sind geklammert. Ich sehe aber die Klammern nicht. Sagen Sie mir die Namen sodass ich so schnell und genau wie möglich verstehe, wer gemeinsam kommt. Ich frage Sie jedes Mal “wer kommt?” und dann sagen Sie mir die Namen.

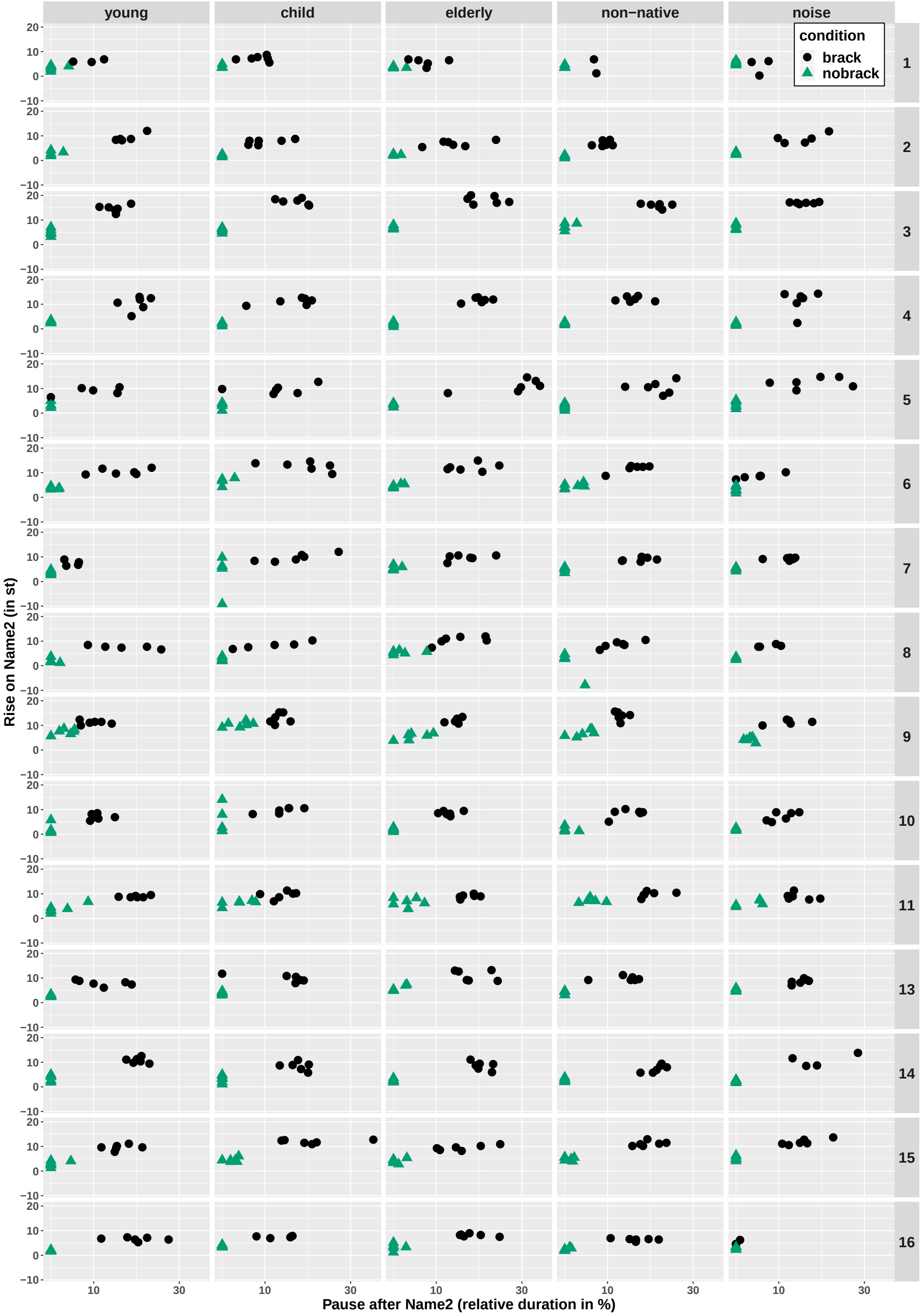
elderly Instruction
Now I am your interlocutor. On the screen you are going to see names that are bracketed. I don’t see the brackets. Say the names (to me) in a way that I can understand as rapidly and accurately as possible who is coming together. I will ask you each time “who is coming?” and then you say the names (to me).

non-native Introduction
Ich bin Zsófi. Ich bin 26 Jahre alt und bin Austauschstudentin an der Universität Potsdam. Ich lerne Deutsch seit einem Jahr. Es macht mir Spaß, aber Deutsch ist gar nicht so einfach. Ich hoffe, dass mein Deutsch schnell besser wird. Ich wohne jetzt in Potsdam in einer WG. Ich mag Potsdam sehr und mache gerne Sport.

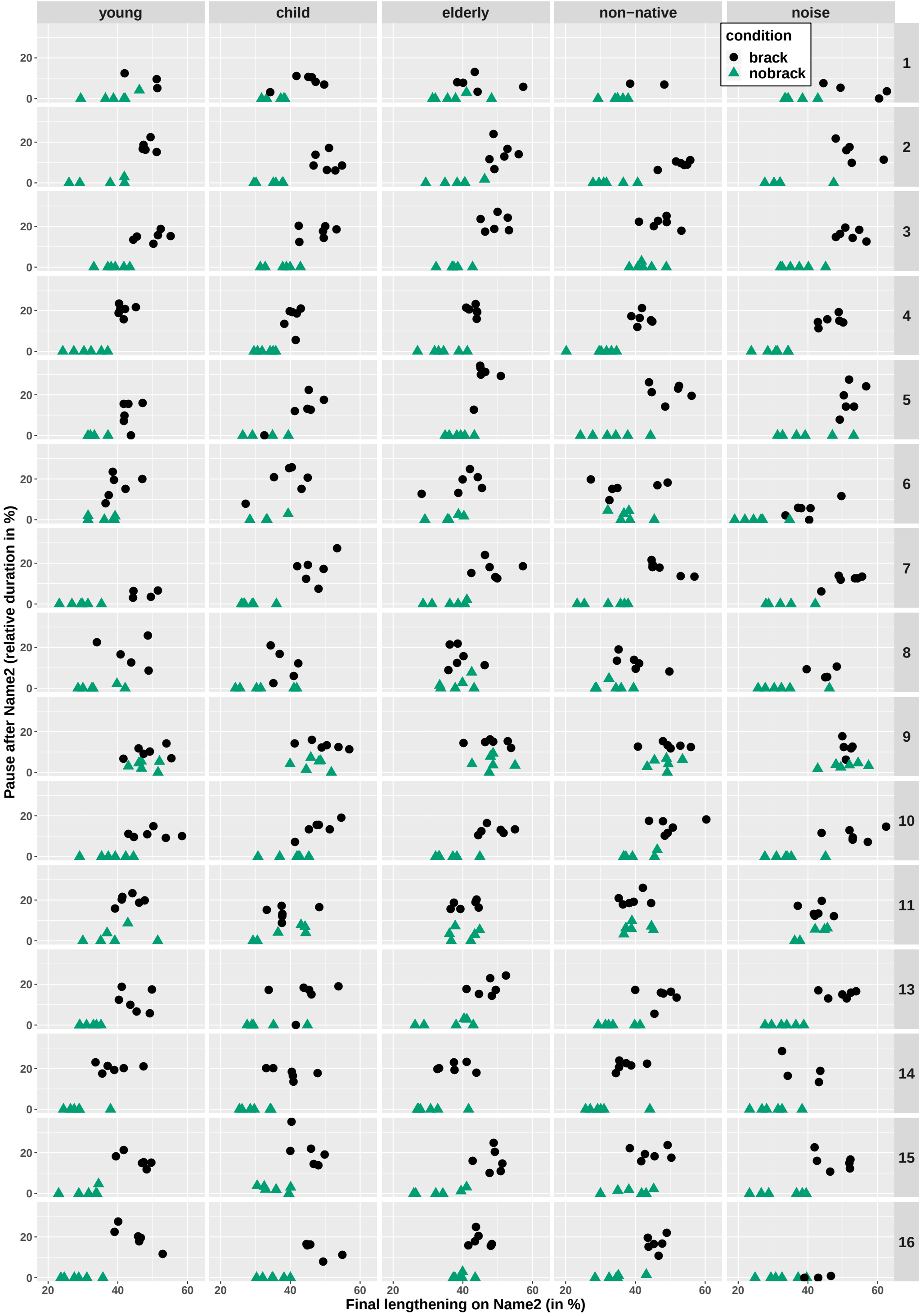
non-native Introduction
I am Zsófi. I am 26 years old and I am an exchange student at the University of Potsdam. I have been studying German for one year. I like it, but German is not easy. I hope that my German will get better soon. I live in a shared flat in Potsdam. I really like Potsdam and I enjoy doing sports.

NON-NATIVE Instruction Ich bin jetzt Ihre Gesprächspartnerin. Auf dem Bildschirm sehen Sie drei Namen mit Klammern. Ich sehe die Namen ohne Klammern. Sagen Sie mir die Namen bitte so, dass ich so schnell und genau wie möglich verstehe, wer gemeinsam kommt. Ich frage jedes Mal “wer kommt?” und dann sagen Sie mir die Namen.	NON-NATIVE Instruction I am now your interlocutor. On the screen you are going to see three names with brackets. I see the names without brackets. Please say the names (to me) in such a way that I can understand as rapidly and accurately as possible who is coming together. I willl ask you each time “who is coming?” and then you say the names (to me).
NOISE Introduction Hallo, ich bin es wieder, Hannah, die Biologiestudentin. Wir sollen das Ganze jetzt nochmal machen, allerdings ist es gerade sehr unruhig und laut, weil irgendwas im Hintergrund rauscht; aber das hören Sie ja selber. Ich sage Ihnen jetzt nochmal, was die Aufgabe ist.	NOISE Introduction Hello, it’s me again, Hannah, the biology student. We are supposed to do the same thing again, however it is currently quite turbulent and noisy, because something in the background is making noise; but you hear it yourself. I tell you the task again.
NOISE Instruction Ich bin jetzt wieder Ihre Gesprächspartnerin. Auf dem Bildschirm sehen Sie gleich jeweils drei Namen mit Klammern. Ich sehe die Namen ohne Klammern. Sagen Sie mir die Namen bitte so, dass ich so schnell und genau wie möglich verstehen kann, wer gemeinsam kommt. Ich frage Sie gleich “wer kommt?” und Sie sagen mir die Namen.	NOISE Instruction I am your interlocutor again. On the screen you are going to see three names with brackets. I see the names without brackets. Please say the names (to me) in such a way that I can understand as rapidly and accurately as possible who is coming together. I will ask you “who is coming?” and you will say the names (to me).

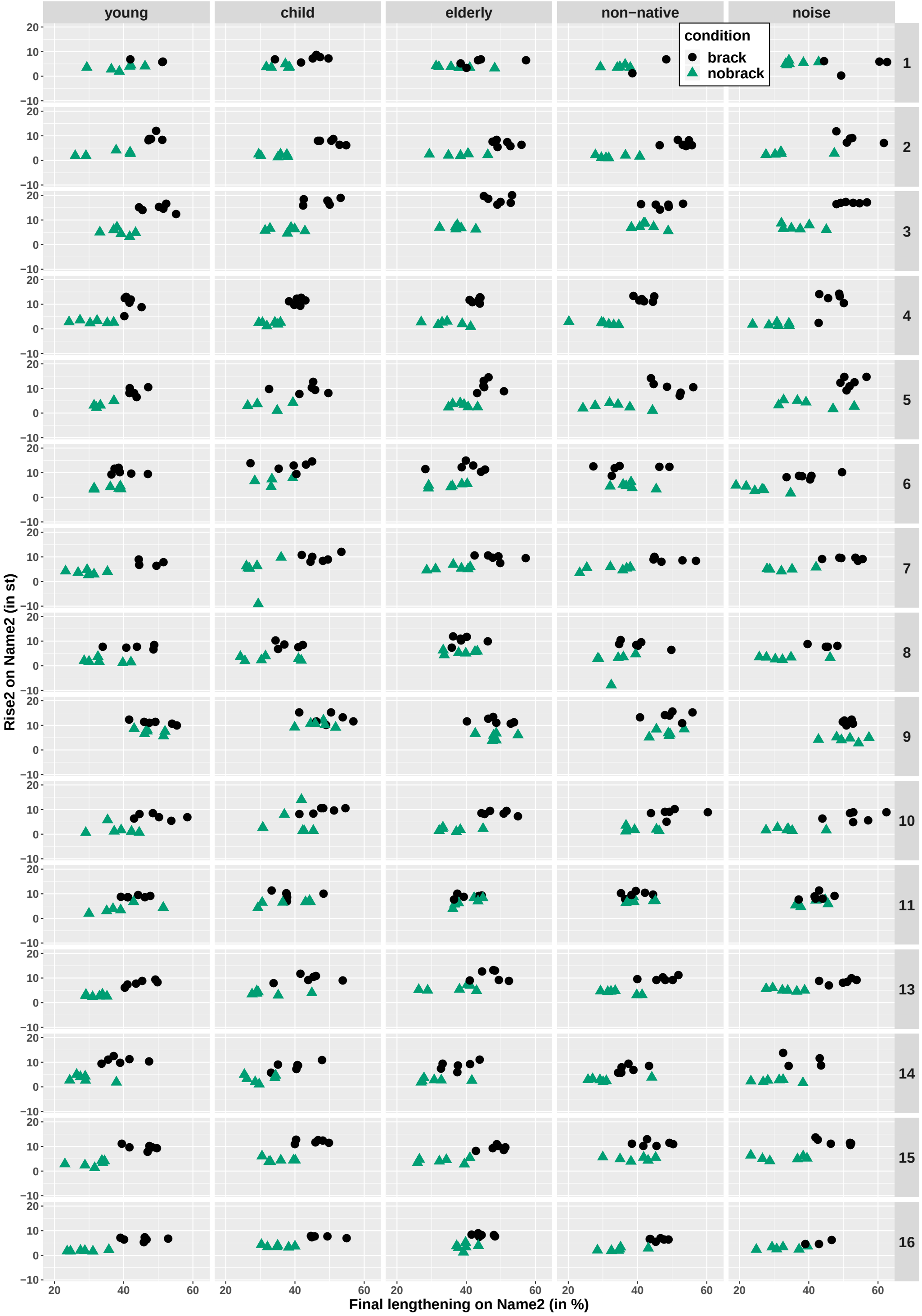
Appendix B: Comparison pause_rise on Name2 with pause plotted on the x-axis and rise on the y-axis. Distribution of datapoints in conditions (black circles: brack, green triangles: nobrack) for each speaker (cf. facets on the right) in each context (cf. facets on the top). The cell in row 4, column 4 (non-native) gives an example of a Cue1-NO_Cue2-NO pattern.



Appendix C: Comparison lengthening_pause on Name2 with final lengthening plotted on the x-axis and pause on the y-axis. Distribution of datapoints in conditions (black circles: brack, green triangles: nobrack) for each speaker (cf. facets on the right) in each context (cf. facets on the top).



Appendix D: Comparison lengthening_rise on Name2 with final lengthening plotted on the x-axis and rise on the y-axis. Distribution of datapoints in conditions (black circles: brack, green triangles: nobrack) for each speaker (cf. facets on the right) in each context (cf. facets on the top). The cell in row 10, column 2 (child) gives an example of a Cue1-PO_Cue2-CO pattern.



6 Study II

Age effects on linguistic prosody in coordinates produced to varying interlocutors: Comparison of younger and older speakers[27]

Abstract

This production study builds on and extends the research on how prosodic cues can be used to resolve syntactic ambiguities. We compared how younger speakers (mean age 25 years, Huttenlauch et al. 2021) and older speakers (mean age 68 years) produced prosodic cues to distinguish between structurally different coordinated three-name sequences without and with internal grouping of the first two names (*Name1 and Name2 and Name3* compared to *(Name1 and Name2) and Name3*). The prosodic cues of interest were variations in f0 (F0 range), duration of segments at the end of the names (final lengthening), and pause insertion. In line with the Proximity/Similarity model by Kentner & Féry (2013), we found that both age groups used all three cues to signal the grouping: Prosodic cues were modified on the group-internal Name1 as well as on Name2 at the right-most element of the group. These prosodic cues were clearly understood by naïve listeners. The study also found that successful prosodic disambiguation was not affected by age-related differences in speech production, such as longer durations or greater variability in the speech of older speakers. Furthermore, we analysed the productions with regard to different contexts, such as addressing interlocutors of different ages and mother tongues, and in noisy environments. We found that both age groups of speakers used the same prosodic cues consistently across all contexts, indicating that the use of prosodic cues to clarify syntactic ambiguities is a stable part of the production process, which we interpret as being in line with models of situational independence of disambiguating prosody (Schafer et al. 2000, Kraljic & Brennan 2005, Speer et al. 2011). Our study provides evidence that the use of these prosodic cues (F0 range, final lengthening, and pause) is a reliable way to clarify ambiguous structures in speech and independent of the speaker's age.

[27] An adapted version of this chapter has been published as Huttenlauch, Clara, Marie Hansen, Carola de Beer, Sandra Hanne & Isabell Wartenburger. 2023. Age effects on linguistic prosody in coordinates produced to varying interlocutors: Comparison of younger and older speakers. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, 157–192. Berlin: Language Science Press. doi:10.5281/zenodo.7777534.

6.1 Introduction

Linguistic prosody, as in prosodic boundaries, can be used to resolve syntactic ambiguities. Such syntactic ambiguities exist in coordinated sequences of more than two elements (e. g., names) since those elements can be grouped internally at different levels. For instance, the three-name sequence *Moni and Lilli and Manu* can describe three individual persons or a group of three persons (i. e., no internal grouping as in (28)) or a group of two persons in addition to one individual person, with two different possibilities for the grouping (i. e., the group can consist of *Moni and Lilli* or of *Lilli and Manu*. (29) gives an example for the internal grouping of *Moni and Lilli* indicated by parentheses). The latter two different groupings correspond to underlying syntactic structures that differ in their direction of embedding. The difference to the first sequence is the depth of embedding. The absence or type of internal grouping as in (28) versus (29) in an answer to the question 'Who will plant a tree?' results in either one, or two, or three planted trees. Prosody, thus, brings the underlying structure to the surface (i. e., disambiguates the otherwise ambiguous surface structure). In this study, we will compare productions of a structure without internal grouping (28) to a structure with internal grouping of the first two elements (29).

(28) Name1 and Name2 and Name3. – without internal grouping

(29) (Name1 and Name2) and Name3. – with internal grouping

Prosodic marking in coordinate sequences

In German, the difference between the two structures (i. e., the resolution of the structural ambiguity) is mainly indicated by one or more of three prosodic cues: F0 change, final lengthening, and pause (Peters et al. 2005, Gollrad et al. 2010, Kentner & Féry 2013, Petrone et al. 2017, for final lengthening see also Schubö & Zerbian 2023). Young speakers have been shown to use these three prosodic cues to clearly mark the internal grouping of coordinated name sequences (Kentner & Féry 2013, Petrone et al. 2017, Huttenlauch et al. 2021). Figure 15 provides visualisations of waveform and spectrogram with F0 contour and segmental annotations of productions without and with internal grouping, respectively, generated using Praat (Boersma & Weenink 2017). The marking of the internal grouping appears as a global and not a local phenomenon, in accordance with the Proximity/Similarity model (Kentner & Féry 2013): Young speakers modified prosodic cues not only at the right edge of the internal group (i. e., on Name2 in the example in (29)), but already earlier in the utterance (i. e., on Name1, see also left and right panel in Figure 15, Kentner & Féry 2013, Huttenlauch et al. 2021). The principle of Proximity relates to the syntactic constituent structure (Kentner & Féry 2013). The proximity of syntactically grouped elements is expressed by a weakening of the prosodic cues (e. g., less final lengthening, lower F0 peak, smaller F0 range) on the leftmost element of two sister elements (e. g., Name1 in (29), Moni in right panel of Figure 15)

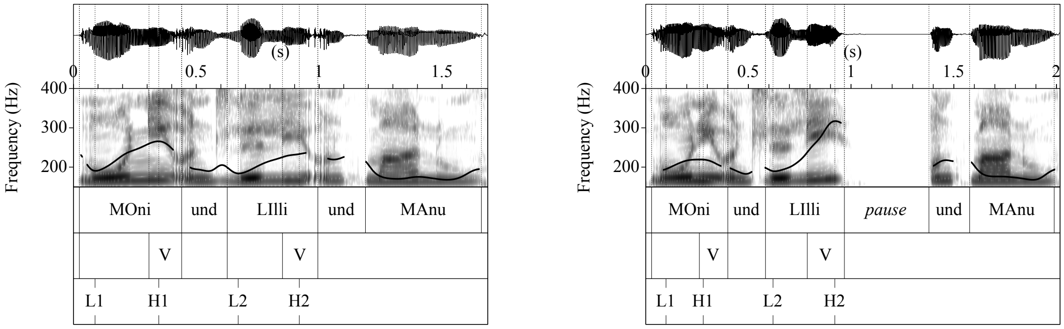


Figure 15: Waveform and spectrogram with F0 contour (black line) of the coordinated name sequence *MOni und LIlli und MAnu* (capital letters correspond to stressed syllable) produced without internal grouping (left) and with internal grouping (right) by a young female speaker. The TextGrid gives an example for the manual annotation of low (L) and high (H) F0 values and the segmentation of the final vowels within Name1 and Name2.

compared to an ungrouped element in the same position (e. g., Name1 in (28), Moni in left panel of Figure 15). The principle of Anti-Proximity predicts a strengthening of the prosodic cues (e. g., more final lengthening, higher F0 peak, larger F0 range, insertion of a pause) on/-after the right-most element of a group than on/after an ungrouped element (e. g., Name2 in (29) versus in (28), Lilli in right versus left panel of Figure 15). The principle of Similarity relates to the depth of syntactic embedding and since it does not apply to our structures we will not discuss it further. In summary, in name sequences with grouping such as (29), the productions of Name1 contain weaker prosodic cues and those of Name2 encompass stronger prosodic cues compared to name sequences without grouping such as (28).

In perception, the early cues on Name1 could reliably be recovered to predict the upcoming structure by more than half of the participants in a two-alternative forced choice decision task with gated stimuli (Hansen et al. 2022). Although all young speakers in Huttenlauch et al. (2021) reliably marked the constituent grouping of coordinated names, they showed inter-speaker variability in how they phonetically realised the prosodic boundary, especially final lengthening was used in a more flexible way than F0 range and pause. Besides prosodic disambiguation, Huttenlauch et al. (2021) investigated the situational (in)dependence of disambiguating prosody by comparing prosodic cues addressed to interlocutors differing in age and mother tongue as well as in the absence/presence of background noise. Despite the phonetic variability in the realisation of prosodic cues between speakers, the data show a rather consistent pattern of prosodic cues across different communicative situations. The latter finding was interpreted as indexing situational independence: Disambiguating prosody seems to be produced automatically by the speakers in a rather invariant manner.

The present study builds on and extends the results on prosodic boundary production of young speakers (Huttenlauch et al. 2021) with productions of older speakers. Data of both age groups were elicited with the same design and materials, which allows for a direct comparison and detailed investigation of age effects. Age has not only been shown to affect language production in terms of word-finding abilities (for a review see Burke & Shafto 2004) but also in terms of altered acoustic characteristics affecting prosody-related features in the tonal and durational domain. Age, thus, has an effect on the same features that are relevant for the realisation of linguistic prosody.[28] Age, therefore, may interact with the modulation of prosodic cues in conveying the intended meaning. In the remaining part of the introduction, we will address age-related changes in the tonal and durational domain in general (Section 6.1 on page 92) and their possible impact on the use of linguistic prosody in particular (Section 6.1 on page 93). Finally, we will present findings on the situational (in)dependence of prosodic cues (Section 6.1 on page 94).

Age-related changes in the tonal and durational domain in general

In the following section, we will summarise previous research on general age-related changes in the tonal and durational domains. It is important to note that studies differ in how they group participants into age ranges and in how many years each age group spans. We will use *young* or *younger speakers* to refer to the age range between 18 and 30 years of age and *older speakers* for ages above 60 years.

In the tonal domain, age effects on fundamental frequency (F0) have been studied for several measures including mean and median F0, the span between minimum and maximum (F0 range), and the variability of those measures captured in standard deviations (*SD*). Here, we focus on the latter two as mean or median F0 are rather uninformative in the context of our study, which focuses on analysing F0 range. So far, results are inconclusive and in part divergent between genders. For F0 range, some studies report no differences between younger and older speakers (Markó & Bóna 2010, Smiljanic & Gilbert 2017), while Dimitrova et al. (2018), Tuomainen & Hazan (2018), and Hazan et al. (2019) observed a larger F0 range for older compared to younger women and Kemper et al. (1998) found a smaller F0 range in older compared to younger speakers irrespective of gender. When it comes to F0 variability, there is evidence for an increase with increasing age (Scukanec et al. 1992, Lortie et al. 2015, Santos et al. 2021). More variability and less stability in older speakers compared to younger speakers was further noticed by several studies looking at more specific measures regarding speech acoustics (including jitter, shimmer, and noise-to-harmonics-ratio; Goy et al. 2013, Lortie et al. 2015, Rojas et al. 2020 among others).

In the durational domain, previous studies observed slower speaking/articulation rates

[28] We are aware of the multitude of non-linguistic information transmitted through prosodic cues including but not limited to the emotional state and background of the speaker. In the context of this study, we are only interested in linguistic prosody.

in older compared to younger speakers (Tuomainen & Hazan 2018, Hazan et al. 2019, Tuomainen et al. 2019, 2021 and references in a review by Tucker et al. 2021: 5), relating this finding mainly to longer syllable or word durations (Scukanec et al. 1996, Harnsberger et al. 2008, Barnes 2013, Dimitrova et al. 2018), longer segment durations (Kemper et al. 1995, Harnsberger et al. 2008, Smiljanic & Gilbert 2017), or an increased number of pauses (Kemper et al. 1998, Dimitrova et al. 2018). However, no evidence for pause duration as a driver of age-related differences in speech rate has been reported so far (Barnes 2013, Smiljanic & Gilbert 2017, Dimitrova et al. 2018).

To sum up, previous researchers provided some evidence for tonal and durational differences between younger and older speakers, indicating increased F0 ranges and durations with increasing age. Since these changes affect the same channel used to convey linguistic meaning, we will address possible interferences in the next paragraph.

Age-related changes in the tonal and durational domain alongside linguistic prosody

We will now turn towards studies that can help to address the question of whether age-related changes in the tonal and durational domain interact with the modulation of disambiguating prosodic cues, as these studies used speech material that explicitly required the use of linguistic prosody. Scukanec et al. (1996) measured the maximal F0 value within the vowel of elicited monosyllabic words in either contrastive or non-contrastive stress position in younger and older female English speakers. Both age groups used F0 in a similar way to mark the focused words (Scukanec et al. 1996: 235). However, independent of the word position in the sentence, older speakers produced higher F0 values than young speakers in words with contrastive stress and lower maximal F0 values in words in non-contrast positions. The authors concluded that, for the analysed data set, age did not influence the productions of "linguistically salient variations in prosodic output" (Scukanec et al. 1996: 238). The difference in the maximal F0 values between words with and without contrastive stress was even larger in older than in young speakers. The same holds true for the durational domain: Even though older speakers produced longer word durations together with larger standard deviations (i. e., more variability), both age groups used duration to linguistically distinguish stressed from unstressed words.

Further evidence that older speakers use lengthening for prosodic disambiguation despite an overall age-related slower speaking rate comes from Tauber et al. (2010) and Barnes (2013) who reported longer durations for older English speakers in disambiguating contexts. Barnes (2013) elicited structurally ambiguous sentences with either high or low attachment of the prepositional phrase (e. g., *The girl hit the boy with the fan*) in younger and older English speakers. Although the study found longer durations of the direct object and the prepositional phrase regardless of target in the productions of older speakers than in the productions of younger speakers, the overall results revealed that both age groups used the prosodic cues mean F0, pause duration, word duration, and mean intensity similarly to disambiguate ambiguous sentences. However, in another task tapping production of lexical stress to differentiate noun-verb pairs with strong-weak and weak-strong stress patterns, "older adults utilised F0 to a significantly greater extent than young adults" (Barnes 2013: 43). Tauber and colleagues elicited structurally ambiguous sentences (e. g., *The lake froze over a month ago*) to explicitly test for age differences in the realisation of disambiguating prosody in English sentences (Tauber et al. 2010). They found that intonational boundaries (defined as pause duration plus duration of the critical word at the boundary) were longer in older than in younger speakers. Notably, both age groups seem to have had difficulties with the task, as the percentage of sentences which were successfully disambiguated via prosody was 66% for older speakers (above chance, $p < .05$) and 59% for the younger age group (not significantly above chance, $p > .06$) (Tauber et al. 2010).

In summary, even though age leads to changes in the tonal and temporal domain in general, there is evidence from English speakers that the modulation of prosody to convey linguistic meaning remains unaffected. Older participants even appear to produce prosodic cues in a more extreme way than younger speakers. To the best of our knowledge, there is no study that addressed age differences in the use of prosody to resolve ambiguities in coordinate structures. If the findings for English ambiguous sentences are transferable to German coordinate structures, we expect that older speakers disambiguate coordinate structures using more extreme prosodic cues than young speakers. This motivates our first research question:

RQ1 Prosodic disambiguation of coordinate name sequences: Do older speakers compared to young speakers show a more extreme use of the three prosodic cues F0 range, final lengthening, and pause on Name1 and Name2 to mark the internal grouping of coordinates in German?

Situational (in)dependence of prosodic cues

In the remaining part of the introduction, we will address the situational (in)dependence of prosodic cues, a second topic investigated in Huttenlauch et al. (2021). It deals with the effects of different types of interlocutors and the absence/presence of noise on the use of disambiguating prosodic cues. Huttenlauch et al. (2021) compared the use of prosodic cues in five *contexts* involving four female interlocutors: a young adult (YOUNG), a child (CHILD), an elderly adult (ELDERLY), and a young non-native speaker of German (NON-NATIVE) and in noise (the young adult with background white noise, NOISE). The productions directed at the young adult native speaker (i. e., the context YOUNG) were taken as a baseline for comparisons. The findings showed stability in the use of prosodic cues for disambiguating the internal structure of coordinates (Huttenlauch et al. 2021). That is, individual speakers produced a limited set of cue patterns with only slight shifts in cue distribution across different contexts. This stability in prosodic patterns for disambiguation irrespective of the context was interpreted in favour of models of situational independence of disambiguating prosody (Schafer et al. 2000, Kraljic & Brennan 2005, Speer et al. 2011). These models predict that disambiguating prosody is produced in an automatic way, for the sake of the speakers themselves, and hence depends neither on the presence or absence of an interlocutor, nor on the type of interlocutor or situational setting (e. g., background noise). Despite arguing for situational independence of disambiguating prosody, Huttenlauch et al. (2021) found slight prosodic modifications in the data that can be attributed to context effects. Similarly, as discussed for the prosodic marking of internal grouping of coordinates in the first part of the introduction, the question arises whether age effects in the tonal and durational domain have an impact on the use of F0 range, final lengthening, and pause when speaking in different contexts and whether we find age effects in the situational (in)dependence of prosodic disambiguation. Research on age effects in speech production to different interlocutors is, to our knowledge, still scarce. In the following, we will briefly summarise existing findings including the context effects found in the productions of young speakers in Huttenlauch et al. (2021).

With regard to addressing a child interlocutor, we will refrain from summarising the immense body of literature treating speech towards preverbal infants since the use of prosody for disambiguation requires that language ability has already been acquired to a certain extent. We are not aware of studies investigating effects of speaker age on prosodic cues uttered towards a child interlocutor. For young speakers, speech towards a child interlocutor has been described as containing an increased F0 range (Biersack et al. 2005, Huttenlauch et al. 2021), lengthened vowels (Biersack et al. 2005), or more pauses (DePaulo & Coleman 1986).

Speech addressing an elderly interlocutor has been explored in data on young and older adult speakers. While younger speakers slowed down their speaking rate by increasing vowel duration and inserting more pauses in speech addressing an elderly interlocutor, older speakers did not do so (Kemper et al. 1995). For older speakers addressing a young interlocutor, however, Kemper and colleagues observed a slower speaking rate than for young speakers. The authors argued that, in comparison to young speakers, older speakers adopt a more simplified speech style including lower speaking rate when addressing a young interlocutor, and thus it is possibly hard for them to slow down even further in order to adapt to an elderly interlocutor (Kemper et al. 1995: 56). Furthermore, young speakers addressing an elderly interlocutor, slowed down their speaking rate with longer pauses, increased final lengthening (Huttenlauch et al. 2021), and increased F0 range or variation in F0 (Thimm et al. 1998, Huttenlauch et al. 2021).

We are not aware of studies investigating effects of speaker age on prosodic cues when addressing a non-native interlocutor. Some studies involving young speakers found no clear differences (DePaulo & Coleman 1986, Uther et al. 2007, Knoll & Scharrer 2007, Knoll et al. 2011, Huttenlauch et al. 2021), while others observed a lowered speech rate due to lengthened pauses (Biersack et al. 2005), a higher mean F0 (Knoll et al. 2015), increased word durations and intensity (Rodriguez-Cuadrado et al. 2018), or an increased F0 range along with segmental modifications described as a more emphatic style (Smith 2007; see Piazza et al. 2021 for a current review on foreigner-directed speech).

Finally, speech in noisy environments compared to silent environments is affected by modulations in several ways. The reported changes are referred to as "Lombard speech" (Lombard 1911 as cited in Zollinger & Brumm 2011) and include decreased speaking rate (due to increased segment or word durations), increased F0 ranges, increased signal amplitude, and spectral changes such as smaller spectral slope (e. g., Junqua 1996, van Summers et al. 1988, Jessen et al. 2003, Zollinger & Brumm 2011, Smiljanic & Gilbert 2017, Tuomainen et al. 2019, 2021). The findings for young speakers in a noisy environment in Huttenlauch et al. (2021) were interpreted as being partly in line with Lombard speech, as they revealed increased final lengthening and decreased pause duration but no changes in F0 range. With respect to age effects in speech adaptation to noise, no age differences were found by Dromey & Scott (2016) and Smiljanic & Gilbert (2017), with the latter reporting an age-independent decrease in speaking rate when noise was present, while Tuomainen et al. (2019) reported a decreased speaking rate only for the older age group.

To summarise, the modifications of prosodic cues in coordinates induced by varying contexts observed by Huttenlauch et al. (2021) were rather small but in line with previous findings. The effect of age on the realisation of prosodic cues in more communicative settings with varying interlocutors is still only scarcely explored. For the reported age-related changes in addressing different interlocutors, the question arises whether they replicate to coordinate structures in German. Given the limited evidence, we keep our second research question rather open:

RQ2 Situational (in)dependence: Do young and older speakers differ in adapting their use of prosodic cues when addressing varying interlocutors?

In the current study, we extend the age range of usually studied participants (in Huttenlauch et al. 2021 19–34 years) to older people aged between 60 and 80 years of age (i. e., comparable to the older age groups in the previously presented literature) and compare the productions of linguistic prosody in young and older adult speakers. Specifically, we explore whether age interacts with the modulation of prosodic cues, especially F0 range, final lengthening, and pause, and whether any such interaction may impact the disambiguation of structurally ambiguous coordinated name sequences and the use of prosodic cues when addressing different interlocutors (i. e., regarding situational (in)dependence of disambiguating prosody).

6.2 Methods and material

Methods, materials, and data of the younger speakers are taken from Huttenlauch et al. (2021) and extended by the data of older speakers.

Participants

Fifteen young monolingual German native speakers (13 female, 1 male, 1 other; age range: 19–34, mean 25.47 years, *SD*: 4.6; see Huttenlauch et al. 2021) and 13 older monolingual German native speakers (9 female, 3 male, 1 no information; age range: 61–80 years, mean: 67.77 years, *SD*: 6.8) were included in the study. Additional five speakers took part in the study, but were discarded due to low task compliance (n = 1), scores below 25 in the Montreal Cognitive Assessment (Nasreddine et al. 2005) (n = 3), or missing data (n = 1). All participants (henceforth *speakers*) were recruited in Potsdam, Germany, and were reimbursed or received course credits (the latter only applies to the young speakers). They were naïve to the purpose of the study and gave written consent to participate. The Ethics Committee of the University of Potsdam approved the procedure of this study (approval number 72/2016). Hearing ability was assessed by a hearing screening using an audiometer (Hortmann DA 324 series) and calculated following the grades of hearing impairment by the WHO as reported in Olusanya et al. (2019). Normal hearing was defined as an average pure-tone audiometry of 25 dB HL or better of 500, 1000, 2000, and 4000 Hz in the better ear. Following this definition, all 15 young speakers and 10 of the older speakers had normal hearing, the remaining speakers showed a slight (n = 2) or moderate impairment (n = 1).

Stimuli

Items As stimuli, we used the same six coordinated name sequences as in Holzgrefe-Lang et al. (2016), Huttenlauch et al. (2021), and Wellmann et al. (2023): Each sequence consisted of three German names coordinated by *und* (English 'and') that appeared in each of two conditions: without internal grouping (30) or with internal grouping of the first two names (31). The grouping of the first two names was visually indicated to the participants by bracketing Name1 and Name2 with parentheses as in (31). The conditions will henceforth be referred to as *brack* for the condition with internal grouping and *nobrack* for the condition without internal grouping. A total of 12 items was used. Young speakers produced each item once per context (see Section 6.2 on page 98), older speakers twice to enlarge the data set and to increase statistical power.

(30) Name1 and Name2 and Name3. Moni und Lilli und Manu.

(31) (Name1 and Name2) and Name3. (Moni und Lilli) und Manu.

	YOUNG *(baseline)*	CHILD	ELDERLY	NON-NATIVE	NOISE
Name:	Hannah	Carlotta	Maria Korbmacher	Zsófi	Hannah + white noise
Age (in years):	24	6	82	26	See YOUNG
Origin:	Eberswalde	Potsdam	NA	NA	
Residence:	Potsdam	Potsdam	Potsdam	Potsdam	
Occupation:	Biology student	School child	Retired school teacher	Exchange student	
Further facts:	Moved to Potsdam for her studies, lives in a shared flat, likes the parks in Potsdam	Likes horse riding, her parents pick her up from school is good at swimming,	Lives for two years in an old-age home with her husband, tends to forget things from time to time	Started to learn German one year ago, lives in a shared flat enjoys doing sports	

Table 11: Fictional names, ages, origins, and further information of the interlocutors present in the five contexts.

The set of coordinates comprised nine different German names in total, all of which were controlled for number of syllables (disyllabic), stress pattern (penultimate), and sonority of the segments (only sonorant material, to facilitate pitch tracking). Six of the names featured the high frontal vowel /i/ in word-final position (Moni, Lilli, Leni, Nelli, Mimmi, and Manni) in order to decrease glottalisation and occurred as Name1 or as Name2. Name3 contained either /u/ or /a/ in word-final position (Manu, Nina, and Lola). Regarding possible collocations of the selected names for each coordinate, there was no particular co-occurrence of two adjacent names (as in, e. g., "Bonnie and Clyde") in the dlexDB corpora (Heister et al. 2011) or in printed sources between 1500 and 2021, as ascertained by the Google Ngram Viewer (Lin et al. 2012).

Contexts Five different communicative contexts (YOUNG, CHILD, ELDERLY, NON-NATIVE, NOISE) were created that differed in the interlocutor and/or the absence/presence of background white noise (see Table 11). Speakers saw their interlocutors on a screen in two short videos each (one with a personal introduction of the interlocutor and one with instructions for the task) to get an audio-visual impression. The young and non-native interlocutors were similar in age to the group of young speakers, the elderly interlocutor was two years older than the oldest speaker in the group of older speakers. A more detailed description of the videos and interlocutors can be found in Huttenlauch et al. (2021).

Procedure

Productions were elicited by means of a referential communication task. Contexts were presented blockwise, always starting with the YOUNG context, which served as a baseline in

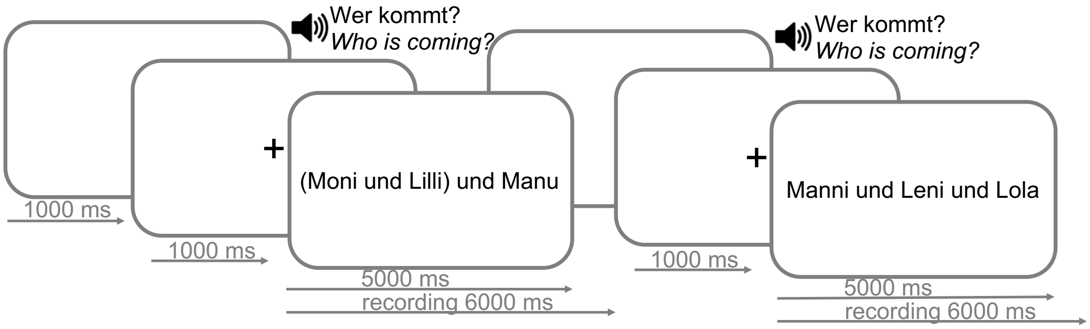


Figure 16: Experimental setting and timing of two trials.

the analysis. The order of the other four contexts was randomised. Each block started with the two video clips of the corresponding interlocutor. Then, for each trial, speakers first saw a fixation cross on the screen accompanied with the auditory presentation of the trigger question *Wer kommt?* ('Who is coming?') via headphones produced by the interlocutor of the current block as a reminder to whom they were talking. After 1000 ms, the fixation cross was replaced by the visual presentation of the name sequence (i. e., the item) in one of the two conditions (see Figure 16). The task was to produce the item in a way that would allow the interlocutor "to understand as rapidly and accurately as possible who is coming together". Recordings took place in a sound-attenuated booth at the University of Potsdam via an Alesis iO/2 audio interface using an AKG HSC271 headset with over-ear headphones and a condenser microphone. The wide screen in the recordings booth had a resolution of 1920 x 1200, stimuli were in Arial, font size 50. The experiment was run from a Dell laptop using Presentation software (Neurobehavioural Systems). Each item was presented in each context once (for young speakers) or twice (for older speakers). Thus, the data set contained 900 individual productions of young speakers (6 name sequences * 2 conditions * 5 contexts * 15 young speakers) and 1560 individual productions of older speakers (6 name sequences * 2 conditions * 5 contexts * 2 repetitions * 13 speakers).

Perception check

After data collection of the production study, all recordings were auditorily presented to naïve listeners who were asked to indicate for each production the perceived condition. To this end they were given two pictograms with three persons each, one pictogram per condition (Figure 17, picture A without and picture B with internal grouping). The aim of the perception check was to assess whether naïve listeners perceive the grouping of the coordinates in the way it was *intended*. By *intended* we refer to the indication of condition which was given to speakers by parentheses around the grouped names in the production study. Obviously, the intention of speakers at the time of the production remains unknown to us.

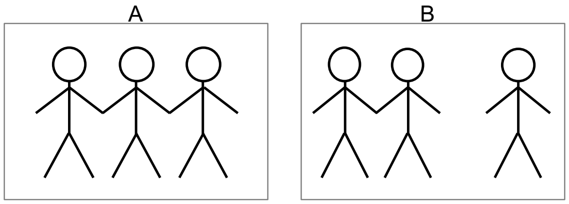


Figure 17: Pictograms used in the perception check depicting the condition without grouping (left panel) and with grouping (right panel).

The data of the young and older age group were rated separately. The recordings were distributed across different lists with 147 to 267 items. Each listener judged one list and each list was judged by seven or eight listeners.

The perception check of the productions of the young speakers was conducted in presence of several listeners in the same room with a paper-and-pen version. Data of 31 listeners (22 female, 9 male; age range: 18–41, mean: 24.1 years, *SD*: 5.8) were analysed. Another 11 listeners took part in the study, but had to be excluded due to technical problems (n = 9), German as a non-native language (n = 1) or a hit-rate 2 *SD* below the mean hit-rate of all listeners (n = 1, see Huttenlauch et al. 2021 for more details).

For the productions of the older speakers, the perception check was transferred onto OpenSesame (Mathôt et al. 2012) and was run as a web-based study on Jatos (Lange et al. 2015) in individual sessions. Data of 49 listeners (29 female, 9 male, 11 other/no information; age range: 18–63, mean: 24.63 years, *SD*: 6.3) were analysed. Another five listeners took part in the study, but had to be excluded due to technical problems.

In the analysis of the perception check, the exclusion threshold for individual productions was set to a hit-ratio 2 *SD* below the mean ratio, as suggested by standard assumptions on the exclusion of data points (e. g., Howell et al. 1998). Hit-ratio was calculated separately for each production as the number of congruent rates (i. e., correct identification of the intended grouping/condition, referred to as *hit-rate*) divided by the number of total rates. Applying this criterion, 36 productions (4%, 11 nobrack, 25 brack) in the group of the young speakers and 66 productions (4%, 39 nobrack, 27 brack) in the group of the older speakers fell below the threshold and were excluded from further analyses. For a more detailed description of procedure and analysis of the perception check see Huttenlauch et al. (2021).

Segmentation and measurements

In addition to the productions excluded based on the perception check, three productions were excluded from analysis in the data set of the older speakers: due to hesitations that made the analysis of condition impossible (n = 2) and due to recording problems (n = 1). The final data set comprised 2355 productions (young: 864, older: 1491). Table 12 provides an overview of how the productions distribute across age groups, conditions, and contexts.

Table 12: Distribution of productions entering statistical analyses across age groups, conditions, and contexts in the final data set.

age group	condition	YOUNG	CHILD	ELDERLY	NON-NATIVE	NOISE
younger	nobrack	87	85	90	90	87
	brack	83	88	89	86	79
older	nobrack	141	148	148	151	153
	brack	151	153	153	148	145

For the extraction of the three prosodic cues under investigation, segment boundaries and pauses were manually annotated in Praat (Boersma & Weenink 2017, version 6.0.32) by following the criteria in Turk et al. (2006). Silent intervals of at least 20 ms duration were considered as pauses (following the procedure in Petrone et al. 2017). F0-minima (L) and F0-maxima (H) on both Name1 and Name2, were manually annotated (example TextGrids are given in Figure 15). The points were set into parts of the signal, where F0 can be reliably measured (i. e., avoiding the edges of segments, glottalised parts in the signal, and parts with other non-modal voice quality). The F0 contour mostly displayed a rising movement on Name1 and Name2, respectively (i. e., L preceded H). Only in a few cases, speakers produced a falling F0 movement on Name1 (young speakers: 88 falls versus 776 rises, older speakers: 108 falls versus 1368 rises) or Name2 (older speakers: 13 falls versus 1458 rises). For some productions in the data of the elderly speakers it was impossible to find reliable locations to annotate either L and/or H points and it was, thus, impossible to measure the F0 range. In those cases, the corresponding item was excluded from the analysis of F0 range for Name1 and/or Name2. This applies to 15 items (1.0% of the productions of older speakers) in the condition without internal grouping and to 20 items (1.3% of the productions of older speakers) with internal grouping. All in all, we aimed for an approach of measuring F0 range that was applicable to the majority of the recordings. For further segmentation criteria see Huttenlauch et al. (2021). For Name1 and Name2 separately, we calculated the three variables F0 range, final lengthening, and pause. The variable F0 range reflects the range between the F0-minimum and the F0-maximum on NameX in semitones (st; calculated as $12 * log_2(F0_H/F0_L)$). The variable final lengthening reflects the duration of the final vowel of NameX divided by the duration of NameX (in %, the final vowel is annotated as V on the second tier of the TextGrid in Figure 15.). The pause variable reflects the duration of a possible pause after NameX divided by the duration of the whole utterance (in %). We chose relative instead of absolute measures as they are independent of individual speech rates and mean fundamental frequency. However, to descriptively assess potential age-related effects, absolute durational measurements were taken into consideration.

Statistical analysis

The workflow of the statistical analyses was similar to that in Huttenlauch et al. (2021), additionally comprising a group comparison between young and older speakers. For each dependent variable (F0 range, final lengthening, pause) on Name1 and Name2, we ran separate linear mixed-effect regression models in R (R Development Core Team 2018)[29]. Each model estimated the difference in the dependent variables between the two age groups (young and older speakers), between the four context comparisons, and between the two conditions (brack and nobrack condition), if applicable. Interactions between context and age group were added to further explore the dependencies of the differences, as well as interactions of context and age group with condition. A maximal model including all main effects and their interactions, as previously described, as well as including a random effects structure with all possible variance components and correlation parameters associated with the four within-subject contrasts (CHILD VS. YOUNG, ELDERLY VS. YOUNG, NON-NATIVE VS. YOUNG, NOISE VS. YOUNG) was always fit first[30]. In order to avoid overfitting of the random effects structure, we followed the approach outlined in Bates et al. (2015a) and conducted an iterative reduction of model complexity. A more detailed explanation of the model reduction, along with all reduced models and the complete model outputs of the fixed effects, can be found on an Open Science Framework project page (https://osf.io/fc8nz) together with the data and code. In the results section, we will only report the statistically significant effects which comprise main effects of condition and/or main effects and interactions of age group.

6.3 Results

In the following, we will first present descriptive results from absolute and relative measurements with a focus on age, including a statistical comparison of the age groups. Hereafter, we will turn towards the results of linear mixed models fit to compare the age groups regarding their use of prosodic cues for disambiguation (RQ1) and regarding their adaptation

[29] cited in original as R Core Team. 2018. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

[30] Prosodic_cue ~ 1 + condition*context*age_group +
(1 + condition +
child_vs_young + elderly_vs_young + nonnat_vs_young + noise_vs_young +
age_group +
condition:age_group +
condition:child_vs_young + condition:elderly_vs_young +
condition:nonnative_vs_young + condition:noise_vs_young +
child_vs_young:age_group + elderly_vs_young:age_group +
nonnative_vs_young:age_group + noise_vs_young:age_group +
condition:child_vs_young:age_group + condition:elderly_vs_young:age_group +
condition:nonnative_vs_young:age_group +
condition:noise_vs_young:age_group | speaker)

Table 13: Descriptive statistics of absolute durational measurements by age group and statistical group comparison.

	Young		Older		Comparison
Measurement (ms)	mean	*SD*	mean	*SD*	*p*
utterance duration	1964.63	292.16	2181.25	444.80	< 0.0001
final vowel duration (Name1)	129.61	40.09	144.68	46.20	< 0.0001
pause duration (after Name2)	172.93	195.24	262.83	330.05	< 0.0001
final vowel duration (Name2)	181.53	59.51	198.24	65.57	< 0.0001

to different interlocutors (RQ2).

Descriptive statistics and statistical age group comparison of absolute durational measurements

In the main section of our analysis, we analysed the use of prosodic cues by measuring the relative duration of speech segments and pauses. This method allowed us to understand how prosodic cues were used, regardless of individual differences in speaking rate or the absolute duration of sounds. Before presenting the relative measurements, we will present some absolute durational measurements to compare the differences between young and older speakers (cf. Table 13). However, we will not include measurements of average F0 by age group because the speaker groups had mixed genders, which could affect our estimation of differences in F0 between the groups.

In our data set we observe longer absolute durations for older as compared to younger speakers for the whole utterance (mean difference of 217 ms), the final vowels of Name1 and Name2 (mean difference of 15 ms and 17 ms, respectively), and the pause after Name2 (mean difference of 89.9 ms). All age group comparisons were statistically significant in linear models with age group as a single sum-contrasted predictor (0.5 for young and -0.5 for older speakers). Moreover, we observe a higher degree of variation (larger *SD*s) for older speakers than for young speakers across all durational measurements.

Descriptive statistics of relative measurements

Relative measurements of F0 range, final lengthening, and pause were used to explore the use of prosodic cues for the disambiguation of coordinates with and without internal grouping. Figure 18 shows a visual description of mean location and spread of F0 range as well as

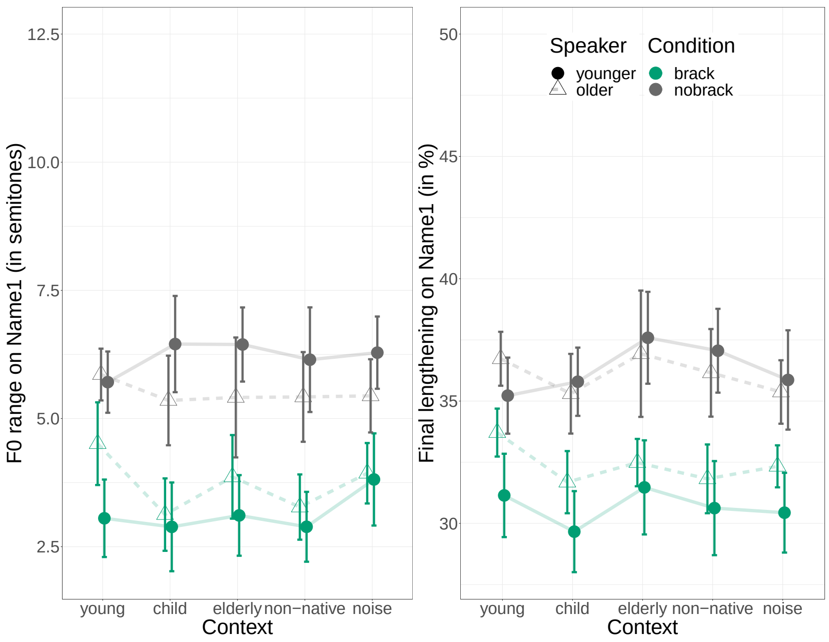


Figure 18: Distribution of raw values of F0 range (left panel) and final lengthening (right panel) on Name1 (y-axis) divided by context (x-axis), condition (colour: grey for nobrack, green for brack), and age group (shape: circles for young speakers, triangles for older speakers). Whiskers show 95% confidence intervals.

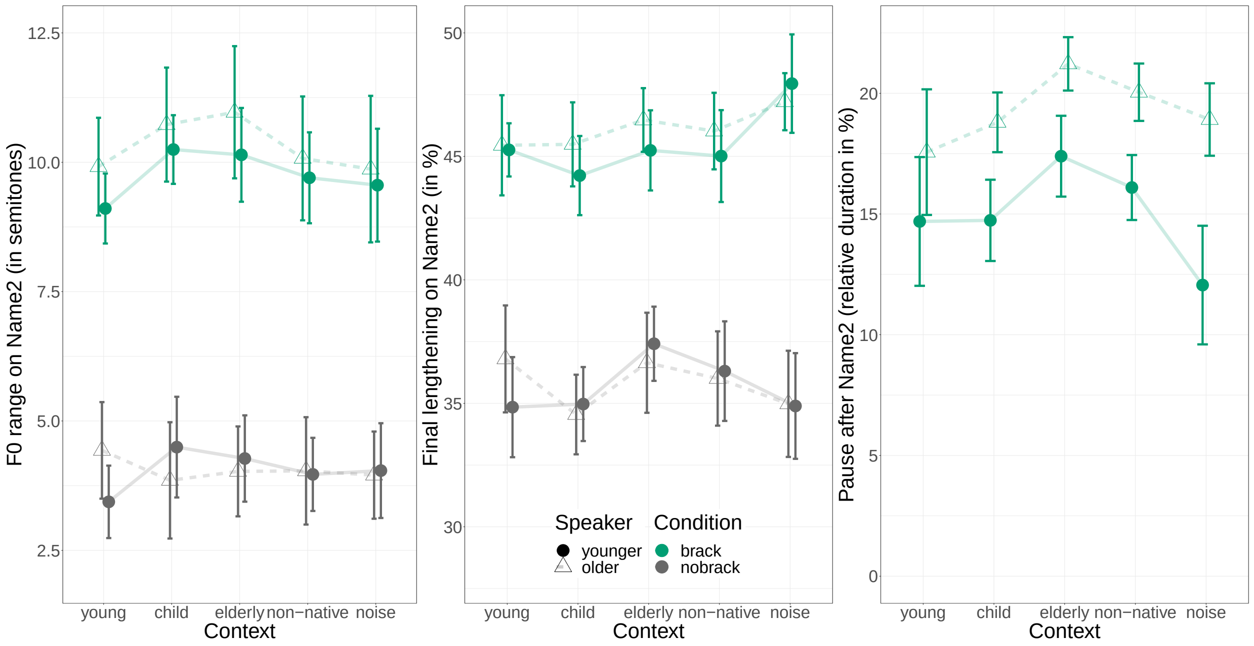


Figure 19: Distribution of raw values of F0 range (left panel), final lengthening (mid panel), and pause (right panel) on Name2 (y-axis) divided by context (x-axis), condition (colour: grey for nobrack, green for brack), and age group (shape: circles for young speakers, triangles for older speakers). Whiskers show 95% confidence intervals.

final lengthening on Name1 by age group, context, and condition. For both cues and for each context, the mean values in the brack condition are lower for younger than for older speakers, while in the nobrack condition in all contexts except YOUNG, the mean values are larger for younger compared to older speakers. Considering these raw data visually, the difference between conditions is larger in the productions of young speakers than in that of older speakers. We did not run statistical analyses and do not report descriptive statistics on pause duration after Name1 since mostly zero values were produced by the participants. That is, a pause after Name1 was only produced in 206 out of 2355 trials in total, 175 times in the nobrack condition and 31 times in the brack condition. Figure 19 shows a visual description of mean location and spread of F0 range, final lengthening, and pause on/after Name2 by age group, context, and condition. There is no apparent visual pattern that would apply to both speaker groups and all three cues. For F0 range and pause in the brack condition, young speakers produced smaller mean values than older speakers. For final lengthening in general and F0 range of the nobrack condition, the values are more mixed between age groups. With regard to the direction of the difference in the degree of F0 range and final lengthening between the brack and nobrack condition, both prosodic cues show smaller values in brack than in nobrack on Name1 and the opposite pattern, larger values in brack than in nobrack, on Name2.

To summarise, a visual inspection of the raw data reveals differences between the two age groups in the amount to which the different prosodic cues were produced in the respective contexts and conditions. Nevertheless, the general patterns for each cue are quite similar across contexts for both, young and older speakers. That is, for instance for F0 range in the brack condition in Figure 19 (left panel, green data points), the connecting lines between contexts have slopes in the same directions between speaker groups and in any case do not cross. We are aware that the descriptive analysis of the data does not allow for any generalisations. In the following sections, we will present the results of the statistical models we ran on each cue and Name individually.

Statistical analyses on Name1

F0 range on Name1

Results for F0 range on Name1 are reported from a reduced model[31] (all final models and code can be found on https://osf.io/fc8nz). Several effects were statistically significant (see

[31] F0_name1 ~ 1 + condition*context*age_group +
(1 + child_vs_young + elderly_vs_young + noise_vs_young +
age_group +
condition:age_group +
nonnative_vs_young:age_group +
condition:child_vs_young:age_group +
condition:nonnative_vs_young:age_group | speaker)

Table 14 and https://osf.io/fc8nz).

Table 14: Selected model estimates and 95% confidence intervals of the fixed effects for F0 range on Name1 including main effect of condition and main effect and interactions of age group.

Predictor	Estimate	95% CI
Intercept	4.666**	(4.060, 5.273)
condition	−1.236**	(−1.559, −0.913)
age group	0.002	(−1.211, 1.216)
condition:age group	−0.593	(−1.239, 0.053)
CHILD VS. YOUNG:age group	1.225**	(0.563, 1.886)
ELDERLY VS. YOUNG:age group	0.757	(−0.259, 1.773)
NON-NATIVE VS. YOUNG:age group	0.928*	(0.217, 1.639)
NOISE VS. YOUNG:age group	1.193*	(0.230, 2.155)
condition:CHILD VS. YOUNG:age group	0.051	(−0.515, 0.616)
condition:ELDERLY VS. YOUNG:age group	−0.013	(−0.450, 0.423)
condition:NON-NATIVE VS. YOUNG:age group	0.130	(−0.404, 0.664)
condition:NOISE VS. YOUNG:age group	0.271	(−0.170, 0.712)

$^{*}p < 0.05$; $^{**}p < 0.01$

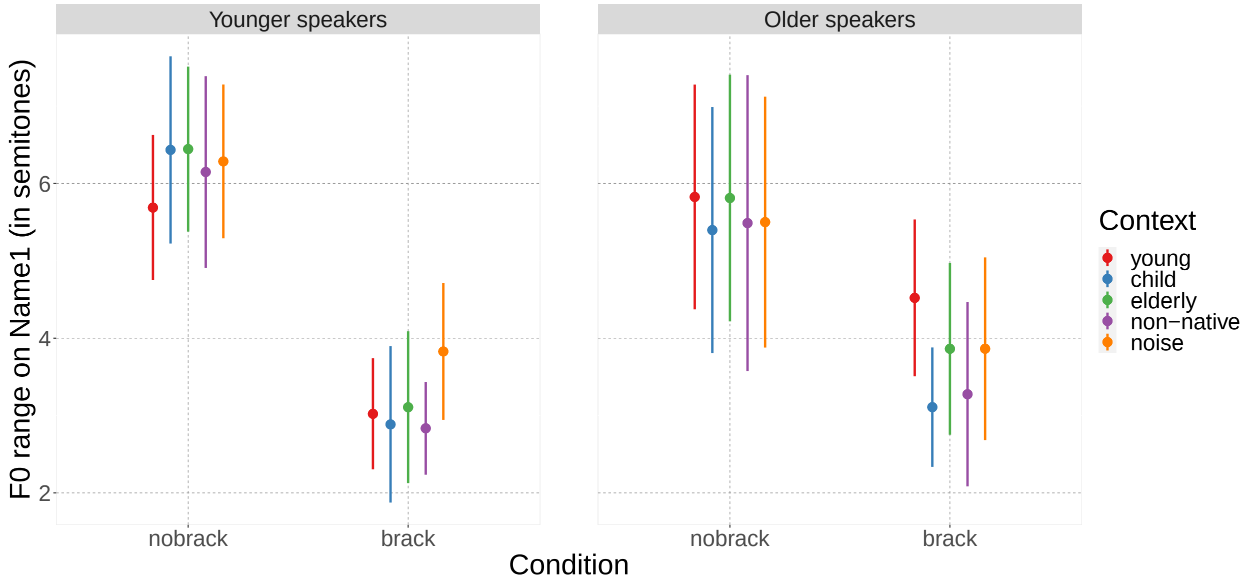


Figure 20: Model predictions for F0 range on Name1 (y-axis) divided by age group (younger speakers left panel, older speakers right panel), condition (x-axis), and context (colour). Whiskers show 95% confidence intervals.

The statistically significant main effect of condition ($\beta = -1.236$, $p < 0.0001$) confirms that F0 range was used for the disambiguation of brack and nobrack on Name1 by speakers of both age groups: The F0 range in the brack condition was decreased by about 2.5 semitones compared to the nobrack condition. With respect to age-related differences in situational (in)dependence, the statistically significant two-way interactions of the context comparisons CHILD vs. YOUNG ($\beta = 1.225$, $p = 0.0003$), NON-NATIVE vs. YOUNG ($\beta = 0.928$, $p = 0.011$), and NOISE vs. YOUNG ($\beta = 1.193$, $p = 0.016$) with age group, respectively, indicate general age-related differences when addressing the child and non-native as compared to the young interlocutor, as well as age-related differences in noisy vs. non-noisy settings with a young interlocutor. In all of the three context comparisons, young speakers increased their F0 range compared to context YOUNG, while older speakers decreased their F0 range. Model predictions for F0 range on Name1 by condition, context, and age group are displayed in Figure 20.

Final lengthening on Name1

Table 15: Selected model estimates and 95% confidence intervals of the fixed effects for final lengthening on Name1 including main effect of condition and main effect and interactions of age group.

Predictor	Estimate	95% CI
Intercept	33.848**	(32.716, 34.980)
condition	−2.366**	(−2.949, −1.784)
age group	−0.794	(−3.058, 1.469)
condition:age group	−1.001	(−2.166, 0.164)
CHILD vs. YOUNG:age group	1.449*	(0.063, 2.834)
ELDERLY vs. YOUNG:age group	1.962	(−0.255, 4.179)
NON-NATIVE vs. YOUNG:age group	1.877*	(0.203, 3.551)
NOISE vs. YOUNG:age group	1.371	(−0.289, 3.032)
condition:CHILD vs. YOUNG:age group	−0.841	(−2.226, 0.545)
condition:ELDERLY vs. YOUNG:age group	−0.361	(−1.740, 1.017)
condition:NON-NATIVE vs. YOUNG:age group	−0.612	(−1.995, 0.771)
condition:NOISE vs. YOUNG:age group	−0.726	(−2.124, 0.672)

$^{*}p < 0.05$; $^{**}p < 0.01$

Results for final lengthening on Name1 are reported from a reduced model[32]. Several effects were statistically significant (see Table 15 and https://osf.io/fc8nz). The statistically

[32] the model can be found on https://osf.io/fc8nz

significant main effect of condition ($\beta = -2.366$, $p < 0.0001$) confirms that final lengthening was used for the disambiguation of brack and nobrack on Name1 by speakers of both age groups: Final lengthening was decreased in the brack condition (where the final vowel span about 31% of the total name duration) as compared to the nobrack condition (where the final vowel span about 36% of the total name duration). With respect to age-related differences in situational (in)dependence, the statistically significant two-way interaction of the context comparison CHILD vs. YOUNG with age group ($\beta = 1.449$, $p = 0.002$) indicates that young speakers, in contrast to older speakers, increased final lengthening when addressing the child compared to the young interlocutor. A similar pattern is predicted by the model for the context comparison NON-NATIVE vs. YOUNG, for which the interaction with age group was statistically significant ($\beta = 1.877$, $p = 0.028$): While final lengthening is increased by young speakers when addressing the non-native as compared to the young interlocutor, final lengthening is decreased by older speakers. Model predictions for final lengthening on Name1 by condition, context, and age group are displayed in Figure 21.

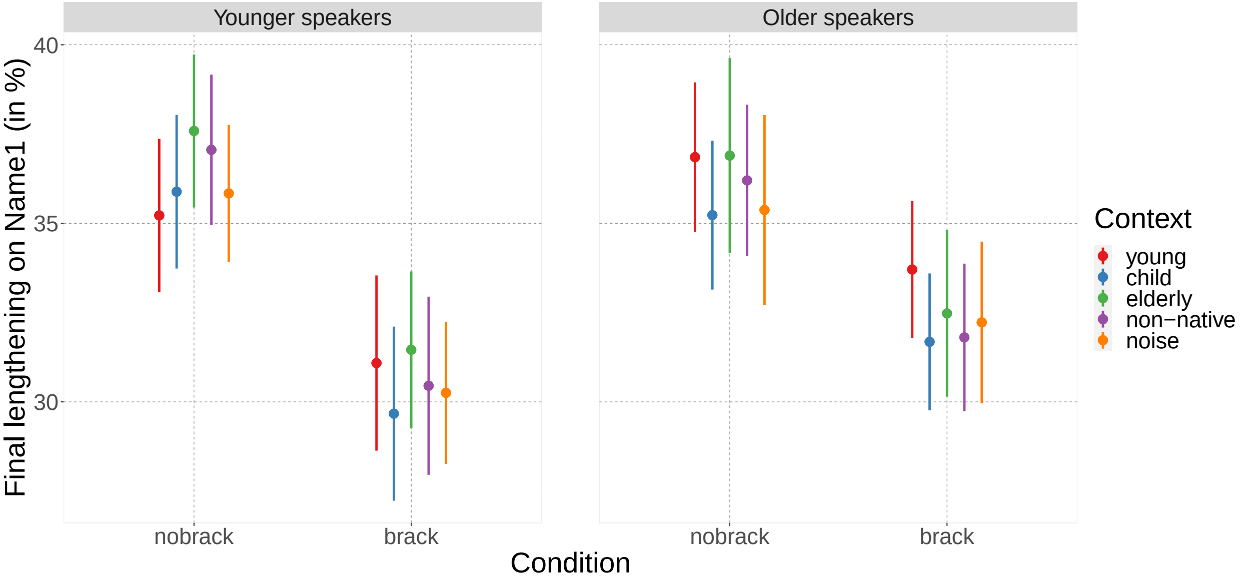


Figure 21: Model predictions for final lengthening on Name1 (y-axis) divided by age group (younger speakers left panel, older speakers right panel), condition (x-axis), and context (colour). Whiskers show 95% confidence intervals.

Statistical analyses on Name2

F0 range on Name2

Results for F0 range on Name2 are reported from a reduced model[33]. Several effects were statistically significant (see Table 16 and https://osf.io/fc8nz). The statistically significant

[33] the model can be found on https://osf.io/fc8nz

Table 16: Selected model estimates and 95% confidence intervals of the fixed effects for F0 range on Name2 including main effect of condition and main effect and interactions of age group.

Predictor	Estimate	95% CI
Intercept	7.097**	(6.370, 7.824)
condition	3.040**	(2.613, 3.468)
condition:age group	−0.345	(−1.200, 0.510)
CHILD vs. YOUNG:age group	0.873*	(0.121, 1.626)
ELDERLY vs. YOUNG:age group	0.573	(−0.413, 1.559)
NON-NATIVE vs. YOUNG:age group	0.636	(−0.384, 1.655)
NOISE vs. YOUNG:age group	0.642	(−0.351, 1.635)
condition:CHILD vs. YOUNG:age group	−0.779*	(−1.419, −0.139)
condition:ELDERLY vs. YOUNG:age group	−0.684	(−1.524, 0.156)
condition:NON-NATIVE vs. YOUNG:age group	−0.306	(−0.968, 0.356)
condition:NOISE vs. YOUNG:age group	−0.442	(−1.193, 0.309)

*$p < 0.05$; **$p < 0.01$

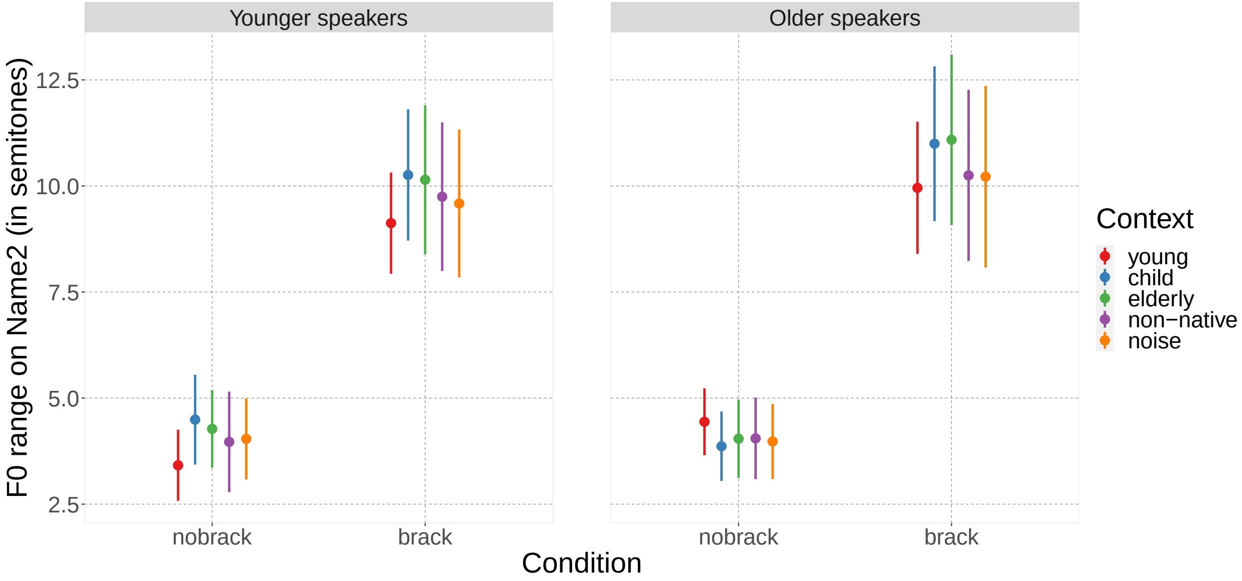


Figure 22: Model predictions for F0 range on Name2 (y-axis) divided by age group (younger speakers left panel, older speakers right panel), condition (x-axis), and context (colour). Whiskers show 95% confidence intervals.

main effect of condition ($\beta = 3.04$, $p < 0.0001$) confirms that F0 range was used for the disambiguation of brack and nobrack on Name2 across both age groups: The F0 range in the brack condition was increased by about six semitones compared to the nobrack condition. With respect to age-related differences in situational (in)dependence, the significant two-way interaction of the context comparison CHILD vs. YOUNG with age group ($\beta = 0.873$, $p = 0.011$) indicates general age-related differences in approaching the child interlocutor compared to the young interlocutor: The F0 range was larger for young speakers than that of older speakers when addressing the child in comparison to the young interlocutor. These age-related patterns diverge even more when context-related prosodic disambiguation is considered and condition is taken into account. The significant three-way interaction of condition, context comparison CHILD vs. context YOUNG, and age group ($\beta = -0.799$, $p = 0.018$) indicates that young speakers increased the F0 range in both conditions, brack and nobrack, when addressing the child as compared to the young interlocutor, while older speakers did so only in the brack condition. In the nobrack condition, however, older speakers decreased the F0 range, resulting in an enhanced difference between the conditions when addressing the child as compared to the young interlocutor. Model predictions for F0 range on Name2 by condition, context, and age group are displayed in Figure 22.

Final lengthening on Name2

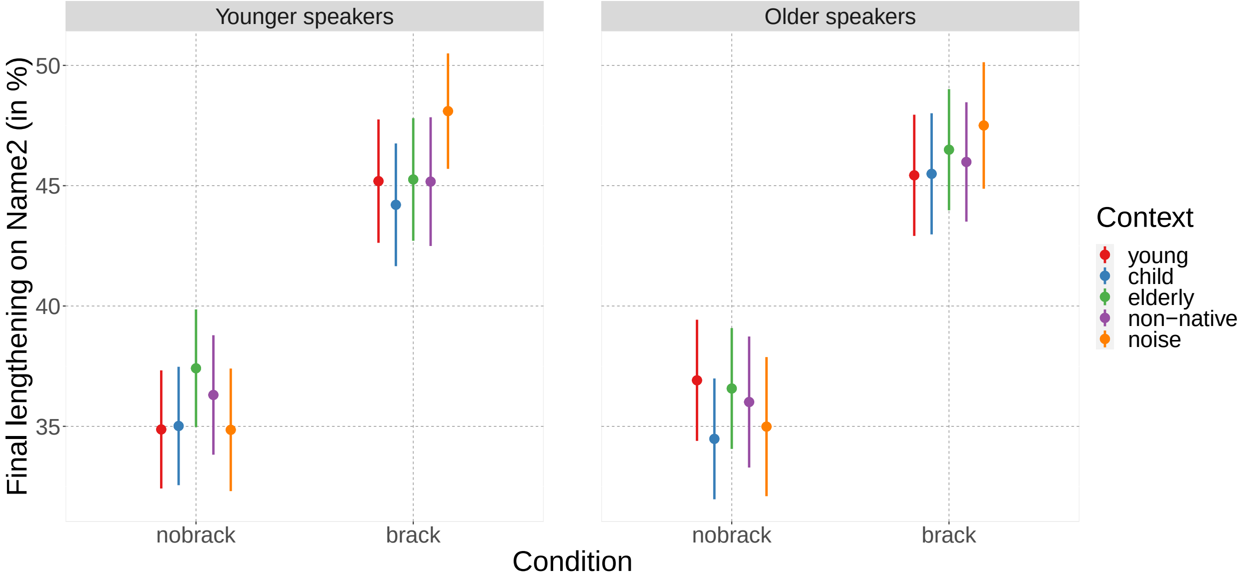


Figure 23: Model predictions for final lengthening on Name2 (y-axis) divided by age group (younger speakers left panel, older speakers right panel), condition (x-axis), and context (colour). Whiskers show 95% confidence intervals.

Table 17: Selected model estimates and 95% confidence intervals of the fixed effects for final lengthening on Name2 including main effect of condition and main effect and interactions of age group.

Predictor	Estimate	95% CI
Intercept	40.813*	(39.477, 42.149)
condition	5.071*	(4.284, 5.858)
condition:age group	−0.248	(−1.822, 1.325)
CHILD vs. YOUNG:age group	0.766	(−0.584, 2.116)
ELDERLY vs. YOUNG:age group	0.943	(−0.399, 2.286)
NON-NATIVE vs. YOUNG:age group	0.880	(−0.467, 2.227)
NOISE vs. YOUNG:age group	1.374	(−0.368, 3.115)
condition:CHILD vs. YOUNG:age group	−1.811**	(−3.161, −0.462)
condition:ELDERLY vs. YOUNG:age group	−1.939**	(−3.281, −0.596)
condition:NON-NATIVE vs. YOUNG:age group	−1.455	(−3.035, 0.125)
condition:NOISE vs. YOUNG:age group	−0.536	(−1.898, 0.827)

$^{*}p < 0.05$; $^{**}p < 0.01$

Results for final lengthening on Name2 are reported from a reduced model[34]. Several effects were statistically significant (see Table 17 and https://osf.io/fc8nz). The statistically significant main effect of condition ($\beta = 5.071$, $p < 0.0001$) confirms that final lengthening was used for the disambiguation of brack and nobrack on Name2 by speakers of both age groups: Final lengthening was increased in the brack condition (the final vowel of Name2 span about 45% of the total duration of Name2) compared to the nobrack condition (the final vowel span about 35% of the total name duration). Regarding age-related differences in prosodic disambiguation and situational (in)dependence, the three-way interaction between condition, the context comparison CHILD vs. YOUNG, and age group ($\beta = -1.811$, $p = 0.009$) indicates that young speakers decreased final lengthening in the brack condition when addressing the child as compared to addressing the young interlocutor, thus decreasing the difference between the conditions. On the contrary, older speakers decreased final lengthening for the same context comparison in the nobrack condition, thus increasing the difference between the conditions. An additional three-way interaction between condition, ELDERLY vs. YOUNG and age group ($\beta = -1.939$, $p = 0.005$) indicates that young speakers increased final lengthening in the nobrack condition when addressing the elderly as compared to the young interlocutor. That is, they reduced the difference between the conditions in context ELDERLY, compared to context YOUNG. Older speakers showed a different behaviour: They increased final lengthening when addressing the elderly as compared to the

[34] the model can be found on https://osf.io/fc8nz

young interlocutor in the brack condition, thus enhancing the difference between the conditions in context ELDERLY. Model predictions for final lengthening on Name2 by condition, context, and age group are displayed in Figure 23.

Pause after Name2

Since the random effects structure of the model analysing pause after Name2 could not be reduced without a significant drop in model fit, results are reported from the maximal model. None of the effects were statistically significant (see Table 18 and https://osf.io/fc8nz).

Table 18: Selected model estimates and 95% confidence intervals of the fixed effects for pause after Name2 including main effects and interactions of age group.

Predictor	Estimate	95% CI
Intercept	19.398**	(14.827, 23.968)
age group	−0.019	(−9.160, 9.122)
CHILD VS. YOUNG:age group	−1.560	(−10.505, 7.386)
ELDERLY VS. YOUNG:age group	−1.202	(−10.113, 7.709)
NON-NATIVE VS. YOUNG:age group	−1.048	(−10.001, 7.906)
NOISE VS. YOUNG:age group	16.472	(−21.986, 54.930)

$^{*}p < 0.05$; $^{**}p < 0.01$

6.4 Discussion

In the current study, we compared the use of prosodic cues produced to disambiguate the internal grouping of coordinated three-name sequences (coordinates) in two conditions, that is, without and with internal grouping of the first two names (nobrack and brack, respectively) between two age groups: young (19–34 years) and older (61–80 years) speakers of German. We concentrated our analysis on the three prosodic cues F0 range, final lengthening, and pause on/after Name1 and Name2. As age affects the stability and variability of tonal and durational features in general, we tested for potential age effects on the modulation of the three prosodic cues for structural disambiguation. Furthermore, we explored whether the situational (in)dependence of disambiguating prosody differs between younger and older speakers, considering their prosodic adaptation to varying contexts. To this end, in both age groups, we elicited coordinates by means of a referential communication task with five contexts: addressing a young adult, a child, an elderly adult, a young non-native adult, and the young adult with background noise.

Looking at the data, we note two things: First of all, descriptively, younger and older speakers produced the three prosodic cues overall quite similarly for prosodic disambiguation and even in the different contexts. This visual observation receives support from the statistical models: For none of the prosodic cues, did the statistical models reveal a main effect for age group. That is, for the use of prosodic cues to mark the internal grouping of coordinates, our data do not provide evidence for a general age-related effect. Second, despite the similarity of the produced prosodic cues, the productions of the older group of speakers are more variable than those of the younger ones, an effect that is evident in larger standard deviations and confidence intervals of the model estimates and the raw data. Increased variability with increased age regarding F0 and durational values is in line with findings of previous studies (Scukanec et al. 1992, 1996, Lortie et al. 2015, Santos et al. 2021, among others).

Regarding our first research question, whether older compared to younger speakers show a more extreme use of F0 range, final lengthening, and pause on Name1 and Name2 to mark the internal grouping, our data do not provide evidence for age-related increases in cue use. In absolute measures, though, older speakers produced longer utterances and longer final vowels on Name1 and Name2 than young speakers, which corresponds to a slower speaking rate since all productions had the same number of syllables. A slower speaking rate is in line with previous findings in the literature (Kemper et al. 1995, Scukanec et al. 1996, Harnsberger et al. 2008, Barnes 2013, Smiljanic & Gilbert 2017, Dimitrova et al. 2018, Tuomainen & Hazan 2018, Hazan et al. 2019, Tuomainen et al. 2019, 2021). Nevertheless, independent of age, speakers in both age groups marked the internal grouping globally in line with the Proximity/Similarity model (Kentner & Féry 2013) using all three cues investigated: In the brack condition, on Name1, speakers of both age groups produced a smaller F0 range and less final lengthening compared to the nobrack condition. This is considered a weakening of the prosodic boundary indicating the sisterhood of the neighbouring element (i. e., Name2 in this case) by Kentner & Féry (2013). On Name2, this pattern was reversed: In the brack condition, speakers of both age groups increased the F0 range and the lengthening of the final segment compared to the nobrack condition and, additionally, inserted a pause after Name2 in the brack condition. This increase of prosodic cues is considered a strengthening of a prosodic boundary (Kentner & Féry 2013). For none of the prosodic cues was the interaction between age group and condition statistically significant. We, thus, did not find support for age-related more extreme use of disambiguating prosodic cues. Across both age groups, the results of the perception checks confirmed that the internal grouping was produced successfully, as the conditions could reliably be recovered by naïve listeners. Only about 4% of the data in each age group led to misunderstandings. That is, despite the variability in the data, speakers of both age groups produced the disambiguating prosodic cues in such a clear way that listeners could correctly resolve the underlying syntactic structure.

Regarding our second research question, whether young and older speakers differ in adapting their use of prosodic cues when addressing varying interlocutors, our data show substantial similarities across age groups. For several model predictions, the estimated means of the non-baseline contexts within one condition deviate in the same direction from the young baseline context in both age groups (cf. brack in Figures 21 and 22). This also explains why only few interactions of context, condition, and age group revealed statistical significance. Nevertheless, there are slight differences between the age groups regarding their adaptations. We will focus our discussion on statistically significant three-way-interactions of age groups, contexts, and condition, as we are mainly interested in the interplay of all three factors.

The two age groups diverged most strongly when addressing a child as compared to a young interlocutor: On Name2, the older speakers produced larger F0 ranges for the child compared to the young interlocutor in condition brack and smaller F0 ranges along with decreased final lengthening in condition nobrack, thus increasing the difference between conditions when addressing the child. Younger speakers, however, rather slightly decreased the difference between brack and nobrack when addressing the child as they reduced final lengthening in the brack condition. This enhanced difference between conditions in older speakers can be interpreted as more adaptation to the child interlocutor in older than in younger speakers. Such an enhanced difference between conditions in older compared to younger speakers also holds true for the context with the elderly interlocutor. Here, the older speakers slightly increased the difference between the conditions by means of an increase in final lengthening on Name2 in the brack condition while the young speakers showed the reverse pattern: They decreased the difference in final lengthening between the conditions by increasing final lengthening in nobrack. Interestingly, from the viewpoint of disambiguation, speakers in both age groups produced a stronger distinction between conditions when addressing their peer compared to addressing a non-age-matched interlocutor: young speakers addressing the young interlocutor and older speakers addressing the elderly interlocutor. We are not aware of any similar findings in the literature. Yet, despite being statistically significant, these differences in adaptation between age groups were in fact quite small in absolute terms, and did not affect the disambiguation of coordinates, as revealed by the perception check (see previously). Together with the large variability in the productions of the older speakers (cf. larger 95% confidence intervals in the Figures with model predictions than for younger speakers), it is questionable whether the effects in the child and elderly contexts compared to the young context are reproducible in the same manner in future studies.

In the remaining two contexts (non-native interlocutor and speech in noise), our data did not demonstrate evidence for differences between the age groups. Given this and given the fact that any context differences across groups did not impact on disambiguation of coordinates in general, regarding our second research question, our data speak in favour of situational independence in both age groups (Schafer et al. 2000, Kraljic & Brennan 2005, Speer et al. 2011). Models of situational independence assume that disambiguating prosody is realised automatically as part of the production process on the side of the speaker and is therefore largely independent of the presence or absence of a listener, the type of listener, or the situational setting. As such, it seems plausible that disambiguating prosody is also independent of the age of the speaker. Our data add to the literature on the effects of different types of interlocutors and the absence/presence of noise on the use of disambiguating prosodic cues the dimension of speaker age. The findings show that situational independence in production of disambiguating prosody holds for older speakers, too, and that prosody production is a stable automatic part of the production process also in older speakers.

Thus, whereas age has frequently been shown to affect other areas of language production (i. e., word-finding abilities, increased phonetic variability, or altered acoustic characteristics), it does not seem to have a (listener-relevant) impact on production of prosodic cues in ambiguous structures. This is in line with an observation by Lortie et al. (2015) regarding a more variable voice in older speakers that did not interact with the ability to control fundamental frequency (participants in their study were asked to produce normal, low, and high frequency voice in sustained vowels). In this sense, our study provides evidence that one important part of the prosody-syntax interface is not affected by age effects: the use of the prosodic cues F0 range, final lengthening, and pause for disambiguation of structurally ambiguous coordinates. Our findings on prosody production in older adults are also of importance in the larger context of investigating linguistic prosody in populations with acquired language and communication disorders resulting from brain lesions (i. e., aphasia or right-hemisphere brain lesions), since participants in these studies are usually older than the typical age groups covered in most studies on healthy prosody processing.

In summary, our data confirm the well-known general age-related changes in absolute durational measures. However, when it comes to the use of tonal and durational prosodic cues to disambiguate the underlying syntactic structure, older speakers modulated duration and F0 range similarly to younger speakers with, if at all, only minimal differences between the the age groups of speakers in our sample. The finding of limited adaptation to different interlocutors favours models of situational independence of disambiguating prosody across both age groups and shows that production of disambiguating prosody at the prosody-syntax interface is unaffected by age.

6.5 Conclusion

In conclusion, young and older speakers in our production study globally marked the internal grouping of coordinated name sequences using F0 range, final lengthening, and pause in a similar way. The modulation of disambiguating prosodic cues seems to be independent of age-related changes in absolute durations. Across both age groups, the use of prosodic cues to resolve the ambiguity in the internal structure of coordinates dominated in comparison to possible prosodic accommodations to the contexts, which we interpret as evidence for situational independence of disambiguating prosody. Prosodic disambiguation thus turns out to be a stable automatic part of the production process, regardless of speaker age.

Acknowledgements

We would like to thank L. Junack for her help with acquisition and pre-processing of the data, as well as for setting up the perception check for the data of the older speakers.

Funding information

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB1287 Project B01 (Project No. 317633480).

7 Study III

Individual differences in early disambiguation of prosodic grouping[35]

Abstract

Prosodic cues help to disambiguate incoming information in spoken language perception. In structurally ambiguous coordinate utterances, such as three-name sequences, the intended grouping is marked by three prosodic cues: F0-range, final lengthening, and pause. To indicate that the first two names are grouped together, speakers typically weaken the durational and tonal cues on the first name, while they are strengthened on the second name, compared to a structure without internal grouping. The current study uses a Gating Paradigm to test whether listeners can decide about the internal grouping of a coordinate structure by already exploiting prosodic information on the first name. 192 stimuli were cut into seven parts (gates) and presented to naïve participants (n = 45) successively (gate by gate) with increasing length of the utterance and amount of prosodic information. In a two-alternative forced choice decision task, accuracy was above chance level after the second name. However, more than half of the participants could already reliably detect grouping patterns after the first name. These inter-individual differences point towards the existence of different subgroups with diverging prosodic parsing strategies. Furthermore, listeners were sensitive to speaker-specific prosodic patterns. Depending on speaker-specific characteristics and individual parsing capacities, it seems possible – at least for a subgroup of listeners – to make predictions about the underlying grouping structure of coordinated name sequences based on early prosodic cues.

7.1 Introduction

Prosody, the modulation of pitch and rhythm, is a pivotal source of information in spoken language comprehension. As prosody accompanies a spoken utterance, it guides the listener along the syntactic structure (e. g., Steinhauer et al. 1999) and the speaker along their mental structure (e. g., Kraljic & Brennan 2005, Speer et al. 2011, Wagner 2005, Watson & Gibson 2004). In more detail, speakers produce prosody for a variety of purposes, including syntactic, lexical, and pragmatic objectives, and thus convey content that is critical for understanding. On the side of the listener, these underlying purposes and meaningful cues must be exploited

[35] An adapted version of this chapter has been published as Hansen, Marie, Clara Huttenlauch, Carola de Beer, Isabell Wartenburger & Sandra Hanne. 2022. Individual differences in early disambiguation of prosodic grouping. *Language and Speech* 0(0). 1–28. doi:10.1177/00238309221127374. This study is written in American English.

in prosodic parsing to identify constituents, structure, and meaning of the utterance (e. g., Clifton et al. 2002).

Prosodic cues demarcate junctures in an utterance, such as the beginning and the end of a discourse segment (Swerts & Geluykens 1993), newly introduced concepts in the discourse (Féry & Kügler 2008), and internal chunks that group semantically and pragmatically related constituents, by means of prosodic boundaries Kentner & Féry (2013).

In this study, we focus on the latter, specifically on boundary-related prosodic markers in German three-name sequences coordinated by *und* ('and') with and without internal grouping of the first two names (see examples (32) and (33) below). These structures are particularly suitable for the investigation of prosodic boundaries: Firstly, linguistic characteristics of coordinated names can easily be controlled when designing the experimental stimuli. Secondly, coordinated name sequences are short and simple – and it has been demonstrated that prosodic cues appear to have larger implications for perception among listeners in shorter constituents than in longer ones (Clifton et al. 2006). In the following examples, the answer to the question *Who is arriving at the station?* may include an internal grouping of two of the persons or may be uttered without such an internal grouping. Example (33) contains no internal grouping and the three persons are possibly all arriving together. In contrast, example (32) contains an internal grouping of the first two constituents (indexed by the brackets), indicating that Moni and Lilli are arriving together while Manu is arriving separately.

Who is arriving at the station?

(32) [Moni und Lilli] und Manu
Internal grouping (bracket)

(33) Moni und Lilli und Manu
No internal grouping (no bracket)

Regarding the production of such prosodic boundaries, Kentner & Féry (2013) developed a model of syntax-prosody mapping, the *Proximity/Similarity Model*, that accounts for the relative strengths of these prosodic boundaries. *Proximity* predicts that a boundary between two names is weakened when they are grouped together in comparison to a structure without internal grouping. That is, the boundary between the first and the second name, *Moni* and *Lilli*, is predicted to be weaker in (32) than in (33). According to *anti-proximity*, a prosodic boundary is strengthened between two names that are of different syntactic levels in comparison to an ungrouped structure. Consequently, the boundary after the second name, *Lilli*, is predicted to be strengthened in (32) compared to (33). This stronger prosodic boundary on the second name indicating a grouping such as in (32) is realized in different languages, including German (Gollrad 2013, Huttenlauch et al. 2021, Kentner & Féry 2013, Peters et al. 2005, Petrone et al. 2017), by a longer duration of the final syllable of the second name (henceforth referred to as final lengthening), a higher rise of the fundamental frequency (F0) on the second name, and a pause right after the second name in comparison to an ungrouped structure. Thus, for coordinated name sequences as in (32), the boundary bearing the most salient cues to syntactic grouping is located at the end of the second name. In this paper, we will call the prosodic cues around the second name *late prosodic cues* to boundaries, as opposed to *early prosodic cues*, which are located before the second name.

In the perception of coordinate structures, pitch, pause, and final lengthening are not equally relevant for ambiguity resolution. In a perception study manipulating late prosodic cues, Gollrad (2013) demonstrated that pitch alone is not sufficient for boundary perception while jointly presented durational cues (pause, final lengthening) may facilitate the parsing process without pitch being present (for similar results from ERP data, see Holzgrefe-Lang et al. 2016). According to Petrone et al. (2017), the pause cue triggers a more categorical shift in prosodic judgments than pitch and final lengthening. These studies focused on late prosodic cues in coordinated name structures on or after the second name such as in (32) compared to (33). However, boundaries are scaled relative to one another and boundary strength is determined not just locally but across the whole utterance (e. g., Clifton et al. 2006, Wagner 2010).

As described above, the *Proximity/Similarity Model* predicts a weakening of the boundary on the first name in (32) vs. (33). These early cues, that is cues on Name1, might already give hints to the grouping of the structure. Corresponding prosodic patterns, in line with proximity/anti-proximity, were observed by Kentner & Féry (2013). Early prosodic cues have also been reported in the study by Huttenlauch et al. (2021), in which participants produced coordinate structures akin to examples (32) and (33) above: differences in cue usage were found not only on the second name (henceforth Name2) but also already on the first name (henceforth Name1). More precisely, pitch was lower and final lengthening was shorter for Name1 in the bracket condition (example (32) than in the no bracket condition (example (33)). Huttenlauch et al. (2021) concluded that speakers used early as well as late prosodic cues to distinguish between conditions.

Several studies have shown that listeners pay attention to more global prosodic features such as phrase length, speech rate, and speaker- as well as language-specific prosodic patterns (e. g., Jun 2003). Moreover, non-local / more distant F0- and durational cues have been shown to influence listeners' perception of segmentation (or the grouping of segments into words, e. g., *foot – notebook – worm / footnote – book – worm*; Brown et al. 2011; Dilley & McAuley 2008). In an ERP study on coordinate structures conducted by Li & Yang (2009), a *closure positive shift*, reflecting the perception of a prosodic boundary, was also elicited for earlier boundaries characterized by more subtle cues. This suggests that listeners are sensitive to early prosodic cues in coordinate utterances. It remains an open question whether such cues are already sufficient for disambiguation. Hence, the overarching research question of this study is: Can listeners exploit early, subtle prosodic cues such as the cues present on Name1 in coordinated name sequences to predict the internal grouping of the utterance?

With respect to the processing of different sentence types, that is questions vs. statements, it has already been shown that listeners can make use of early prosodic cues for disambiguation (Face & D'Imperio 2005 for Spanish; Petrone & Niebuhr 2014 for German; van Heuven & Haan 2002 for Dutch; Vion & Colas 2006 for French). Thus, we predict that listeners may be able to use early cues for disambiguation in other linguistic contexts as well. We will make use of non-manipulated productions of coordinate structures from the study by Huttenlauch et al. (2021) to investigate our research question. If early prosodic cues are sufficient to predict the underlying syntactic structure, future studies should pay considerably more attention to these cues in (perception) experiments on disambiguating prosody, as opposed to investigating the influence of the local, late boundary cues only. We employ a Gating Paradigm (see section 7.2) to investigate how listeners make use of gradually increasing amounts of prosodic information.

Besides investigating the use of early cues for disambiguation in perception, we focus on inter-individual differences in the production of prosodic boundaries in coordinate structures and their influence on perception. Several studies have observed variability in cue combinations and attunement of cues on the second constituent, with pause being the most stable cue that was produced (Peters et al. 2005, Gollrad 2013, Kentner & Féry 2013, Petrone et al. 2017, Huttenlauch et al. 2021). The general variability in the usage of cues that was found in these studies points towards a considerable inter-individual range of degrees of freedom in cue realization.

Similar to production, variability can also be found in perception: Cangemi et al. (2015) investigated production and perception of question-answer pairs with fictional name targets in different linguistic focus structures: broad, narrow, and contrastive focus. Answer sentences were structurally identical, with the respective focus structure being signaled (or disambiguated) by means of prosody. Listeners had to match the target sentences (recorded in a preceding production part with different participants) to one of the different focus conditions. They varied in their decoding of prosodic contrasts across speakers and, additionally, in their decoding of prosodic contrasts produced by particular speakers. The authors suggest an individual-specific network of phonological knowledge that leads to speaker- and listener-specific differences in the identification of prosodic contrasts. When it comes to online processing of prosody, Kim 2019[36] observed differences in prosodic cue usage for the perception of disambiguating boundaries over the course of listeners' fixations to the target picture in a Visual World Paradigm: some listeners looked to the correct target earlier than others, depending on the available prosodic cues, which varied between conditions.

Individual differences in prosody perception were also found in studies on boundary and prominence annotations conducted by means of *rapid prosody transcription* (RPT): Cole

[36] Correction: the reference should be Kim (2020).

et al. (2017) found a few “super annotators” in each of their participant groups, despite considerable differences in group characteristics (spoken dialect, lab-based vs. crowd-sourced settings). The concept of “super annotators” refers to a high agreement rate in prosodic marking within *Tones and Break Indices* (TOBI) annotations (Silverman et al. 1992) between a naïve annotator and trained TOBI annotators. However, Bishop et al. (2020) point out that the existence of “super annotators” might be due to chance, as they only found “one or two” among 158 naïve participants (Bishop et al. 2020: 7). In another RPT study, Roy et al. (2017) report subsets of participants who made use of global cues such as intensity, while others used these cues to a minor extent. Overall, effects differed substantially across the annotators, with trained annotators performing better than untrained ones. Among the untrained annotators in this study, only a subset seemed to use the same prosodic cues as the trained annotators did. The authors suggest that these individual differences are driven either by differences in sensitivity to a cue or by differences in the sensitivity to contextual factors that predict the occurrence of a prosodic boundary/a prominent feature. These assumptions are supported by Baumann & Winter (2018), who interpret their findings on prosodic prominence as some listeners paying more attention to pitch-related features and less to semantic-syntactic and lexical features while others show the reverse pattern. These findings do not necessarily generalize to boundary perception but can be seen as another indication of different listener groups.

If major differences exist between listeners regarding prosody processing styles (e. g., Yu 2013), cochlear responses to tonal aspects in the speech signal (Ladd 2008), or communicative skills that are required for prosody perception (e. g., Jun & Bishop 2015), it is important to capture these individual differences. When averaging over the entire group of participants, an important part of the information about prosodic processing is lost. Therefore, in this study, we will explore individual differences in the use of early prosodic cues for syntactic disambiguation.

7.2 Aims and hypotheses of the current study

This study aims to investigate listeners’ ability to exploit early prosodic cues for the detection of the intended internal grouping in German coordinated name sequences. For this purpose, we use the Gating Paradigm, a specific experimental setup introduced by Grosjean (1980). The Gating Paradigm is a method in which a whole stimulus is cut into several parts (gates) and the participants are presented with gated snippets of successively increasing length.

Gating studies have been used successfully for research on spoken word recognition. Specifically, in the domain of prosody and listeners’ predictions based on prosodic parsing, the Gating Paradigm has been used to gain insights into listeners’ exploitation of pitch accents (Cutler & Otake 1999), the prediction of sentence length (Grosjean 1983, 1996, O’Brien et al. 2013) and sentence continuation (Hughes & Szczepek Reed 2011), and the intonation of questions (Petrone & Niebuhr 2014), as well as for assessing speech segmentation among native listeners in comparison to second language learners (Field 2008). It is noteworthy that Beach (1991) used a version of the Gating Paradigm (short vs. long sentence beginning conditions; stimulus example: *Jay believed...* vs. *Jay believed the gossip...*, Beach 1991: 4) to show that listeners made use of prosody to distinguish direct object and sentence complement syntactic structures before the complete sentence information was available. In Allopenna et al. (1998), a Gating Paradigm was used along with eye tracking to investigate continuous prosodic mapping. In the current study, we applied the Gating Paradigm to investigate listeners' prosodic parsing of the coordinate stimuli produced by four of the participants in the study of Huttenlauch et al. (2021). To this end, the coordinated three-name sequences were cut into seven gates (g1–g7), where the first gate comprised the first syllable of Name1. With each gate, the subsequent syllable was added in a cumulative manner. Thus, g1 included the first syllable of Name1, g2 comprised the first and second syllable of Name1 (that is, the complete first constituent), g3 comprised the first constituent and the conjunction, and so forth (a more detailed description is given in section 7.3 on page 123).

After each gate, participants had to decide if the structure belonged to the condition with or without grouping. The following research questions (RQ) were addressed:

RQ 1: At which gate can listeners reliably predict the structure of German coordinated name sequences with or without internal grouping of the first two names?

As in the grouping condition (bracket) the most alerting prosodic cues occur at or after the final syllable of Name2 (Huttenlauch et al. 2021), listeners were predicted to reliably detect the internal grouping at g5 (i. e., after Name2). If cues that are located at earlier points in the utterance can already serve as reliable markers for grouping patterns, listeners' decisions about internal grouping should already be above chance level in early gates (i. e., before Name2).

RQ 2: Are there individual differences among listeners with respect to prosodic parsing capacities?

As described above, previous evidence from perception experiments usually mirrors mechanisms that were found in production. Different individuals naturally exhibit some degree of variability in production (e. g., Huttenlauch et al. 2021) and variability has also been observed in perception (Cangemi et al. 2015). To our knowledge, listener variability has not been investigated in coordinate structures before. This second research question was rather exploratory and thus not tied to specific predictions.

7.3 Methods and procedures

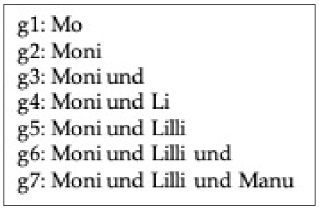
Participants

A total number of 45 adults participated in this study (39 female, 6 male, mean age = 22.37, *SD* = 3.42, age range = 18–30 years). All of them were monolingual native speakers of German without self-reported neurologic or psychiatric symptoms, language impairments, and hearing or vision problems. All participants were students at the University of Potsdam and were recruited via an online participant database. They received course credits or monetary compensation for participation. Written informed consent was obtained from all participants prior to the study. They were naïve to the purpose of the study. The procedure for this study was approved by the Ethics Committee of the University of Potsdam (approval number 72/2016).

Stimuli

Structure of the source material The gated stimuli were based on non-manipulated recordings of coordinate structures taken from a production study by Huttenlauch et al. (2021). We will refer to these original recordings of coordinate structures as the *source material* in the following. The source material appeared in two grouping conditions, one with internal grouping of the first two names as in (32), and one without internal grouping as in (33). The same coordinate structures had been previously used in production and perception studies (e. g., Holzgrefe-Lang et al. 2016, Huttenlauch et al. 2021) and consisted of six items that all had the same structure: a sequence of three German names coordinated by *und* (‘and’). All names were disyllabic, stressed on the penultimate syllable, and ended either in an /i/ (Moni, Lilli, Leni, Nelli, Mimmi, and Manni) in the position of Name1 and Name2 or in /u/ or /a/ (Manu, Nina, and Lola) as Name3. We controlled for frequency effects of adjacent names: The occurrence of all possible adjacent name combinations was non-frequent in the dlexDB corpora (Heister et al. 2011) as well as in printed sources covering the years 1500 to 2021 accessed in an online-search using the Google Ngram Viewer (Lin et al. 2012).

Of all the 15 speakers analyzed by Huttenlauch et al. (2021), four speakers were selected on the basis of a perception check conducted in the same study. This perception check had been carried out to confirm that the internal grouping of constituents produced by the speakers following the instructions in the production experiment was congruent with the structure perceived by naïve listeners (n = 31 in Huttenlauch et al. 2021). In contrast to the procedure of the current study, participants in the perception check listened to the complete productions. For each production, they were asked to identify the grouping condition (internal grouping vs. no internal grouping) and choose between two pictograms, one depicting two persons grouped together and the third person standing alone (as in Figure 28a in section 7.3



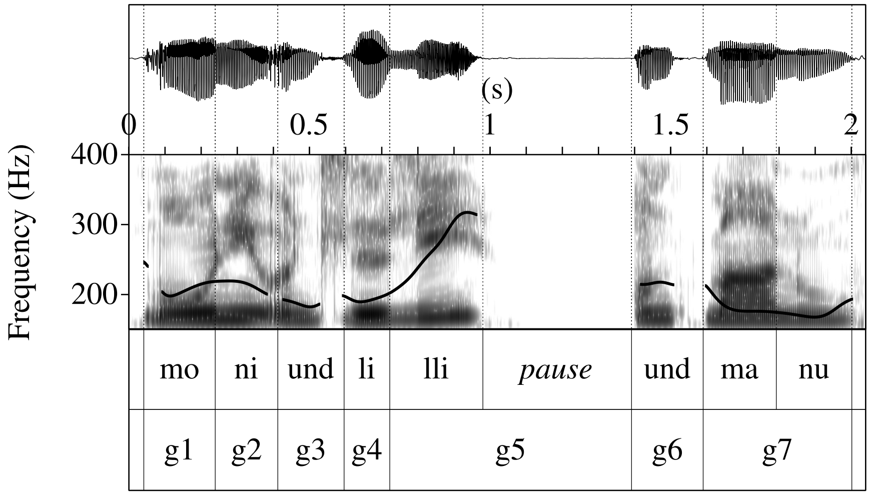
g1: Mo
g2: Moni
g3: Moni und
g4: Moni und Li
g5: Moni und Lilli
g6: Moni und Lilli und
g7: Moni und Lilli und Manu

Figure 24: Example of the segmental material in each of the seven gates (g1–g7) that were cut from a complete coordinate three-name sequence (cf., g7).

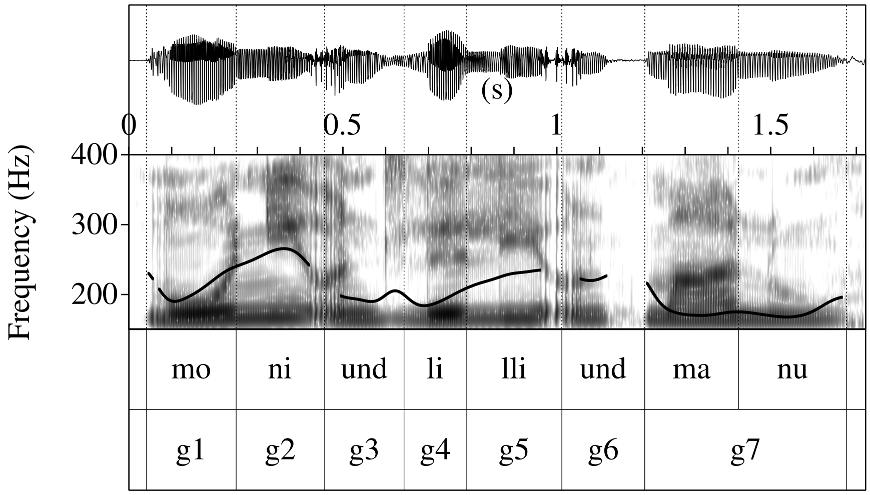
on page 129, referring to internal grouping) and one with three persons grouped together (as in Figure 28b in section 7.3 on page 129, referring to no internal grouping). In the analysis, the ratio of the number of congruent responses to the number of total responses (referred to as rating accuracy) was calculated. The 48 productions of the four speakers selected for the current study (6 name sequences * 2 conditions * 4 speakers) had achieved a slightly higher rating accuracy (mean per speaker > 98%) than the productions of the remaining eleven speakers (mean: 94%). We interpreted high ratings as indicating that the intended structure could reliably be recovered by naïve listeners when listening to the complete coordinate structure. The four selected speakers (speaker IDs 6, 10, 11, and 16) all identified as female and had a mean age of 24 years (*SD*: 4.24, range 21–30).

Creation of the gated stimuli For the current study, the 48 recordings were each cut into seven parts (gates, g1–g7), yielding a total number of 336 gated stimuli. Ascending gate numbers represent longer utterance durations and an increasing amount of prosodic information (see Figure 24, Figure 25a and Figure 25b showing the position of the gates in the utterance). As of g7, the corresponding recording comprised the whole utterance (i. e., a complete coordinated three-name sequence). For the cutting procedure, the segment boundaries and pauses, as previously labeled by Huttenlauch et al. (2021) according to the criteria of Turk et al. (2006) in the software *Praat* (Boersma & Weenink 2017), were used.

Descriptive visualization of speaker-specific cue use In the following, we will provide a short description of the prosodic nature of the source material. We will mainly focus on the three prosodic cues that have commonly been investigated in previous studies, including Huttenlauch et al. (2021), as indicators of internal grouping: F0 movement, final lengthening, and pause after Name1 and Name2. F0 movement captures the distance between the F0-minimum and the F0-maximum in semitones separately on Name1 and Name2. Final lengthening gives the duration of the final vowel of a name relative to the duration of the whole name (in percent), again, separately for Name1 and Name2, and the variable pause contains the duration of a possible pause following Name1 and Name2 relative to the duration of the whole utterance (in percent). In the bracket version of the source material, the prosodic cues on Name1 are expected to be smaller as compared to the no bracket version,



(a)



(b)

Figure 25: (a) Oscillogram/spectrogram with F0 contour (solid line), names, and corresponding gate numbers for a example stimulus for the bracket condition. (b) Oscillogram/spectrogram with F0 contour (solid line), names, and corresponding gate numbers for a example stimulus for the no bracket condition.

while on Name2, the prosodic cues are expected to be larger as compared to the no bracket version.

For all productions of the four speakers selected as source material for the gated stimuli for the current study, the two grouping conditions could reliably be differentiated by naïve listeners (cf., 7.3). However, this does not rule out the possibility that there are inter-individual differences due to speaker-specific use of prosodic cues. This was also confirmed by the analysis in Huttenlauch et al. (2021), which is why we describe the prosodic cues of the source material of the stimuli in a speaker-specific manner.

Figure 26 shows the distributions of the three cues (rows) in raincloud plots (Allen et al. 2019) separately for Name1 (left column) and Name2 (right column) and individually for the four speakers (y-axis). The pause after Name1 is not visualized as most productions lack a pause at this position. The figure depicts values for individual utterances (black dots for the bracket condition, grey dots for the no bracket condition) together with the density distribution and a box plot within cue, condition, and speaker. Overall, black and grey dots as well as the density show a larger overlap within single cues and speakers on Name1 (especially for final lengthening) as compared to Name2. We interpret an overlap as an indication that the corresponding cue was not used distinctively between the bracket and the no bracket condition. Thus, final lengthening on Name1 was not used to clearly distinguish between conditions. In contrast, for F0-range on Name1, two speakers show less overlap (i. e., 6 and 11) than the other two. The figure suggests that the former two speakers systematically used F0 on Name1 to differentiate the bracket from the no bracket condition. On Name2, there are more cases where the two conditions show no overlap (e. g., F0-range in speakers 16, 11, and 6, pause in all speakers). In sum, on Name2 more cues diverge between conditions than on Name1. Nevertheless, as mentioned in the introduction, the analysis by Huttenlauch et al. (2021) also revealed reliable cues on the group level on Name1.

The three described prosodic cues are relative measures that unfold over time or in relation to the surrounding speech material. This makes it difficult to determine a specific point in time where a cue is located or takes effect. The association of a prosodic cue with a specific gate is, thus, always a simplification. We associate g2 and g5 with the cues final lengthening and F0-range. In the case of F0-range, the F0-minima are largely located on g1 and g4, while g2 and g5 bear the end positions of the movement. It is, therefore, possible that there are already perceivable differences on the preceding gates. The production of a silent pause, however, is only perceivable in the presence of following speech material (i. e., on g3 and g6). Thus, simplified, the gate corresponding to the boundary at the group edge on Name2 in the bracket condition is g5. Finally, it should be noted that additional cues may be present in the utterance which have, so far, not been investigated.

For an additional visualization, we extracted F0 values of the source material in order to be able to consider the F0 contour of the whole utterance in its continuous nature, using a customized praat script that combines the procedures of Mausmooth (Cangemi 2016) and

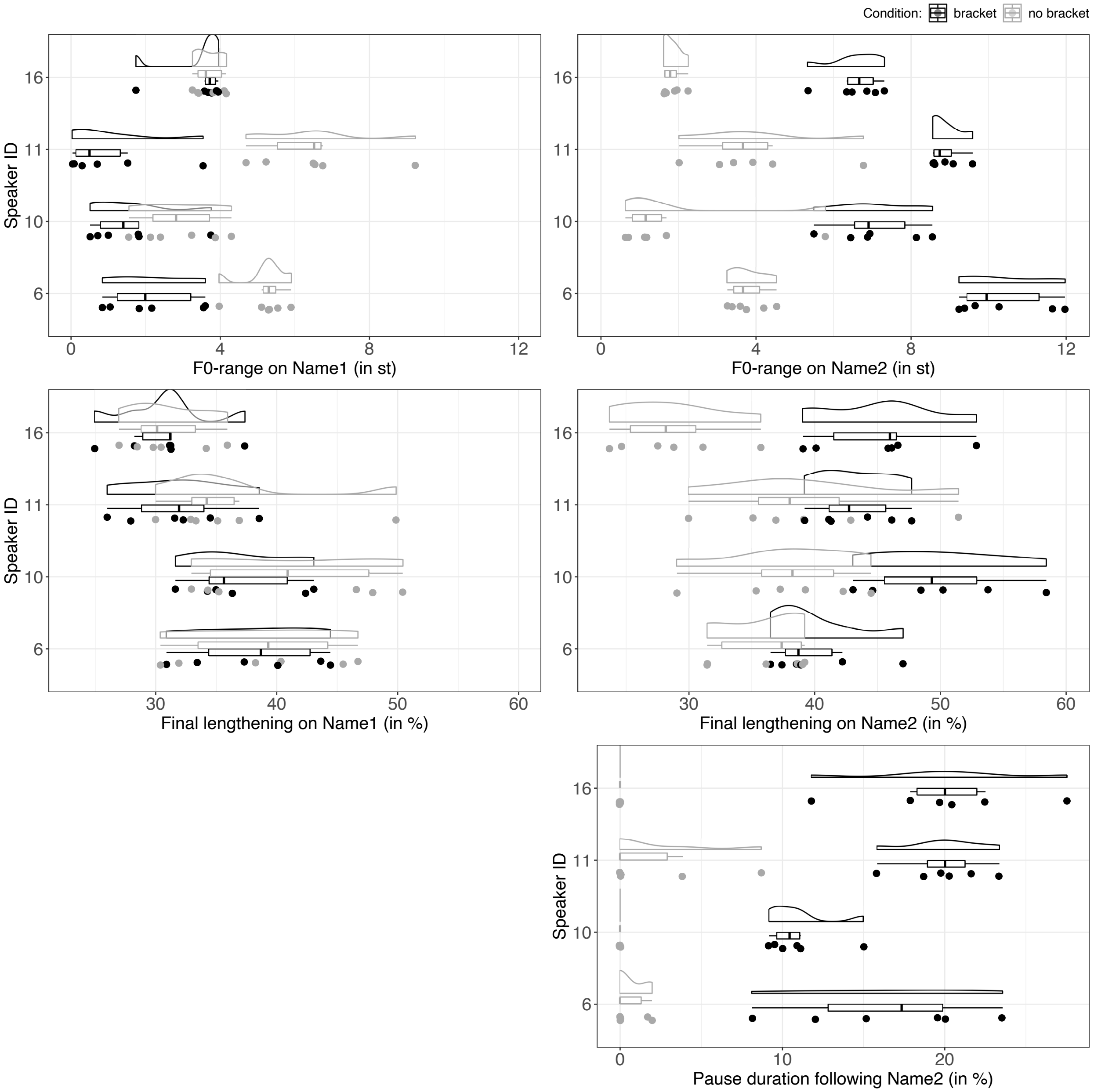


Figure 26: Distribution of the three prosodic cues F0-range (upper row), final lengthening (mid row), and pause (bottom row) on Name1 (left column) and on Name2 (right column) by condition (black–bracket, gray–no bracket) separated for speakers (y-axis).

Note. The raincloudplot combines the probability distribution (density), central tendency measure (boxplot with median), and raw jittered data points per condition. The pause after Name1 is not visualized as most productions lack a pause at this position.

ProsodyPro (Xu 2013). Unreliable pitch points were removed manually before smoothing the pitch contour with a bandwidth of 10 Hz. After interpolation of pitch points, the contour was smoothed again with a bandwidth of 15 Hz following the procedure in Cangemi (2016). A total of 140 F0 values (10 per each segment in the names and 10 per coordination) were extracted and converted into semitones (st) relative to 1 Hz following Hazan et al. (2016) to facilitate a comparison independent of pitch height. Intervals labeled as pauses were not considered. Figure 27 shows, thus, the smoothed F0 contours, plotted separately for each speaker and condition (bracket and no bracket productions on top of each other). As time on the x-axis is normalized and pauses are excluded, the figure does not contain information in the durational domain. The (normalized) time domains of the seven gates are given by vertical lines. Considering the black mean lines (solid for bracket and dashed for no bracket) and the shaded standard deviations in the time domains of g1 and g2, differences between speakers become apparent, since the two lines neatly overlap for speaker 16, but start to diverge on g2 for the other three speakers.

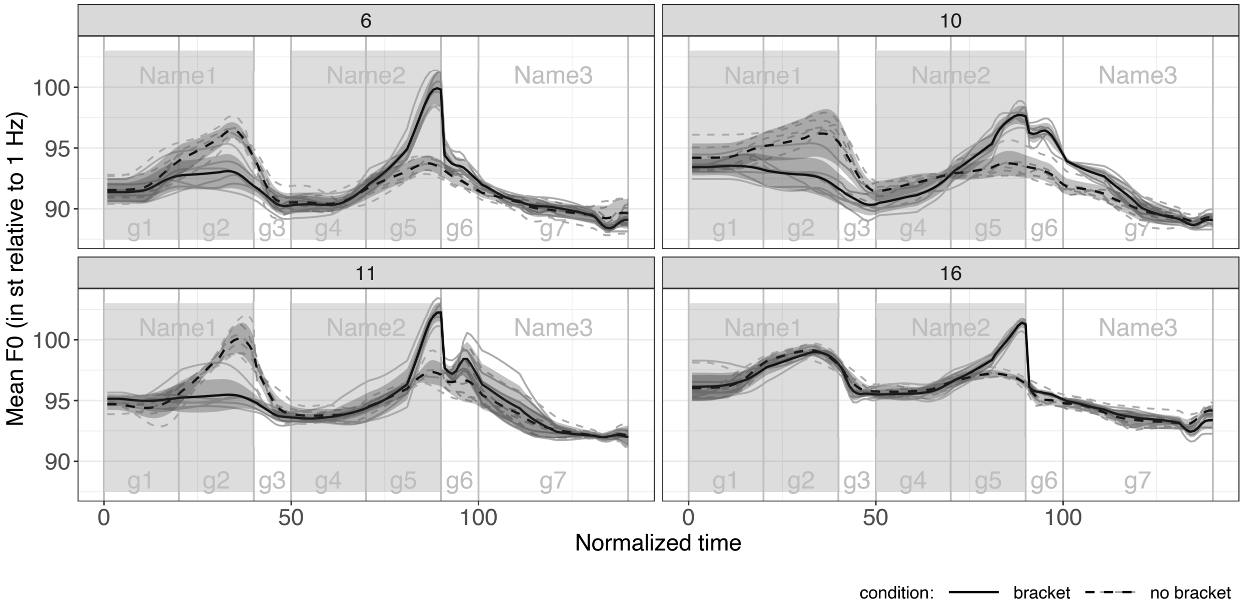
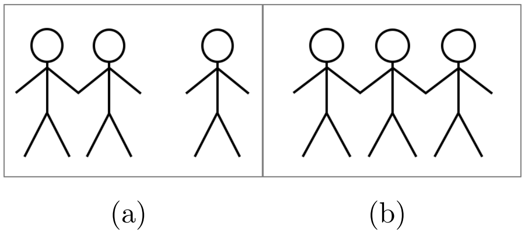


Figure 27: Time normalized smoothed F0-contours for the bracket (solid) and the no bracket (dashed) condition, separated for speakers (panels).

Note. Black lines show means per speaker and condition, gray lines show individual productions. Shaded bands indicate standard deviations. Time domains of names and gates are given by vertical lines and/or shading. Note that all gates start at the beginning of the utterance; the vertical line to the right of a gate, indicates the right edge of this gate.

Experimental procedure

The experiment took place in the Acoustics Laboratory at the University of Potsdam. Participants were tested one by one with a single session lasting about 60 minutes, of which the actual experiment took about 30 minutes. Participants were seated in a sound-attenuated booth in front of a flat-panel display with 1920 x 1200 resolution. They received the instructions in verbal and in written form and were given the opportunity to ask questions before and after the practice phase. The practice phase was run prior to the test phase and consisted of two gated utterances, one for each condition, thus 14 audio snippets in total. The stimuli presented in the practice phase had been produced by a different, randomly chosen speaker and had also been verified regarding the identifiability of the respective condition in the perception check in Huttenlauch et al. (2021).



(a) (b)

Figure 28: (a) Pictogram (bracket condition) indicating button press in the experiment. (b) Pictogram (no bracket condition) indicating button press in the experiment.

In the practice and test phases, the gated audio stimuli were presented via a HSC 271 headset (produced by AKG Acoustics). Randomization of source stimuli but not gated stimuli (meaning that the ascending gates in each test item were always from the same individual uncut source stimulus) was implemented for test items by means of eight different randomization lists. Scripts were written and run in *Open Sesame* (Mathôt et al. 2012), version 3.3.6, logging all data that *Open Sesame* gathered during the experiment and selecting the variables relevant for analysis after the experiment while dropping redundant columns. *Open Sesame* was executed on a Dell laptop that was located outside the sound booth and connected to all technical devices used in the experiment via an Alesis io12 interface.

The experimental task was a forced-choice decision task with two alternatives in which participants had to assign each gated stimulus to one condition, no bracket or bracket. Answers were given via button press on a Cedrus RB-840 button box, using the left and right index fingers, after the stimulus onset. To avoid time delays while answering, participants were advised to place their fingers on the buttons again after each trial. Answers were supposed to be given as fast as possible. Thus, it was possible to already give an answer before the end of the auditory stimulus presentation. One trial consisted of the auditory presentation of a gated stimulus while showing a fixation cross on the screen, followed by the visual presentation of two pictograms after 1000 milliseconds, each referring to one of the two conditions (see Figure 28a and Figure 28b). In four of the randomization lists, the bracket option was localized on the left side of the screen and the no bracket option on the right, while in the other four lists the pictograms were switched. After each given answer, participants had to rate their confidence for the given answer on a seven-point scale using the respective number bars on a keyboard (1 corresponded to completely unsure, 4 corresponded to somewhat sure, 7 corresponded to completely sure). The confidence rating (CR) was followed by a blank screen that lasted for 2000 milliseconds before the start of the next trial.

Statistical analysis

All calculations were executed using the software *RStudio*, version 1.3.1056 (R Development Core Team 2018). Visualizations were also generated in *RStudio*, using the package *ggplot2* (Wickham 2016), version 3.3.3. From the variables logged by *Open Sesame* during the experiment, the following variables were selected and used for analyses as outcome variables, predictor variables, or random effects: response accuracy (correct/incorrect), condition (no bracket/bracket), confidence rating (CR), speaker (6, 10, 11, 16), item, and subject. The data analysis was carried out in the Frequentist Framework, using Linear Mixed Models (*LMM*) and Generalized Linear Mixed Models (*GLMM*) for significance testing. A conservative alpha level of .05 was predefined. For model implementation, the functions *lmer* and *glmer* from the package *lme4* (Bates et al. 2015b)[37], version 1.1-26, were used. For all predictor variables that were used in the analyses outlined in the following, the significance of the predictor was evaluated in model comparisons using the *anova* function from the package car (Fox & Weisberg 2018), version 3.0-10. Likewise, the best applicable model complexity was assessed, further taking into account the *Akaike information criterion* (Akaike 1974) as well as the *Bayesian information criterion* (Schwarz 1978). Predictor contrasts were tied to research questions and predetermined predictions, or to certain exploratory questions that were specified before running the model analysis, and were coded using the R package *MASS* (Venables & Ripley 2013), version 7.3-53. Random effects were determined as proposed by Barr et al. (2013). For logistic regressions, extracted model estimates were transformed from log odds to percentage proportions prior to interpretation. All reported results lie within the respective 95% Confidence Interval (*CI*).

Analysis of response accuracy

Change range and significance tests Using a Binomial Sign Test, the accuracy score that indicates a robust performance above chance within one gate was calculated. This was

[37] cited in original as Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker. 2014. Fitting linear mixed-effects models using lme4. *ArXiV Preprint ArXiv:1406.5823* doi:10.18637/jss.v067.i01

done using the function *binom.test* from the R base package *stats*, with a one-sided test (alternative "greater"). For an additional reassurance of robustness, we checked whether a performance above chance was constant for successive gates within participants. Significance of observed differences between gates was calculated using a *GLMM*. Following the first research question, *gate* was included as a predictor coded with a Sliding Difference Contrast. As a result, the linear model successively compared the levels of the factor *gate* against each other – g2 was compared to g1, g3 to g2, and so forth. The full model further comprised random effects of *gate* with correlating varying intercepts and slopes by subjects and items. *Condition* was not included as a predictor since the predictions for the related research question were not specific to the no bracket or bracket condition.

Post hoc ratings of response patterns and subgroup analysis After data collection, the data were visualized and explored to get an overview of the distributions and to check for outliers and unexpected or interesting patterns. Two distinct response patterns were apparent when looking at the visualizations of the given responses per participant. In order to find out whether participants could potentially be grouped according to their response patterns, we let six individuals with a background in experimental linguistics match the visualizations per participant (as in Figure 29a and Figure 29b in section 7.4 on page 136, but unsorted) to one of the proposed subgroups. The two recognizable subgroups were described to the raters in a neutral way that did not include any hypothesized background assumptions, thus, the descriptions define patterns resulting from response behaviors (pressing one or both buttons) and do not make any claims about response decisions (see below). There was also the opportunity to assign a participant to an alternative third group (*Neither of the above* (n)), if the response pattern did not fit either description.

The descriptions of the different response patterns given to the raters read as follows:

Group 1: **Waiting pattern** (w):

Participants in this group stuck to one button during the first gates for the vast majority of trials. This results in a response pattern with one condition mostly at an accuracy of 1 (i. e., correct) and the other at 0 (i. e., incorrect). At higher gate numbers, participants used both buttons and the overall accuracy increases.

Group 2: **Identification pattern** (i)

Participants in this group used both buttons right from the beginning and throughout the experiment. This results in a response pattern with both conditions distributed across accuracy 1 and 0.

Group 3: **Neither of the above** (n)

For participants in this group, it is impossible to deduce a certain response pattern. They neither have a visible tendency to fit in group 1 nor group 2.

Following the results of the rating, participants were categorized into subgroups.

To complement the percentage values of the agreement and assess rating reliability while accounting for the possibility of guessing, Fleiss Kappa was computed. Free-marginal kappa was chosen here, since the raters were not restricted regarding their distributions of cases into categories (Randolph 2005). The additional variable *subgroup* was added to the results data frame and was used as a predictor variable, with *gate* nested under it in the maximal *GLMM*. Random intercepts and slopes for *gate* and *subgroup* by items were also included. Since subgroups are linked to subjects, no random intercepts and slopes by subjects were included. The applied Sum Contrast compared each factor level of *subgroup* to the grand mean. Thus, this model was testing for statistically significant differences between the subgroups within each gate.

Exploratory analysis of accuracy by speaker Different speakers naturally exhibit individual patterns of prosodic cue usage. To explore differences between speakers and the performance within each speaker for each gate, a *GLMM* was run using *speaker* (i. e., the person who had produced the coordinate structure in Huttenlauch et al. (2021) as a sum-contrasted predictor variable nested under *gate*. This model compared factor levels of *speaker* with a reference level (speaker 6) and the gate-wise increase in performance by speaker. The choice of a (in this case arbitrary) reference level was required to make a speaker-comparison possible. The full model also contained random effects for speaker, with correlated varying intercepts and slopes by subjects. Since items are linked to speakers, varying intercepts and slopes were only calculated by subjects.

Familiarization effects Experimental tasks are often different from natural processing – so is a forced-choice decision task with gated stimuli. To check for familiarization effects, a unique variable for *familiarity* was created, based on the possibility that participants might have undergone adaptation to the task or learning over the course of the experiment as indicated by increasing accuracy scores. For this, the very first ten (out of 192) coordinate structures each participant encountered (split into seven gates, thus equivalent to the first 70 trials) were categorized as *unfamiliar*. All following trials (n = 1274) were categorized as *familiar*, in the sense of post-familiarization. A *GLMM* was set up evaluating *familiarity* as a sum-contrasted predictor of *accuracy*, including varying intercepts and slopes for *familiarity* by subjects. Additionally, a model with *familiarity* nested under *subgroup* was compared to the model solely including *familiarity*. The full model included varying intercepts and slopes for *subgroup* by items and for *familiarity* by subjects, including correlation parameters. Since the speaker that each participant encountered initially varied according to randomization lists, an interaction of *familiarity* and *speaker* was also tested.

Analysis of confidence ratings As another complementary analysis, CRs were analyzed using a *GLMM* with the potential predictors *accuracy, gate, condition, speaker*, and *subgroup*. The full model comprised a random structure with correlated varying intercepts and slopes for all significant predictors by subjects and items.

7.4 Results

Two participants had to be excluded from analysis due to performance at chance level or below at g7 (where the whole utterance was presented) or accuracy scores of more than two Standard Deviations (*SDs*) below the group mean at g7. The observed performance in these two individuals indicates a lack of ability to identify internal grouping correctly after all prosodic information was given – or possibly a lack of motivation for correct task execution. A remaining number of 43 participants were included for data analysis (37 female, 6 male, mean age = 22.14, *SD* = 2.83, age range = 18–30 years).

Descriptive statistics

Table 19 provides an overview of means and *SDs* by gate for accuracy and CRs. Accuracy increases with higher gates while *SDs* decrease. CRs also increase (that is, confidence in given responses increases) with higher gate numbers.

Table 19: Means and *SDs* of accuracy (proportion correct) and confidence ratings (CRs, 1–7) by gate (across all 43 participants). * accuracy above chance.

	Gate 1	Gate 2	Gate 3	Gate 4	Gate 5	Gate 6	Gate 7
				Accuracy			
Mean	0.57	0.62	0.67*	0.69*	0.87*	0.96*	0.97*
SD	0.50	0.48	0.47	0.46	0.33	0.20	0.17
				Confidence ratings			
Mean	1.37	1.87	2.54	3.17	5.23	6.42	6.77
SD	0.82	1.19	1.54	1.77	1.75	1.14	0.76

Statistical analyses of response accuracy

Response accuracy in relation to chance and additional check of robustness The accuracy value at which performance was robustly considered above chance is 0.65, resulting in g3 being the gate where performance exceeds the chance range at the group level. About 65% of participants already scored above chance that early (see Table 20). At g5, nearly all participants scored above chance.

Table 20: Number and percent of participants (n = 43) with accuracy above chance by gate.

	Gate 1	Gate 2	Gate 3	Gate 4	Gate 5	Gate 6	Gate 7
				Accuracy			
n	9	18	28	30	42	43	43e
%	20.93	41.86	65.12	69.77	97.67	100	100

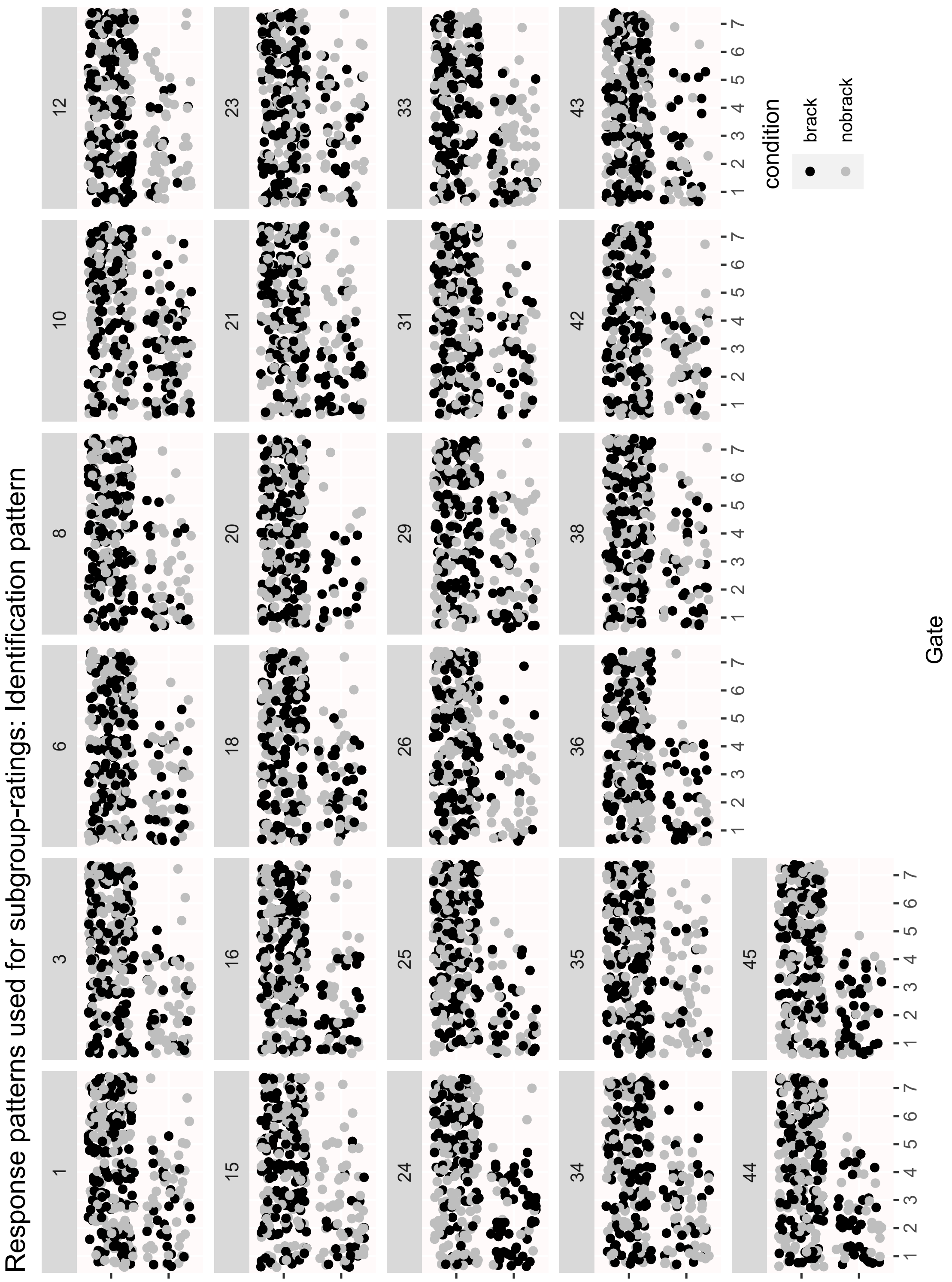
The additional sanity check (robustness of an above-chance score in subsequent gates per participant) revealed a less robust performance within participants for the first gate than for all following gates. That is, at g2, two out of nine participants that scored above chance at g1 no longer showed performance above chance. As of g3, however, the performance of all participants who scored above chance was constant for subsequent gates.

Generalized Linear Mixed Models The variable *gate* was a significant predictor of accuracy ($p < .0001$) and was included in the GLMM. Since the full model did not converge, it was reduced. The most complex converging model was a zero correlation parameter (zcp) model with varying intercepts and slopes for *gate* by *subjects* and *items*. Fixed effects are displayed in Table 21. Statistically significant differences between gates were found for g5 compared to g4, g6 compared to g5, and g7 compared to g6.

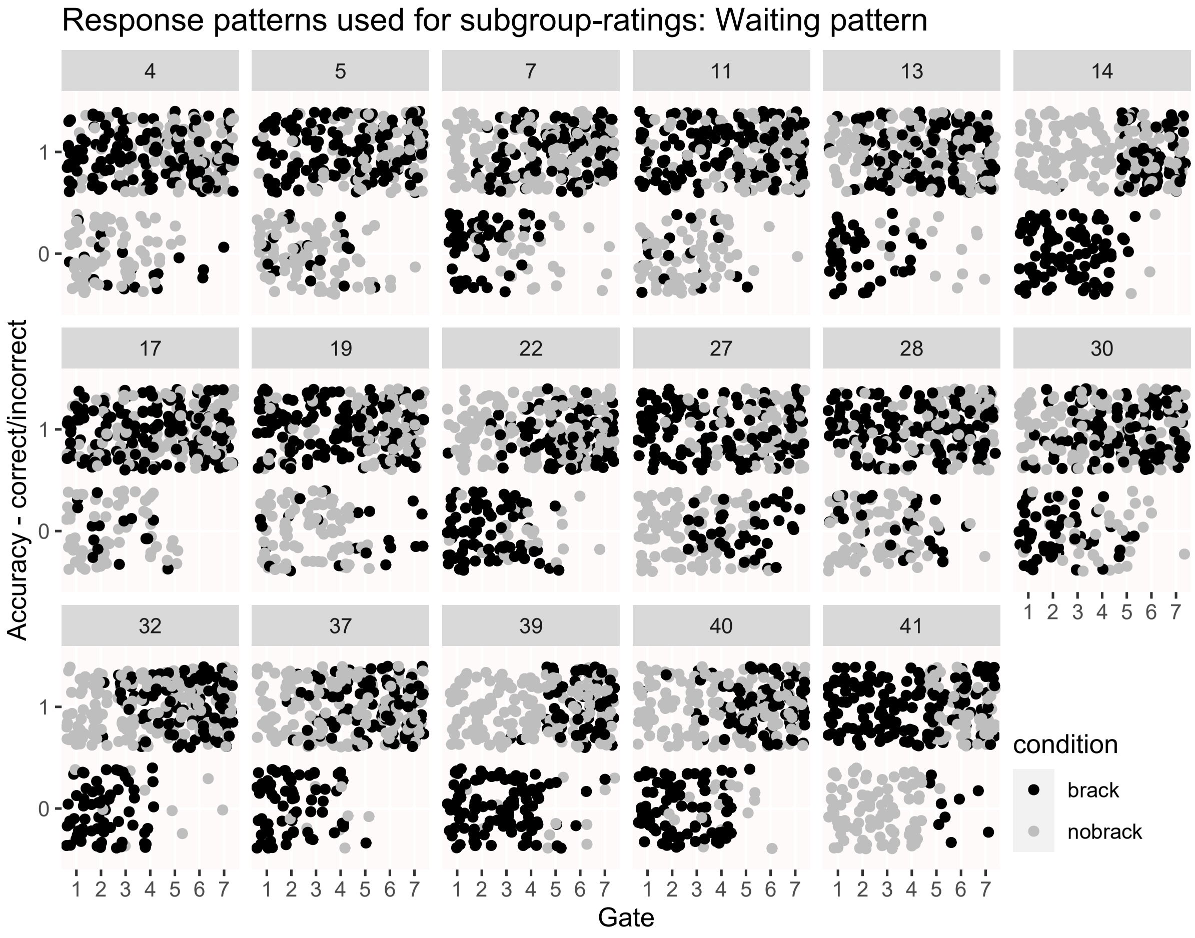
Table 21: Fixed effects of the model on accuracy by gate. A Sliding Difference Contrast was used to successively compare adjacent factor levels. Estimates are presented as in the original model output (log odds) as well as in percentage increase from gate to gate. Statistically significant effects are marked in bold ($p < .05$).

Effect	%	Log odds	*SE*	*Z*	*p*
Intercept	87.361	1.933	0.094	20.635	<**.0001**
Gate 2 vs 1	2.669	0.267	0.162	1.651	0.099
Gate 3 vs 2	2.785	0.28	0.162	1.726	0.084
Gate 4 vs 3	1.073	0.101	0.161	0.627	0.053
Gate 5 vs 4	9.252	1.417	0.212	6.684	<**.0001**
Gate 6 vs 5	9.502	1.496	0.255	5.866	<**.0001**
Gate 7 vs 6	6.365	0.771	0.344	2.24	<**.025**

Ratings of answer patterns Interrater agreement was 76.89% with a free-marginal kappa value of 0.65 (95% *CI*: 0.54, 0.77). 26 participants were assigned to the identification pattern subgroup and 17 to the waiting pattern subgroup (see Figure 29a and Figure 29b). None of the participants was assigned to the third option (neither of the above). It is noteworthy


Response patterns used for subgroup-ratings: Identification pattern
1
3
6
8
10
12
15
16
18
20
21
23
24
25
26
29
31
33
34
35
36
38
42
43
44
45
Accuracy - correct/incorrect
1
0
Gate
1 2 3 4 5 6 7
condition
● brack
● nobrack

(a)



(b)

Figure 29: (a) Scatterplots of response patterns of the participants assigned to the identification subgroup. The numbers refer to the participant IDs. (b) Scatterplots of response patterns of the participants assigned to the waiting subgroup. The numbers refer to the participant IDs.

that all participants who already scored above chance at g1 (see Table 20) were rated as belonging to the identification pattern subgroup.

Subgroup analysis There was strong evidence for *subgroup* as a predictor of *accuracy* ($p < .0001$). The maximal model including varying intercepts and slopes for *gate* and *subgroup* by items did not converge, thus the results from a model with correlated varying intercepts but no slopes are reported (see Table 22). Figure 30 additionally shows accuracy per gate and subgroup.

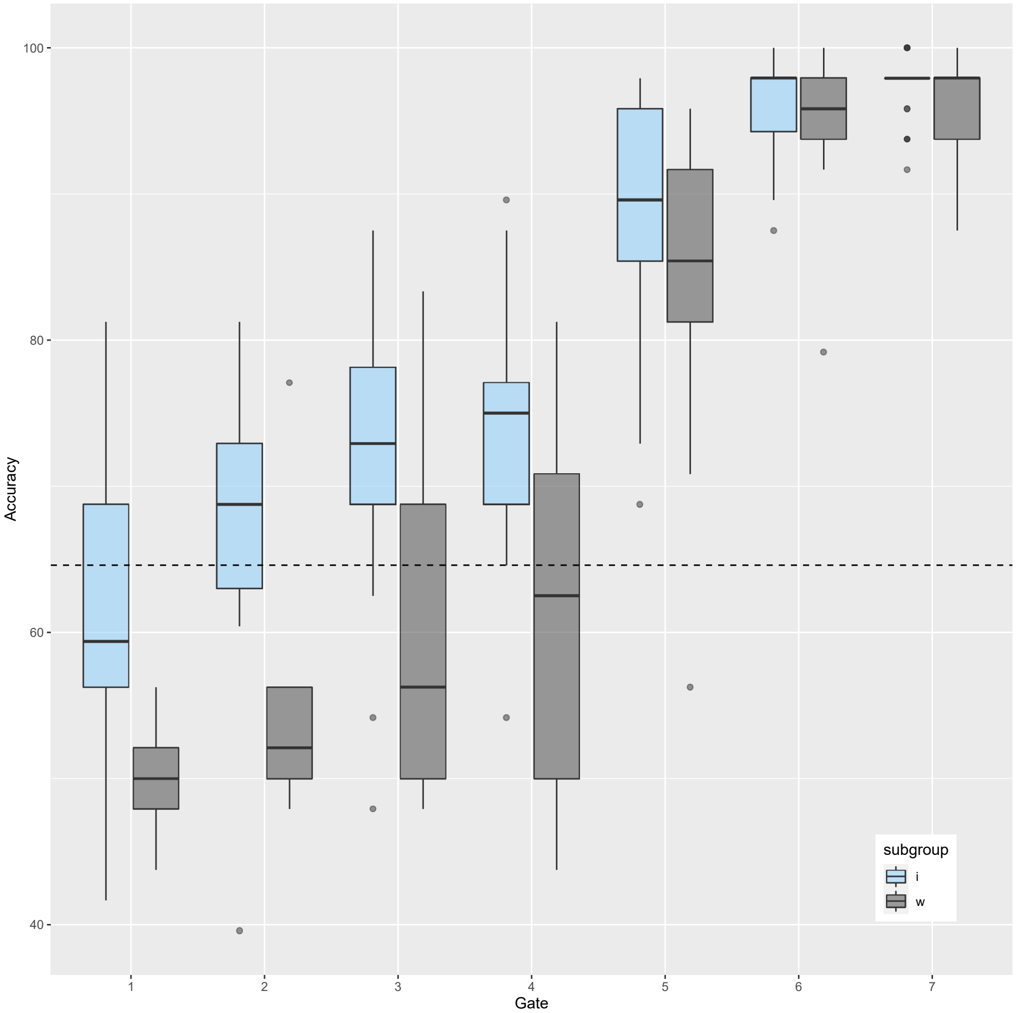


Figure 30: Boxplots of accuracy (in %) by gate and subgroup, identification pattern group (i)/waiting pattern group (w).

The model revealed that up to g5, accuracy was significantly higher for the identification pattern subgroup compared to the subgroup of participants classified as showing the waiting pattern. There were no significant differences in accuracy between the subgroups for g6 and g7.

Table 22: Fixed effects of the model including *subgroup* (identification pattern (i) /waiting pattern (w)) nested under *gate* as a predictor of *accuracy*. Estimates refer to the intercept and are presented as in the original model output (log odds) as well as in percent difference between the two subgroups. Significant effects are marked in bold ($p < .05$).

Effect	%	Log odds	*SE*	*Z*	*p*
Intercept	84.663	1.708	0.055	31.239	<**.0001**
Gate 1: (i)	6.216	0.239	0.046	5.146	<**.0001**
Gate 2: (i)	8.8	0.337	0.048	7.028	<**.0001**
Gate 3: (i)	8.435	0.324	0.051	6.371	<**.0001**
Gate 4: (i)	8.534	0.327	0.07	6.328	<**.0001**
Gate 5: (i)	4.651	0.179	0.084	2.544	**0.011**
Gate 6: (i)	2.791	0.107	0.118	0.908	0.364
Gate 7: (i)	3.794	0.146	0.141	1.033	0.302

Accuracy by speaker There was no evidence for *speaker* as a single predictor of accuracy in a model comparison ($p = .427$). A model including *speaker* nested under *gate* attained a significantly better fit to the data ($p < .0001$). The full model including random slopes did not converge. The results were thus extracted from a model with correlated varying intercepts. For the productions of speakers 6 and 11, the model revealed a significant increase in listeners' accuracy with increasing gates: g2 compared to g1 (3.615%, $p = .03$; 5.581%, $p = .0002$), g3 to g2 (3.709%, $p = .03$; 4.663%, $p = .004$), g5 to g4 (9.859%, $p < .0001$; 8.642%, $p < .0001$), and g6 compared to g5 (12.019%, $p < .0001$; 0.452%, $p = .002$). For speakers 10 and 16, a significant increase in listeners' accuracy was found for g5 compared to g4 (9.326%, $p < .0001$; 12.745%, $p < .0001$) as well as for g6 compared to g5 (10.838%, $p < .0001$; 7.571%, $p < .0001$).

Figure 31 complements the analysis results. Thus, for productions stemming from two speakers (speaker 6, speaker 11), accuracy already improved significantly early. Only later, that is with higher gates, did listeners' performance also increase for speaker 10 and speaker 16, with a more pronounced effect for speaker 16.

Familiarization effects *Familiarity* was a significant predictor of *accuracy* in the model comparison ($p < .0001$). There was no evidence for an interaction between *familiarity* and *speaker* ($p = .193$). The GLMM including *familiarity* nested under *subgroup* fit the data significantly better than the model including *familiarity* as a single predictor ($p < .0001$). Results are reported from the full model. The effect of familiarity was significant ($p = .002$), with a 3.194 percent higher accuracy for familiar than for unfamiliar items. Participants with an identification pattern outperformed participants with a waiting pattern in both the familiarization phase, by 12.457 percent, and the post-familiarization phase, by 13.702 percent.

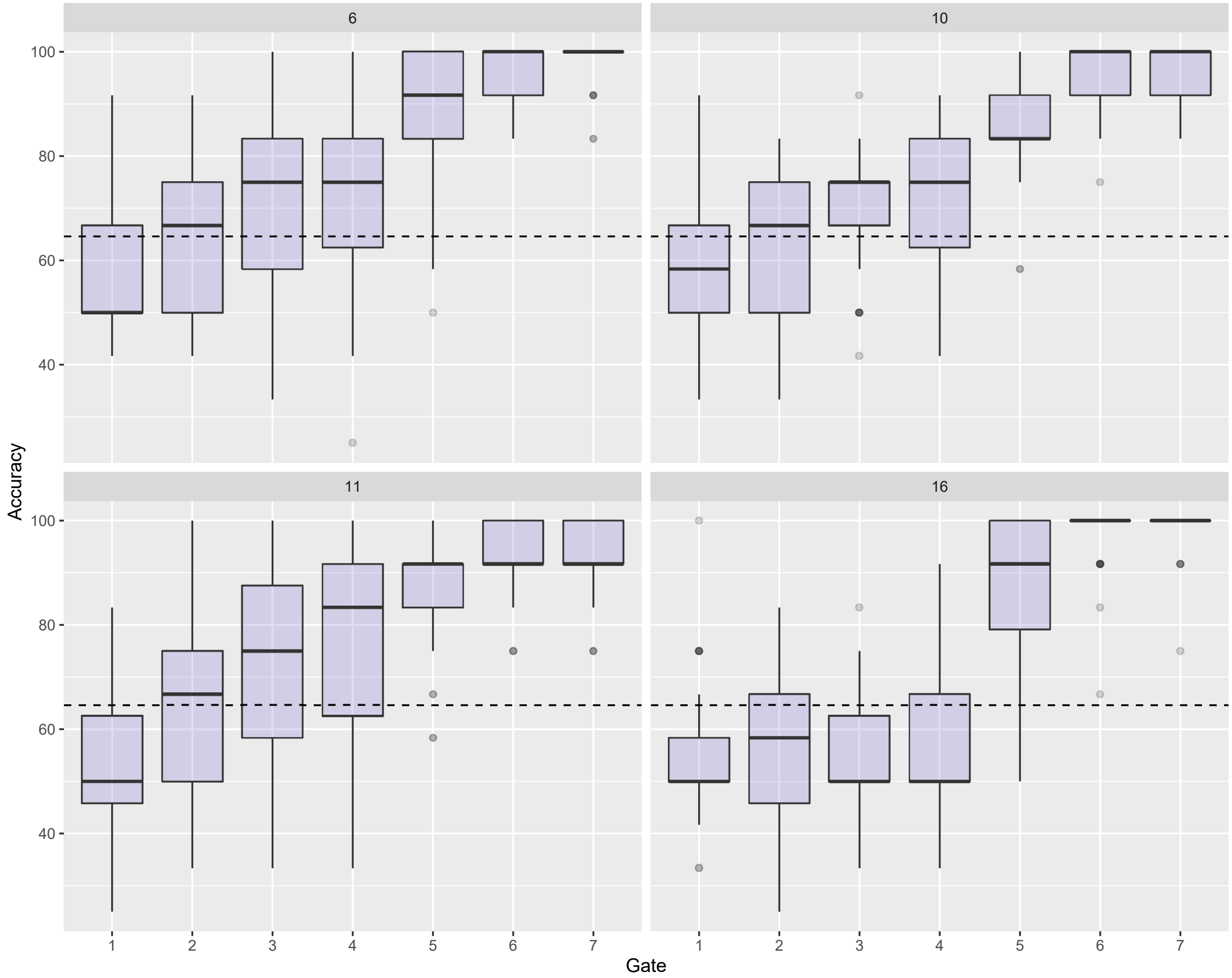


Figure 31: Boxplots of accuracy (in %) by gate and speakers (6, 10, 11, 16).

Confidence ratings

Except for *condition* ($p = 0.361$) and *speaker* ($p = 0.361$), all other tested predictor variables were significant ($p < .0001$ for *gate*, $p < .0001$ for *accuracy*, $p = .013$ for *subgroup*) in the model comparison. Due to a convergence failure, the full model was reduced. The most complex model that converged included correlated varying intercepts by subjects and items. The model predicted correct answers to be linked to higher CR scores than incorrect answers (4.371%, $p < .0001$). The effect of *subgroup* was also significant ($p = .012$), with participants of the identification pattern subgroup scoring 12.491 percent higher in CRs than participants of subgroup (w). Comparisons between gates were highly significant for g2 compared to g1 (5.336%, $p < .0001$), g3 to g2 (4.529%, $p < .0001$), g4 to g3 (3.822%, $p < .0001$), g5 to g4 (5.485%, $p < .0001$), as well as for g6 compared to g5 (5.391%, $p < .0001$). Figure 32 shows the increasing proportion of high CRs in higher gates.

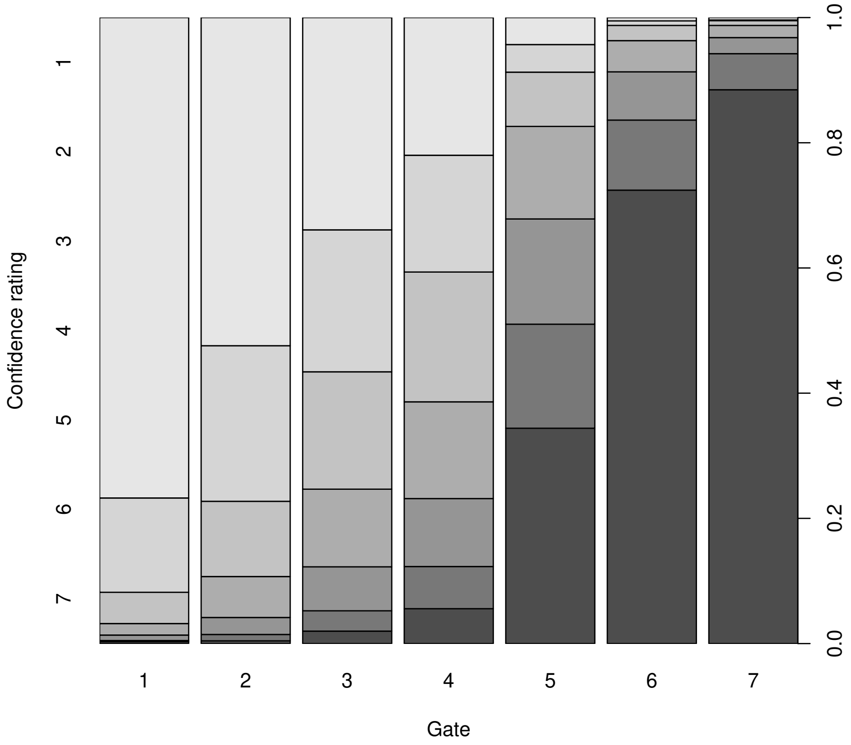


Figure 32: Bar plot of CRs (ratings from one to seven, in proportions) by gate. *Note.* Darker colors represent higher confidence (for instance, at g7, the proportion of participants that rated their confidence very high, was about .9, as opposed to g1, where the proportion of participants who rated their confidence very low was about .8).

Summary of results

This study investigated listeners' ability to exploit early prosodic cues in coordinated three-name sequences to identify the internal grouping of the constituents. At the group level, accuracy exceeded the chance range at g3. Gate-wise comparisons of accuracy were significant for g5 compared to g4, g6 to g5, and g7 to g6. Ratings of the response patterns of participants revealed two subgroups: participants with a waiting pattern primarily stuck to one response button (i. e., one choice) during the first gates up to g5 (Name2), while participants with an identification pattern used both response buttons right from the beginning. *Subgroup* also was a significant predictor of *accuracy*: the identification pattern subgroup significantly outperformed the waiting pattern subgroup at all gates up to g5 (Name2). Similarly, the identification pattern subgroup already scored above chance at g2 (and the following gates) while the waiting pattern subgroup only exceeded the chance range at g5. Also, all participants who already scored above chance at the first gate (n = 9) belong to the identification pattern subgroup.

The experimental stimuli stemmed from four different speakers and thus *speaker* was included as predictor within each gate. For productions of speaker 6 and speaker 11, accuracy already increased significantly early, for all gate comparisons starting with g2 compared to g1. For speaker 10 and speaker 16, an increase in accuracy was not statistically significant until later gates, starting with g5 as compared to g4. Although the visualizations of the speaker specific cue use in the source material are not precisely part of the statistical analyses, we relate our findings to their features for an easier interpretation. In Figure 26, the mean F0-range on Name1 does not overlap between bracket and no bracket conditions for speakers 6 and 11 but they do overlap for the remaining two speakers. The time-normalized F0 contours in Figure 27 show that the F0 contours of the bracket and the no bracket condition start to diverge on Name1 (g2) not only for speakers 6 and 11, but also for speaker 10, though not for speaker 16.

Regarding a possible familiarization, accuracy slightly increased across participants after the first 70 trials, but there was no interaction of *familiarity* and *speaker*. Again, the identification pattern subgroup outperformed the waiting pattern subgroup in both phases, the familiarization phase (that is, the first 70 trials) and the post-familiarization phase. CRs increased gradually across gates (see descriptive statistics in Table 19 as well as Figure 32), which was confirmed by the corresponding *GLMM* – all gate comparisons up to g6 were statistically significant. Confidence in correct trials was rated higher than confidence in incorrect trials. Furthermore, participants assigned to the identification pattern subgroup had higher confidence in their answers than participants with a waiting pattern. *Speaker* and *condition* (bracket/no bracket) did not significantly contribute to explaining variance in the model addressing CRs.

7.5 Discussion

This study was designed to gain insights into the role of early, scarcely investigated prosodic cues in the perception of coordinated three-name sequences with and without internal grouping of the first two names. "Early" refers to the location of the cues, namely on/after the first name (Name1), that is, before the most salient prosodic cues on/after the second name (Name2), which is at the group edge. The overarching question was whether these early cues can be used to predict the syntactic structure of the evolving utterance. More precisely, at which gate are listeners able to reliably distinguish between sequences with or without internal grouping (RQ 1)? A second aim was to explore variability in listeners' respective parsing capacities (RQ 2). Stimuli consisted of three-name sequences that were cut into seven parts (gates, g1–g7) and that were presented to participants with successively increasing length and thus, an increasing amount of prosodic information. The analysis of response accuracy of a two-alternative forced choice decision task was complemented by an analysis of the individual confidence ratings (CRs).

In general, the findings are in line with the prediction that listeners can reliably detect the internal grouping after Name2: at the related gate (g5), almost all participants' performances (97.67%) exceeded the chance range. One participant did not score above chance until g6, possibly due to the cutting of the stimuli: the pause cue is only reliably perceivable in the following gate (g6), since silence at the end of g5 is indistinguishable from the end of the gated recording. The same holds for the pause cue that is present at g2 (Name1) – it will only be reliably perceivable at g3. Additional prosodic information may possibly be located at the coordinating conjunction *und* (English 'and'), which is present at both g3 and g6.

We will now discuss our findings with respect to our research questions. Regarding RQ 1, the processing of early prosodic cues, group level performance was above chance at g3, thus, a reliable detection of internal grouping was already possible shortly after Name1. Gate 3 corresponds to the snippet containing the first name and the following coordinating *und*. Therefore, it is the part of the utterance, where the *Proximity/Similarity Model* by Kentner & Féry (2013) predicts prosodic differentiation between structures with and without internal grouping and where differences in the use of prosodic cues had been observed by production studies (Huttenlauch et al. 2021, Kentner & Féry 2013). The visual inspection of F0 movement in the source material matches this prediction for three out of four speakers: diverging F0 contours are observable as of the second syllable of Name1 (g2) for speakers 6, 11 and 10. Our results suggest that listeners in the current study were able to exploit these early cues for disambiguation. Furthermore, with respect to RQ 2, the study results indicate individual differences among listeners: at least 20 percent of the participants were able to make reliable decisions about the internal grouping even earlier than the group mean, namely already at the first two gates corresponding to Name1 (20.93% of the listeners at g1, 41.86% at g2 – see Table 20) and more than half of the listeners made a reliable decision at the gate before Name2 (g3, 65.12%). Note that performance was constantly above chance for subsequent gates among these participants, hence we consider the above-chance performance to be quite robust.

Overall, the observation of variability between listeners was underlined by the finding of two subgroups: 26 listeners were classified into an identification pattern subgroup and 17 into a waiting pattern subgroup through a rating of their response patterns. For the identification pattern subgroup, the clear above-chance performance at g2 and the fact that all participants who could already reliably judge the internal grouping at g1 belong to this subgroup indicate a high prosodic parsing capacity. However, for the waiting pattern subgroup, it is not clear whether their chance performance up to g5 is due to varying prosodic parsing *capacities* or varying *strategies* for task completion. The former assumption is supported by the finding of listener variability in prosodic parsing that has been observed in different experimental tasks and with different speech materials by Cangemi et al. (2015), Cole et al. (2017), and Roy et al. (2017). The analysis of CRs suggests that participants in the identification pattern subgroup were also more confident about their given answers than the waiting pattern subgroup.

This may be interpreted as an indication of enhanced parsing skills in the identification pattern subgroup and further confirms the existence of clearly distinct differences between the individuals in the two subgroups. These differences are also supported by the fact that accuracy remained significantly higher for participants with an identification pattern than for the waiting pattern subgroup up to g5, where we find the late prosodic cues at the second syllable of Name2. Furthermore, the effect size for differences between subgroups per gate is the highest at g2, where early prosodic information related to Name1 is located. The statistically significant differences at early gates (g1, g2, g3) for the subgroups are presumably related to individual differences with respect to sensitivity to F0 cues. The visual inspection of F0 movement (see Figure 26 in section 7.3 on page 127) as produced by the speakers of our stimuli suggests systematic use of F0 to distinguish between no bracket and bracket conditions. These individual differences are in line with the listener-specific attention to pitch-related features described by Baumann & Winter (2018) for prosodic prominence.

Now, we will discuss the exploratory analyses on speaker-related processing. Identifiability of internal grouping seems to depend not only on parsing capacities or internal strategies of the listener but also on the cues that are produced by the speaker. This assumption is based on the complementary analysis of accuracy by speaker: for two out of four speakers, the statistical models predict a significant improvement in accuracy already at early gates, before Name2. These findings are in line with previous findings on speaker-dependent accuracy in prosodic parsing tasks (Cangemi et al. 2015, Swerts & Geluykens 1994). Figures 26 and 27, which show descriptive visualizations of the three prosodic cues investigated by Huttenlauch et al. (2021), reveal differences in the cue use between the individual speakers which go along with the response behavior of the listeners. For speakers 6, 10, and 11 in Figure 27, means of the smoothed time normalized F0-contours and the SDs of the bracket versus the no bracket condition diverge from each other on the second syllable of Name1 (i. e., the time domain of g2), while they completely overlap in the productions of speaker 16. For the former three speakers, the listeners' mean accuracy was already above the chance range at g2, while it was below for the latter (cf. Figure 31). Although it needs to be clarified whether the difference on Name1 produced by speakers 6, 10, and 11 is audible, the observation provides an indication that F0 is used by listeners to predict the internal grouping structure. In contrast, a visual consideration of final lengthening on Name1 (in Figure 26) does not reveal clear differences between conditions in any of the speakers. Of course, we are aware that caution is needed when drawing conclusions based on the visual inspection of graphs. Thus, future research should statistically verify this issue further.

Interestingly, speaker-related differences in disambiguation are not mirrored by the analysis of participants' confidence ratings at the first gates containing early cues; that is, participants were not more confident in their decisions on speakers 6, 10, or 11 than on speaker 16. Thus, listeners were probably not aware of the fact that certain speakers supplied obviously more "useful" cues than others.

Lastly, we will discuss another exploratory analysis we ran to account for the rather artificial nature of the experimental task: we investigated the influence of familiarity with the task. The analysis revealed a mild improvement in participants' performances after a familiarization phase of the first 70 trials. The superior performance of the identification pattern subgroup is present in both phases, suggesting that participants of this subgroup did not acquire their superior parsing capacities over the course of time but brought them with them. At the same time, the waiting pattern subgroup could not benefit more than the identification pattern subgroup from the familiarization with the task. After all, familiarizing with the task did not seem to have a decisive influence on the skills that were used to solve it. As we could not find an interaction of familiarity and speaker, our results tend to support a syntagmatic process in which listeners identify prosodic features by means of variation in the local context, as opposed to changes perceived in relation to a speaker-specific prosodic space.

With respect to our overarching research question, predictions about the syntactic structure of the whole name sequence seem to be possible based on early prosodic cues on Name1, and about 65 percent of the listeners are sensitive to this early information. Listeners additionally face the difficulty of compensating for a high degree of individual speaker differences. Rapid integration of incoming prosodic information into the parsing process may be an especially rewarding effort in structures of higher complexity than coordinated name sequences.

In any case, the findings of this study underline the global nature of prosodic boundaries as they are already indicated by earlier cues in an utterance which can be effectively used by (at least some) listeners for syntactic parsing. Especially regarding F0 as a cue to internal grouping, it seems necessary to consider the whole time course over which it unfolds, as at least some speakers modulate F0 right at the beginning of the utterance to distinguish between bracket and no bracket conditions and it seems that some listeners use this information for disambiguation. Boundary phenomena, thus, should not be investigated solely as local phenomena, detached from the whole prosodic context, but in a more global manner.

For further investigation of the processing of early cues, a study using the Visual World Paradigm would be a valuable method. By using eye tracking, results from gated stimuli could be compared to those from ungated stimuli (as in Allopenna et al. 1998), to determine how cue exploitation in the Gating Paradigm compares to processing in a more natural setting. This would also allow to corroborate our findings on individual variability in listeners' integration of prosodic markers for ambiguity resolution over the course of prosodic parsing. It would also be interesting to test if our results can be replicated with a wider range of productions and/or productions from more natural settings. As demonstrated by Clifton et al. (2006), among listeners, prosodic cues appear to have larger implications for perception in shorter than in longer constituents among listeners. Furthermore, prosody was observed to be especially crucial in disambiguating utterances which are different in interpretation with respect to intended grouping (Watson & Gibson 2004). Moreover, it would be interesting to investigate the perception-production link and to see if individuals who produce stronger early prosodic cues perform better at perceiving/exploiting these cues than individuals who do not produce clear early prosodic cues.

7.6 Conclusion

The results of this study strongly indicate variability among listeners regarding prosodic parsing: some listeners were already able to correctly predict at the first name whether it belongs to a three-name sequence with or without internal grouping of the first two names. This suggests that these listeners were sensitive to prosodic cue information that is located earlier in the name sequence than the prosodic cues at the end of the grouping (referred to as later cues on the second name). Other listeners were not able to correctly identify the prosodic pattern until the end of the second name. In addition to individual parsing capacities, listeners' responses showed sensitivity to speaker-specific variability that matches the individual differences in prosodic cues observed for the speakers the productions stemmed from. The speakers whose productions received the highest accuracy ratings at early gates show visible differences in F0 on the first name between conditions. As we did not specifically analyze possible facilitation effects of specific prosodic cues for perception, statistical verification of this observation remains for future research. Overall, the data support the notion that prosodic marking of internal grouping is not a local phenomenon but rather unfolds globally over the course of an utterance – and that early prosodic cues provide meaningful information which can be exploited for ambiguity resolution, at least by a subset of listeners.

Acknowledgements

We are very grateful to the comments of two anonymous reviewers and the associate editor who helped to improve the paper considerably.

Data availability statements

Data and code are available via SAGE.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 317633480 – SFB 1287.

8 Study IV

Individual variability in prosodic marking of locally ambiguous sentences

A shorter version of this chapter has been presented at the international conference Speech Prosody 2022 and parts of the following text are taken literally from the corresponding publication (Huttenlauch et al. 2022) and an associated supplement (https://osf.io/gychu/).

Abstract

The German case marking system contains case syncretisms, which, along with a relatively free word order, can lead to sentences with local ambiguities, for instance, SVO and OVS sentences with string-identical noun phrases (NPs) in sentence-initial position. Prosodic marking constitutes one possibility for ambiguity resolution. Comprehension studies showed that listeners are sensitive to f0-manipulations on NP1 of such sentences (e. g., different pitch accents). Here, we investigated which prosodic cues speakers use when they are asked to provide disambiguation in their productions as early as possible. We elicited productions of German SVO and OVS verb-second main clauses, that begin with a case-ambiguous NP1 and are string-identical up to the post-verbal unambiguous NP2. We focused our analysis on the f0-contours in the ambiguous part of the sentences. Overall, there was no consistent f0-pattern that distinguished SVO from OVS sentences. However, analyses with Generalised Additive Mixed Models revealed distinctive f0-contours on the individual level with later and higher f0-peaks on NP1 in SVO vs. OVS sentences. We found that at least some speakers systematically distinguish word order in locally case-ambiguous structures by prosodic cues (f0, silent intervals). The variability in our data suggests to consider the individual level when dealing with specific tasks.

8.1 Introduction

In German, despite its rich morphological case system for marking grammatical function, the surface form of noun phrases (NPs) can be ambiguous. For instance, for NPs[38] involving feminine and neuter nouns, respectively, the surface form is identical in nominative and accusative case. Case is marked on the determiner: *die* for feminine NPs and *das* for neuter NPs both in nominative as well as in accusative case, respectively. Such NPs are case-ambiguous. Furthermore, German allows for a relatively free word order: In addition to subject-verb-object (SVO) sentences, the non-canonical word order of object-verb-subject (OVS) is also possible. Thus, if the determiner *die* or *das* is part of an NP at the beginning

[38] We do not distinguish between determiner phrase (DP) and noun phrase (NP) in this work.

of a sentence, the syntactic function of that NP as well as its thematic role remains open: it is ambiguous between subject and direct object as well as between agent and patient. Therefore, the word order configuration could potentially be both, SVO or OVS. If the ambiguity gets resolved at later points in the sentence (e. g., by a case-marked post-verbal NP or by verb inflection), the sentence is called temporarily or locally ambiguous (see (34) and (35)). Besides morphological case markers, prosody, verb semantics, and (visual) context can resolve or influence such thematic role-assignment ambiguities. Sentences, in which NP2 is also case-ambiguous between nominative and accusative case are globally ambiguous (36). Although grammatically possible, OVS sentences in German are rather infrequent (< 4% in several corpora on written sentences), with an even lower rate of OVS sentences with an accusative object (< 1%) than OVS sentences with a dative object (Bader & Häussler 2010). Possibly, the larger number of dative objects is influenced by passivised ditransitive verbs that license object-first word order (Bader & Häussler 2010).

(34) Das Kamel tritt nun den Tiger. (locally ambiguous: SVO)
the.NOM/ACC camel kicks now the.ACC tiger
'The camel now kicks the tiger.'

(35) Das Kamel tritt nun der Tiger. (locally ambiguous: OVS)
the.NOM/ACC camel kicks now the.NOM tiger
'The tiger now kicks the camel.'

(36) Das Kamel tritt nun die Gazelle. (globally ambiguous)
the.NOM/ACC camel kicks now the.NOM/ACC gazelle
'The camel now kicks the gazelle/the gazelle now kicks the camel.'

For comprehension of spoken locally ambiguous sentences, studies have reported a strong SVO bias in German (Hanne et al. 2015, Henry et al. 2017): Listeners expect NP1 to be the subject, thus agent, and NP2 to be the object, thus patient, of the sentence. These studies explored language comprehension using the visual-world paradigm (Tanenhaus et al. 1995, Allopenna et al. 1998, and a review by Huettig et al. 2011), where participants' eye movements are recorded while participants see objects or a scene on the screen and are presented with an auditory stimulus (e. g., an instruction such as "Click on the candle" or a sentence such as *Das Kamel tritt nun den Tiger.*). The eye movements of a listener are systematically related to their linguistic processing (Allopenna et al. 1998). If hearing a sentence beginning such as *The bird eats*, listeners were shown to look to a potential patient on the screen, for instance a worm, and that, in German, eye movements can be modulated by case-marking of NP1, marking it as either agent or patient of the sentence (Kamide et al. 2003). Eye movements, thus, reflect the expected continuation of the sentence in such a setting[39].

[39] There is no universally applicable interpretation of eye movements as they depend, for example, on the type of the task, as discussed in Huettig et al. 2011.

Using eye tracking in the visual-world paradigm, Weber et al. (2006) showed that prosody can weaken the SVO bias in comprehension. In the study by Weber et al. (2006), participants were auditorily presented with locally ambiguous sentences like “Die.nom/acc Katze jagt womöglich den.acc Vogel.” (‘The.nom/acc cat chases possibly the.acc bird.’) and “Die.nom/acc Katze jagt womöglich der.nom Hund.” (‘The.nom/acc cat chases possibly the.nom dog.’) together with a visual scene, depicting a cat, a dog, a bird, and an unrelated object. For the SVO sentence, the bird is the target and the dog is the foil picture, for the OVS sentence, the dog is the target and the bird the foil picture. SVO sentences were presented with a rising L*+H pitch accent (late-peak) on NP1 and a high H* nuclear pitch accent on the verb, while OVS sentences had a rising L+H* nuclear pitch accent (medial peak) on NP1; both sentences ended in a low L-% boundary tone (the accents are coded following the GToBI model following Grice & Baumann 2002, Grice et al. 2005). As the default position for the nuclear accent in West-Germanic languages is the last argument of the verb (Ladd 2008), in this case NP2, both sentences had a non-default nuclear accent placement (Weber et al. 2006). Listeners interpreted case-ambiguous NP1s more often as subject when the sentence beginning carried an L*+H pitch accent on NP1 and an H* pitch accent on the verb. No such SVO preference was found in sentences with L+H* pitch accent on NP1. However, no intonation condition resulted in a preference for OVS compared to SVO sentence reading. Nevertheless, the results suggest, that intonation plays a role in early disambiguation.

In a similar vein, also using the visual-world paradigm, however for case-unambiguous SVO and OVS sentences, Henry et al. (2017) reported an additive use of morphological and prosodic cues. They presented participants with sentences in four conditions: (i) a case-only SVO sentence “Der.nom Hahn frisst gleich die.nom/acc Blume” (The.nom chicken eats soon the.nom/acc flower.), (ii) the case-marked SVO sentence with an H+L* nuclear pitch accent on NP2 and a non-specified pitch movement on NP1, (iii) a case-only OVS sentence “Den.acc Hahn frisst gleich der.nom Fuchs.” (The.acc chicken eats soon the.nom fox. ‘The fox will soon eat the chicken.’), and (iv) the case-marked OVS sentence with an L+H* pitch accent on NP1 (Henry et al. 2017: 5). All sentences ended in an L-% boundary tone. The visual scene, similar to Weber et al. (2006), depicted NP1, a potential agent, a potential patient, and a distracting unrelated object. In their eye tracking experiment, the presence of both, morphological and prosodic cues, facilitated prediction of the upcoming structure observed by an increase in speed and accuracy in looks to a target object.

In a comparable study, 5-year-old German-learning children have been shown to rely on prosody to determine thematic roles in ambiguous sentences (Grünloh et al. 2011). The material contrasted case-marked OVS sentences (e. g., “Den.acc Hund *VERB* der.nom Löwe.” The.acc dog VERB the.nom lion. ‘The lion is VERBING the dog.’) with globally ambiguous sentences (e. g., “Die.nom/acc Katze *VERB* die.nom/acc Kuh.” ‘The.nom/acc cat is VERBING the.nom/acc cow.’). As verbs they used novel verbs describing transitive actions. Similar to the study by Henry et al. (2017), four conditions were created: case-marked OVS sentences and case-ambiguous sentences each with prosodic marking and in a de-accented version. Prosodic marking is described as "Contrastive Intonation condition" with a "strong, rising L+H* pitch accent on the first noun phrase" (Grünloh et al. 2011: 399), thus, in parallel to the pitch accents described for OVS sentences in the studies by Weber et al. (2006) and Henry et al. (2017). The task was to decide between two small video sequences showing two animals enacting one of the four previously introduced transitive actions that only differed in which of the two animals has the agent and which the patient role. Similar to an adult control group, children used the pitch accent on NP1 as cue for a patient-first sentence, thus overriding the SVO bias. In contrast to the adults, children needed prosody in combination with case marking to correctly choose the video depicting the patient-first role order (OVS). Both age groups judged the ambiguous sentences in $>$ 90% of the responses as SVO, irrespective of the presence of the L+H* pitch accent. In a second study, each trial was preceded by a discourse context naming a wrong patient in an SVO sentence and embedding the target sentence in a correcting answer (e. g., context: "Der.NOM Löwe *VERB* den.ACC Frosch! corrective answer: Nicht den.ACC Frosch *VERB* der.NOM Löwe, sondern **den.acc Hund *VERB* der.nom Löwe**." 'The.NOM lion is VERBING the.ACC frog. Not the.ACC frog that is.VERBING the.NOM lion, but the.ACC dog is VERBING the.NOM lion.') (Grünloh et al. 2011: 408f.). Together with a context, children's pointings to patient-first scenes increased, especially in the presence of contrastive intonation. Grünloh et al. (2011) conclude, that "prosody has the power to work against this word order bias and that the information in the sound stream seems to be sufficiently rich to allow children to abstract participant roles" (Grünloh et al. 2011: 415).

In summary, studies on language comprehension showed that listeners can make use of prosodic cues (e. g., manipulated f0-contours) for thematic role-assignment (which can serve for disambiguation of locally ambiguous sentences) even before unambiguous morphological cues are accessible (Weber et al. 2006). Moreover, prosodic cues increased speed and accuracy of prediction in language comprehension in the presence of morphological cues (Henry et al. 2017).

For production, Weber et al. (2006) reported variability in the f0-contours produced on locally ambiguous OVS sentences with intonation phrase breaks and silent intervals following NP1. It remains open, however, whether there is a clearly differential pattern in f0-contours for SVO and OVS sentences within and across participants or whether the prosodic marking of such sentences is subject to individual variability. In the present study, we analysed productions of locally ambiguous SVO and OVS sentences with a main focus on f0 along with silent intervals. Our research questions read as follows:

RQ1: How do speakers use prosodic cues (f0, silent intervals) when asked to provide disambiguation as early as possible in productions of locally ambiguous sentences?

RQ2: Do we find differences in the f0-contours in the ambiguous part of the sentence (on NP1 and the following verb) between SVO and OVS sentences within and across speakers?

We focused our analysis on the f0-trajectory in the ambiguous part of the sentences, as comprehension studies have manipulated f0 in this region. We expected large inter-individual variability in the productions, since the task was rather specific and variability, especially in productions of OVS structures, has been reported previously (e. g., Weber et al. 2006).

8.2 Methods and procedure

Participants

Sixteen native speakers of German (12 identified as female, 4 identified as male; mean age 24 years, *SD*: 3.1, age range: 20–30 years) were included in the production study. Two additional speakers took part in the study, but were discarded due to artefacts in the recordings. All participants received course credits or monetary reimbursement for their participation and were naïve to the purpose of the study. The procedure of this study has been approved by the Ethics Committee of the University of Potsdam (approval number 72/2016) and each speaker gave written consent to participate. All speakers had normal hearing defined as an average pure-tone audiometry of 25 dB HL or better for 500, 1000, 2000, and 4000 Hz in the better ear assessed using an audiometer (Hortmann DA 324 series) and calculated following the classification of hearing impairment by the WHO as reported in Olusanya et al. (2019).

Materials

Stimuli Items consisted of locally ambiguous and semantically reversible German verb-second main clauses (see (34) and (35)) and corresponding black-and-white line-drawings depicting agent, patient, and the action (see Figure 33). The final stimulus material consisted of 20 different transitive verbs and each verb appeared in two word order conditions: (34) with subject-verb-object word order (SVO) and (35) with object-verb-subject word order (OVS). Thus, a total of 40 items was used. In order to control for semantic reversibility and sentence-picture correspondence, the stimulus material was created and validated in three steps involving two pre-studies, which are described in the following. The pre-studies included a larger number of sentences, in order to be able to select the ones with the highest ratings. Further the pre-studies included globally ambiguous sentences.

In the **first step** of the pre-studies for the creation of the stimuli, we constructed possible sentences using 29 transitive German verbs and two noun phrases (NP). We constructed semantically reversible sentences in SVO and OVS word order that were either locally (34, 35) or globally (36) ambiguous (see (37) for an example of an unambiguous SVO sentence). All sentences had the same structure: NP1.sg.neut + verb.3.p.sg.pres + adverb *nun* (Engl. 'now') + NP2.sg.masc.

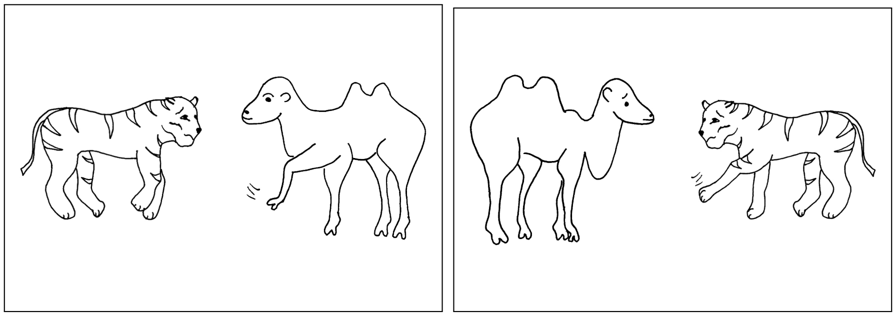


Figure 33: Example of black-and-white line-drawings used as visual stimuli for SVO (left) and OVS (right) word order. The corresponding SVO sentence is *Das.*NOM/ACC *Kamel tritt den.*ACC *Tiger.* (the.NOM/ACC camel kicks the.ACC tiger.), the corresponding OVS sentence is *Das.*NOM/ACC *Kamel tritt der.*NOM *Tiger.* (the.NOM/ACC camel kicks the.NOM tiger.).

(37) Der Elefant tritt nun den Tiger. (unambiguous: SVO)
the.NOM elephant kicks now the.ACC tiger
'The elephant now kicks the tiger.'

Animate nouns of three different categories were used as NP1 and NP2: persons/humans, animals, and fairy tale characters. Only nouns of the same category were used within the same sentence. For the locally ambiguous sentences, a neuter NP1 and a masculine NP2 were combined, for the globally ambiguous sentences, a feminine NP was used instead of the masculine one. The neuter determiner *das* and the feminine determiner *die* are case-ambiguous between nominative and accusative case while the masculine determiner is case-unambiguous with *der* in nominative and *den* in accusative case. The masculine determiner of NP2 in the locally ambiguous sentences, thus, morpho-syntactically disambiguated the syntactic and thematic roles of both NPs as either the subject/agent or the object/patient of the action. No such disambiguation is provided by the feminine determiners of NP2 and the disambiguity between syntactic and thematic roles of both NPs is maintained until the end of the sentence. For each verb and ambiguity condition, four sentences were created with at least one sentence of each noun category. The nouns in both positions (NP1 and NP2) were balanced for their mean frequency, their number of phonemes, and number of syllables in dlexDB (Heister et al. 2011). The nouns in NP1 can be divided into three categories on the basis of syllable structure and stress pattern: (i) monosyllabic nouns (e. g., HUHN, Engl. 'hen'), (ii) disyllabic nouns with penultimate stress (e. g., MAEDchen, Engl. 'girl', capital letters correspond to stressed syllable), and (iii) disyllabic nouns with ultimate stress (e. g., phanTOM, Engl. 'phantom'). A total of 232 sentences was created (116 locally ambiguous sentences: 29 verbs $*$ 4 versions of a sentence + 116 globally ambiguous sentences: 29 verbs $*$ 4 versions of a sentence).

In the **second step**, the semantic reversibility of the thematic roles in the sentences, that is, an equal likelihood for both nouns of the sentence to take the role of agent or patient of the action, was assessed in a rating study on the reversibility of the sentences. The rating study encompassed the 116 locally ambiguous and the 116 globally ambiguous sentences and included 72 German native speakers (59 identified as female, 13 identified as male; mean age: 22 years (*SD*: 5.01), age range: 17–49 years)[40]. In this section, we mainly focus on the locally ambiguous sentences. However, as the selection process of the stimuli for the production study contained also globally ambiguous sentences, they are included for the sake of completeness. The rating study was conducted in a group session in a pen-and-paper version. The task was to rate the plausibility of the subject of the sentence to be the agent of the action in that particular sentence. The ratings were given on a five-point Likert scale ranging from "not plausible at all" *1* to "rather plausible" *3* to "very plausible" *5*. All sentences were either locally (34, 35) or globally (36) ambiguous. Sentences were distributed across two lists. In list version (a), the globally ambiguous sentences had the feminine noun in the first position and the locally ambiguous sentences started with the neuter noun. In list version (b), the globally ambiguous sentences started with the neuter noun and the other sentences with the masculine noun (not being locally ambiguous anymore). Thus, there were two versions of each sentence, one with each noun in the first position. Each list contained 232 potential stimulus sentences and 25 filler sentences with implausible agent-patient relation (e.g., *The grandpa now baptises the phantom.*). The items in the lists were pseudo-randomised with no direct repetition of the same nouns and verbs in consecutive trials and not more than four sentences in a row starting with the same determiner.

In the analysis, we assessed the pairwise difference between ratings of version (a) and version (b) of each sentence. If the comparison revealed no statistically significant difference, a sentence pair was counted as semantically reversible and kept for further steps. In an additional step, we controlled for the appearance of the same NP across sentences and reduced the maximum appearance to four times. This process resulted in 24 sentence pairs, for which black-and-white line-drawings were created, one per thematic role assignment (henceforth referred to as pictures, see Figure 33).

In the **third step**, we assessed the comprehension agreement for the pictures in a sentence-picture matching task. This procedure was similar to the one described in previous studies (e. g., Hanne et al. 2015). The sentence-picture matching task was realised with 41 participants (36 identified as female, 5 identified as male; mean age: 22 years (*SD*: 3.8), age range: 18 – 31 years). The study was conducted in a group session in a pen-and-paper version. In each trial, the target and the foil picture (in the foil picture, thematic roles were reversed) were presented visually (see Figure 33) and either the SVO version or the OVS version of the sentence was read aloud by the experimenter with a neutral prosody (i. e., not distinguishing between SVO and OVS sentences). For the globally ambiguous sentences, both readings are string-identical. In order to evoke the reading of NP2 as agent, we used a

[40] Another four participants took part in the study but had to be excluded from the analysis due to missing values (n = 2) and German as non-native language (n = 2).

sentence with a passive construction (38).

(38)	Das	Kamel	wird	von	der	Gazelle	getreten.	(passive)
	the.NOM/ACC	camel	is	by	the.DAT	gazelle	kicked	

‘The camel is being kicked by the gazelle.’

The task was to mark the corresponding picture on an answer sheet. There were 48 items (24 in each ambiguity condition). Items were pseudo-randomised with at least five items between the presentation of the same verb, and at least three items before the repetition of the same noun in NP1.

In the analysis, the agreement between SVO and OVS version of each sentence and the respective picture was checked. One item was excluded due to imbalance in NP1 between locally and globally ambiguous sentences (*treten*) and three items were turned into practice items (*schieben, stechen, ziehen*). The final set consisted of 20 test item pairs and three practice item pairs in each ambiguity condition. NP1 and verb were parallelised across locally and globally ambiguous sentences.

Contexts For conducting the production study of the locally ambiguous sentences in a semi-interactive setting, two different interlocutors (referred to as contexts) were created that were virtually present during the task: a young and an elderly female native speaker of German. Speakers got an audio-visual impression of their interlocutors from two short videos each presented on a screen. The first video contained a personal introduction of the interlocutor (see Table 23 for a summary of the information provided by each interlocutor) and the second video contained instructions for the task. The two contexts were part of five contexts used in production studies to elicit coordinates (studies I and II).

Table 23: Fictional names, ages, origins, and further information of the interlocutors present in the two contexts.

	YOUNG *(baseline)*	ELDERLY
Name:	Hannah	Maria Korbmacher
Age (in years):	24	82
Origin:	Eberswalde	NA
Residence:	Potsdam	Potsdam
Occupation:	Biology student	Retired school teacher
Further information:	Moved to Potsdam for her studies, lives in a shared flat, likes the parks in Potsdam	Lives for two years in an old-age home with her husband, tends to forget things from time to time

Experimental procedure

Productions were elicited by means of a referential communication task. Speakers were seated in front of a wide screen (resolution 1920 x 1200) in a sound-attenuated recording booth at the University of Potsdam wearing an AKG HSC 271 headset with over-ear headphones and a condenser microphone. On the screen, speakers were presented first with an audio-visual impression of the corresponding interlocutor in two short videos (one video with a personal introduction of the interlocutor and another one with instructions for the task). The young interlocutor (baseline) was always presented first and the elderly interlocutor in a second block. Each trial started with a fixation cross on the screen and was replaced after 1000 ms by the target and the foil picture with the corresponding sentences printed below the pictures in text font Arial and font size 30. After a preview time of 4000 ms, the target picture was highlighted with a green frame along with the auditory presentation of the question *Was sehen Sie?* (‘What do you see?’) via headphones. The question was asked by the interlocutor and was intended to trigger the production (i. e., to simulate question-answer dyads) and as a reminder of the current interlocutor’s identity. The task was to produce the target sentence in a way that would allow the interlocutor to identify the target picture “as rapidly and accurately as possible”; the speakers were told that the interlocutor would see the same pictures (without the sentences) and had to find the matching one. The experiment started with a practice phase including six trials followed by the test phase. Productions were recorded via an Alesis iO|2 interface (at 48 kHz). The experiment was run from a Dell laptop using the software Presentation (Neurobehavioural Systems). Speakers produced each item twice, once addressing the young and once the elderly interlocutor.

Data treatment

In total, 1280 sentences were produced (20 verbs $*$ 2 word order conditions $*$ 2 interlocutor contexts $*$ 16 speakers). In the present analysis, we focused on the productions in the baseline context (n = 640). In case the productions contained hesitations (i. e., slips of the tongue, restarts), the part with the hesitation was cut out if the remaining part constituted a fluent utterance (n = 6). For three productions (0.5%) the hesitations could not be cut out and the complete productions were excluded from the subsequent analyses of f0.

In the remaining 637 productions, constituent boundaries (NP1, verb + adverb, NP2) were segmented in Praat (Boersma & Weenink 2017) following standard segmentation criteria (Turk et al. 2006). Silent intervals preceding stops were segmented as part of the following constituent and labelled as closure, while silent intervals preceding non-stops were segmented between constituents and labelled as pauses. As NP2 always started with the lenis plosive /d/, we could not reliably distinguish between a pause and the silence of the closure at this position (see Lehiste 1973b: 1230 on notes on the impossibility of determining the duration of initial voiceless plosives). The same problem applies partly to the location preceding the verb, as six of the verbs began with a plosive. F0-values were extracted with a customised praat script combining the procedures of Mausmooth (Cangemi 2015) and ProsodyPro (Xu 2013). For each soundfile, a pitch object was created and unreliable pitch points were removed manually. The pitch contour was smoothed with a bandwidth of 10 Hz before interpolation of pitch points and with a bandwidth of 15 Hz afterwards following the procedure in Cangemi (2015). Twenty f0-values per constituent (only the intervals that contained speech were considered) were extracted and converted into semitones (st) relative to 1 Hz following Hazan et al. (2016) in order to ease comparison independent of pitch height. The smoothed f0-contours were plotted with centering at the onset of the verb separately for each speaker and item (SVO and OVS production on top of each other). In a separate step, interval durations were extracted automatically using a Praat script. We collapsed durations of intervals labelled as pauses and closures. The durations of the silent intervals preceding the verb (following NP1) and preceding NP2 (following *nun*) were plotted separately for speaker and word order condition in violin plots (Figure 37). F0-contours and silent intervals were analysed separately.

8.3 Analyses and results

In accordance to our rather open research questions, we chose an exploratory analysis procedure consisting of a combination of visual and statistical inspection of the data in several steps. All data and code are available on OSF: https://osf.io/gychu/.

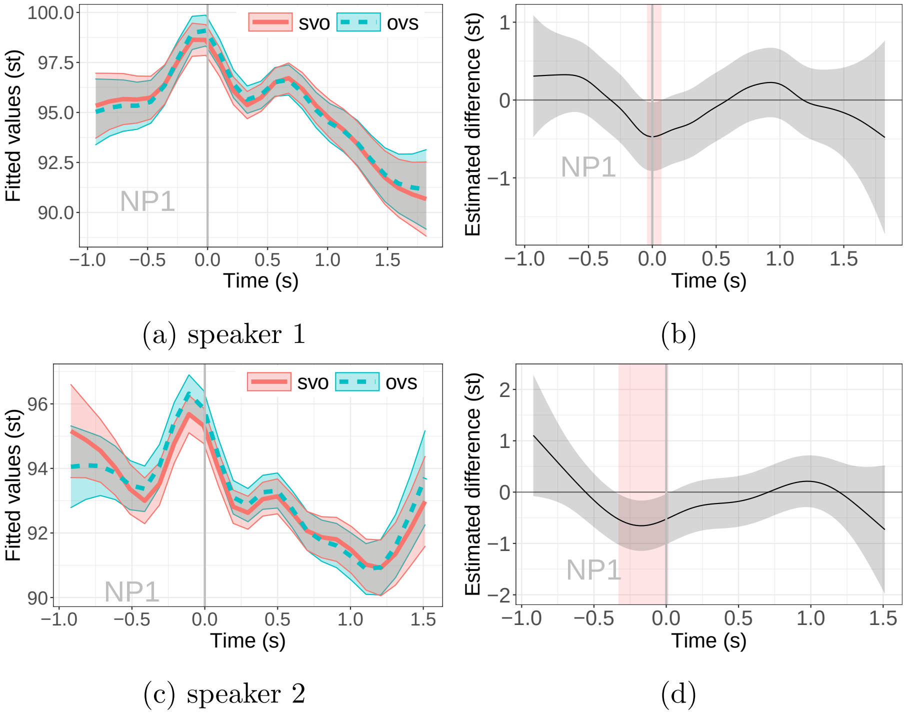
We started with a visual inspection of the f0-contours within speakers. For most speakers and items, f0-contours of SVO and OVS sentences overlapped quite neatly, thus, showing no visible difference between word order conditions in the ambiguous part of the sentence. Yet, for some speakers, consistent differences in the f0-contours between conditions were noticeable. Across speakers, different f0-contours were produced.

To statistically corroborate the observations from visual analysis in f0-contours between SVO and OVS sentences, we fitted Generalised Additive Mixed Models (GAMMs, Wood 2017, Baayen et al. 2017, Sóskuthy 2017, Wieling 2018) in R (R Development Core Team 2018) using the R package *mgcv* (Wood 2017, 2011). GAMMs allow to model time-varying data with non-linear patterns controlling for random-effects and autocorrelation (Baayen et al. 2017, Wieling 2018, Chuang et al. 2020, Sóskuthy 2021) and have been successfully used in previous analyses of f0-contours (Chuang et al. 2020, Zahner et al. 2020, Sóskuthy 2021).

Since we were interested in distinctive f0-contours for word order condition within speakers, we decided to fit an individual model for each speaker. The f0 time series (in st, 60 measurements per production, i. e., 20 per constituent) were entered as response variable to the model and word order (as ordered factor) as a predictor (parametric difference term). To directly test whether the difference between SVO and OVS sentences is significant, we added a reference and a difference smooth separately to the model (s(time, bs = “tp”, k = 10) + s(time, by = condition.ord, bs = “tp”, k = 10)) (Sóskuthy 2017, Wieling 2018, Sóskuthy 2021). We compared the model to a nested model without the difference terms using the function compareML() in the R package *itsadug* (van Rij et al. 2020) to check whether their inclusion was justified. Model complexity was increased stepwise with model comparisons at each step. The number of basis functions was doubled from k = 10 (default) to 20 if the function gam.check() revealed a low *p*-value and a k-index <1. The final models all included an AR(1) error model to correct for autocorrelation in the residuals (cf. Wieling 2018, Chuang et al. 2020) and a random smooth for item (s(time, item, bs = “fs”, m = 1)).

For significance testing and interpretation of the effect of word order condition on the produced f0-contours, we checked the summary statistics and the difference curves plotted with the plot_diff() function of the R package *itsadug* (van Rij et al. 2020). Visualisation is crucial for interpreting results of GAMMs (Wieling 2018, Sóskuthy 2021). The difference curve visualises the comparison between the non-linear smooths of the two condition levels (here SVO minus OVS) with a pointwise 95% confidence interval (CI). Values above zero indicate larger f0-values in SVO and values below zero indicate larger f0-values in OVS. The difference between the two conditions is significant if the pointwise 95% CI of the difference curve does not include zero. Across all items and speakers, NP2 started on average 0.66 s (*SD*: 0.16) after verb onset. We, therefore, consider the time window between the onset of NP1 and 0.66 s (onset of NP2) as a rough approximation of the ambiguous part of the sentence. Note, that for some areas of the utterance, the fitted values are less reliable. This is the case at the beginning and at the end of the utterance (productions differed in their duration, leading to fewer and less reliable f0-measures at the outer extremes; utterance-final glottalisation) and at the onset of the verb (as silent intervals labeled as pauses were dismissed in the extraction of f0-values and the interpolated contour might be disturbed). We are aware that interpretation needs caution as the report of significant intervals of any minimal duration (Sóskuthy 2021) and the decision where to look for differences strongly influence the results (Roettger 2019).

Across speakers, the difference curves of individual speakers showed a diverse pattern. Overall, for ten speakers, the respective models revealed significant differences between SVO and OVS word order within the time window of syntactic ambiguity (i. e., preceding NP2, cf. panels on the right in Figure 34). Note that these differences do not exclude possible further differences during NP2 (as can be seen in the panels (34f), (34j), (34n), (34p), and (34t) in Figure 34). For other two speakers, the estimated differences were located more than 0.66 s after the onset of the verb, thus during the case-unambiguous NP2 (cf. panels at the right in Figure 35) and, for the remaining four speakers, no significant differences were attested throughout the whole utterance (cf. panels on the right in Figure 36). In Figure 36, the pointwise 95% CI includes zero (on the y-axis) throughout the complete utterances, indicating, that the non-linear smooths of the two conditions (panels on the left in Figure 36) do not differ significantly.



(a) speaker 1

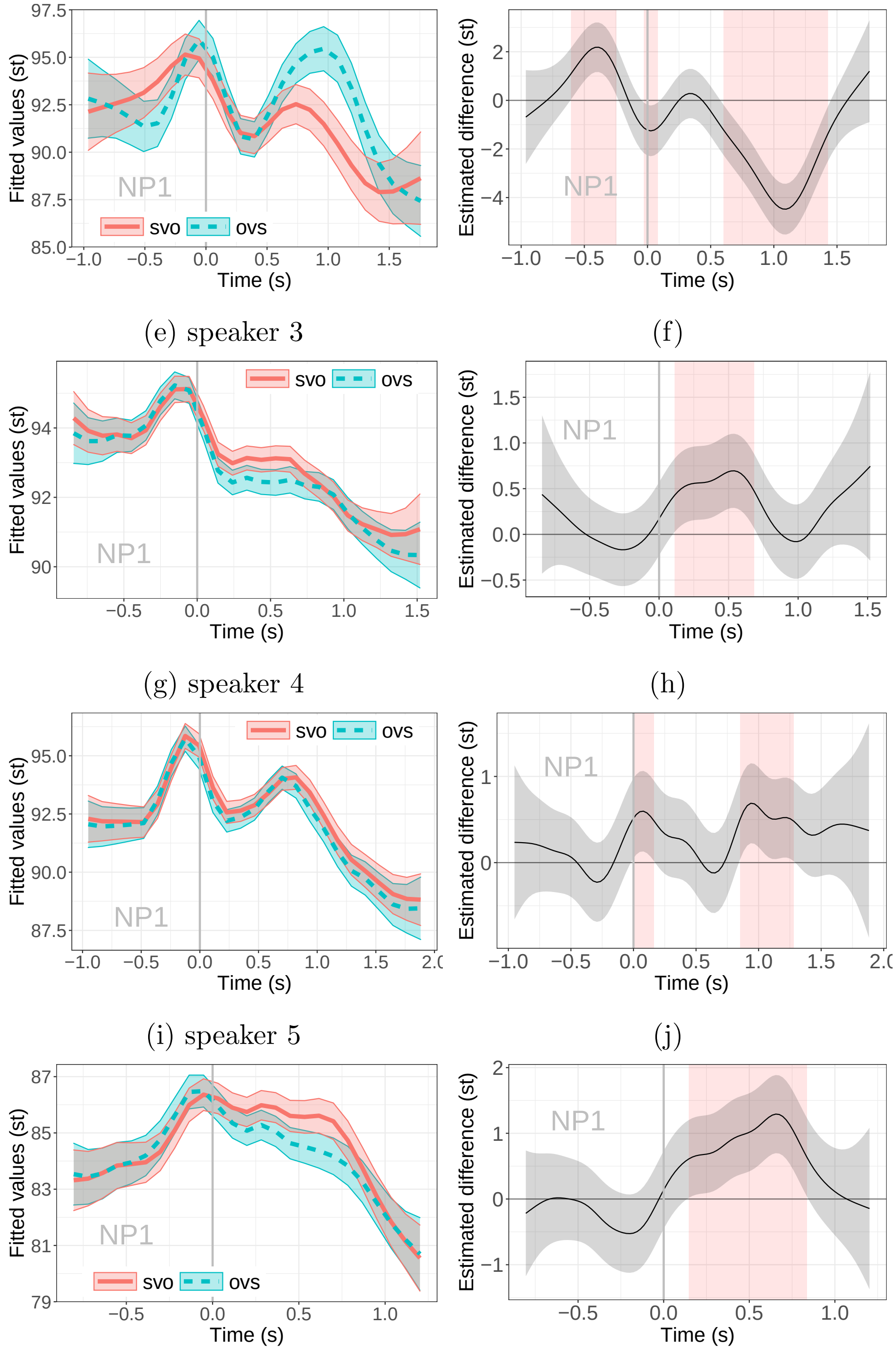
(b)

(c) speaker 2

(d)

Figure 34: Predicted difference in ambiguous part: f0-values (st) predicted by the GAMMs for SVO (red solid line) and OVS (blue dashed line) (left) and predicted differences (SVO-OVS, right) for ten speakers (rows). Time (s) is centered to the offset of NP1/onset of the verb (vertical line). Shaded bands indicate pointwise 95% confidence intervals (CI). The difference is significant if the CI excludes zero (indicated by red shading). Figure continues on next pages.

A closer look at the regions of significant difference between f0-contours preceding NP2 in Figure 34 revealed variability across speakers as to location and duration of the differences predicted by the model (shaded areas), including locations we previously discussed as less reliable: for instance the very beginning of the utterance (n = 2, cf. (34r) and (34t) in Figure 34) and relatively local around the onset of the verb (n = 2, cf. (34b) and (34j) in Figure 34). The two speakers (1 and 5) for whom the model predicted differences around the onset of the verb, produced silent intervals between NP1 and the verb (cf. Figure 37), which makes the model predictions less reliable at this location, as the durations of silent intervals were excluded from the time-normalised f0 data and the f0-contours were interpolated. The same applies to speaker 3 (34f in Figure 34), however, we would not discard this speaker at this point of the analysis, as the model predicted also a difference between the curves within the time domain of NP1. For other speakers in Figure 34, NP1 was not included in the significant time window, as contours diverged only after the offset of NP1 within the verb (n = 1, cf. (34h) in Figure 34) and divergence even extended into NP2 (n = 1, cf. (34l) in



(e) speaker 3

(f)

(g) speaker 4

(h)

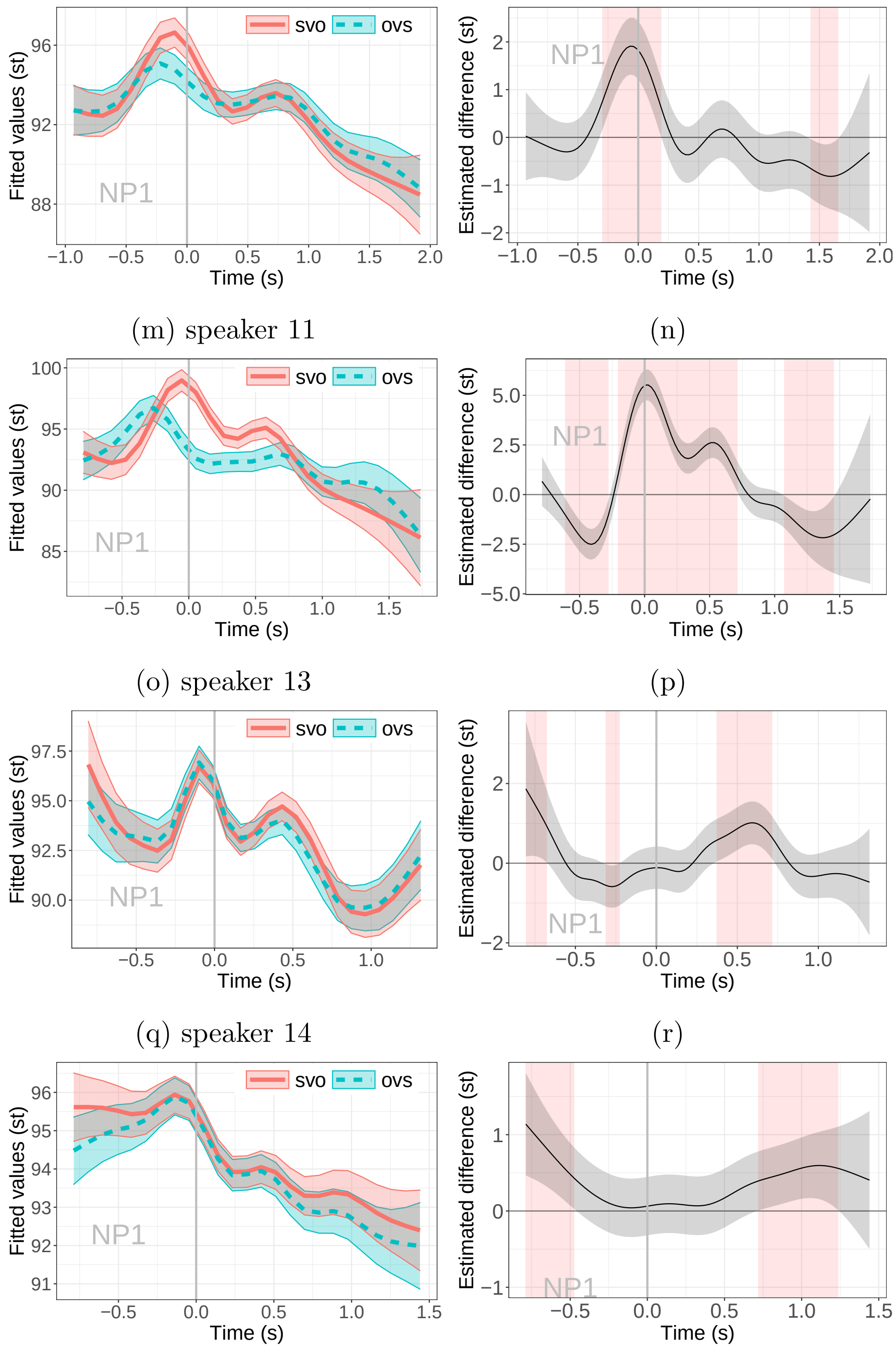
(i) speaker 5

(j)

(k) speaker 9

(l)

Figure 34: continued Figure: Predicted difference in ambiguous part. Complete caption is given on previous page. Figure continues on next page.



(m) speaker 11

(n)

(o) speaker 13

(p)

(q) speaker 14

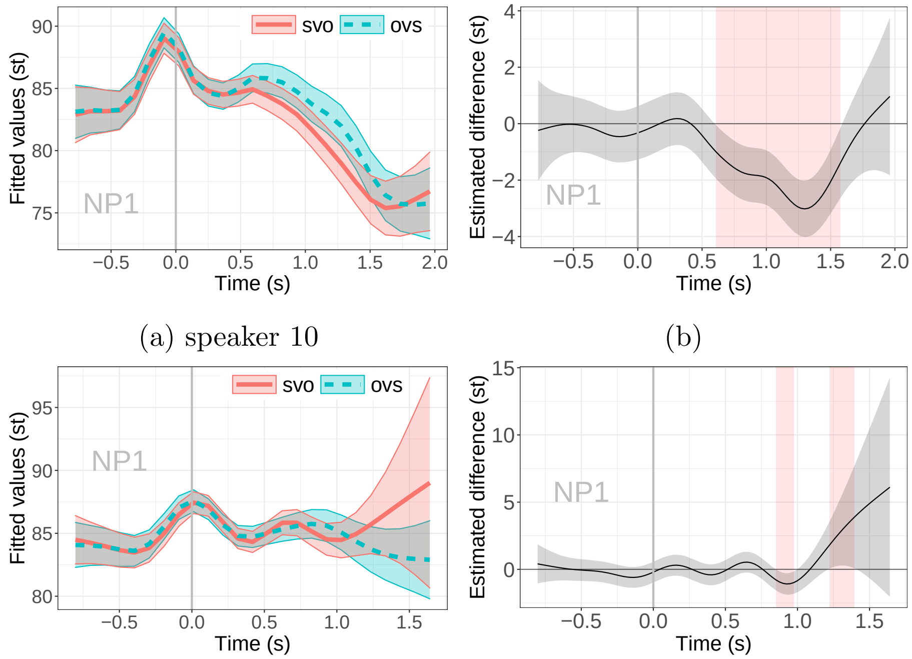
(r)

(s) speaker 17

(t)

Figure 34: continued Figure: Predicted difference in ambiguous part. Complete caption is given on first page of the figure.

Figure 34).



(a) speaker 10

(b)

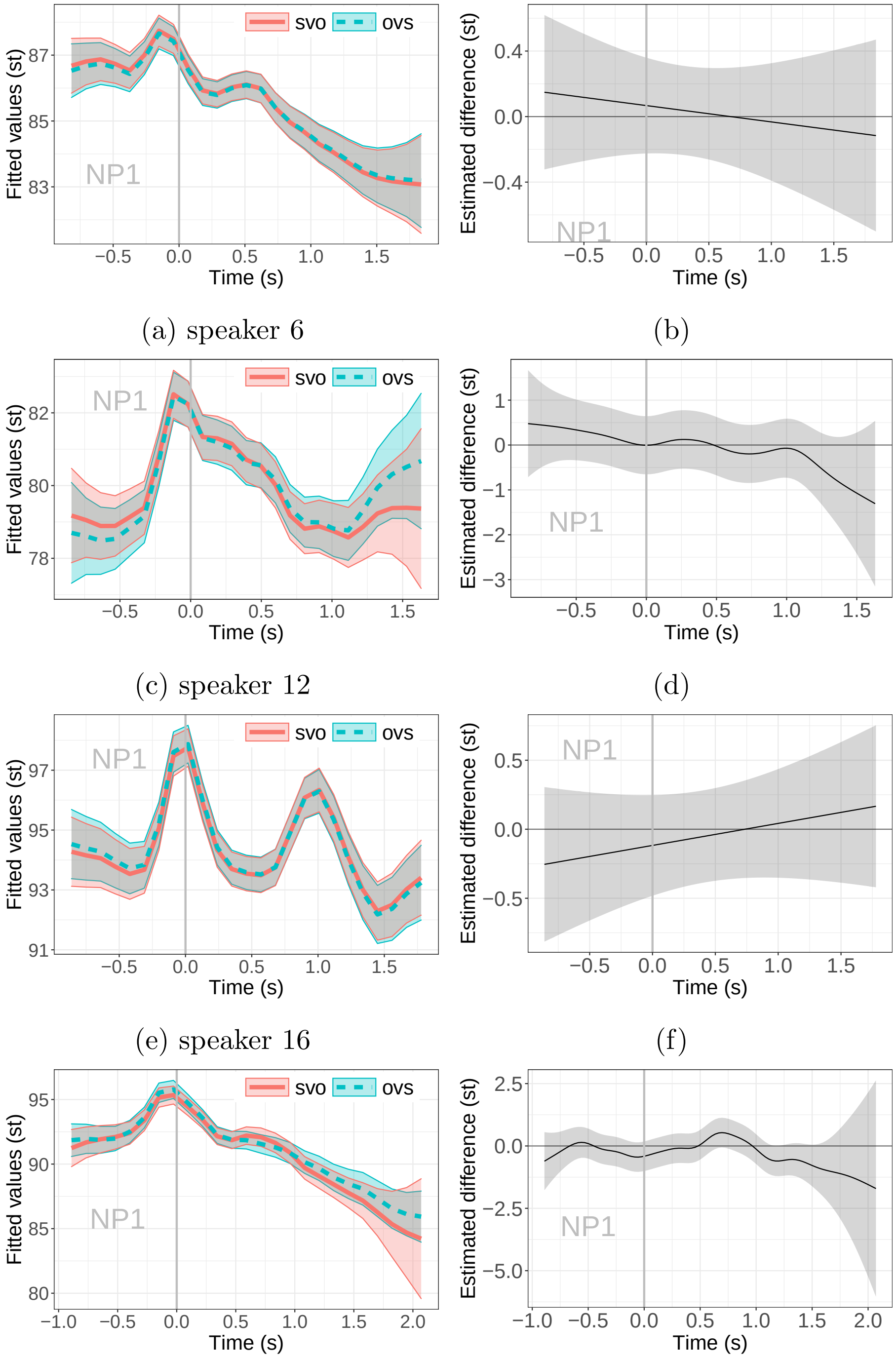
(c) speaker 15

(d)

Figure 35: Predicted difference in unambiguous part: f0-values (st) predicted by the GAMMs for SVO (red solid line) and OVS (blue dashed line) (left) and predicted differences (SVO-OVS, right) for two speakers (rows). Time (s) is centered to the offset of NP1/onset of the verb (vertical line). Shaded bands indicate pointwise 95% confidence intervals (CI). The difference is significant if the CI excludes zero (indicated by red shading).

Overall, the time windows of the differences predicted by the model had durations between one hundred and more than nine hundred milliseconds. The difference between conditions was mainly (not exclusively) positive, indicating larger f0-values in SVO than in OVS sentences. A positive difference between SVO and OVS during NP1 was for instance predicted for speakers 3, 11, and 13, while a negative difference was predicted for speaker 2. For speaker 13, the model predicted a negative difference between SVO and OVS sentences within NP1 followed by a positive difference towards the end of NP1 and the beginning of the verb. The f0-trajectories predicted for SVO and OVS (cf. (34o) in Figure 34) diverge not only in peak height but also with regard to the contour. Moreover, the parts where differences in f0-trajectories are predicted extend over 330 ms within NP1 and over 900 ms including the part of the verb. We run a post-hoc analysis on the data of this individual speaker that will be described in a separate section.

With respect to the f0-trajectories on the verb (corresponding to the time domain between 0 and approximately 0.66 on the x-axis in the Figures 34 and 35), the data show variability between individuals in the f0-contour predicted by the GAMMs (left panels in Figures 34



(a) speaker 6

(b)

(c) speaker 12

(d)

(e) speaker 16

(f)

(g) speaker 18

(h)

Figure 36: No predicted difference: f0-values (st) predicted by the GAMMs for SVO (red solid line) and OVS (blue dashed line) (left) and predicted differences (SVO-OVS, right) for four speakers (rows). Time (s) is centered to the offset of NP1/onset of the verb (vertical line). Shaded bands indicate pointwise 95% confidence intervals (CI). The difference is significant if the CI excludes zero.

and 35) but consistency in the direction of the difference predicted between word order conditions: For all speakers the model predicted a positive difference (i. e., higher f0-values in SVO compared to OVS sentences, cf. (34h), (34l), (34p), (34r)). The time point of 0.66 s after verb onset is just a rough approximation as verbs differed in number of syllables and segments and speakers differed in articulation rate. For some of the speakers, the difference predicted between SVO and OVS sentences continues into the time domain of NP2. This is particularly striking in the case of speaker 9 (cf. (34l)), as the two f0-trajectories are predicted to differ for almost 700 ms. For speaker 13 and 14 (cf. (34p), and (34r), respectively) the difference is predicted for more than 917 ms and 342 ms, respectively.

For the sake of completeness, we also mention the cases for which the f0-trajectories were predicted to diverge after the approximate verb offset at 0.66 s and thus after the disambiguation through the case-unambiguous NP2. In this last part of the sentences, positive as well as negative differences were predicted. For two speakers, positive differences were predicted, thus, higher f0-values for SVO compared to OVS sentences (cf. (34j) and (34t), while for three speakers, negative difference were predicted, thus, lower f0-values for SVO compared to OVS sentences (cf. (34f), (34p), and (35b)).

Silent intervals

With regard to silent intervals, we also started with a visual inspection of the data within speakers (cf. Figure 37). Overall, durations were rather small, ranging for many speakers between 0 and 100 ms. For comparison, the mean closure duration for the German lenis plosiv /d/ in word-initial position is 63 ms (*SD*: 13.2) according to (Kuzla & Ernestus 2011: 6). As for the f0-contours, the data contained large variability: For ten speakers, we appraised no considerable production of silent intervals neither preceding the verb nor preceding NP2, while three speakers produced silent intervals in one but not the other position, and three speakers produced silent intervals in both positions. The main interest was whether speakers differed the production of silent intervals between word order conditions.

To statistically corroborate the observations from visual analysis, we used a series of intra-individual Wilcoxon Signed Rank Tests (Siegel 1956), since this non-parametric procedure for comparing two groups with non-independent data points can handle non-normally distributed data (which was the case in our data on silent intervals). We used the corresponding base R function to test for statistical significance of the difference in the durations of the silent intervals between SVO and OVS in both positions of interest in the utterance. The results are included in Figure 37 with p-values $> .05$ given as n.s. (non significant).

The tests revealed statistically significant differences between the conditions in both positions of interest for speaker 13 (preceding the verb: $V = 69$, p =0.02; preceding NP2: $V = 2$, $p = 0.03$), and in the position preceding NP2 solely for speaker 3 ($V = 131$, $p = 0.01$) (cf. corresponding facets in Figure 37). For speaker 3, the mean duration of the silent intervals

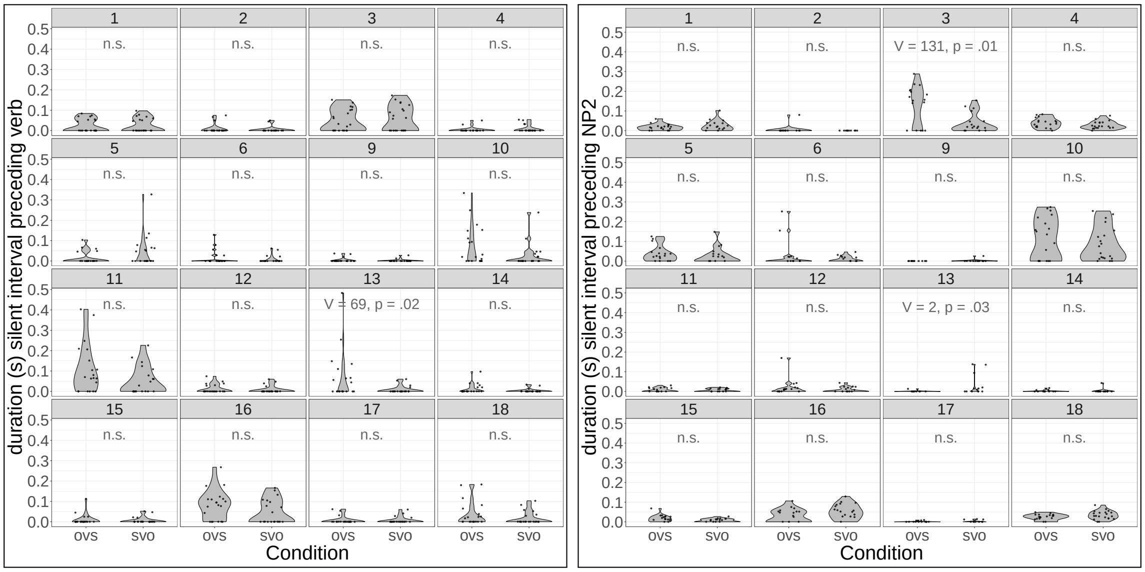


Figure 37: Durations (in s) of silent intervals preceding the verb (left panel) and NP2 (right panel) in individual utterances (dots) with SVO (right) and OVS (left) word order separately for each speaker (facets). Density distribution is given with shaded areas. Results of intra-individual Wilcoxon Signed Rank Tests (n.s. for *p*-values > .05).

preceding NP2 was longer in OVS (0.12 s) than in SVO (0.04 s), while for speaker 13, there were silent intervals in SVO (0.02 s) but none in OVS. Contrary to the silent intervals preceding NP2 for speaker 13, the mean duration of the silent intervals preceding the verb was longer in OVS (0.07 s) than in SVO (0.01 s).

Post-hoc analysis of the f0-contours of an individual speaker

In this section, we give a closer description of the two f0-contours that speaker 13 (20 years, identified as female, born and raised in Brandenburg area) used distinctively on SVO and OVS sentences[41]. The f0-contours underwent several steps of analysis. In a first step, we automatically extracted for each constituent the duration, the minimum and maximum f0-values and calculated time (*risetime*) and slope of the rising f0-movement. In paired t-test comparisons between SVO and OVS productions, risetime and slope revealed statistically significant differences in the productions of this speaker.

For a closer description of the two f0-contours on NP1, the f0-movement was annotated relative to the stressed vowel on NP1 adapting the acoustic annotation in Braun (2006). In comparison to automatically extracted f0-minima and maxima as described in the method section, manual annotations allow to quantify the f0-movement, in our case especially risetime, f0-range, and slope, with respect to segmental landmarks, in our case the stressed

[41] These f0-contours were selected as examples for the recording of stimuli to be used in an eye tracking study.

vowel. We annotated the following landmarks: (i) the start of the vowel in the stressed syllable, (ii) the end of the vowel in the stressed syllable, (iii) the f0-minimum preceding the rise, and (iv) the f0-maximum after the end of the rise. In addition to risetime (in seconds, s), f0-range (in semitones, st: $f0_{range}(st) = 12 * log_2(\frac{f0_{max}(Hz)}{f0_{min}(Hz)})$) and slope (st/s), we calculated the alignment (in s) of the f0-minimum relative to the onset of the stressed vowel (*alignL*, positive values indicate that the f0-minimum was located after the vowel onset, negative values indicate that the f0-minimum was located before the vowel onset) and the alignment (in s) of the f0-maximum relative to the end of the stressed vowel (*alignH*, here, negative values indicate that the f0-maximum was located after the vowel offset and positive values indicate that the f0-maximum was located before the vowel offset)[42].

We assessed the difference between SVO and OVS sentences separately for each variable calculating separate Wilcoxon Signed Rank Tests (Siegel 1956) implemented in base R (R Development Core Team 2018). As described for the silent intervals, this non-parametric procedure for comparing two groups with non-independent data points can be applied on non-normally distributed data (which was the case in our data). The tests revealed statistically significant differences between the conditions for four of the five variables: alignH ($V = 34$, $p = 0.008$), risetime ($V = 177$, $p = 0.008$), f0-range ($V = 209$, p <0.001), and slope ($V = 193$, $p = 0.001$). For alignL, the difference between SVO and OVS was not significant ($V = 107$, $p = 1.0$). Thus, the position of the f0-minimum of the rise on NP1 was similar between SVO and OVS sentences, while the position of the f0-maximum differed between conditions and, hence, also the risetime and the slope. In both conditions, the f0-rise on NP1 had its minimum around 130 ms before the beginning of the stressed vowel (alignL −0.136 s for SVO and −0.132 s for OVS)[43]. In SVO sentences, the f0-peak was located after the end of the stressed vowel (−0.033 s) while in OVS sentences, the f0-peak was preceding the end of the stressed vowel (0.050 s). The f0-rise in SVO compared to OVS sentence was on average 74 ms longer (0.281 s vs. 0.207 s), 4.5 st wider (8.57 st vs. 4.07 st), and 9.99 st/s steeper (30.90 st/s vs. 20.91 st/s). The perceptual impression was a rising stressed syllable on NP1 in the majority of SVO and a high toned stressed syllable with a final fall in OVS sentences. In SVO, the pitch stayed on a high level until the end of NP1 and descended to a lower level only during the following verb while in OVS sentences, the f0-contour fell to a lower level at the end of NP1. On NP2, the speaker mostly produced an early peak on the noun (i. e., a high pitched determiner and a lower pitched noun), however, we did not quantify this impression.

A visual impression of the produced f0-contours is given in Figure 38. Since the number of segments in and around the stressed syllable affect the intonation contour, we separated

[42] Formulae used for the calculation of alignment: $alignL = tL - tV_{start}$, $alignH = tV_{end} - tH$ (t stands for time stamp, V_{start} for the beginning of the stressed vowel, and V_{end} for the end of the stressed vowel).

[43] We use mean values per condition for descriptive purpose. We are aware, that the Wilcoxon Signed Rank Test is not comparing the condition means.

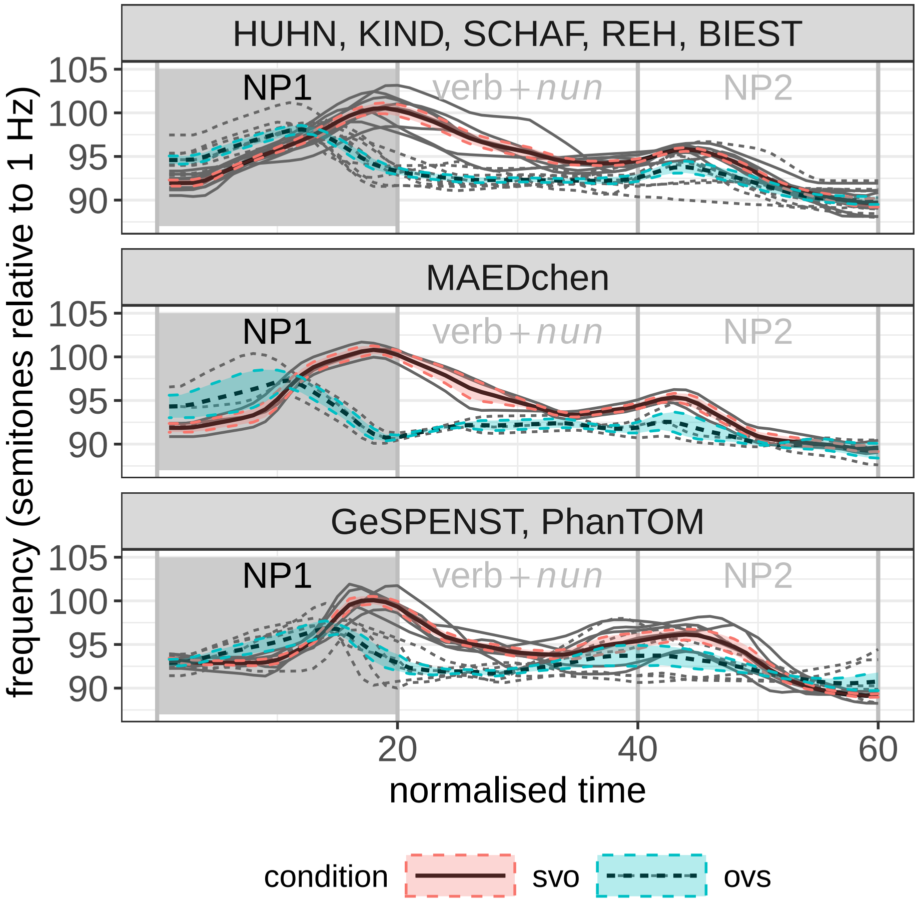


Figure 38: Selected time-normalised f0-contours (st relative to 1 Hz) of individual productions (grey) and means (black) with standard error (shaded bands) from speaker 13 for SVO (red, solid lines) and OVS (blue, dashed lines) separated for number of syllables and stress pattern (top: monosyllabic, mid: disyllabic strong-weak, bottom: disyllabic weak-strong), capital letters correspond to stressed syllable. The time domain of NP1 is shaded in grey.

the items by number of syllables and stress pattern of the noun in NP1 for visualisation: (i) monosyllabic nouns (n = 10), (ii) disyllabic nouns with a strong-weak stress pattern (n = 4), and (iii) disyllabic nouns with a weak-strong stress pattern (n = 6). Furthermore, we excluded productions that deviated from the two f0-contours we aimed to describe (n = 3 in SVO, n = 2 in OVS). The group of monosyllabic nouns consisted, thus, of eight SVO and nine OVS productions (see grey lines in top panel of Figure 38), the group of disyllabic nouns with strong-weak stress pattern of three productions per word order condition (see grey lines in mid panel of Figure 38), and the group of disyllabic nouns with weak-strong stress pattern of six productions per word order condition (see grey lines in bottom panel of Figure 38). F0-values were extracted with the same customised praat script described in the method section combining the procedures of Mausmooth (Cangemi 2015) and ProsodyPro

(Xu 2013). The only difference is, that the f0-contours are time-normalised with 20 time steps in each constituent, thus, durational differences between utterances are not visible. All f0-contours in Figure 38 are in semitones relative to 1 Hz (Hazan et al. 2016).

With regard to durational differences, this speaker produced longer preverbal silent intervals in OVS than SVO sentences and silent intervals preceding NP2 in SVO, as decribed in the section on silent intervals. For the mean durations of NP1 in OVS (0.605 s) and SVO (0.558 s) sentences, however, the Wilcoxon Signed Rank Test with condition revealed no statistically significant difference ($V = 27157$, $p = 0.32$).

8.4 Discussion

With respect to RQ1, the results reveal that most of our speakers did not use the investigated prosodic cues (f0-trajectories, presence/durations of silent intervals between constituents) distinctively at all in SVO vs. OVS sentences. Only a minority of speakers used f0 and/or silent intervals. The prosodic marking of OVS sentences with longer preverbal silent intervals compared to SVO sentences is partly in line with Weber et al. (2006) who reported silent intervals for OVS sentences in the productions of individual speakers in their pilot recordings. However, only one speaker in our data set produced silent intervals with distinctive durations between the two word order conditions (speaker 13). Overall, the investigated prosodic cues in our study show large inter-individual variability that impedes general conclusions. However, the fact that, intra-individually, the speakers were very stable in their prosodic realisations supports accounts, which point towards the importance of individual patterns (Cangemi et al. 2015).

Concerning RQ2, intra-individual analyses indicated that there are statistical differences in the f0-contours between SVO and OVS sentences, suggesting that some speakers in our study indeed prosodically disambiguated the two word order conditions early in the sentence (i. e., before the disambiguating NP2) by means of f0. Nevertheless, the majority did not prosodically disambiguate at all or only later in the sentence pointing to the fact that a clear prosodic differentiation in f0-contour for these types of SVO and OVS sentences does not exist in German. For two speakers, the predicted f0-trajectories showed larger values in SVO than OVS sentences in the time domain of NP1 (cf. speaker 11 (34m) and speaker 13 (34o) in Figure 34) along with a positive predicted difference between the two curves, indicating that these two speakers produced a clearly higher f0-peak on NP1 in SVO than in OVS sentences. The post-hoc analyses of the f0-contours of speaker 13 confirmed the later peak in SVO than in OVS sentences, along with a wider f0-range and a steeper slope of the rise. The distinct contours on NP1 show some similarities with the pitch accents used in the stimuli of previous comprehension studies on SVO and OVS sentences (Weber et al. 2006, Grünloh et al. 2011, Henry et al. 2017). Nevertheless, on the basis of our measurements, we do not postulate that we found the two phonological pitch accents L+H* and L*+H. The assignment of the two labels requires more than acoustic measurements (cf. Zahner-Ritter et al. 2022 for a fine grained description of distinguishing between the two pitch accents in German). For the remaining speakers, GAMMs were fitted on the f0-values of all sentences irrespective of syllable structure and stress pattern. We are, hence, rather cautious in drawing further conclusions about the alignment of the tonal movement with the segmental structure. Clearly, more detailed analyses are necessary in order to allow fine-grained comparisons of our data to previously used stimuli in comprehension studies (such as Henry et al. 2017, Weber et al. 2006).

The finding of large variability between speakers is in line with our expectations. However, we did not expect to find such a high degree of intra-individual consistency in the prosodic contours of locally ambiguous SVO and OVS sentences (i. e., no distinction between the two word orders). One possible reason could be individual difficulties of some speakers to get the non-canonical OVS-reading of the sentences (as reported besides the recordings). This is not surprising given the low frequency of sentences with OVS word order (Bader & Häussler 2010). Further, the large consistency within speakers nourishes the assumption that there is no clear underlying prosodic concept to distinguish between the two word orders. The difficulty of the task might be reflected by a paused speaking manner, introducing silent intervals between constituents; a pattern, which was apparent in the productions of several speakers. However, the durations of these silent intervals were overall within the range of mean durations of German stops (Kuzla & Ernestus 2011) and their difference between SVO and OVS sentences only revealed statistical significance for two speakers.

Furthermore, since the studied sentences contained only temporary ambiguities, which were resolved at the post-verbal NP, there was, strictly speaking, no necessity to resolve the ambiguity via other means (e. g., prosody). This is in no way to say that the speakers did not make an effort to prosodically distinguish SVO from OVS sentences early. If the temporary ambiguity is the reason why speakers did not prosodically distinguish between SVO and OVS sentences, we would expect them to do so if faced with globally ambiguous sentences (cf. example in (36)). In a pilot study with seven speaker-listener pairs (13 individuals identified as female, 1 individual identified as male; mean age 25, *SD*: 6.3, age range: 18–41 years), we elicited productions of globally ambiguous sentences and listener responses in a pen-and-paper version. Listeners and speakers had large difficulties to get the OVS interpretation of the sentences. Similar to the procedure of the present study, the speakers were given two pictures with a corresponding sentence (Figure 39). The two pictures differed with respect to the mapping of thematic roles on the two NPs of the sentence. The listeners were given the same pictures but without the sentence. The listeners asked *Was siehst du?* ('What do you see?'), and the speakers' task was to produce the sentence in a way the listeners could mark the corresponding picture. Test items consisted of eleven sentences in SVO word order and ten sentences in OVS word order preceded by two practice items, one per word order condition. We only analysed the responses of the listeners. Overall, the results show a strong bias towards an SVO interpretation: In five speaker-listener pairs, all sentences were interpreted as SVO, independent of the intended word order, that is, the listener marked for all sentences the picture depicting NP1 as agent of the action. Thus, all OVS sentences were interpreted as SVO. For one pair, nine out of the ten OVS sentences were interpreted as SVO, while one sentence was correctly interpreted as OVS. For the remaining pair, six OVS sentences were interpreted as SVO and two SVO sentences as OVS, thus 8 errors in total. For comparison, six speaker-listener pairs (9 individuals identified as female, 3 individuals identified as male; mean age 24.3, *SD*: 3.1, age range: 19–30 years) were recorded with locally ambiguous sentences in a similar setting. The results also showed a strong SVO bias: In all pairs, listeners interpreted in SVO sentences (n = 10) NP1 as agent (no errors). For the OVS sentences (n = 11) the results were more variable: no errors in one pair, two errors in one pair, three errors in three pairs, and six errors in the remaining pair. For the locally ambiguous sentences, listeners could use the case-unambiguous NP2 to resolve the ambiguity created by NP1. Nevertheless, the results show an SVO bias, as more SVO responses were given irrespective of the sentence produced. For the globally ambiguous sentences, no morpho-syntactic disambiguation was provided and listeners had to rely on other cues possibly provided by the speakers. The weak performance of listeners suggests that speakers had a hard time to prosodically distinguish between SVO and OVS sentences and that speakers did not provide sufficient cues for the listeners to disambiguate the sentences at least no cues, the listeners could use. This speaks more in favour of a difficulty to get the non-canonical reading of the sentences in OVS word order independent of a case-unambiguous NP2 than missing effort at the side of the speakers.

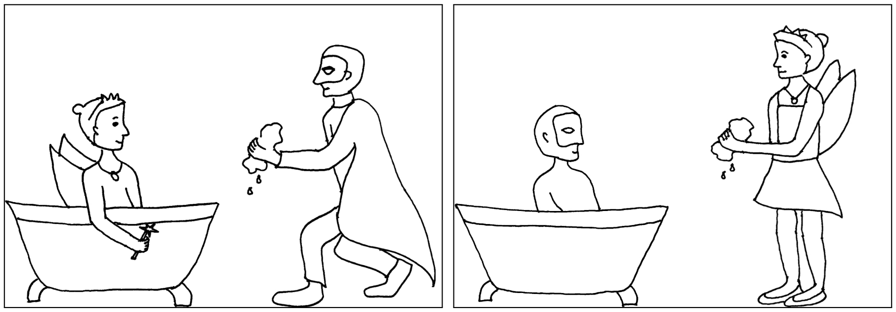


Figure 39: Example of black-and-white line-drawings used as visual stimuli for SVO (left) and OVS (right) word order of the globally ambiguous sentence *Das*.NOM/ACC *Phantom badet die*.NOM/ACC *Fee.* (the.NOM/ACC phantom bathes the.NOM/ACC fairy.).

We do not claim that the investigated prosodic cues are the only means to achieve prosodic disambiguation of locally ambiguous sentences. Future research should examine additional cues (e. g., constituent duration, timing and alignment of f0-movement, intensity) and include the comprehension side.

8.5 Conclusion

We analysed productions of locally case-ambiguous sentences with SVO and OVS word order, in which the syntactic ambiguity gets morphologically resolved by the case-unambiguous post-verbal NP2. We explored whether speakers prosodically distinguish the two word order conditions already earlier in the sentence, that is during the morpho-syntactically ambiguous region. We focused our analysis mainly on f0 as this cue was reported to facilitate disambiguation in comprehension. For most speakers, we found no differences in prosodic cues between SVO and OVS sentences. Nevertheless, since some individual speakers produced systematically distinctive f0-contours, silent intervals, or both, we argue for considering the individual level in order to acknowledge variability in prosodic realisations of the syntactic structure of sentences.

Additional files

Additional files to this article can be found in Appendix A.

Acknowledgements

We thank L. Junack for the drawings and her help with segmenting the data and A. Hofmann for her help with testing participants and annotation. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 317 633 480 - SFB 1287.

Appendix A: Wording of instruction in the young and elderly contexts to elicit locally ambiguous sentences. The interlocutors introduced themselves using the same text as in the studies of coordinates, given in Appendix A of study I on page 83.

German original	Translation into English
young Instruction Ich bin jetzt Ihre Gesprächspartnerin. Auf dem Bildschirm sehen Sie gleich jeweils zwei Bilder mit zwei Sätzen. Ich sehe nur die Bilder. Bitte sagen Sie mir den Satz so, dass ich so schnell und genau wie möglich verstehe, welches Bild gemeint ist. Ich frage Sie gleich immer was Sie sehen und Sie sagen mir den Satz.	young Introduction I am your interlocutor now. On the screen you will see two pictures and two sentences. I see only the pictures. Please tell me the sentence so that I understand as rapidly and as accurately as possible which picture is meant. I always ask you what you see and you tell me the sentence.
elderly Instruction Ich bin jetzt Ihre Gesprächspartnerin. Auf dem Bildschirm sehen Sie zwei Bilder und zwei Sätze. Ich sehe nur die Bilder. Sagen Sie mir den Satz so, dass ich schnell und genau verstehe, welches Bild gemeint ist. Ich frage Sie jedes Mal was Sie sehen und dann Sie sagen mir den Satz.	elderly Instruction I am your interlocutor now. On the screen you will see two pictures and two sentences. I see only the pictures. Tell me the sentence in such a way that I rapidly and accurately understand which picture is meant. I ask you every time what you see and then you tell me the sentence.

9 General Discussion

> It is regrettable that very few studies in this area have made any attempt to establish whether different productions of an ambiguous sequence do indeed contain different prosodic patterns. Clearly, listeners need an actual difference in prosodic realization if they are to distinguish local ambiguities via prosody; examination of the prosodic characteristics of each version of such a structure is thus crucial. (Cutler et al. 1997: 167)

The present work contributes to the systematic study of variability in disambiguating prosody in production and comprehension providing a combined investigation of the use of prosody in structural ambiguities and in different conversational contexts. Both, the resolution of structural ambiguities and the adaptation to conversational contexts, can be transmitted through the channel of prosody. In addition, the thesis considers the individual level how speakers and listeners deal with prosodic means in ambiguous structures. The complex prosodic signal, here specifically the domains of duration and fundamental frequency (*f0*), allows for variability at different levels. Prosodic disambiguation was studied with a focus on German name sequences of three names (*coordinates*) in two conditions: without and with internal grouping of the first two names (*Name1* and *Name2*, respectively) in two production studies (studies I and II) and one comprehension study (study III): *Name1 and Name2 and Name3* vs. *(Name1 and Name2) and Name3*. The studies of coordinates were complemented with production data of locally ambiguous sentences with a case-ambiguous first noun phrase (*NP1*) in subject-verb-object (*SVO*) and object-verb-subject (*OVS*) word order (study IV). Five conversational contexts were created (cf. Table 24, *contexts*). The contexts involved interlocutors in three age groups (child, young adult, elderly adult) who have German as their first language (L1) and were presented without background white noise, the young adult with German as their L1 and presented with background white noise, and a young adult with an L1 other than German. The interlocutors were virtually present during the elicitation of productions. They presented themselves in short, pre-recorded videos and engaged with the speakers in question-answer dyads.

Context was used as a within-subject variable in the production elicitation. We analysed the prosodic behaviour in response to the different contexts in a group of young adult speakers (*intra-group level*), within and between individual young adult speakers (*intra-individual level* and *inter-individual level*), and between the group of young and a group of older adult speakers (*inter-group level*). For comprehension, variability was analysed within a group of young adult listeners and between individual young adult listeners. An overview of the studies and at which levels they addressed variability is given in Table 25.

The thesis addresses three aims. The **first aim** is to improve our understanding of the form of prosodic grouping in the disambiguation of coordinates. Within this aim, we replicate the involvement of f0-movement, final lengthening, and pause in prosodic grouping,

Table 24: Schematic representation of the five situational contexts along the dimensions of age and L1 of the interlocutor, and the absence/presence of background white noise in the communication situation used for the elicitation in the production. Small capitals mark the name of each context. *Note*: Studies I and II involve all five contexts, study IV the YOUNG context.

		age		
background white noise	L1	child	young adult	elderly adult
absent	not German	–	NON-NATIVE	–
absent	German	CHILD	YOUNG	ELDERLY
present	German	–	NOISE	–

extending the findings to productions of older adult speakers and comprehension of young adults. We then argue that the distribution of these cues indicates global marking of internal grouping as opposed to local marking at the group edge.

Table 25: Distribution of the four studies of the thesis with regard to different levels at which variability was analysed in production and comprehension.

	group level		individual level	
	production	comprehension	production	comprehension
between	**inter-group**		**inter-individual**	
	study II		study I, study IV	study III
within	**intra-group**		**intra-individual**	
	study I, study II	study III	study I	

The **second aim** is to deepen our knowledge of the relationship between prosody and syntax by investigating whether the close link between prosody and syntax is maintained in different conversational situations or whether the aforementioned disambiguating prosodic cues are modified when speakers address different interlocutors.

The **third aim** is to discuss possible generalisations of the findings on prosodic grouping. Within this aim, we discuss structured variability and how it supports a phonological category of grouping. Further, we discuss whether a relative character of boundary strength conflicts with reliable decoding of early cues. We end with a return to the starting point exploring prosodic disambiguation in another syntactically ambiguous structure.

We will start with a summary of the main results with a focus on variability at the individual level before addressing the aims of the thesis.

9.1 Summary of major results

Study I focused on productions of young adult speakers, while study II compared the results of the young adults with productions of older adult speakers. The results of both studies showed that the three prosodic cues mentioned in the literature as main cues of prosodic boundaries play a role in the marking of the right edge of the name group as well as group-internally: f0-range, final lengthening, and an additional pause. The production of prosodic cues for disambiguation appeared to be unaffected by age-related differences in absolute durations and a generally larger variability in the productions of older speakers. The results of the two studies show smaller f0-ranges and less final lengthening on Name1 in the condition with internal grouping than in the condition without internal grouping. Contrary, there was a larger f0-range, more final lengthening and a pause on Name2 in the condition with internal grouping than in the condition without internal grouping. These findings are clearly consistent with the predictions of the model of Kentner & Féry (2013), which predicts proximity for Name1 and anti-proximity for Name2. The prosodic differences on Name1 motivated the question addressed in study III whether, in comprehension, listeners are able to use the prosodic marking on Name1 to decode the structure of the upcoming coordinate (i. e., the absence/presence of an internal grouping). Comprehension data from a gating paradigm in study III showed that this is the case for more than half of the listeners who were able to successfully predict the absence/presence of internal grouping already after hearing Name1. Overall, the results support a global view of prosodic grouping, where a group of names is separated from the names outside the group and shows proximity (cohesion) in the inside by weakening the group-internal boundary.

With regard to the contexts, both speaker age groups in studies I and II produced rather stable patterns across contexts. At the inter-group level (study II), our data showed no statistically significant main effect for age group, although, in absolute measures, the data supported previous findings of longer absolute durations in older speakers compared to younger adult speakers. Contrary to our expectations based on previous literature, we did not find evidence for an age-related increased use of prosodic cues for disambiguation. We will summarise the results of study IV in the following section on inter-individual differences.

Inter-individual level In the following we are going to summarise the results of individual variability in the different studies for the inter-individual and the intra-individual level separately. Regarding the exploration of inter-individual variability, studies I and IV provide insights from production partly complemented with insights from comprehension (study III, complementing study I). The two production studies differed in the elicited stimuli, coordinates in study I and locally ambiguous sentences in study IV. The analysis of study I concentrated on the three prosodic cues f0-range, final lengthening, and pause that are commonly used for prosodic grouping (Wagner & Watson 2010, Kentner & Féry 2013, Cole 2015, Petrone et al. 2017). In study IV, we analysed possible silent intervals between constituents (preceding the verb and preceding NP2) and the complete f0-contours in the ambiguous part of the sentence as differences in f0-contour were reported for similar structures (Weber et al. 2006, Henry et al. 2017). One could ask why we did not use the same measures on both materials, for instance comparing the f0-range on the names in coordinates with the NPs in the locally ambiguous sentences. The target words (names) in study I were homogeneous in number of syllables and syllable structure consisting exclusively of CVCV syllables. The NPs in study IV were more variable with respect to number of syllables, syllable structure, and stress pattern. This difference in structure makes the use of f0-range, a measure that does not consider the durational level of the f0 movement, less reliable for comparisons. We will briefly present the results of the three studies before providing a joint discussion.

The analysis of inter-individual variability in the productions of coordinates in study I revealed differences as to how speakers combined the prosodic cues to disambiguate between conditions and to how distinctive the cues were used. The vast majority of the speakers used at least two of the three cues distinctively between the conditions without and with internal grouping. This is in line with data on British English on lists of three nouns either combined to a list of two items or a list of three items: In that study, the majority of speakers used more than one cue and differed individually as to which cues they combined (Peppé et al. 2000: 323). Peppé et al. (2000) reported lengthening and silent pauses as more reliable cues than rising f0 or pitch reset. In contrast, young adults in our data used mostly f0-range of the rise and pause to distinguish between the conditions while final lengthening was used more variably: for some speakers the conditions were clearly distinguished by final lengthening, for others only partially or without distinction. The differences in cue combinations between the two languages are an interesting topic to pursue in future studies. At the same time, they have to be taken with caution, as Peppé et al.'s (2000) results are based on perceptual judgements of production data and not fully verified by instrumental measurement (Peppé et al. 2000: 323).

With respect to the production of locally ambiguous sentences (study IV), f0-trajectories and silent intervals showed large inter-individual variability. Only a few individuals used f0 and/or silent intervals distinctively between conditions, producing larger f0-values in the ambiguous part of the SVO sentence than in the OVS sentence and in one case longer preverbal silent intervals in OVS than in SVO. Variability is observable in that not all individuals produced the same shape of contours. Speakers differed inter-individually with regard to the shapes of f0-contours they produced on the locally ambiguous sentences, producing for instance different types and numbers of pitch accents. For the majority of the speakers, however, the f0-trajectories of SVO and OVS sentences overlapped, showing intra-individually large consistency across word order conditions and no disambiguation involving tonal cues. Similar findings were reported by Ouyang & Kaiser (2015) for sentences differing in focus and givenness in American English: "between-subject variability and within-subject consistency were observed in both the shapes of f0 contours and the ranges of f0 values" (Ouyang & Kaiser 2015: 166). A possible reason for the large consistency between SVO and OVS sentences could be individual difficulties with the non-canonical OVS reading. Additionally, the resolution of the temporary ambiguity in the second part of the sentence makes an early prosodic disambiguation, strictly speaking, not necessary. Further, the consistency across word order conditions can be viewed as an indicator for the absence of clearly differing categories or abstract concepts for the prosodic distinction between SVO and OVS sentences, at least for the non-canonical word order. We will elaborate on this in the last section of the discussion.

Regarding the comprehension of coordinates (study III), individual listeners showed different response behaviours that could be classified into two subgroups: a “waiting pattern” subgroup and an “identification pattern” subgroup. Listeners in the “identification pattern” subgroup were able to reliably distinguish between conditions using early prosodic cues (present on Name1). Accuracy of listeners in the “waiting pattern” subgroup exceeded chance range only after listening to Name2. Inter-individual variability in the data set was restricted as only these two response patterns were taken by raters in the classification task (a third option “neither of the above” was given but never chosen). The confidence ratings of the listeners in the “identification pattern” subgroup were higher (indicating more confidence) than in the “waiting pattern” subgroup, which is in line with observations by Price et al. (1991) who report that “subjects were rarely confident and incorrect” (Price et al. 1991: 2960). The results strongly indicate differences in the parsing of prosodic cues between individuals, which was also reported by Cangemi et al. (2015).

The results of all three studies provide clear evidence for inter-individual variability in the use of prosodic cues in the presence of structural ambiguities. While in comprehension, individual response patterns could be assigned to two groups, in production, no clear-cut patterns in behaviours beyond the individual were apparent.

Intra-individual level Intra-individual variability was analysed for the productions of coordinates by young adult speakers (study I) concerning the question whether individual speakers used and combined the investigated prosodic cues differently dependent on the contexts. The data showed a rather stable pattern across contexts for almost half of the speakers using the same cues in similar combinations across contexts, thus showing strong similarity. With regard to individual cues, final lengthening, again, was used less distinctively, while mostly f0-range and pause were used to distinguish clearly between the grouping conditions independent of context. Overall, intra-individually, speakers produced prosodic disambiguation between coordinates without and with internal grouping in a consistent manner and without large differences between contexts.

Taking together the inter- and intra-individual level, we observed that young adult speakers differed between each other in how they used and combined prosodic cues when faced with structural ambiguities in production. At the same time, the inter-individual differences came along with intra-individual consistency. We found this pattern disregarding whether speakers successfully disambiguated between structures (e. g. coordinates) or not (e. g. locally ambiguous sentences). For the locally ambiguous sentences, the consistency applies irrespective of the word order condition, while for the coordinates, all speakers successfully disambiguated between the two conditions and showed consistency in to which degree they use the prosodic cues: Within prosodic cues, they varied little, with most variability in final lengthening. Further, speakers used inter-individually different cue combinations for the prosodic grouping of coordinates. In conclusion, our data provide evidence for inter-individual variability in the production of prosodic cues along with intra-individual stability.

For comprehension, the results of study III show differences at the group level of young adult listeners in the amount of prosodic information needed to predict the intended grouping. Some listeners were able to reliably predict the upcoming structure already after hearing Name1, while others were able to do so only after hearing Name2. The data provide evidence for inter-individual differences in the use of prosodic information for correctly identifying the upcoming structure. Further, at the group level, we found different response patterns for the productions of individual speakers. It remains open for future research to analyse the individual level of the listener with regard to speaker-specific prosodic cues. The findings of Cangemi et al. (2015), reporting inter-individual patterns in the responsiveness to different speaker profiles and cues, strongly suggest that it might be worth looking at the individual level as well. In their study, listeners were presented with fictitious names embedded in a carrier phrase (*Melanie will Dr.* ***Bahber*** *treffen* 'Melanie wants to meet Dr. **Bahber**', target word in bold) produced with either broad, narrow, or contrastive focus by different speakers. Listeners' task was to match the productions to the correct focus context. Productions were analysed with regard to several prosodic cues, namely peak alignment, peak height, duration of the target word, duration of the first word, and number of prenuclear accents. Speakers varied with regard to which cues or cue combinations they used to produce the different focus contrasts. Additionally, listeners differed inter-individually in how reliably they decoded the intonational contrasts produced by individual speakers. Cangemi et al. (2015) argue for an interaction between speaker- and listener-specific behaviour in that some listeners are particularly good at decoding the structures encoded by certain speakers but that there are no individual speakers that are more intelligible to all listeners overall. Taking together the results by Cangemi et al. (2015) and our study, the further exploration of the individual level appears promising. However, clear effects at the group level are not necessarily reflected for each individual participant, as reported for production data by Niebuhr et al. (2011). Nevertheless, the obvious inter-individual differences in parsing strategies in our data suggest that there are also inter-individual differences in listeners of speaker-specific cue use.

We now turn towards the discussion of the three aims of this thesis.

9.2 Aim 1: The form of prosodic grouping of coordinates

The **first aim** is to improve our understanding of the form of prosodic grouping studied in the distribution of the three prosodic cues, f0-movement, final lengthening, and pause, involved in ambiguity resolution in the case of coordinates in German focusing on two sub-points. With the **first sub-point** we aim to replicate the involvement of the three prosodic cues, f0-range, final lengthening, and pause, in the ambiguity resolution of coordinates and to extend them to older adult speakers. With the **second sub-point** we aim to deepen the insights of the distribution of prosodic cues within the utterance addressing the question whether the cues are globally or locally used in production and comprehension.

Three prosodic cues are involved in the ambiguity resolution of coordinates: f0-range, final lengthening, and pause. We have been able to replicate the involvement of the three cues in prosodic grouping in productions of young adult speakers. The use of these three cues in prosodic grouping is in line with previous findings on German coordinates (Wellmann et al. 2012, Kentner & Féry 2013, Holzgrefe-Lang et al. 2016, Petrone et al. 2017, among others). Besides the replication of previous findings, the thesis contributes new insights by extending the investigation of prosodic cue production to a non-young speaker group with an average age of 68 years. Our data showed that prosodic grouping was largely unaffected by age-related changes (despite differences in absolute measures in the durational domain[44]). Both age groups used the same three prosodic cues, f0-range, final lengthening, and pause, to disambiguate between coordinates without and with internal grouping. Young and older adult speakers showed a comparably effective use of linguistic prosody in that both age groups successfully disambiguated between conditions as shown by the low number of misperceived items in the perception check (4% of the productions of both groups). Grouping, thus, could be reliably recovered from the prosodic form by naïve listeners in 96% of the cases in the productions of young and older adult speakers, respectively. Similarly high accuracy rates for the identification of the intended grouping of ambiguous sequences were reported for English name sequences (94.4% Lehiste 1973b: 1231) and British English noun sequences (92.6% Peppé et al. 2000: 322). For German sequences of four names with different levels of embedding, the overall accuracy was 71% (Kentner & Féry 2013: 302)[45].

Finding the three prosodic cues in production leads to the question of whether these three prosodic cues are also used for the perceptual distinction between the grouping conditions. We answer this question positively based on exploratory post-hoc visual inspections of productions that were not perceived as intended: Misperceived productions often went along with one of the three prosodic cues falling within the distribution of the values of the other condition. This is exemplified in Figure 40 for the productions of an individual

[44] We did not analyse for age-related changes in f0 in our data due to imbalanced distribution of speakers that identified as female and male.

[45] I mention the accuracy values in Kentner & Féry (2013) for the sake of completeness. However, their data cannot be directly compared to ours, as in our data three names were combined to two different sequences (i. e., two answer options in comprehension), while in the data by Kentner & Féry (2013), four names were combined in six different versions (i. e., six answer options).

speaker in a single context (12 datapoints: 6 name sequences $*$ 2 conditions) and the cues final lengthening (x-axis) and the f0-range of the rise (y-axis). The datapoints included in the analysis ($n = 9$) are in solid colour, while datapoints of productions excluded after the perception check are in lighter shaded colour ($n = 3$, grey circles). A closer look reveals that the excluded productions were from the condition with internal grouping (circles) and that two of them had a percentage of final lengthening that was similar in value to the productions without internal grouping (green triangles). Thus, at the acoustic level, those productions showed a percentage of final lengthening on Name2 that was more similar to the percentages of final lengthening in other items produced without internal grouping than those with internal grouping by this speaker (i. e., the grey dots were closer to the green triangles than to other black dots.).

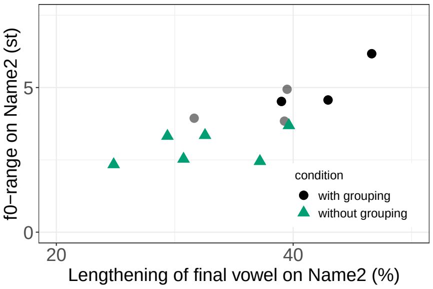


Figure 40: Distribution of two prosodic cues on Name2 produced by an individual speaker (ID 16) in a single context (NOISE) in the conditions without internal grouping (green triangles) and with internal grouping (black circles). Final lengthening is plotted on the x-axis and f0-range on the y-axis. The solid coloured datapoints were included in the analysis, the lighter shaded datapoints had been excluded following the perception check.

The same phenomenon is visible in a study by de Beer et al. (2022) on production data of people with either left- or right-hemispheric damage producing similar coordinate name sequences without and with internal grouping as elicited in study I and II. The authors descriptively compared the use of f0-range, final lengthening, and pause between accurately and inaccurately[46] produced coordinates. The data show a clear pattern: In the correctly identified productions, the cues in the two conditions did not overlap, while for the incorrectly identified productions the distributions overlapped considerably between the condition without and with internal grouping (de Beer et al. 2022: 4789). That is, in those productions, in which neurotypical listeners had difficulty identifying the intended grouping condition, at least one cue overlapped between conditions. These exploratory observations need to be proven by experimental and statistical consolidation in future studies. Nevertheless, our observations suggest the assumption that in single productions overlapping cues can lead

[46] Production accuracy had been assessed by a group of neurotypical naïve listeners (de Beer et al. 2022: 4783).

to a confusion in recovering the intended internal grouping. Further support for the use of f0-movement, final lengthening, and pause in the perceptual chunking of speech is provided by a comprehension study showing that German speaking listeners use the three cues to identify boundaries in an unfamiliar language (Estonian, Ots & Taremaa 2023).

Prosodic cues are globally distributed within the utterance and not restricted to specific locations. In the second sub-point regarding the form of prosodic grouping, we concentrated on the distribution of the prosodic cues within the name sequence, specifically on the two names, which were grouped together in one of the conditions (Name1 and Name2). The production data (study I and II) showed that two of the three prosodic cues also play a major role on Name1: f0-movement and final lengthening. This applied equally to productions of young and older adult speakers. The finding of name group-internal prosodic weakening along with prosodic strengthening at the group edge as predicted by proximity and anti-proximity in the model by Kentner & Féry (2013) support the view that prosodic grouping is not only a matter of separation from surrounding material, demarcating the group from the non-group material, but at the same time a matter of cohesion of sister elements within a group (Cutler et al. 1997). In other words, prosodic grouping is not exclusively indicated by a prosodic boundary at the right edge of the group and, thus, not locally bound to the group edge and the point of structural ambiguity, but rather globally marked. There are different ways used to describe this dichotomy: non-local or global as opposed to local, distal as opposed to proximate. Besides for data on German coordinates (Kentner & Féry 2013, Petrone et al. 2017), the interplay of weakening and strengthening of prosodic cues at different locations within a structure was also reported on British English data by Peppé et al. (2000). Speakers were asked to produce three nouns that either form a list of three items (*ice, cream, and fruits*) or of two items (*ice-cream and fruits*). The authors observed that

> it is not simply a case of lengthening one or the other part of the utterance, or inserting silence at one particular point. For all prosodic elements, the relationship between exponency on the two parts was important. For instance, an apparent lengthened vowel in the first noun [...] did not signal a 3-list [...] unless the second noun was shorter. (Peppé et al. 2000: 322)

The non-locality of the prosodic grouping is further in line with accounts that consider prosodic marking as a phenomenon with a relative character (Clifton et al. 2002, Wagner 2005, Frazier et al. 2006, Cole 2015). In this view, prosodic boundaries are considered in relation to other boundaries with which they are in a strength relation. These other boundaries can be at the same position in a comparable structure (as for coordinates without compared to with grouping) or they can be preceding or following the boundary at issue in the same structure. An example for the latter are ambiguous lexical structures (e. g., *tie* *murder bee* vs. *timer derby*) embedded in longer utterances. So called distal (or non-local) prosodic patterns (regularity in f0 and/or timing cues) in the utterance were shown to affect whether listeners perceived a final monosyllabic or disyllabic word (Morrill et al. 2014). The necessity to interpret local prosodic events with respect to the surrounding material is also reported by Schafer et al. (2000). They found that the difference between early and late closure readings is not only marked by prosodic boundaries at the point of syntactic ambiguity but already earlier in the utterance.

The observation of global, name group-internal, prosodic marking is closely related to the question whether the acoustical difference between conditions on Name1 is observable in comprehension, in other words whether listeners can employ these "early" cues to predict the upcoming structure. In the words of Cutler et al. (1997):

> Does prosodic information serve to resolve ambiguity, such that sentences which admit of more than one interpretation when they are written are effectively unambiguous when spoken? And is prosodic information consulted "on-line" in order to select between alternative syntactic analyses which present themselves, albeit temporarily during the processing even of an unambiguous sentence? (Cutler et al. 1997: 185)

The results of study III partly answer this question in the affirmative: The young adult listeners classified in the "identification pattern" subgroup were able to correctly identify the upcoming structure already after listening to Name1 and, thus, before the prosodic cues at the group edge were available (Name2). The overall good performance of the listeners in the later gates (following gate5) is not surprising, as the stimuli had been selected because of their high accuracy rates in the perception check included in study I. The prosodic difference in f0-range and final lengthening on Name1 between a structure without and with internal grouping was thus not only measurable at the phonetic/acoustic level, but could also be used for prediction during processing by at least some listeners. Further, some speakers seem to be more successful in providing early cues than others, as listeners' response accuracy differed between speakers. This was also supported by descriptive visual inspections of the cue distribution on Name1 showing more distinct cue use in some speakers and less in others.

The finding of a subgroup of listeners that is able to use early cues in Name1 to reliably predict the upcoming structure has implications for methodological considerations on the investigation of boundary perception, especially for the investigation of individual cues. In order to be able to tear apart the influence of individual cues, studies use stimuli that are acoustically manipulated to control the presence or absence of cues or cue combinations and/or their degree. Previous studies on infant and adult processing of prosodic boundaries in German manipulated prosodic cues directly at the boundary/right group edge taking a natural production of a coordinate without internal grouping as a base (cf. Holzgrefe et al. 2013, Holzgrefe-Lang et al. 2016, 2018 for studies measuring event-related potentials (ERP) and Wellmann et al. 2012, Holzgrefe-Lang et al. 2016, Wellmann et al. 2023 for behavioural studies). In studies collecting electrophysiological measures, such as ERP, it is desirable to locate the difference between conditions (the absence/presence of prosodic cues) to a certain time in the stimulus in order to determine a time window to look for a brain response that can be associated with it. A natural production of a coordinate without internal grouping that contains boundary cues that were manipulated on Name2 also contains potentially misleading cues on Name1. Listeners are sensitive to these early cues as shown by data of ERP and behavioural measurements. If on a recording of an ungrouped coordinate, in addition to a locally added f0 change on Name2, f0 was flattened on Name1, this stimulus elicited a larger closure positive shift (CPS, indicating the processing of a prosodic boundary) and larger mean proportions of boundary judgements compared to a stimulus with f0 manipulation on Name2 only (Holzgrefe-Lang et al. 2016).

To sum up, our results and previous findings on the production and comprehension of coordinates without and with internal grouping show that Name1 in the group carries a lot of information about the grouping structure already. The marking of prosodic grouping is not restricted to a specific location but appears as a more widespread phenomenon. This has implications for the analysis of production and comprehension data. Therefore, material for comprehension studies should be carefully selected accordingly. The cues build up over time of a coordinate and speakers and listeners differ in the amount of information they produce and use to reliably mark and predict the upcoming structure. For some speakers, some listeners are able to decode these early cues effectively and use them to predict the upcoming structure. Prosodic grouping is a global phenomenon involving prosodic cues at distal (non-local) positions. Even these prosodic cues at distal positions are an important source of information and can be sufficient for predicting the upcoming structure.

9.3 Aim 2: Situational independence of disambiguating prosody

The **second aim** of the thesis is to deepen our knowledge of the relationship between prosody and syntax by exploring whether the close link between prosody and syntax is maintained in different conversational contexts or whether the aforementioned disambiguating prosodic cues are modified when speakers address different interlocutors with possibly different needs. In our data (study I and II), coordinates in all five contexts (directed at a young adult, a child, an elderly adult (all L1 speakers of German) a non-native young adult, and the young adult in a noisy environment) showed only small differences in some contexts. The two grouping conditions were equally well prosodically disambiguated irrespective of the context. Conversely, there was no context in which speakers disambiguated substantially less clearly. There were no contexts from which a particularly large number of items were excluded from the analysis after the perception check. We conclude from this that the production of clear disambiguating prosody was not modified when addressing different interlocutors. The small measurable differences between contexts were not recoverable in comprehension and, more importantly, did not affect successful prosodic disambiguation. We interpret this finding as situational independence of disambiguating prosody and as informative for the nature of the relationship between prosody and syntax in production.

The relationship between prosody and syntax is visible and rather direct for the coordinates, as the syntactic and semantic structure (the grouping) was represented in the prosodic structure as evidenced by successful prosodic disambiguation. This holds, irrespective of speaker age, for productions of young and older adult speakers. Further, it appears as stable since the analyses regarding different levels of variability taken together revealed only small differences between conversational contexts at the intra-group as well as at the intra-individual level. At the intra-individual level, young adult speakers only varied scarcely in their use of cues and cue combinations between contexts and, more importantly, these adaptations did not impact the successful marking of prosodic grouping, prosodic grouping was largely unaffected by contexts. Furthermore, the poor recognition of contexts in the perception check indicates that there were no individual speakers who performed significantly better in providing prosodic adaptation to the contexts than others. Intra-group and intra-individual level mirror each other: In the present data, small differences across contexts at the group level go along with small differences at the individual level. This also implies that prosodic disambiguation was consistent across contexts. We interpret this as a close relation between prosody and syntax.

In our data, prosodic disambiguation was produced irrespective of interpreted possible needs of the addressed interlocutor or the presence of background white noise. This is in line with findings on English coordinates that were prosodically disambiguated independent of type of interlocutors (Wagner 2005). We interpret this in favour of models of situational independence of disambiguating prosody (Schafer et al. 2000, Kraljic & Brennan 2005, Speer et al. 2011). Prosodic disambiguation appears as an automatic part of the production process. If disambiguating prosodic cues would be produced only in cases in which the linguistic and extra-linguistic contexts of the utterance lack sufficient cues for disambiguation, they would be rather unreliable for comprehension. The results of our comprehension data clearly suggest that listeners make use of the disambiguating prosodic cues to distinguish between conditions without and with internal grouping. This was further supported by descriptive post-hoc observations of production data: in cases when prosodic cues overlapped between grouping conditions, conditions were less reliably distinguished in those productions.

Regarding the small or even absent context effects, we will discuss three possible limitations. First, an explanation for the small context effects in our production data could lie in the choice of prosodic cues that were analysed. The possibility that speakers produced additional prosodic cues to adapt to the contexts cannot be disregarded (e. g., hyperarticulated vowels measured by mean F1/F2 as reported for non-native directed speech by Knoll et al. 2015). This may be particularly relevant to the NOISE context as this was identified best in the perception check. Other additional cues reported for speech in noise include increased intensity and spectral changes (e. g., van Summers et al. 1988, Junqua 1996, Davis et al. 2006, Lu & Cooke 2008, Landgraf et al. 2017). Half of the productions (65 out of 119) for which at least 75% of the listeners correctly identified the context had been produced in the NOISE context, which corresponds to a third of all productions in this context (192 productions in total). Interestingly, at the same time, the largest number of excluded items after the perception check for young adult speakers (failing the distinction of grouping condition) belonged to the NOISE context (15 of a total of 38 excluded items had been produced in the NOISE context, followed in number by 10 items excluded from the YOUNG context). We, thus, find the largest number of misperceived grouping in the context with background white noise (NOISE) followed in number by the context that always was presented first (YOUNG). For the context presented first, this could be due to adaptation to the task. For the NOISE context, the presence of background white noise possibly affected the speaker's cognitive resources used for planning and articulation, which get reflected in the weaker prosodic output. This would support the notion that disambiguating prosody is related to speech planning and articulation and therefore rather produced "for" the speaker than as beneficial for the interlocutor (Bard et al. 2000, Schafer et al. 2000, Kraljic & Brennan 2005, Speer et al. 2011). In the case of the background white noise, creating a more difficult communication situation, speakers did not statistically increase the prosodic disambiguation, but rather sometimes failed to produce the difference in grouping condition reliably[47].

Second, the small context effects might be a result of the design we used to collect the production data: Interlocutors were only auditorily present during data recording and speakers received no feedback on their performance, neither a request for repetition nor an approving conversational sound. It is possible that a misunderstanding or a request for repetition would have triggered more accommodations in the speech addressed to the interlocutors. However, a repetition possibly conveys the notion of insistence that would add another layer of pragmatic information that is transmitted through the channel of prosody. Insistence (tested in second compared to first instances of calling someone) was reported to come along with longer syllable durations in Catalan vocatives directed at subordinates (Borràs-Comes et al. 2015: 74) and larger f0-ranges in Colombian Spanish vocatives (Huttenlauch 2016). In addition, Borràs-Comes et al. (2015) and Huttenlauch (2016) reported that different intonation contours predominated in insistent calls compared to initial calls. It remains open for investigation whether a request for repetition would trigger increased prosodic accommodation to the addressed interlocutors leading to larger differences between contexts or whether insistence causes prosodic modifications independent of interlocutors. A counterexample for

[47] However, this only applies to barely 8% of the productions (7.8% corresponding to 15 out of 192) and only to the data of young adult speakers, whereas for the older speakers the numbers of excluded items did not differ much between the contexts (20 in YOUNG, 14 in NOISE, 13 in NON-NATIVE, 11 in CHILD and ELDERLY, respectively).

more effort at the side of the speaker in cases of misunderstandings of the listener were found by Bard et al. (2000) in a map task. Participants in the role of information givers gave less clearly produced instructions in a repetition of the task even though their interlocutor was new to the task (a different one than in the first round).

Returning to the small context effects, a further argument that the setup limits context effects was reported by Xie et al. (2021). They discussed repetitive productions of a single construction with no feedback as possible limitations for the productions of English statements compared to questions as "likely to reduce overall variability in production and potentially promoting entrenchment of intonation contours across items" (Xie et al. 2021: 20). The authors further discussed that if applicable, "the current set of production data would underestimate the amount of within-talker variability compared to true underlying distributions to be observed in a more naturalistic form of language use" (Xie et al. 2021: 20). As the study by Xie et al. (2021) did not include different contexts, the comparison is not straight forward. However, if such limitations apply to simpler setups, they should apply to more complex setups too. If this is the case, we would expect more intra-speaker variability in a recording setting with fillers and other distracting tasks.

Third, it is further possible that the lack of context effects in our data was not due to the recording method but due to the short target utterances (eight syllables) that offered limited space for prosodic modifications. The findings of phonetic differences between interlocutors by another study speaks against the setup as limiting factor (Pescuma et al. 2023). The study used a similar design with pre-recorded interlocutors with different attire and hair styles illustrating different degrees of formality, mentioning that their participants did not report back a lack of naturalness in the communicative situation (Pescuma et al. 2023: 18f. project C02). In our study and in the study in Pescuma et al. (2023), speakers sat in front of a screen, in our case in a relatively small recording booth and in the other study at a table with the screen at the other end of the table. The conversational situations in the latter study differed in formality (formal: talking to a boss or professor, informal: talking to a neighbour or a fellow student). The two studies differed in the task and the speech materials recorded. In our case speakers were asked to read aloud visually presented name sequences in a way that the interlocutor could retrieve the internal structure (i. e., read speech material controlled for segmental composition, number of syllables, and stress pattern). The task in the other study was to either request an extension of a deadline or a pay raise from the interlocutor, or to talk with them about the city or a restaurant (i. e., spontaneous speech in a face-threatening situation and in a more relaxed situation; with less control about the speech material at the side of the researchers). In a preliminary analysis, the authors reported phonetic differences between tasks and interlocutor's formality, for instance, higher and more variable f0 and more dispersed vowels in the formal situation (Pescuma et al. 2023: 19). Besides the difference in speech materials and type of speech between the studies, data in our case were elicited previous to the Covid-19 pandemic, which came along with a strong increase in online-meetings, for instance via zoom, while in the other study participants were already more familiar to speaking in front of a screen.

To sum up, we observed small context effects in the productions of prosodic cues used to disambiguate coordinates in our data, which went along with a consistent prosodic disambiguation. We interpret this consistency in the prosodic grouping as support for models in favour of situational independence of prosodic disambiguation and in favour of a close link between syntax and prosody. The internal structure of coordinates was disambiguated irrespective of the type of addressee or the absence/presence of background white noise and the prosodic disambiguation is interpreted as part of the production process rather than dependent on the situation.

Summarising the findings of aim 1 and aim 2, the prosodic cues, f0-range, final lengthening, and pause, are involved in the resolution of structural ambiguities in a global manner. Prosodic disambiguation is not restricted to the point of syntactic ambiguity (the group edge at Name2) but builds up along the utterance. Prosodic groups are characterised by prosodic cohesion (proximity). Such non-local prosodic cues are not only found in coordinates but also reported for other syntactic structures (Schafer et al. 2000, Morrill et al. 2014, among others). Prosodic disambiguation in coordinates was produced irrespective of the age of the speakers (young and older adults) and independent of context (type of interlocutor or presence of background noise). We found a strong link between syntax and prosody. In the next section, we consider variability at the individual level and discuss how it provides support for an abstract concept of grouping and whether this can be generalised to other structures.

9.4 Aim 3: Generalisations of grouping and beyond coordinates

The **third aim** is to discuss possible generalisations of the findings on prosodic grouping. The aim is divided into three sub-points. In the **first sub-point**, we discuss structured variability and how it supports a phonological category of grouping. In the **second sub-point**, we discuss whether a relative character of the strength of prosodic cues in grouping conflicts with reliable decoding of early cues. In the **third sub-point**, we come back to the starting point exploring prosodic disambiguation in another syntactically ambiguous structure.

A phonological category of grouping. The productions of the group of young adult speakers showed a clear prosodic distinction between the two grouping conditions of coordinates (between *Caro and Toni and Jana* and *(Caro and Toni) and Jana*) using three prosodic cues: f0-movement, final lengthening, and pause[48]. At the inter-individual level,

[48] The same holds for the group of older adult speakers. As this section will deal with the data of young adult speakers, we will not further mention the older speakers.

the data of the young adult speakers showed variability in how these three cues were combined and to which degree they were used for prosodic grouping. Overall, the variability in cues and cue combinations that we observe between and also within speakers can be described as gradual in opposition to categorical. All speakers reliably marked the distinction between the grouping conditions with at least one of these prosodic cues investigated and most of the speakers used at least two of these cues. Differences between speakers relate to the degree of distinction between conditions of single cues. The variability appears in a structured way and not random. This is also reflected in the fact the individual differences in prosodic grouping did not lead to difficulties in recovering the grouping as shown by overall high accuracy rates in the perception checks. The inter-individual variability in cue combinations is in line with data on British English on lists of three nouns either forming a list of two items or a list of three items: In that study, the majority of speakers used more than one cue and differed individually as to which cues they combined (Peppé et al. 2000: 323). The inter-speaker variability in how the three cues are used and combined speaks in favour of some flexibility with regard to the prosodic realisation of the grouping (Wagner 2005). This flexibility seems to lie within certain limits, which do not hamper discrimination in comprehension. Along these lines of structured variability, we can think of a division between an abstract conceptual level and a concrete phonetic level. The abstract concept (here grouping) would be common to different speakers, while the concrete level offers space for flexibility in the phonetic realisation. On prosodic grouping, Cutler & Isard (1980) described the abstract concept with “the intention of marking off a syntactic unit by assigning it a grouping of its own” (Cutler & Isard 1980: 260). More in general on constituent boundaries in prosodic structure, Ladd (2008) wrote that they “are in the first instance abstractions, not actual phonetic events” (Ladd 2008: 9). With regard to the concrete and variable realisations of this abstract concept, for successful communication, it is necessary that the abstract concept is recoverable from the concrete realisation in comprehension. This prerequisite is satisfied for the grouping conditions tested in the coordinates in this thesis. Listeners in the perception check of study I and II and in the gating study in study III were able to distinguish between grouping conditions. We thereby assume a phonological category of grouping. We can speculate that the availability of multiple cues for the phonetic realisation of the phonological category offers an advantage. Multiple marking of the grouping with cues from the durational and tonal domains can possibly maintain recognition, even if individual domains are only conditionally available to listeners (for instance due to limited resolution of fundamental frequency in cochlear implants, hearing impairments).

Prosodic grouping or chunking of elements into groups of two (binary grouping) is also observable even if not induced by experimental manipulation (e. g., production of telephone numbers Baumann & Trouvain 2001). The abstract concept of grouping is not limited to prosodic grouping but found in other cognitive processes. A study exploring kinematic grouping in action sequences report on grouping mechanisms that can be compared to the prosodic ones (Hilton et al. 2022). Hilton et al. (2022) report on durational cues that correspond to final lengthening and pause in a study on action sequences such as *slide, lift, and roll*. Participants in their study were asked to perform three actions "in a continuous action sequence, either with or without boundaries between" the second and the third action (Hilton et al. 2022: 1421) similar to the production of coordinates studied in this thesis. The results show longer durations of the second action and a delayed onset of the third action in sequences with a boundary than in sequences without a boundary (Hilton et al. 2022). Longer durations by slowing down is also common in music to indicate the end of groups (ritardando, comparable to final lengthening in speech). Grouping is also used in music to create rhythm and expectation (cf. Huron 2006, Jackendorff 2009, Reich & Rohrmeier 2014). Huron (2006) relates the strength of expectation to perceptual grouping in music with the relative absence of expectation marking the boundaries of perceptual chunks (Huron 2006: 157). The strength of expectation in this sense can be compared to the ideas of proximity or cohesion. For future research it would be interesting to study, whether group-internal weakening of boundaries observed in prosodic grouping is also observable in grouping outside speech. One promising tool to approach this question could be the notion of expectation discussed by Huron (2006). We conclude that grouping is a common phenomenon in an even more general sense, as the chunking of complex processes into smaller parts facilitates processing (cf. Frazier et al. 2006, Jackendorff 2009).

Does a relative character of the strength of prosodic cues in grouping conflict with reliable decoding of early cues? We already mentioned the notion of a relative character of cue or boundary strength in prosodic disambiguation of ambiguous structures. The strength of cues or boundaries can be compared at two levels: at a syntagmatic (at two positions within the same structure) and at a paradigmatic level (at the same position between two structures). Exemplified for coordinates, the syntagmatic level corresponds to a comparison between the boundaries on Name1 and Name2 with the same index in Table 26. The paradigmatic level corresponds to a comparison between the boundaries on Name1$_i$ and Name1$_k$ or between Name2$_i$ and Name2$_k$ in Table 26[49]. The descriptions of boundary rank (Wagner 2005) and proximity (Kentner & Féry 2013) refer to the syntagmatic level, while the paradigmatic level is used in the analysis described by Kentner & Féry (2013) and applied in studies I and II.

In a strict interpretation of a paradigmatic relationship, a structure receives its identity in comparison to another structure (e. g., whether Name1$_i$ belongs to a grouped or an ungrouped structure can be determined in comparison to Name1$_k$). Applied to the comprehension of ambiguous structures, this would suggest that at least two different versions need to be available for the successful decoding of a structure. However, this is not compatible with

[49] The same two levels can be applied to other types of ambiguous sentences such as sentences with high vs. low attachment.

Table 26: Visualisation of the syntagmatic (rows) and paradigmatic (columns) level in a sequence of names without internal grouping (upper row) and with internal grouping (bottom row). Index *i* and *k* mark the syntagmatic relationship. The digits 1, 2, and 3 mark the paradigmatic relationship.

	paradigmatic		
syntagmatic	Name1$_i$	Name2$_i$	Name3$_i$
	(Name1$_k$	Name2$_k$)	Name3$_k$

reality, because we recognise the grouping structure even when we encounter it in isolation or at the first mention. Since different structures are not always available for comparison, boundary comparisons on a syntagmatic level seem more appropriate for comprehension. For production, Peppé et al. (2000) stated the following regarding the distinction between *ice, cream, and fruits* and *ice-cream and fruits*: "an apparent lengthened vowel in the first noun [...] did not signal a 3-list [...] unless the second noun was shorter" (Peppé et al. 2000: 322). This would suggest that listeners have to wait until the second noun in order to be able to decode the number of items in a list. Exemplified for coordinates, an f0 peak on Name1 followed by Name2 with a lower f0 peak would indicate a structure without internal grouping, while Name2 with a higher f0 peak would indicate a structure with internal grouping. However, the results of study III show that some listeners were able to distinguish between the two structures already after hearing Name1. Does that mean that the results conflict with the notion of a relative character of prosodic marking? One could argue that, strictly speaking, the study is not comparable with a case of a structure in isolation as listeners were confronted with two different grouping conditions and had to decide between them. With the exception of the first trial, listeners had trials to compare the current trial with. Listeners could compare between different trials and learn the distinction. Would that mean that the result of study III is due to a learning effect? A closer look at the data (study III) reveals only a small increase in accuracy between the first ten stimuli[50] and the remaining trails, but the group of listeners, which was able to differentiate between conditions based on prosodic information on Name1, was evident also in the beginning.

The results of an ERP study by Holzgrefe et al. (2013) comparing brain responses to structures differing in grouping of the first and the last two names (i. e., *(Caro and Toni) and Jana* vs. *Caro and (Toni and Jana)*) can be interpreted as support for a relative character in a syntagmatic relationship and that the strength of a boundary can only be evaluated in relation to a preceding boundary. For the second structure, no CPS (indicating the processing of a prosodic boundary) was elicited on the prosodic cues on *Caro*, while in the first structure a CPS indicated the processing of a prosodic boundary on *Toni*. The authors concluded that the prosodic cues present on the first name were not interpreted

[50] Corresponding to 70 trials, as each stimulus was cut into 7 gates.

as cues to a phrase boundary and that the elicitation of a CPS needs preceding prosodic context. A limiting factor of this study may be that participants were not encouraged to resolve the structural ambiguity. Holzgrefe et al. (2013) speculate that a task directed at the ambiguity resolution could elicit a CPS also in response to the first name. The result of the successful processing of the prosodic cues in our behavioural study III can be interpreted as support for this assumption[51]. Nevertheless, the objection regarding the relative character mentioned for study III applies again. Based on studies with repeated measures, we cannot exclude the possibility that listeners compare between trials of different conditions. One way of excluding a paradigmatic comparison between boundaries and evaluating listeners' responses to a structure in isolation would be a single-trial experiment (i. e., every participant responds to a single item, cf. Laurinavichyute & von der Malsburg 2022 on semantic and agreement attraction in sentence comprehension). Such a study could be implemented using the gated stimuli of study III. Each participant would respond to the seven gates of one coordinate either without or with internal grouping. Reliable prediction of the upcoming structure after hearing Name2 would support the idea that boundaries are processed relative to preceding material. Reliable prediction of the upcoming structure earlier in the utterance could mean that no previous knowledge about the speaker or similar structures is necessary to correctly understand a prosodically disambiguated structure.

Coming back to the starting point: Do we find clear prosodic disambiguation in another syntactically ambiguous structure? Lets come back to the starting point of this thesis: the exploration of the use of prosodic means to resolve structural disambiguities. So far, we discussed the specific case of internal grouping of coordinates. We observed structured variability in the use of three prosodic cues used to differentiate between two meanings (conditions) of the syntactically ambiguous coordinate. Despite gradual inter-individual variability, a categorical distinction in comprehension was possible and we generalised to a phonological category of grouping. We will now consider another structure with syntactic ambiguity: locally ambiguous sentences with SVO and OVS word order (study IV). Do we find a stable prosodic strategy for the prosodic disambiguation between SVO and OVS sentences that can be interpreted in comprehension and shows inter-individual variability in realisation? In study IV, speakers did not consequently use early prosodic means to distinguish between the two word order conditions. Only two out of 16 speakers produced different f0 contours on SVO and OVS sentences in the ambiguous part of the sentence. In comparison to the phonological category of grouping, which was realised consistently at the surface by all speakers, although with differences in the phonetic realisation, there seems to be no category for the prosodic disambiguation of locally ambiguous sentences with a case-ambiguous NP1. Further, in study IV, the inter-individual differences in how speakers

[51] Again, we are comparing between a right branching structure and a structure without internal grouping, but the results should be similar.

realised the sentences, were rather categorical than gradual. Speakers produced different types and numbers of pitch accents and not only different degrees of a feature (e. g., size of the f0 range). Overall, we cannot find a common set of cues that is varied in degree at the individual level, which makes the variability unstructured.

We can think of two possibilities why speakers did not produce clear prosodic distinctions between the two word order conditions. First, strictly speaking, there is no need for prosodic disambiguation in locally ambiguous sentences. The ambiguous role assignment of NP1 is disambiguated by the case-unambiguous NP2. If the presence of morpho-syntactic disambiguation is the reason for lacking prosodic disambiguation, we would expect speakers to prosodically distinguish between globally ambiguous sentences. This was tested in a pilot study with speaker-listener pairs, which is described in study IV. The results showed that speakers as well as listeners had difficulties with the OVS interpretation of globally ambiguous sentences. The vast majority of sentences were interpreted in SVO word order. Thus, the absence of morpho-syntactic cues did not provoke prosodic disambiguation.

The second reason might be the strong bias towards an SVO interpretation. SVO word order is overall more frequent than OVS word order in German (Bader & Häussler 2010), which is also reflected in a strong bias towards interpreting a case-ambiguous NP1 as the agent rather than the patient of an action (Weber et al. 2006, Hanne et al. 2015, among others). Possibly, speakers had difficulties in retrieving the OVS interpretation, which impeded them from marking it prosodically. This bias is not present in coordinates with different groupings, at least in our data. In the coordinated name sequences used in our studies, all possible adjacent name combinations were non-frequent (as assessed in the dlexDB corpora by Heister et al. 2011 and in printed sources covering the years 1500 to 2021 in an online-search using the Google Ngram Viewer by Lin et al. 2012). This means that there were no name combinations such as *Bonnie and Clyde* or *Hänsel and Gretel* that frequently go together. The non-frequency of name combinations in our material suggests the absence of a bias towards either a structure without or with internal grouping. The absence of clear prosodic distinction between the two word order conditions can be interpreted as a hint that there is no common prosodic category for the marking of OVS word order in German. Further, the ambiguity cannot be clearly represented in different tree structures. The ambiguous sentence beginning correspond to the same structure: an NP with a VP sister, with the NP differing in case marking between SVO and OVS word order.

Nevertheless, we found a distinct f0 pattern in the production data. The two word order conditions differed in height and position of the f0 peak produced on NP1 (with a higher and later peak on SVO than OVS sentences). This clear pattern raised the question of the existence of an abstract concept to distinguish between the two word orders. Such an abstract concept should be recoverable by listeners. To follow up this hypothesis, we tested listeners in a forced-choice decision task with new recordings of the same locally ambiguous sentences as in study IV, which reproduced the f0 distinction on NP1 described in study IV

(Schneider 2022). Sentences were recorded in two prosody conditions: a so called natural condition (based on the measures of the model speaker) and an exaggerated condition with an increased f0-range in comparison to the natural condition. Listeners were presented with the productions up to the disambiguating NP2 (i. e., only the ambiguous part of the sentence). The task was to select between an SVO and an OVS sentence ending (NP2 with accusative case disambiguating NP1 as agent for SVO reading and NP2 with nominative case disambiguating NP1 as patient for OVS reading) presented in written form on screen. Listeners received feedback in form of a happy or a sad smiley. Mean response accuracy was higher for SVO than for OVS structures and higher in the exaggerated than in the natural prosody condition. However, overall, mean response accuracy was below 66% in the natural prosody condition[52] and the sensitivity to discriminate between the two word order conditions (assessed with d'-analyses, Jang et al. 2009) was poor to moderate. The results of the study by Schneider (2022) showed no reliable assignment of the two f0 contours of the sentence beginnings to SVO and OVS interpretations.

With regard to the question of whether we find also a stable prosodic strategy for prosodic disambiguation between SVO and OVS word order, the weak performance in production when using f0 to differentiate between word order and the only moderate sensitivity in discriminating between the two f0 contours and mapping them to the word order conditions in comprehension do not support the tested f0 contours as realisations of prosodic categories for discriminating between SVO and OVS word order that would be common to German language users. As the absence of evidence is not evidence of absence, the results do not allow to reject the hypothesis of the existence of such a prosodic category for the marking of word order in general.

To sum up, we investigated prosodic means to resolve structural ambiguities. Based on inter-individual variability in the prosodic marking of a contrast between two interpretations of a syntactically ambiguous structure and the fact that the two interpretations could be inferred by listeners from individually varying concrete realisations we have postulated a phonological category of grouping. This was the case for coordinate name structures without or with internal grouping of the first two names, which were prosodically disambiguated by inter-individually varying combinations of f0-range, final lengthening, and pause. Grouping conditions could be recovered by naïve listeners. For locally ambiguous sentences in SVO and OVS word order, most speakers showed f0 contours varying in number and type of pitch accents, however not between word order conditions. The measurable difference between conditions present in the productions of one speaker was not clearly recovered by listeners. Our data provide evidence for an abstract concept of prosodic grouping but not of prosodic marking an OVS sentence as opposed to an SVO sentence.

[52] Mean response accuracy was calculated with the subject means of correct responses. For the exaggerated prosody condition mean response accuracy was below 72%.

10 Conclusion

Strings of words can correspond to more than one interpretation or underlying structure, which makes them ambiguous. A special case are coordinated sequences of more than two elements: The elements can be differently grouped within the sequence, which results in different unambiguous structures. This ambiguity can be resolved by prosodic means.

We found that different age groups of adult speakers reliably use fundamental frequency, final lengthening, and pause to prosodically disambiguate between two different conditions (without and with internal grouping of the first two names in a sequence of three names). Prosodic disambiguation builds up during the utterance and appears as a global phenomenon as opposed to a local phenomenon bound to the group edge. The first two names in a structure with internal grouping show prosodic cohesion between them (proximity) and the second name is prosodically set off from the third name (anti-proximity). Our findings support the existence of a phonological category of prosodic grouping that allows for individual variability at the phonetic realisation. Prosodic grouping has a relative character in that the boundary strength between elements in a group and at the group edge are determining but not the presence of a specific type of boundary. Regardless of the variability in the productions of individual speakers, naïve listeners were able to recover the different conditions in the productions of both age groups of speakers. Some listeners were able to predict the upcoming structure already after hearing the first name, before the point of syntactic ambiguity at the second name. We further found support for a strong link between prosody and syntax: speakers reliably produced prosodic disambiguation irrespective of the conversational situation. Our findings support models in favour of situational independence of disambiguating prosody. For another ambiguous structure that arises from case-ambiguity and allows for string-identical sentence beginnings of subject-verb-object and object-verb-subject word order, we did not find a clear prosodic pattern to resolve the local ambiguity.

To conclude, the chunking of elements into smaller groups is not restricted to prosodic grouping and can be found in different aspects of life including action sequences (Hilton et al. 2022) and music (Huron 2006, Jackendorff 2009). Chunking of complex processes into smaller groups facilitates processing (Frazier et al. 2006, Jackendorff 2009). This is not restricted to the field of prosody. It is a more general phenomenon. It remains for the future to investigate whether the observation of boundary weakening between elements that belong to the same group found in our studies on prosody can be transferred to grouping outside the domain of prosody. Huron (2006) discussed the notion of expectation in the context of music and describes a boundary as absence of expectation that evokes a sense of closure (Huron 2006: 157). The notion of expectation is comparable to the notion of cohesion described by Cutler et al. (1997) that results in the addition of a node to the current constituent (Cutler et al. 1997: 169).

References

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6). 716–723. doi:10.1109/tac.1974.1100705. [Cit. on p. 130]

Allbritton, David W., Gail McKoon & Roger Ratcliff. 1996. Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(3). 714–735. doi:10.1037/0278-7393.22.3.714. [Cit. on pp. 32, 35, 52, 54, and 57]

Allen, Micah, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marhall, Jordy van Langen & Rogier A. Kievit. 2019. Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research* 4. 63. doi:10.12688/wellcomeopenres.15191.2. [Cit. on p. 126]

Allopenna, Paul D., James S. Magnuson & Michael K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38(4). 419–439. doi:10.1006/jmla.1997.2558. [Cit. on pp. 122, 144, and 147]

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson & Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech* 34(4). 351–366. doi:10.1177/002383099103400404. [Cit. on p. 40]

Arvaniti, Amalia. 2019. Crosslinguistic variation, phonetic variability, and the formation of categories in intonation. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th ICPhS*, 1–6. Canberra, Australia: Australasian Speech Science and Technology Association Inc. [Cit. on pp. 16 and 26]

Baayen, Harald, Douglas Bates, Reinhold Kliegl & Shravan Vasishth. 2015. RePsychLing: Data sets from Psychology and Linguistics experiments. `https://github.com/dmbates/RePsychLing`. [Cit. on p. 67]

Baayen, Harald, Shravan Vasishth, Reinhold Kliegl & Douglas Bates. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206–234. doi:10.1016/j.jml.2016.11.006. [Cit. on pp. 16 and 155]

Bader, Markus & Jana Häussler. 2010. Word order in German: A corpus study. *Lingua* 3. 717–762. doi:10.1016/j.lingua.2009.05.007. [Cit. on pp. 147, 167, and 190]

Bard, Ellen Gurman, Anne H. Anderson, Catherine Sotillo, Matthew Aylett, Gwyneth Doherty-Sneddon & Alison Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language* 42. 1–22. doi:110.1006/jmla.1999.2667. [Cit. on pp. 183 and 184]

Barnes, Daniel R. 2013. *Age-related changes to the production of linguistic prosody*. Purdue, IN: Purdue University Master thesis. `https://docs.lib.purdue.edu/open_access_theses/17`. [Cit. on pp. 28, 93, 94, and 113]

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. doi:10.1016/j.jml.2012.11.001. [Cit. on p. 130]

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & Harald Baayen. 2015a. Parsimonious mixed models. *arXiv preprint* `https://arxiv.org/abs/1506.04967v2`. [Cit. on pp. 67 and 102]

Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker. 2015b. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi: 10.18637/jss.v067.i01. [Cit. on pp. 67 and 130]

Baumann, Stefan & Jürgen Trouvain. 2001. On the prosody of German telephone numbers. In *Seventh European Conference on Speech Communication and Technology*, `https://asa.isca-speech.org/archive_v0/archive_papers/eurospeech_2001/e01_0557.pdf`. [Cit. on p. 186]

Baumann, Stefan & Bodo Winter. 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics* 70. 20–38. doi: h10.1016/j.wocn.2018.05.004. [Cit. on pp. 121 and 143]

Beach, Cheryl M. 1991. The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language* 30(6). 644–663. doi:10.1016/0749-596X(91)90030-N. [Cit. on p. 122]

de Beer, Carola, Andrea Hofmann, Frank Regenbrecht, Clara Huttenlauch, Isabell Wartenburger, Hellmuth Obrig & Sandra Hanne. 2022. Production and comprehension of prosodic boundary marking in persons with unilateral brain lesions. *Journal of Speech, Language, and Hearing Research* 65(12). 4774–4796. doi:10.1044/2022_JSLHR-22-00258. [Cit. on p. 178]

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2). 145–204. [Cit. on pp. 52 and 57]

van den Berg, Rob, Carlos Gussenhoven & Toni Rietveld. 1992. Downstep in Dutch: Implications for a model. In Gerard J. Docherty & D. Robert Ladd (eds.), *Papers in laboratory phonology II: Gesture, segment, prosody*, vol. 335, 359. Cambridge: Cambridge University Press. [Cit. on p. 50]

Biersack, Sonja, Vera Kempe & Lorna Knapton. 2005. Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech. In *Proceedings of InterSpeech*, 2401–2404. Lisbon, Portugal. doi:10.21437/Interspeech.2005-46. `http://www.interspeech2005.org/`. [Cit. on pp. 2, 33, 54, 78, 95, and 96]

Bishop, Jason, Grace Kuo & Boram Kim. 2020. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics* 82. 100977. doi:10.1016/j.wocn.2020.100977. [Cit. on p. 121]

Blum-Kulka, Shoshana, Juliane House & Gabriele Kasper. 1989. Investigating cross-cultural pragmatics: An introductory overview. In Shoshana Blum-Kulka, Juliane House & Gabriele Kasper (eds.), *Cross-cultural pragmatics: Requests and apologies*, 1–34. Norwood, NJ: Ablex. [Cit. on pp. 27 and 32]

Boersma, Paul & David Weenink. 2017. Praat: Doing phonetics by computer. `www.praat.org`. [Cit. on pp. 60, 66, 90, 101, 124, and 154]

Bögel, Tina. 2015. *The syntax-prosody interface in lexical functional grammar*. Konstanz: University of Konstanz Doctoral thesis. `http://nbn-resolving.de/urn:nbn:de:bsz:`

352-0-403020. [Cit. on pp. 8, 11, 13, and 15]

Borràs-Comes, Joan, Rafèu Sichel-Bazin & Pilar Prieto. 2015. Vocative intonation preferences are sensitive to politeness factors. *Language and Speech* 58(1). 68–83. doi: 10.1177/0023830914565441. [Cit. on pp. 27, 28, and 183]

Braun, Bettina. 2006. Phonetics and phonology of thematic contrast in German. *Language and Speech* 49(4). 451–493. doi:10.1177/00238309060490040201. [Cit. on p. 163]

Brown, Meredith, Anne Pier Salverda, Laura C. Dilley & Michael K. Tanenhaus. 2011. Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review* 18(6). 1189–1196. doi:10.3758/s13423-011-0167-9. [Cit. on p. 119]

Bulgarelli, Federica & Elika Bergelson. 2022. Talker variability shapes early word representations in English-learning 8-month-olds. *Infancy* 27(2). 341–368. doi:10.1111/infa.12452. [Cit. on p. 1]

Burke, Deborah M. & Meredith A. Shafto. 2004. Aging and language production. *Current Directions in Psychological Science* 13(1). 21–24. doi:10.1111/j.0963-7214.2004.01301006.x. [Cit. on p. 92]

Bußmann, Hadumod. 2008. *Lexikon der Sprachwissenschaft [Encyclopaedia of Linguistics]*. Stuttgart: Kröner. [Cit. on pp. 6 and 7]

Cangemi, Francesco. 2009. Phonetic detail in intonation contour dynamics. In Stephan Schmid, Michael Schwarzenbach & Dieter Studer (eds.), *La dimensione temporale del parlato (Atti del V Convegno Nazionale AISV)*, 325–334. EDK Editore. [Cit. on p. 16]

Cangemi, Francesco. 2015. mausmooth. `http://ifl.phil-fak.uni-koeln.de/sites/linguistik/Phonetik/pdf-publications/2015/cangemi2015mausmooth.pdf`. [Cit. on pp. 155 and 165]

Cangemi, Francesco. 2016. *mausmooth: Eyeballing made easy* [Poster presentation]. 7th Conference on Tone and Intonation in Europe. `http://ifl.phil-fak.uni-koeln.de/sites/linguistik/Phonetik/pdf-publications/2015/cangemi2015mausmooth.pdf`. [Cit. on pp. 126 and 128]

Cangemi, Francesco, Martina Krüger & Martine Grice. 2015. Listener-specific perception of speaker-specific production in intonation. In Susanne Fuchs, Daniel Pape, Caterina Petrone & Pascal Perrier (eds.), *Individual differences in speech production and perception*, vol. 3 Speech Production and Perception, 123–145. Frankfurt am Main: Peter Lang. [Cit. on pp. 31, 34, 120, 122, 142, 143, 166, 175, and 176]

Chuang, Yu-Ying, Janice Fon & R. Harald Baayen. 2020. Analyzing phonetic data with generalized additive mixed models. *PsyArXiv preprint: psyarxiv.com/bd3r4* doi:10.31234/osf.io/bd3r4. [Cit. on pp. 16, 155, and 156]

Clifton, Charles, Jr., Katy Carlson & Lyn Frazier. 2002. Informative prosodic boundaries. *Language and Speech* 45(2). 87–114. doi:10.1177/00238309020450020101. [Cit. on pp. 29, 53, 77, 118, and 179]

Clifton, Charles, Jr., Katy Carlson & Lyn Frazier. 2006. Tracking the what and why of speakers' choices: Prosodic boundaries and the length of constituents. *Psychonomic Bulletin & Review* 13(5). 854–861. doi:10.3758/bf03194009. [Cit. on pp. 118, 119, and 144]

Cole, Jennifer. 2015. Prosody in context: A review. *Language, Cognition and Neuroscience*

30(1). 1–31. doi:10.1080/23273798.2014.963130. [Cit. on pp. 1, 4, 5, 11, 14, 16, 18, 19, 173, and 179]

Cole, Jennifer, Timothy Mahrt & Joseph Roy. 2017. Crowd-sourcing prosodic annotation. *Computer Speech and Language* 45. 300–325. doi:10.1016/j.csl.2017.02.008. [Cit. on pp. 120 and 142]

Cutler, Anne, Delphine Dahan & Wilma van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40(2). 141–201. doi:10.1177/00238309970400. [Cit. on pp. 4, 5, 21, 22, 26, 171, 179, 180, and 192]

Cutler, Anne & Stephen D. Isard. 1980. The production of prosody. In Brian Butterworth (ed.), *Language production*, vol. 1 Speech and talk, 245–269. London: Academic Press. [Cit. on pp. 4, 11, 20, 30, and 186]

Cutler, Anne & Takashi Otake. 1999. Pitch accent in spoken-word recognition in Japanese. *The Journal of the Acoustical Society of America* 105(3). 1877–1888. doi:10.1121/1.426724. [Cit. on p. 121]

Davis, Chris, Jeesun Kim, Katja Grauwinkel & Hansjörg Mixdorff. 2006. Lombard speech: Auditory (A), visual (V) and AV effects. In *Proceedings of Speech Prosody 2006*, 248–252. Dresden, Germany. [Cit. on pp. 34, 55, 80, and 183]

DePaulo, Bella M. & Lerita M. Coleman. 1986. Talking to children, foreigners, and retarded adults. *Journal of Personality and Social Psychology* 51(5). 945–959. doi:10.1037/0022-3514.51.5.945. [Cit. on pp. 2, 33, 54, 78, and 95]

Dilley, Laura C. & J. Devin McAuley. 2008. Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language* 59(3). 294–311. doi:10.1016/j.jml.2008.06.006. [Cit. on p. 119]

Dimitrova, Snezhina, Bistra Andreeva, Christoph Gabriel & Jonas Grünke. 2018. Speaker age effects on prosodic patterns in Bulgarian. In *Proceedings of Speech Prosody 2018*, 709–713. Poznán, Poland. doi:10.21437/SpeechProsody.2018-144. [Cit. on pp. 92, 93, and 113]

Dorokhova, Lydia & Mariapaola D'Imperio. 2019. Rise dynamics determines tune perception in French: The case of questions and continuations. In *Proceedings of the 19th ICPhS*, 691–695. Canberra, Australia: Australasian Speech Science and Technology Association Inc. [Cit. on p. 16]

Dromey, Christopher & Sarah Scott. 2016. The effects of noise on speech movements in young, middle-aged, and older adults. *Speech, Language and Hearing* 19(3). 131–139. doi:10.1080/2050571X.2015.1133757. [Cit. on p. 96]

Enzinna, Naomi & Sam Tilsen. 2019. Boards for automated referential communication task. Tech. rep. Cornell Working Papers in Phonetics and Phonology. doi:10.5281/zenodo.3724619. [Cit. on pp. 39 and 40]

Face, Timothy L. & Mariapaola D'Imperio. 2005. Reconsidering a focal typology: Evidence from Spanish and Italian. *Italian Journal of Linguistics / Rivista di linguistica* 17(2). 271–289. `https://hal.science/hal-00380688`. [Cit. on p. 120]

Féry, Caroline & Gerrit Kentner. 2010. The prosody of embedded coordinations in German and Hindi. In *Proceedings of Speech Prosody 2010*, Chicago, IL. [Cit. on pp. 11, 19, and 22]

Féry, Caroline & Frank Kügler. 2008. Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics* 36(4). 680–703. doi:10.1016/j.wocn.2008.05.001. [Cit. on p. 118]

Field, John. 2008. Revising segmentation hypotheses in first and second language listening. *System* 36(1). 35–51. doi:10.1016/j.system.2007.10.003. [Cit. on p. 122]

Folk, Laura & Florian Schiel. 2011. The Lombard effect in spontaneous dialog speech. In *Proceedings of InterSpeech*, 2701–2704. doi:10.21437/Interspeech.2011-690. [Cit. on pp. 34 and 55]

Fox, John & Sanford Weisberg. 2018. *An R companion to applied regression.* London: SAGE Publications Sage UK. [Cit. on p. 130]

Frazier, Lyn, Katy Carlson & Charles Clifton, Jr. 2006. Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences* 10(6). 244–249. doi:10.1016/j.tics.2006.04.002. [Cit. on pp. v, 50, 77, 179, 187, and 192]

Fuchs, Susanne, Daniel Pape, Caterina Petrone & Pascal Perrier. 2015. *Individual differences in speech production and perception*, vol. 3. Frankfurt am Main: Peter Lang. doi:10.3726/978-3-653-05777-5. [Cit. on p. 26]

Garnier, Maëva, Lucie Bailly, Marion Dohen, Pauline Welby & Hélène Lœvenbruck. 2006. An acoustic and articulatory study of Lombard speech: Global effects on the utterance. In *Proceedings of Interspeech 2006*, 2246–2249. Pittsburgh, PA. doi:10.21437/Interspeech.2006-323. [Cit. on pp. 34 and 55]

Gollrad, Anja. 2013. *Prosodic cue weighting in sentence comprehension.* Potsdam: University of Potsdam Doctoral thesis. `https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/8195/file/gollrad_diss.pdf`. [Cit. on pp. 118, 119, and 120]

Gollrad, Anja, Esther Sommerfeld & Frank Kügler. 2010. Prosodic cue weighting in disambiguation: Case ambiguity in German. In *Proceedings of Speech Prosody 2010*, Chicago, IL. [Cit. on pp. 13, 50, and 90]

Goy, Huiwen, David N. Fernandes, M. Kathleen Pichora-Fuller & Pascal van Lieshout. 2013. Normative voice data for younger and older adults. *Journal of Voice* 27(5). 545–555. doi:10.1016/j.jvoice.2013.03.002. [Cit. on p. 92]

Grice, Martine & Stefan Baumann. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 191. 267–298. [Cit. on pp. 15, 66, and 148]

Grice, Martine & Stefan Baumann. 2007. An introduction to intonation – functions and models. In Jürgen Trouvain & Ulrike Gut (eds.), *Non-native prosody: Phonetic description and teaching practice*, 25–52. Berlin: De Gruyter Mouton. doi:10.1515/9783110198751.1.25. [Cit. on pp. 4, 5, 11, 14, 15, 16, 17, 18, and 20]

Grice, Martine, Stefan Baumann & Ralf Benzmüller. 2005. German intonation in autosegmental-metrical phonology. In Sun-Ah Jun (ed.), *Prosodic typology: The phonology of intonation and phrasing*, 55–83. Oxford: Oxford University Press. [Cit. on pp. 15, 66, and 148]

Grosjean, François. 1980. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics* 28(4). 267–283. doi:10.3758/bf03204386. [Cit. on pp. 47 and 121]

Grosjean, François. 1983. *How long is the sentences? Prediction and prosody in the on-line* *processing of language*. Berlin: Walter de Gruyter. [Cit. on p. 121]

Grosjean, François. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes* 11(1–2). 107–134. doi:10.1080/016909696387231. [Cit. on p. 121]

Grosjean, François, Jean-Yves Dommergues, Etienne Cornu, Delphine Guillelmon & Carole Besson. 1994. The gender-marking effect in spoken word recognition. *Perception & Psychophysics* 56(5). 590–598. doi:10.3758/BF03206954. [Cit. on p. 47]

Grosjean, François & Cendrine Hirt. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes* 11(1-2). 107–134. doi:10.1080/016909696387231. [Cit. on p. 48]

Grünloh, Thomas, Elena Lieven & Michael Tomasello. 2011. German children use prosody to identify participant roles in transitive sentences. *Cognitive Linguistics* 22(2). 393–419. doi:10.1515/cogl.2011.015. [Cit. on pp. 148, 149, and 166]

Gussenhoven, Carlos. 2004. *The phonology of tone and intonation*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511616983. [Cit. on p. 5]

Hanne, Sandra, Frank Burchert, Ria De Bleser & Shravan Vasishth. 2015. Sentence comprehension and morphological cues in aphasia: What eye-tracking reveals about integration and prediction. *Journal of Neurolinguistics* 34. 83–111. doi:10.1016/j.jneuroling.2014.12.003. [Cit. on pp. 147, 152, and 190]

Hansen, Marie, Clara Huttenlauch, Carola de Beer, Isabell Wartenburger & Sandra Hanne. 2022. Individual differences in early disambiguation of prosodic grouping. *Language and Speech* 0(0). 1–28. doi:10.1177/00238309221127374. [Cit. on pp. 11, 91, and 117]

Harnsberger, James D., Rahul Shrivastav, W. S. Brown Jr., Howard Rothman & Harry Hollien. 2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice* 22(1). 58–69. doi:10.1016/j.jvoice.2006.07.004. [Cit. on pp. 93 and 113]

Häussler, Jana & Markus Bader. 2012. Grammar- versus frequency-driven syntactic ambiguity resolution: The case of double-object constructions. In Monique Lamers & Peter de Swart (eds.), *Case, word order and prominence: Interacting cues in language production and comprehension*, 273–301. Heidelberg: Springer Dordrecht. [Cit. on p. 13]

Hazan, Valerie & Rachel E. Baker. 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America* 130(4). 2139–2152. doi:10.1121/1.3623753. [Cit. on p. 40]

Hazan, Valerie, Outi Tuomainen, Jeesun Kim & Chris Davis. 2019. The effect of visual cues on speech characteristics of older and younger adults in an interactive task. In *Proceedings of the 19th ICPhS*, 815–819. Canberra, Australia: Australasian Speech Science and Technology Association Inc. [Cit. on pp. 92, 93, and 113]

Hazan, Valerie, Outi Tuomainen & Michèle Pettinato. 2016. Suprasegmental characteristics of spontaneous speech produced in good and challenging communicative conditions by talkers aged 9–14 years. *Journal of Speech, Language, and Hearing Research* 59(6). S1596–S1607. doi:10.1044/2016_JSLHR-S-15-0046. [Cit. on pp. 15, 128, 155, and 166]

Heister, Julian, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken & Reinhold Kliegl. 2011. dlexDB: Eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDB: A lexical database for the psychological and linguistic research]. *Psychologische Rundschau* 62(1). 10–20. doi:10.1026/0033-3042/a000029. [Cit. on pp. 60, 98, 123, 151, and 190]

Henry, Nick, Holger Hopp & Carrie N. Jackson. 2017. Cue additivity and adaptivity in predictive processing. *Language, Cognition and Neuroscience* 32(10). 1229–1249. doi: 10.1080/23273798.2017.1327080. [Cit. on pp. 21, 147, 148, 149, 166, 167, and 173]

Herrmann, Annika. 2016. Wortakzent und Intonation in Gebärdensprachen. In Ulrike Domahs & Beatrice Primus (eds.), *Handbuch Laut, Gebärde, Buchstabe*, 245–263. Berlin: De Gruyter. doi:10.1515/9783110295993-014. [Cit. on p. 4]

van Heuven, Vincent J. & Judith Haan. 2002. Temporal distribution of interrogativity markers in Dutch: A perceptual study. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology 7*, 61–86. Berlin, New York: De Gruyter Mouton. doi:10.1515/9783110197105.1.61. [Cit. on p. 120]

Hilton, Matt, Isabell Wartenburger, Julius Verrel & Birgit Elsner. 2022. Pre-boundary lengthening and pause signal boundaries in action sequences. In *Proceedings of the 44th Annual Conference of the Cognitive Sciences*, vol. 44, 1420–1426. https://escholarship.org/uc/item/5ks5h1h4. [Cit. on pp. 187 and 192]

Holzgrefe, Julia, Caroline Wellmann, Caterina Petrone, Hubert Truckenbrodt, Barbara Höhle & Isabell Wartenburger. 2013. Brain response to prosodic boundary cues depends on boundary position. *Frontiers in Psychology* 4. 421. doi:10.3389/fpsyg.2013.00421. [Cit. on pp. 17, 21, 180, 188, and 189]

Holzgrefe-Lang, Julia. 2017. *Prosodic phrase boundary perception in adults and infants*. Potsdam: University of Potsdam Doctoral thesis. https://publishup.uni-potsdam.de/frontdoor/index/index/docId/40594. [Cit. on pp. 11 and 50]

Holzgrefe-Lang, Julia, Caroline Wellmann, Barbara Höhle & Isabell Wartenburger. 2018. Infants' processing of prosodic cues: Electrophysiological evidence for boundary perception beyond pause detection. *Language and Speech* 61(1). 153—-169. doi:10.1177/0023830917730590. [Cit. on p. 180]

Holzgrefe-Lang, Julia, Caroline Wellmann, Caterina Petrone, Romy Räling, Hubert Truckenbrodt, Barbara Höhle & Isabell Wartenburger. 2016. How pitch change and final lengthening cue boundary perception in German: Converging evidence from ERPs and prosodic judgements. *Language, Cognition and Neuroscience* 31(7). 904–920. doi: 10.1080/23273798.2016.1157195. [Cit. on pp. 11, 17, 21, 60, 97, 119, 123, 177, 180, and 181]

Howell, David C., Marylène Rogier, Vincent Yzerbyt & Yves Bestgen. 1998. *Statistical methods in human sciences*. New York, NY: Wadsworth. [Cit. on pp. 47, 64, and 100]

Huettig, Falk, Joost Rommers & Antje S. Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica* 137(2). 151–171. doi:10.1016/j.actpsy.2010.11.003. [Cit. on p. 147]

Hughes, Rebecca & Beatrice Szczepek Reed. 2011. Learning about speech by experiment: Issues in the investigation of spontaneous talk within the experimental research paradigm. *Applied Linguistics* 32(2). 197–214. doi:10.1093/applin/amq044. [Cit. on p. 121]

Huron, David. 2006. *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: The MIT Press. [Cit. on pp. v, 187, and 192]

Huttenlauch, Clara. 2016. *The purpose shapes the vocative: Prosodic analysis of productions of Colombian Spanish*. University of Konstanz: Department of Linguistics Master thesis. [Cit. on p. 183]

Huttenlauch, Clara, Carola de Beer, Sandra Hanne & Isabell Wartenburger. 2021. Production of prosodic cues in coordinate name sequences addressing varying interlocutors. *Laboratory Phonology* 12(1). 1–31. doi:10.5334/labphon.221. [Cit. on pp. 11, 22, 33, 34, 49, 89, 90, 91, 92, 94, 95, 96, 97, 98, 100, 101, 102, 118, 119, 120, 122, 123, 124, 126, 129, 132, 142, and 143]

Huttenlauch, Clara, Marie Hansen, Carola de Beer, Sandra Hanne & Isabell Wartenburger. 2023. Age effects on linguistic prosody in coordinates produced to varying interlocutors: Comparison of younger and older speakers. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, vol. 12, 157–192. Berlin: Language Science Press. doi:10.5281/zenodo.7777534. [Cit. on pp. 11 and 89]

Huttenlauch, Clara, Marie Hansen, Carola de Beer, Isabell Wartenburger & Sandra Hanne. 2022. Individual variability in prosodic marking of locally ambiguous sentences. In *Proceedings of Speech Prosody 2022*, 165–169. Lisbon, Portugal. doi:10.21437/SpeechProsody.2022-34. [Cit. on pp. 12 and 146]

Hwang, Jiwon, Susan E. Brennan & Marie K. Huffman. 2015. Phonetic adaptation in non-native spoken dialogue: Effects of priming and audience design. *Journal of Memory and Language* 81. 72–90. doi:10.1016/j.jml.2015.01.001. [Cit. on p. 40]

Höhle, Barbara, Tom Fritzsche, Natalie Boll-Avetisyan, Marc A. Hullebus & Adamantios Gafos. 2021. Respect the surroundings: Effects of phonetic context variability on infants' learning of minimal pairs. *Journal of the Acoustical Society of America: Express Letters* 1(2). 024401. doi:10.1121/10.0003574. [Cit. on p. 1]

Höhle, Barbara, Tom Fritzsche, Katharina Meß, Mareike Philipp & Adamantios Gafos. 2020. Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Developmental Science* 23(5). e12950. doi:10.1111/desc.12950. [Cit. on p. 1]

Ip, Martin Ho Kwan & Anne Cutler. 2022. Juncture prosody across languages: Similar production but dissimilar perception. *Laboratory Phonology* 13(1). 1–49. doi:10.16995/labphon.6464. [Cit. on p. 11]

Jackendorff, Ray. 2009. Parallels and nonparallels between language and music. *Music perception* 26(3). 195–204. doi:10.1525/mp.2009.26.3.195. [Cit. on pp. v, 187, and 192]

Jang, Yoonhee, John T. Wixted & David E. Huber. 2009. Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General* 138(2). 291–306. [Cit. on p. 191]

Jannedy, Stefanie & Melanie Weirich. 2014. Sound change in an urban setting: Category instability of the palatal fricative in Berlin. *Laboratory Phonology* 5(1). 91–122. doi:10.1515/lp-2014-0005. [Cit. on p. 32]

Jessen, Michael, Olaf Köster & Stefan Gfroerer. 2003. Effect of increased vocal effort on average and range of fundamental frequency in a sample of 100 German-speaking male subjects. In *Proceedings of the 15th ICPhS*, 1623–1626. Barcelona, Spain. https://ww w.internationalphoneticassociation.org/icphs/icphs2003. [Cit. on pp. 34, 55, and 96]

Jun, Sun-Ah. 2003. Prosodic phrasing and attachment preferences. *Journal of Psycholinguistic Research* 32(2). 219–249. doi:10.1023/A:1022452408944. [Cit. on p. 119]

Jun, Sun-Ah & Jason Bishop. 2015. Priming implicit prosody: Prosodic boundaries and individual differences. *Language and Speech* 58(4). 459–473. doi:10.1177/0023830914563368. [Cit. on p. 121]

Junqua, Jean-Claude. 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America* 93(1). 510–524. doi:10.1121/1.405631. [Cit. on pp. 34 and 55]

Junqua, Jean-Claude. 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication* 20(1). 13–22. doi:10.1016/S0167-6393(96)00041-6. [Cit. on pp. 34, 55, 80, 96, and 183]

Kamide, Yuki, Christoph Scheepers & Gerry T. M. Altmann. 2003. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research* 32. 37–55. doi:10.1023/A:1021933015362. [Cit. on p. 147]

Kempe, Vera, Sonja Schaeffler & John C. Thoresen. 2010. Prosodic disambiguation in child-directed speech. *Journal of Memory and Language* 62(2). 204–225. doi:10.1016/j.jml.2009.11.006. [Cit. on pp. 33 and 54]

Kemper, Susan, Patrice Ferrell, Tamara Harden, Andrea Finter-Urczyk & Catherine Billington. 1998. Use of elderspeak by young and older adults to impaired and unimpaired listeners. *Aging, Neuropsychology, and Cognition* 5(1). 43–55. doi:10.1076/anec.5.1.43.22. [Cit. on pp. 78, 79, 92, and 93]

Kemper, Susan, Dixie Vandeputte, Karla Rice, Him Cheung & Julia Gubarchuk. 1995. Speech adjustments to aging during a referential communication task. *Journal of Language and Social Psychology* 14(1). 40–59. doi:10.1177/0261927X95141003. [Cit. on pp. 2, 33, 54, 78, 93, 95, and 113]

Kentner, Gerrit & Caroline Féry. 2013. A new approach to prosodic grouping. *The Linguistic Review* 30(2). 277–311. doi:10.1515/tlr-2013-0009. [Cit. on pp. iv, 3, 9, 11, 14, 17, 19, 22, 23, 25, 36, 49, 50, 55, 56, 58, 77, 81, 89, 90, 113, 118, 119, 120, 142, 173, 177, 179, and 187]

Kentner, Gerrit, Isabelle Franz, Christine A. Knoop & Winfried Menninghaus. 2023. The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase. *Journal of Phonetics* 97. 101225. doi:10.1016/j.wocn.2023.101225. [Cit. on p. 19]

Kerkhofs, Roel, Wietske Vonk, Herbert Schriefers & Dorothee J. Chwilla. 2008. Sentence processing in the visual and auditory modality: Do comma and prosodic break have parallel functions? *Brain Research* 1224. 102–118. doi:10.1016/j.brainres.2008.05.034. [Cit. on p. 5]

Kim, Jiseung. 2019. Individual differences in the production of prosodic boundaries in American English. *The Journal of the Acoustical Society of America* 145(3). 1933. doi:10.1121/1.5102039. [Cit. on p. 120]

Kim, Jiseung. 2020. *Individual differences in the production and perception of prosodic* *boundaries in American English.* University of Michigan: College of Literature, Science, and the Arts Doctoral dissertation. `https://deepblue.lib.umich.edu/bitstream/handle/2027.42/162927/jiseungk_1.pdf?sequence=1`. [Cit. on p. 120]

Kimball, Amelia E. & Jennifer Cole. 2016. Pitch contour shape matters in memory. In *Proceedings of Speech Prosody 2016*, 1171–1175. Boston, MA. doi:10.21437/SpeechProsody.2016-241. [Cit. on p. 17]

Knoll, Monja & Lisa Scharrer. 2007. Acoustic and affective comparisons of natural and imaginary infant-, foreigner- and adult-directed speech. In *Proceedings of Interspeech*, 1414–1417. Antwerp, Belgium. doi:10.21437/Interspeech.2007-29. [Cit. on p. 95]

Knoll, Monja, Lisa Scharrer & Alan Costall. 2011. "Look at the shark": Evaluation of student- and actress-produced standardised sentences of infant- and foreigner-directed speech. *Speech Communication* 53(1). 12–22. doi:10.1016/j.specom.2010.08.004. [Cit. on p. 95]

Knoll, Monja Angelika, Melissa Johnstone & Charlene Blakely. 2015. Can you hear me? Acoustic modifications in speech directed to foreigners and hearing-impaired people. In *Proceedings of InterSpeech*, 2987–2990. Dresden, Germany. doi:10.21437/Interspeech.2015-618. [Cit. on pp. 96 and 182]

Kohler, Klaus J. 1983. Prosodic boundary signals in German. *Phonetica* 40(2). 89–134. doi:10.1159/000261685. [Cit. on pp. 18 and 19]

Kraljic, Tanya & Susan E. Brennan. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology* 50(2). 194–231. doi:10.1016/j.cogpsych.2004.08.002. [Cit. on pp. 35, 52, 53, 57, 89, 95, 114, 117, 182, and 183]

Kuzla, Claudia & Mirjam Ernestus. 2011. Prosodic conditioning of phonetic detail in German plosives. *Journal of Phonetics* 39(2). 143–155. doi:10.1016/j.wocn.2011.01.001. [Cit. on pp. 162 and 167]

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojensen Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. doi:10.18637/jss.v082.i13. `http://CRAN.R-project.org/package=lmerTest`. [Cit. on p. 67]

Ladd, D. Robert. 1986. Intonational phrasing: The case for recursive prosodic structure. *Phonology* 3. 311–340. `https://www.jstor.org/stable/4615402`. [Cit. on p. 50]

Ladd, D. Robert. 2008. *Intonational phonology.* Cambridge: Cambridge University Press. 2nd ed. [Cit. on pp. 15, 66, 121, 148, and 186]

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to field work and instrumental techniques* chap. Acoustic analysis of phonation types, 169–181. Oxford: Blackwell Publishing. [Cit. on p. 5]

Landgraf, Rabea, Gerhard Schmidt, Johannes Köhler-Kaeß, Oliver Niebuhr & Tina John. 2017. More noise, less talk: The impact of driving noise and in-car communication systems on acoustic-prosodic parameters in dialogue. In *43. Deutsche Jahrestagung für Akustik (DAGA)*, 1485–1488. Kiel. [Cit. on pp. 34, 55, 80, and 183]

Lange, Kristian, Simone Kühn & Elisa Filevich. 2015. "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS one* 10(6). e0130834. doi:10.1371/journal.pone.0130834. [Cit. on p. 100]

Laurinavichyute, Anna & Titus von der Malsburg. 2022. Semantic attraction in sentence comprehension. *Cognitive Science* 46(2). e13086. doi:10.1111/cogs.13086. [Cit. on p. 189]

Lehiste, Ilse. 1973a. Phonetic disambiguation of syntactic ambiguity. *Glossa* 7(2). 107–121. [Cit. on pp. 9 and 14]

Lehiste, Ilse. 1973b. Rhythmic units and syntactic units in production and perception. *The Journal of the Acoustical Society of America* 54(5). 1228–1234. doi:10.1121/1.1914379. [Cit. on pp. 6, 9, 20, 21, 26, 30, 54, 154, and 177]

Lehiste, Ilse, Joseph P. Olive & Lynn A. Streeter. 1976. Role of duration in disambiguating syntactically ambiguous sentences. *The Journal of the Acoustical Society of America* 60(5). 1199–1202. doi:10.1121/1.381180. [Cit. on pp. 7, 9, and 14]

Leinonen, Eeva & Carolyn Letts. 1997. Referential communication tasks: Performance by normal and pragmatically impaired children. *European Journal of Disorders of Communication* 32(2s). 53–65. doi:10.1111/j.1460-6984.1997.tb01624.x. [Cit. on p. 39]

Li, W. & Y. Yang. 2009. Perception of prosodic hierarchical boundaries in Mandarin Chinese sentences. *Neuroscience* 158(4). 1416–1425. doi:10.1016/j.neuroscience.2008.10.065. [Cit. on p. 119]

Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman & Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics ACL*, 169–174. Jeju, Republic of Korea. [Cit. on pp. 98, 123, and 190]

Lindgren, Astrid. 1988. *Wir Kinder aus Bullerbü.* Hamburg: Oetinger. Illustrated by Ilon Wikland, translated by Else von Hollander-Lossow. [Cit. on p. 11]

Lombard, Étienne. 1911. Le signe de l'élévation de la voix. *Annales des Maladies de l'Oreille et du Larynx* 37. 101–109. [Cit. on pp. 55 and 96]

Lortie, Catherine L., Mélanie Thibeault, Matthieu J. Guitton & Pascale Tremblay. 2015. Effects of age on the amplitude, frequency and perceived quality of voice. *AGE* 37(6). 1–24. doi:10.1007/s11357-015-9854-1. [Cit. on pp. 92, 113, and 115]

Lu, Youyi & Martin Cooke. 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America* 124(5). 3261–3275. doi:10.1121/1.2990705. [Cit. on pp. 34, 55, 80, and 183]

Lüdeling, Anke, Artemis Alexiadou, Aria Adli, Karin Donhauser, Malte Dreyer, Markus Egg, Anna Helene Feulner, Natalia Gagarina, Wolfgang Hock, Stefanie Jannedy, Frank Kammerzell, Pia Knoeferle, Thomas Krause, Manfred Krifka, Silvia Kutscher, Beate Lütke, Thomas McFadden, Roland Meyer, Christine Mooshammer, Stefan Müller, Katja Maquate, Muriel Norde, Uli Sauerland, Stephanie Solt, Luka Szucsich, Elisabeth Verhoeven, Richard Waltereit, Anne Wolfsgruber & Lars Erik Zeige. 2022. Register: Language users' knowledge of situational-functional variation. *Register Aspects of Language in Situation* 1(1). 1–58. doi:10.18452/24901. [Cit. on p. 32]

Luo, Jinhong, Steffen R. Hage & Cynthia F. Moss. 2018. The Lombard effect: From acoustics to neural mechanisms. *Trends in Neurosciences* 41(12). 938–949. doi:10.1016/j.tins.2018.07.011. [Cit. on pp. 34 and 55]

Mann, Henry B. & Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*

18(1). 50–60. doi:10.1214/aoms/1177730491. [Cit. on p. 68]

Markó, Alexandra & Judit Bóna. 2010. Fundamental frequency patterns: The factors of age and speech type. In *Proceedings of the Workshop "Sociophonetics, at the crossroads of speech variation, processing and communication"*, 45–48. Pisa, Italy. [Cit. on p. 92]

Mathôt, Sebastiaan, Daniel Schreij & Jan Theeuwes. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44(2). 314–324. doi:10.3758/s13428-011-0168-7. [Cit. on pp. 100 and 129]

MATLAB. 2019. Version 9.6.0.1135713 Update 3. The MathWorks Inc. [Cit. on p. 68]

McGraw, Kenneth O. & SP Wong. 1992. A common language effect size statistic. *Psychological Bulletin* 111(2). 361–365. doi:10.1037/0033-2909.111.2.361. [Cit. on p. 68]

Morrill, Tuuli H., Laura C. Dilley & J. Devin McAuley. 2014. Prosodic patterning in distal speech context: Effects of list intonation and f0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics* 46. 68–85. doi:10.1016/j.wocn.2014.06.001. [Cit. on pp. 180 and 185]

Napoleão de Souza, Ricardo. 2023. Segmental cues to IP-initial boundaries: Data from English, Spanish, and Portuguese. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, vol. 12, 35–86. Berlin: Language Science Press. doi:10.5281/zenodo.7777528. [Cit. on p. 14]

Nasreddine, Ziad S., Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings & Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 53(4). 695–699. doi:10.1111/j.1532-5415.2005.53221.x. [Cit. on p. 97]

Nespor, Marina & Irene Vogel. 1986. *Prosodic phonology.* Dordrecht: Foris. [Cit. on pp. 29, 50, and 53]

Neurobehavioural Systems. 2018. Version 20.1. `https://www.neurobs.com/`. [Cit. on pp. 99 and 154]

Niebuhr, Oliver, Mariapaola D'Imperio, Barbara Gili Fivela & Francesco Cangemi. 2011. Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category. In *Proceedings of the 17th ICPhS*, 120–123. [Cit. on pp. 30, 31, and 176]

O'Brien, Mary Grantham, Carrie N. Jackson & Alison Eisel Hendricks. 2013. Making use of cues to sentence length in L1 and L2 German. *Linguistic Approaches to Bilingualism* 3(4). 448–477. doi:110.1075/lab.3.4.03obr. [Cit. on p. 121]

Olusanya, Bolajoko O., Adrian C. Davis & Howard J. Hoffman. 2019. Hearing loss grades and the international classification of functioning, disability and health. *Bulletin of the World Health Organization* 97(10). 725–728. doi:10.2471/BLT.19.230367. [Cit. on pp. 97 and 150]

van Ommen, Sandrien, Natalie Boll-Avetisyan, Saioa Larraza, Caroline Wellmann, Ranka Bijeljac-Babic, Barbara Höhle & Thierry Nazzi. 2020. Language-specific prosodic acquisition: A comparison of phrase boundary perception by French- and German-learning infants. *Journal of Memory and Language* 112. 104108. doi:10.1016/j.jml.2020.104108. [Cit. on pp. 17, 18, and 21]

Ots, Nele & Piia Taremaa. 2023. Chunking an unfamiliar language: Results from a perception study of German listeners. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, vol. 12, 87–117. Berlin: Language Science Press. doi:10.5281/zenodo.7777530. [Cit. on p. 179]

Ouyang, Iris Chuoying & Elsi Kaiser. 2015. Individual differences in the prosodic encoding of informativity. In Susanne Fuchs, Daniel Pape, Caterina Petrone & Pascal Perrier (eds.), *Individual differences in speech production and perception*, vol. 3 Speech Production and Perception, 147–188. Frankfurt am Main: Peter Lang. [Cit. on pp. 26, 29, 30, and 174]

Paschen, Ludger, Susanne Fuchs & Frank Seifart. 2022. Final lengthening and vowel length in 25 languages. *Journal of Phonetics* 94. 101179. doi:10.1016/j.wocn.2022.101179. [Cit. on p. 18]

Peppé, Sue, Jane Maxim & Bill Wells. 2000. Prosodic variation in Southern British English. *Language and Speech* 43(3). 309–334. doi:10.1177/00238309000430030501. [Cit. on pp. 10, 11, 26, 27, 28, 29, 31, 32, 38, 46, 174, 177, 179, 186, and 188]

Peppé, Sue J. E. 2009. Why is prosody in speech-language pathology so difficult? *International Journal of Speech-Language Pathology* 11(4). 258–271. doi:10.1080/17549500902906339. [Cit. on pp. 4 and 5]

Pescuma, Valentina N., Dina Serova, Julia Lukassek, Antje Sauermann, Roland Schäfer, Aria Adli, Felix Bildhauer, Markus Egg, Kristina Hülk, Aine Ito, Stefanie Jannedy, Valia Kordoni, Milena Kuehnast, Silvia Kutscher, Robert Lange, Nico Lehmann, Mingya Liu, Beate Lütke, Katja Maquate, Christine Mooshammer, Vahid Mortezapour, Stefan Müller, Muriel Norde, Elizabeth Pankratz, Angela G. Patarroyo, Ana-Maria Pleşca, Camilo G. Ronderos, Stephanie Rotter, Uli Sauerland, Gohar Schnelle, Britta Schulte, Gediminas Schüppenhauer, Bianca Maria Sell, Stephanie Solt, Megumi Terada, Dimitra Tsiapou, Elisabeth Verhoeven, Melanie Weirich, Heike Wiese, Kathy Zaruba, Lars Erik Zeige, Anke Lüdeling & Pia Knoeferle. 2023. Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods. *Frontiers in Psychology* 13. 1–31. doi:10.3389/fpsyg.2022.964658. [Cit. on pp. 32, 33, and 184]

Peters, Benno. 2006. *Form und Funktion prosodischer Grenzen im Gespräch [Form and function of prosodic boundaries in conversation]*. Kiel: Christian-Albrechts-Universität zu Kiel Doctoral thesis. https://macau.uni-kiel.de/receive/diss_mods_00002078. [Cit. on p. 19]

Peters, Benno, Klaus J. Kohler & Thomas Wesener. 2005. Phonetische Merkmale prosodischer Phrasierung in Deutscher Spontansprache [Phonetic cues of prosodic phrasing in German spontaneous speech]. In Klaus J. Kohler, Felicitas Kleber & Benno Peters (eds.), *Prosodic structures in German spontaneous speech*, vol. 35, 143–184. AIPUK, IPDS Kiel. [Cit. on pp. 17, 19, 20, 21, 50, 56, 90, 118, and 120]

Petrone, Caterina & Oliver Niebuhr. 2014. On the intonation of German intonation questions: The role of the prenuclear region. *Language and Speech* 57(1). 108–146. doi:10.1177/0023830913495651. [Cit. on pp. 48, 120, and 122]

Petrone, Caterina, Hubert Truckenbrodt, Caroline Wellmann, Julia Holzgrefe-Lang, Isabell Wartenburger & Barbara Höhle. 2017. Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. *Journal of Phonetics* 61. 71–92.

doi:10.1016/j.wocn.2017.01.002. [Cit. on pp. 11, 14, 17, 18, 19, 20, 24, 25, 31, 50, 53, 54, 56, 90, 101, 118, 119, 120, 173, 177, and 179]

Pfau, Roland & Josep Quer. 2010. Nonmanuals: Their grammatical and prosodic roles. In Diane Brentari (ed.), *Sign languages*, 381–402. Cambridge: Cambridge University Press. [Cit. on p. 4]

Piazza, Giorgio, Clara D. Martin & Marina Kalashnikova. 2021. The acoustic features and didactic function of foreigner directed speech: A literature review. *PsyArXiv Preprints* doi:10.31234/osf.io/scnke. [Cit. on pp. 2 and 96]

Pierre, Thomas St., Katherine S. White & Elizabeth K. Johnson. 2023. Who is running our experiments? The influence of experimenter identity in the marshmallow task. *Cognitive Development* 65. doi:10.1016/j.cogdev.2022.101271. [Cit. on p. 39]

Price, Patti J., Mari Ostendorf, Stefanie Shattuck-Hufnagel & Cynthia Fong. 1991. The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America* 90(6). 2956–2970. doi:10.1121/1.401770. [Cit. on pp. 7, 21, and 175]

R Development Core Team. 2018. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`. [Cit. on pp. 66, 102, 130, 155, and 164]

Randolph, Justus J. 2005. *Free-marginal multirater kappa (multirater κ_{free}): An alternative to Fleiss' fixed-marginal multirater kappa* [Paper presentation]. In *Proceedings of the joensuu learning and instruction symposium*, Joensuu, Finland. `https://files.eric.ed.gov/fulltext/ED490661.pdf`. [Cit. on p. 132]

Reich, Uli & Martin Rohrmeier. 2014. Batidas latinas: On rhythm and meter in Spanish and Portuguese and other forms of music. In Javier Caro Reina & Renata Szczepaniak (eds.), *Syllable and word languages*, 391–420. Berlin: Walter de Gruyter. doi:10.1515/9783110346992.391. [Cit. on p. 187]

van Rij, Jacolien, Martijn Wieling, R. Harald Baayen & Hedderik van Rijn. 2020. itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 2.4. [Cit. on p. 156]

Rodriguez-Cuadrado, Sara, Cristina Baus & Albert Costa. 2018. Foreigner talk through word reduction in native/non-native spoken interactions. *Bilingualism: Language and Cognition* 21(2). 419–426. doi:10.1017/S1366728917000402. [Cit. on p. 96]

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology* 10(1). 1–27. doi:10.5334/labphon.147. [Cit. on pp. 14, 15, 20, 46, and 156]

Rojas, Sandra, Elaina Kefalianos & Adam Vogel. 2020. How does our voice change as we age? A systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research* 63(2). 533–551. doi:10.1044/2019_JSLHR-19-00099. [Cit. on p. 92]

Rost, Gwyneth C. & Bob McMurray. 2009. Speaker variability augments phonological processing in early word learning. *Developmental Science* 12(2). 339–349. doi:10.1111/j.1467-7687.2008.00786.x. [Cit. on p. 1]

Rost, Gwyneth C. & Bob McMurray. 2010. Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy* 15(6). 608–635. doi:10.1111/j.1532-7078.2010.00033.x. [Cit. on p. 1]

Roy, Joseph, Jennifer Cole & Timothy Mahrt. 2017. Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology* 8(1). 1–36. doi:10.5334/labphon.108. [Cit. on pp. 20, 31, 32, 121, and 142]

Ryan, Ellen Bouchard, Mary Lee Hummert & Linda H. Boich. 1995. Communication predicaments of aging: Patronizing behavior toward older adults. *Journal of Language and Social Psychology* 14(1). 144–166. doi:10.1177/0261927X95141008. [Cit. on p. 78]

Santos, Aline Oliveira, Juliana Godoy, Kelly Silverio & Alcione Brasolotto. 2021. Vocal changes of men and women from different age decades: An analysis from 30 years of age. *Journal of Voice* doi:10.1016/j.jvoice.2021.06.003. [Cit. on pp. 92 and 113]

Schad, Daniel J., Sven Hohenstein, Shravan Vasishth & Reinhold Kliegl. 2018. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *arXiv preprint arXiv:1807.10451* . [Cit. on p. 67]

Schafer, Amy J., Shari R. Speer, Paul Warren & S. David White. 2000. Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research* 29(2). 169–182. doi:10.1023/A:1005192911512. [Cit. on pp. 35, 52, 57, 89, 95, 114, 180, 182, 183, and 185]

Schneider, Kathleen. 2022. *Variability in prosodic cue comprehension: Local ambiguity resolution in German SVO and OVS sentences*. Humboldt-University zu Berlin: Berlin School of Mind and Brain Master thesis. [Cit. on p. 191]

Schubö, Fabian & Sabine Zerbian. 2023. The patterns of pre-boundary lengthening in German. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, vol. 12, 1–34. Berlin: Language Science Press. doi:10.5281/zenodo.7777526. [Cit. on pp. 18, 19, and 90]

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2). 461–464. doi:10.1214/aos/1176344136. [Cit. on p. 130]

Scukanec, Gail P., Linda Petrosino & Roger D. Colcord. 1996. Age-related differences in acoustical aspects of contrastive stress in women. *Folia Phoniatrica et Logopaedica* 48(5). 231–239. doi:10.1159/000266414. [Cit. on pp. 28, 93, and 113]

Scukanec, Gail P., Linda Petrosino & Michael P. Rastatter. 1992. Fundamental frequency variability in elderly women during production of stressed and unstressed words. *Perceptual and Motor Skills* 74(3_suppl). 1091–1095. doi:10.2466/pms.1992.74.3c.1091. [Cit. on pp. 92 and 113]

Seidl, Amanda, Kristine H. Onishi & Alejandrina Cristia. 2014. Talker variation aids young infants' phonotactic learning. *Language Learning and Development* 10(4). 297–307. doi:10.1080/15475441.2013.858575. [Cit. on p. 1]

Siegel, Sidney. 1956. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill. [Cit. on pp. 162 and 164]

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert & Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Proceedings of ICSLP*, 867–870. Banff, Canada. doi:10.21437/ICSLP.1992-260. [Cit. on pp. 15, 66, and 121]

Smiljanic, Rajka & Rachael C. Gilbert. 2017. Acoustics of clear and noise-adapted speech in children, young, and older adults. *Journal of Speech, Language, and Hearing Research*

60(11). 3081–3096. doi:10.1044/2017_JSLHR-S-16-0130. [Cit. on pp. 2, 92, 93, 96, and 113]

Smith, Caroline L. 2007. Prosodic accommodation by French speakers to a non-native interlocutor. In *Proceedings of the 16th ICPhS*, 1081–1084. Saarbrücken, Germany. [Cit. on pp. 2, 33, 54, 78, and 96]

Snedeker, Jesse & John Trueswell. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language* 48(1). 103–130. doi: 10.1016/S0749-596X(02)00519-3. [Cit. on pp. 35, 52, 53, and 57]

Sóskuthy, Márton. 2017. Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339* . [Cit. on pp. 16, 155, and 156]

Sóskuthy, Márton. 2021. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics* 84. 101017. doi:10.1016/j.wocn.2020.101017. [Cit. on pp. 16, 155, and 156]

Speer, Shari R., Paul Warren & Amy J. Schafer. 2011. Situationally independent prosodic phrasing. *Laboratory Phonology* 2(1). 35–98. doi:10.1515/LABPHON.2011.002. [Cit. on pp. 29, 35, 50, 51, 52, 53, 57, 89, 95, 114, 117, 182, and 183]

Stanton, Bill J., Leah H. Jamieson & George D. Allen. 1988. Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 331–334. New York, NY: IEEE. doi:10.1109/ICASSP.1988.196583. [Cit. on pp. 34 and 55]

Steinhauer, Karsten, Kai Alter & Angela D. Friederici. 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience* 2(2). 191–196. doi:10.1038/5757. [Cit. on pp. 21 and 117]

Šturm, Pavel & Jan Volín. 2023. Occurrence and duration of pauses in relation to speech tempo and structural organization in two speech genres. *Languages* 8(1). 23. doi:10.3390/languages8010023. [Cit. on p. 19]

van Summers, W., David B. Pisoni, Robert H. Bernacki, Robert I. Pedlow & Michael A. Stokes. 1988. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America* 84(3). 917–928. doi:10.1121/1.396660. [Cit. on pp. 2, 34, 55, 80, 96, and 183]

Swerts, Marc & Ronald Geluykens. 1993. The prosody of information units in spontaneous monologue. *Phonetica* 50(3). 189–196. doi:10.1159/000261939. [Cit. on p. 118]

Swerts, Marc & Ronald Geluykens. 1994. Prosody as a marker of information flow in spoken discourse. *Language and Speech* 37(1). 21–43. doi:10.1177/002383099403700102. [Cit. on p. 143]

Taglicht, Josef. 1998. Constraints on intonational phrasing in English. *Journal of Linguistics* 34(1). 181–211. doi:10.1017/S0022226797006877. [Cit. on pp. 22, 23, 55, 56, and 77]

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard & Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217). 1632–1634. doi:10.1126/science.7777863. [Cit. on p. 147]

Tauber, Sarah K., Lori E. James & Paula M. Noble. 2010. The effects of age on using prosody to convey meaning and on judging communicative effectiveness. *Psychology and Aging* 25(3). 702–707. doi:10.1037/a0019266. [Cit. on pp. 28, 93, and 94]

Thimm, Caja, Ute Rademacher & Lenelis Kruse. 1998. Age stereotypes and patronizing messages: Features of age-adapted speech in technical instructions to the elderly. *Journal of Applied Communication Research* 26(1). 66–82. doi:10.1080/00909889809365492. [Cit. on pp. 2, 33, 54, 78, and 95]

Torrey, Cristen, Susan R. Fussell & Sara B. Kiesler. 2005. Appropriate accommodations: Speech technologies and the needs of older adults. In *Caring machines: AI in eldercare: Papers from the AAAI fall symposium*, 99. Arlington, VA. [Cit. on p. 78]

Trouvain, Jürgen, Camille Fauth & Bernd Möbius. 2016. Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech. In *Proceedings of Speech Prosody 2016*, 31–35. Boston, MA. doi:10.21437/SpeechProsody.2016-7. [Cit. on p. 19]

Trouvain, Jürgen & Martine Grice. 1999. The effect of tempo on prosodic structure. In *Proceedings of the 14th ICPhS*, 1067–1070. San Francisco, CA. [Cit. on p. 30]

Tucker, Benjamin V., Catherine Ford & Stephanie Hedges. 2021. Speech aging: Production and perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 12(5). e1557. doi:10.1002/wcs.1557. [Cit. on p. 93]

Tuomainen, Outi & Valerie Hazan. 2018. Investigating clear speech adaptations in spontaneous speech produced in communicative settings. In Mária Gósy & Tekla Etelka Gráczi (eds.), *Challenges in analysis and processing of spontaneous speech*, 9–25. Budapest, Hungary: Research Institute for Linguistics of the Hungarian Academy of Sciences. doi:10.18135/CAPSS.9. [Cit. on pp. 92, 93, and 113]

Tuomainen, Outi, Valerie Hazan & Linda Taschenberger. 2019. Speech communication in background noise: Effects of aging. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th ICPhS*, 805–809. Canberra, Australia: Australasian Speech Science and Technology Association Inc. [Cit. on pp. 93, 96, and 113]

Tuomainen, Outi, Linda Taschenberger, Stuart Rosen & Valerie Hazan. 2021. Speech modifications in interactive speech: Effects of age, sex and noise type. *Philosophical Transactions of the Royal Society B* 377(1841). 20200398. doi:10.1098/rstb.2020.0398. [Cit. on pp. 93, 96, and 113]

Turk, Alice. 2009. Is prosody the music of speech? Advocating a functional perspective. *International Journal of Speech-Language Pathology* 11(4). 316–320. doi:10.1080/17549500903003086. [Cit. on p. 4]

Turk, Alice, Satsuki Nakai & Mariko Sugahara. 2006. Acoustic segment durations in prosodic research: A practical guide. In Stefan Sudhoff, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schließer (eds.), *Methods in empirical prosody research*, vol. 3 Language, Context, and Cognition, 1–28. Berlin: Walter de Gruyter. doi:10.1515/9783110914641.1. [Cit. on pp. 66, 101, 124, and 154]

Turk, Alice E. & Stefanie Shattuck-Hufnagel. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 35(4). 445–472. doi:10.1016/j.wocn.2006.12.001. [Cit. on pp. 18, 19, and 21]

Uther, Maria, Monja A. Knoll & Denis Burnham. 2007. Do you speak E-NG-LI-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication* 49(1). 2–7. doi:10.1016/j.specom.2006.10.003. [Cit. on p. 95]

Van Engen, Kristin J., Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim & Ann R. Bradlow. 2010. The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53(4). 510–540. doi:10.1177/0023830910372495. [Cit. on p. 40]

Vanrell, Maria del Mar, Ingo Feldhausen & Llüisa Astruc. 2018. The Discourse Completion Task in Romance prosody research: Status quo and outlook. In Ingo Feldhausen, Jan Fliessbach & Maria del Mar Vanrell (eds.), *Methods in prosody: A Romance language perspective*, 191–227. Berlin: Language Science Press. doi:10.5281/zenodo.144134. [Cit. on p. 32]

Varadarajan, Vaishnevi S. & John HL Hansen. 2006. Analysis of Lombard effect under different types and levels of noise with application to in-set speaker ID systems. In *Proceedings of InterSpeech*, 937–940. Pittsburgh, PA. [Cit. on pp. 34, 55, and 79]

Venables, William N. & Brian D. Ripley. 2013. *Modern applied statistics with S-PLUS*. New York, NY: Springer. doi:10.1007/978-0-387-21706-2. [Cit. on p. 130]

Vion, Monique & Annie Colas. 2006. Pitch cues for the recognition of yes-no questions in French. *Journal of Psycholinguistic Research* 35(5). 427–445. doi:10.1007/s10936-006-9023-x. [Cit. on p. 120]

Wagner, Michael. 2005. *Prosody and recursion*. Cambridge, MA: Massachusetts Institute of Technology Doctoral thesis. http://hdl.handle.net/1721.1/33713. [Cit. on pp. 9, 11, 22, 23, 29, 35, 50, 53, 55, 56, 57, 59, 77, 81, 117, 179, 182, 186, and 187]

Wagner, Michael. 2010. Prosody and recursion in coordinate structures and beyond. *Natural Language & Linguistic Theory* 28(1). 183–237. doi:10.1007/s11049-009-9086-0. [Cit. on pp. 11, 22, 23, 55, 56, 77, and 119]

Wagner, Michael & Duane G. Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes* 25(7). 905–945. doi:10.1080/01690961003589492. [Cit. on pp. 4, 5, 11, 14, and 173]

Watson, Duane & Edward Gibson. 2004. The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes* 19(6). 713–755. doi:10.1080/01690960444000070. [Cit. on pp. 22, 55, 77, 117, and 144]

Weber, Andrea, Martine Grice & Matthew W. Crocker. 2006. The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition* 99(2). B63–B72. doi:10.1016/j.cognition.2005.07.001. [Cit. on pp. 21, 31, 148, 149, 150, 166, 167, 173, and 190]

Wellmann, Caroline, Julia Holzgrefe, Hubert Truckenbrodt, Isabell Wartenburger & Barbara Höhle. 2012. How each prosodic boundary cue matters: Evidence from German infants. *Frontiers in Psychology* 3. 580. doi:10.3389/fpsyg.2012.00580. [Cit. on pp. 11, 17, 18, 21, 177, and 181]

Wellmann, Caroline, Julia Holzgrefe-Lang, Hubert Truckenbrodt, Isabell Wartenburger & Barbara Höhle. 2023. Developmental changes in prosodic boundary cue perception in German-learning infants. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, vol. 12, 119–156. Berlin: Language Science Press. doi:10.5281/zenodo.7777532. [Cit. on pp. 11, 21, 97, and 181]

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. Switzerland: Springer. doi:10.1007/978-3-319-24277-4. [Cit. on p. 130]

Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70. 86–116. doi:10.1016/j.wocn.2018.03.002. [Cit. on pp. 16, 155, and 156]

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B Statistical Methodolody* 73(1). 3–36. doi:10.1111/j.1467-9868.2010.00749.x. [Cit. on p. 155]

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. Boca Raton, FL: CRC press. 2nd ed. [Cit. on pp. 16 and 155]

Xie, Xin, Andrés Buxó-Lugo & Chigusa Kurumada. 2021. Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition* 211. 104619. doi:10.1016/j.cognition.2021.104619. [Cit. on pp. 1, 26, 30, 31, and 184]

Xu, Yi. 2013. ProsodyPro – a tool for large-scale systematic prosody analysis. In Brigitte Bigi & Daniel Hirst (eds.), *Tools and resources for the analysis of speech prosody*, 7–10. Aix-en-Provence, France: Laboratoire Parole et Langage. `http://www2.lpl-aix.fr/~trasp/Proceedings/19724-trasp2013.pdf`. [Cit. on pp. 128, 155, and 166]

Yu, Alan Chi Lun. 2013. Individual differences in socio-cognitive processing and the actuation of sound change. In Alan Chi Lun Yu (ed.), *Origins of sound change: Approaches to phonologization*, 201–227. Oxford: Oxford University Press. [Cit. on p. 121]

Yule, George. 1997. *Referential communication task*. New York: Routledge. doi:10.4324/9781315044965. [Cit. on p. 39]

Zahner, Katharina, Manluolan Xu, Yiya Chen, Nicole Dehé & Bettina Braun. 2020. The prosodic marking of rhetorical questions in standard Chinese. In *Proceedings of Speech Prosody 2020*, 389–393. Tokyo, Japan. doi:10.21437/SpeechProsody.2020-80. [Cit. on pp. 16 and 155]

Zahner-Ritter, Katharina, Marieke Einfeldt, Daniela Wochner, Angela James, Nicole Dehé & Bettina Braun. 2022. Three kinds of rising-falling contours in German *wh*-questions: Evidence from form and function. *Frontiers in Communication* 7. 838955. doi:10.3389/fcomm.2022.838955. [Cit. on p. 167]

Zhang, Xinting. 2012. *A comparison of cue-weighting in the perception of prosodic phrase boundaries in English and Chinese*. Michigan, MI: University of Michigan Doctoral thesis. [Cit. on pp. 10, 11, 17, and 18]

Zollinger, Sue Anne & Henrik Brumm. 2011. The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* 148(11). 1173–1198. doi:10.1163/000579511X605759. [Cit. on pp. 2, 34, 55, and 96]