Pitch accent type affects stress perception in German: Evidence from infant and adult speech processing

Dissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie (Dr.phil.)

title page

vorgelegt von
Katharina Zahner

an der

Geisteswissenschaftliche Sektion
Fachbereich Linguistik

Konstanz, 2019
Tag der mündlichen Prüfung: 05. April 2019

1. Referentin: Prof. Dr. Bettina Braun
2. Referentin: Prof. Dr. Janet Grijzenhout
3. Referentin: JunProf. Dr. Adriana Hanulíková
This thesis focuses on explicit segmentation in infants and lexical activation in adults – processes of word recognition that are crucial for the respective target group (e.g., Cutler, 2012). Specifically, we examine how pitch accent type affects stress processing in segmentation and lexical activation. Pitch accents generally mark words and stressed syllables therein as prominent at the utterance level, mostly for information-structural purposes (e.g., Ladd, 2008). Broadly speaking, pitch accents differ in where – in regard to the stressed syllable – the fundamental frequency (f0) peak is realized, i.e., the f0 peak can precede or follow the stressed syllable, or it can be realized on the stressed syllable. Consequently, utterance-level intonation guides the status of f0 on a stressed syllable (either high-pitched or low-pitched, or rising or falling) and renders f0 an unreliable cue to the position of lexical stress for listeners. Using the Head-Turn Preference Paradigm with infants and the Visual-World Eye-Tracking Paradigm with adults, this thesis studied how (phonological) alignment differences in different pitch accent types affect lexical stress processing in German infants and adults.

Our experimental results showed that both German infants and adults were influenced by the position of the f0 peak in regard to the stressed syllable when processing lexical stress. In particular, German infants extracted trochaic units only when the stressed syllable was high-pitched. When naturally produced, infants furthermore mistook high-pitched unstressed syllables as word onset cues; the isolated f0 cue (resynthesized), however, did not lead to mis-segmentation. German adults, in turn, were influenced by different pitch accent types, such that f0 peaks on unstressed initial syllables led to the temporary activation of stress competitors with initial stress. To account for the observed effects of pitch accent type on stress-based segmentation and lexical activation, we proposed two underlying mechanisms in the thesis: Option 1 is that high-pitched syllables stand out perceptually (salience account). Option 2 is that listeners learned to associate high-pitched syllables with metrical stress, because of a frequent occurrence of H*-accents in the ambient language (frequency account). In an exposure-test paradigm (exposure phase with low-pitched accents and subsequent eye-tracking study) with German adults, we put the frequency account to test and examined whether the weighting of the f0 cue for stress processing is affected by the frequency of occurrence of high- and low-pitched stressed syllables in the immediate input. Results showed a reduced competitor activation, indicating that the frequency with which different pitch accent types occur in spoken communication modulate the cue weights for acoustic cues to stress, here high f0.

In conclusion, this dissertation is relevant to both psycholinguistic research and the interface between phonetics and phonology. From a psycholinguistic
perspective, it contributes towards unravelling the influence of intonation on lexical processing by showing that high f0 guides the perception of lexical stress in infant metrical segmentation and adult lexical activation. The perception of metrical strength relations in a word can be shifted if the f0 peak and the metrically stressed syllable are not aligned. Hence, pitch accent type influences lexical processing in intonation languages, in which pitch is not contrastively used. Furthermore, the manipulation of the occurrence frequency of different pitch accent types, in particular, allows for important theoretical conclusions at the phonetics-phonology interface. In this regard, our findings speak in favour of a phonological basis of the association between high f0 and lexical stress. More precisely, we argue that it is not (only) the phonetic cues (here the acoustic cue high f0) that guide the perception of lexical stress. Rather, the learned association between high f0 and lexical stress generates expectations on which cues make a syllable appear to be stressed.
Acknowledgements

Along my way of writing this thesis, I came across many people who supported or inspired me in one way or another. Here, I would like to say thank you, since this thesis would not have been possible without your support!

First and foremost, my gratitude goes to my supervisor Bettina Braun. The present work would definitely not be the same without her critical feedback, her enlightening comments and suggestions. I benefitted a lot from her expertise in phonetics and psycholinguistics, and her knowledge about experimental research and statistical analyses. Thank you, Bettina, for all your support and patience, and especially for your positive and motivating attitude that always kept me going! Thanks also to my second supervisor, Janet Grijzenhout, who gave me a very warm welcome in the Babylab and who offered her invaluable expertise in language acquisition and theoretical phonology so many times. Dank je wel, Janet! Adriana Hanulíková was kind enough to join the board of reviewers for this thesis as an external referee. Thank you for all the helpful comments and ideas on this thesis! Thank you all!

I would like to particularly thank Muna Schönhuber and Sophie Kutscheid for joining me in the adventure to study stress processing in infants and adults. Thanks for all your input and time! I am also very grateful to Jana Neitsch and another anonymous speaker for lending me their voices in many recording sessions. Thanks for your patience and endurance, and especially for your aptitude in intonation. From the AG Braun and the Babylab team, I would like to thank the research assistants and interns for help with recruitment, testing, or annotation: Sarah Ahmad, I. A., Andrea Beeken, Carolin Bernhardt, Phoebe Braunwarth, Nathalie Czeke, Stephanie Gustedt, Clara Huttenlauch, Angela James, Jasmin Rimpler, Nicole Saks, Annika Schilk, Helena Schlipf, Johanna Schnell, Stefan Volk, and Maximilian Wiesner. I would like to say thanks to my peer PhD students and post-docs in the department for many coffee breaks and chats that made the way so much nicer: Tina Bögel, Marieke Einfeldt, Farhat Jabeen, Katerina Kalouli, Mariya Kharaman, Sophie Kutscheid, Monika Lindauer, Jana Neitsch, Alexandra Rehn, Muna Schönhuber, Giusy Turco, Talina Weber, and Daniela Wochner! We had great times together! From the administrative side of the department of linguistics, I want to express my cordial thanks to Achim Kleinmann for all his technical support, to Barbara Werner, Anita Mademann and Tania Simeoni for help with organizational issues, and to Carmen Kelling for always having an open ear!

This thesis benefitted a lot from feedback and comments at various conferences and other occasions. In this regard, I profited from the SpeechNet BaWü group
initiated by Andrea Weber, Bettina Braun, Adriana Hanulíková, and Sabine Zerbian, and their collaboration with Anne Cutler’s group at the MARCS Institute in Sydney. Thanks to SpeechNet for your support and organizing great events! Thank you, Anne, for having me in your lab at MARCS and for your valuable comments and feedback on my work and on ideas for future work! Thank you also to Jenny Yu and Heather Kember for collaborating on “Herr Schiefer” but also for enriching my stays at MARCS. I am also grateful to Ann-Kathrin Grohe for advice on the design of the exposure-test paradigm (and for being a perfect travel mate), to Antje Schweitzer for comments on the materials for the ratings, to Nathaniel Phillips for support with plotting, to Vincent Porretta for support with plotting and GAMMs, and to Barbara Samlowski for retrieving syllable frequencies. I highly appreciate feedback to this work and discussion thereof (in research colloquia or at conferences) by Bistra Andreeva, Titia Benders, Natalie Boll-Avetisyan, Nicole Dehé, Carsten Eulitz, Claudia Friedrich, Claudia Männel, Pilar Prieto, Giusy Turco. I would particularly like to thank some peer PhD students who went this way with me. I am very thankful to Clara Huttenlauch, Iris Hüscher, Laura Hahn, and Jeanin Jügler for talking about work but also for having many good moments at conferences.

∞

For financial support additional thanks are owed to the Cusanuswerk, which provided me with a 3-year-PhD scholarship! This thesis would surely not have been possible without all the infants and adults participating in the studies. Particularly, I would like thank all parents who made the effort and brought their infants to the lab. Here, I would also like to mention all the playgroups and midwifes for support in recruiting. Finally, I owe a great deal to Angela James, Jenny Yu, Lea Jansson, and Nathi Czeke for reading through this thesis (even last-minute!).

∞

Life would be nothing without good company. In this regard, I would like to thank my friends, those from “home home” who I have known for ages and those who I have met during the time in Konstanz. Thanks for bringing all the “non-linguistic” topics in my life! Finally, I would like to thank my family, my parents Maria and Heinz, and my brother Matthias, who gave me all the support I could wish for! Thank you for sparking and sharing the interest in language! Last but definitely not least, I would like to thank René – simply for everything! I could not have done this without you!

∞

Thank you all who supported me along the way!
Contents

Thesis abstract .............................................................................................................. i
Acknowledgements ...................................................................................................... iii
Contents ......................................................................................................................... v
List of Figures ................................................................................................................ vii
List of Tables ................................................................................................................ ix
List of Abbreviations ................................................................................................... x

1 General introduction ................................................................................................. 1
  1.1 Rationale of the thesis ........................................................................................... 2
  1.2 Outline of the thesis ............................................................................................. 5

2 Intonation and lexical stress ..................................................................................... 7
  2.1 Intonation ............................................................................................................. 8
    2.1.1 Models to describe (German) intonation ......................................................... 10
    2.1.2 Pitch accent types as phonological categories ............................................... 15
    2.1.3 Excursus: Communicative functions of pitch accent types ......................... 20
  2.2 Lexical stress ....................................................................................................... 23
    2.2.1 Physical correlates of lexical stress ................................................................ 23
    2.2.2 Stress as an abstract notion ........................................................................... 27
  2.3 Interim summary ................................................................................................. 36

3 The processing of stress and intonation ................................................................. 39
  3.1 F0 as a stress cue? .............................................................................................. 40
    3.1.1 F0 and the perception of phrasal prominence ............................................... 40
    3.1.2 F0 and the perception of lexical stress .......................................................... 41
  3.2 Adult spoken word recognition ............................................................................ 47
    3.2.1 Mapping the speech signal onto words ......................................................... 47
    3.2.2 Models of spoken word recognition .............................................................. 52
    3.2.3 Stress and intonation in spoken word recognition ......................................... 54
  3.3 Infant word segmentation .................................................................................... 64
    3.3.1 Finding words in speech ............................................................................... 64
    3.3.2 Cues and strategies in infant word segmentation ......................................... 68
    3.3.3 Stress and intonation in infant segmentation .............................................. 70
  3.4 Interim summary ................................................................................................. 77

4 Pitch accent type affects stress-based segmentation in German infants 79
  4.1 Introduction .......................................................................................................... 80
  4.2 Experiment 1 ....................................................................................................... 88
    4.2.1 Methods .......................................................................................................... 88
    4.2.2 Results and interim discussion ..................................................................... 92
    4.2.3 Experiment 1b: Acoustic matching? .............................................................. 96
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.4</td>
<td>Experiment 1c: Alternation?</td>
<td>97</td>
</tr>
<tr>
<td>4.3</td>
<td>Experiment 2</td>
<td>100</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Methods</td>
<td>100</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Results and interim discussion</td>
<td>104</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Experiment 2b: Isolating the f0 cue</td>
<td>107</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>111</td>
</tr>
<tr>
<td>5</td>
<td>Pitch accent type affects lexical activation in German adults</td>
<td>115</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>116</td>
</tr>
<tr>
<td>5.2</td>
<td>Experiment 3</td>
<td>120</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Methods</td>
<td>120</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Analysis and results</td>
<td>126</td>
</tr>
<tr>
<td>5.3</td>
<td>Discussion</td>
<td>135</td>
</tr>
<tr>
<td>6</td>
<td>Addressing mechanisms for the processing of f0 as a stress cue</td>
<td>139</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>140</td>
</tr>
<tr>
<td>6.2</td>
<td>Experiment 4</td>
<td>145</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Methods</td>
<td>145</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Analysis and results</td>
<td>148</td>
</tr>
<tr>
<td>6.3</td>
<td>Discussion</td>
<td>156</td>
</tr>
<tr>
<td>7</td>
<td>General discussion</td>
<td>163</td>
</tr>
<tr>
<td>7.1</td>
<td>Summary of experimental results</td>
<td>164</td>
</tr>
<tr>
<td>7.2</td>
<td>Selected theoretical aspects and future lines of research</td>
<td>167</td>
</tr>
<tr>
<td>7.2.1</td>
<td>F0 as a stress cue?</td>
<td>167</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Implications for infant word segmentation</td>
<td>173</td>
</tr>
<tr>
<td>7.2.3</td>
<td>Implications for adult word recognition</td>
<td>177</td>
</tr>
<tr>
<td>7.3</td>
<td>Conclusion</td>
<td>183</td>
</tr>
<tr>
<td>Zusammensfassung</td>
<td>185</td>
<td></td>
</tr>
<tr>
<td>Bibliography</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>Appendix A: Experiment 1 – Materials</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>Appendix B: Experiment 2 – Materials</td>
<td>227</td>
<td></td>
</tr>
<tr>
<td>Appendix C: Experiment 3 – Materials and analyses</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>Appendix D: Experiment 3 – Post-hoc reading study</td>
<td>239</td>
<td></td>
</tr>
<tr>
<td>Appendix E: Experiment 4 – Materials in exposure-phase</td>
<td>243</td>
<td></td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Caption</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1:</td>
<td>Visualization of an early peak and a medial peak in the notation of KIM.</td>
<td>12</td>
</tr>
<tr>
<td>Figure 2:</td>
<td>Finite state model of intonation according to Pierrehumbert (1980, p. 29).</td>
<td>14</td>
</tr>
<tr>
<td>Figure 3:</td>
<td>Visualization of the three different pitch accent types under investigation in GToBI and KIM.</td>
<td>16</td>
</tr>
<tr>
<td>Figure 4:</td>
<td>Overview of the manipulation procedure in the categorical perception experiments in Kohler (1991c).</td>
<td>17</td>
</tr>
<tr>
<td>Figure 5:</td>
<td>Distribution of lexical stress placement in three West-Germanic languages, English, German, and Dutch, adapted from Cutler (2012, p. 53).</td>
<td>30</td>
</tr>
<tr>
<td>Figure 6:</td>
<td>Mapping process between the acoustic signal and the lexical level in the episodic and abstractionist framework.</td>
<td>49</td>
</tr>
<tr>
<td>Figure 7:</td>
<td>Sound pressure wave and spectrogram of a German utterance.</td>
<td>64</td>
</tr>
<tr>
<td>Figure 8:</td>
<td>Potential re-interpretation of the word-prosodic structure under the assumption that f0 is a sufficient cue to lexical stress.</td>
<td>87</td>
</tr>
<tr>
<td>Figure 9:</td>
<td>Example sound pressure wave, spectrogram, and pitch track for three intonation conditions in Experiment 1.</td>
<td>89</td>
</tr>
<tr>
<td>Figure 10:</td>
<td>Schematic illustration of the experimental setup (HPP booth) in the BSL.</td>
<td>91</td>
</tr>
<tr>
<td>Figure 11:</td>
<td>Means of difference in looking time to novel and familiar items in Experiment 1.</td>
<td>93</td>
</tr>
<tr>
<td>Figure 12:</td>
<td>Example sound pressure wave, spectrogram, and pitch track in Experiment 1c.</td>
<td>98</td>
</tr>
<tr>
<td>Figure 13:</td>
<td>Example sound pressure wave, spectrogram and pitch track for two intonation conditions in Experiment 2.</td>
<td>102</td>
</tr>
<tr>
<td>Figure 14:</td>
<td>Means of difference in looking time to novel and familiar items in Experiment 2.</td>
<td>105</td>
</tr>
<tr>
<td>Figure 15:</td>
<td>Example sound pressure wave, spectrogram, and pitch track in Experiments 2 and 2b.</td>
<td>109</td>
</tr>
<tr>
<td>Figure 16:</td>
<td>Percentage of erroneous responses to syllable 1 and syllable 3 in the WSW-targets for offline stress judgements in Zahner et al. (2019).</td>
<td>119</td>
</tr>
<tr>
<td>Figure 17:</td>
<td>Example sound pressure wave, spectrogram, and pitch track for an experimental trial in Experiment 3.</td>
<td>123</td>
</tr>
<tr>
<td>Figure 18:</td>
<td>Experimental setup in the Phonlab with SR Eyelink 1000 Plus in a desktop mount system.</td>
<td>125</td>
</tr>
</tbody>
</table>
Figure 19: Exemplar screen for an experimental trial in Experiment 3. 125
Figure 20: Evolution of fixations to words on screen in Experiment 3. 128
Figure 21: Competitor fixations across intonation condition in Experiment 3. 128
Figure 22: Auto-correlation graph for the basic model for competitor fixations in Experiment 3. 130
Figure 23: GAMM difference curve in competitor fixations in early-peak condition minus medial-peak condition in Experiment 3. 132
Figure 24: GAMM difference curve in target fixations in early-peak condition minus medial-peak condition in Experiment 3. 134
Figure 25: Example sound pressure wave, spectrogram, and pitch track for exemplar units from the exposure phase in Experiment 4. 146
Figure 26: Exemplar screens for exposure-test paradigm in Experiment 4. 147
Figure 27: Evolution of fixations to words on screen in Experiment 4. 149
Figure 28: Auto-correlation graph for the basic model for competitor fixations in Experiment 4. 149
Figure 29: GAMM difference curve in competitor fixations in early-peak condition minus medial-peak condition in Experiment 4. 150
Figure 30: GAMM difference curve in target fixations in early-peak condition minus medial-peak condition in Experiment 4. 152
Figure 31: GAMM difference curve of the difference in competitor fixations in early-peak vs. medial-peak condition across the two experiments (Experiment 3 vs. Experiment 4). 154
Figure 32: GAMM difference curve of the difference in target fixations in early-peak vs. medial-peak condition across the two experiments (Experiment 3 vs. Experiment 4). 156
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Caption</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1:</td>
<td>Summary of communicative functions of different pitch accent types.</td>
<td>22</td>
</tr>
<tr>
<td>Table 2:</td>
<td>Overview of different notions of prosodic prominence.</td>
<td>37</td>
</tr>
<tr>
<td>Table 3:</td>
<td>Overview of studies on infant stress-based segmentation.</td>
<td>74</td>
</tr>
<tr>
<td>Table 4:</td>
<td>Overview of description of experimental stimuli in studies summarized in Table 3.</td>
<td>83</td>
</tr>
<tr>
<td>Table 5:</td>
<td>Acoustic realization of WSW-target words in the familiarization phase for the three intonation conditions in Experiment 1.</td>
<td>90</td>
</tr>
<tr>
<td>Table 6:</td>
<td>Overview of average looking times to novel and familiar test lists in Experiment 1.</td>
<td>93</td>
</tr>
<tr>
<td>Table 7:</td>
<td>Acoustic realization of WSW-target words in the familiarization phase for the flipped medial-peak condition in Experiment 1c.</td>
<td>98</td>
</tr>
<tr>
<td>Table 8:</td>
<td>Acoustic realization of WWS-target words in the familiarization phase for the two intonation conditions in Experiment 2.</td>
<td>103</td>
</tr>
<tr>
<td>Table 9:</td>
<td>Overview of average looking times to novel and familiar test lists in Experiment 2.</td>
<td>104</td>
</tr>
<tr>
<td>Table 10:</td>
<td>Acoustic realization of WSW-targets in two intonation conditions in Experiment 3 (naturally recorded, before PSOLA-resynthesis).</td>
<td>122</td>
</tr>
<tr>
<td>Table 11:</td>
<td>Final GAMM for competitor fixations in Experiment 3.</td>
<td>131</td>
</tr>
<tr>
<td>Table 12:</td>
<td>Final GAMM for target fixations in Experiment 3.</td>
<td>134</td>
</tr>
<tr>
<td>Table 13:</td>
<td>Final GAMM for competitor fixations in Experiment 4.</td>
<td>150</td>
</tr>
<tr>
<td>Table 14:</td>
<td>Final GAMM for target fixations in Experiment 4.</td>
<td>151</td>
</tr>
<tr>
<td>Table 15:</td>
<td>Final GAMM of combined analysis for competitor fixations in Experiments 3 and 4.</td>
<td>153</td>
</tr>
<tr>
<td>Table 16:</td>
<td>Final GAMM of combined analysis for target fixations in Experiments 3 and 4.</td>
<td>155</td>
</tr>
<tr>
<td>Table 17:</td>
<td>Distribution of contours in German adult-directed speech, adapted from Peters et al. (2005).</td>
<td>159</td>
</tr>
<tr>
<td>Table 18:</td>
<td>Distribution of GToBI pitch accent types in the KIDS corpus.</td>
<td>160</td>
</tr>
</tbody>
</table>
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADS</td>
<td>Adult-directed speech</td>
</tr>
<tr>
<td>AM</td>
<td>Autosegmental-metrical</td>
</tr>
<tr>
<td>AmE</td>
<td>American English</td>
</tr>
<tr>
<td>AusE</td>
<td>Australian English</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>AuToBI</td>
<td>Automatic Tones and Break Indices</td>
</tr>
<tr>
<td>BE</td>
<td>British English</td>
</tr>
<tr>
<td>BSL</td>
<td>BabySpeech Lab (at the University of Konstanz)</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>CV</td>
<td>Consonant-vowel (sequence)</td>
</tr>
<tr>
<td>CMFP</td>
<td>Cross-modal fragment priming</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalography</td>
</tr>
<tr>
<td>ERP</td>
<td>Event-related potentials</td>
</tr>
<tr>
<td>f0</td>
<td>Fundamental frequency</td>
</tr>
<tr>
<td>GAMM</td>
<td>General additive mixed model</td>
</tr>
<tr>
<td>glmer</td>
<td>General linear mixed effects model (logistic)</td>
</tr>
<tr>
<td>GToBI</td>
<td>German Tones and Break Indices</td>
</tr>
<tr>
<td>HPP</td>
<td>Head-Turn Preference Procedure</td>
</tr>
<tr>
<td>IDS</td>
<td>Infant-directed speech</td>
</tr>
<tr>
<td>ISI</td>
<td>Inter-stimulus-interval</td>
</tr>
<tr>
<td>IPA</td>
<td>International Phonetics Alphabet</td>
</tr>
<tr>
<td>KIDS</td>
<td>Konstanz prosodically infant-directed speech Corpus</td>
</tr>
<tr>
<td>KIM</td>
<td>Kieler Intonation Model</td>
</tr>
<tr>
<td>lmer</td>
<td>Linear mixed effects model</td>
</tr>
<tr>
<td>MSS</td>
<td>Metrical Segmentation Strategy</td>
</tr>
<tr>
<td>n.s.</td>
<td>Not significant</td>
</tr>
<tr>
<td>o.p.b.</td>
<td>Occurrences per billon</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>o.p.m.</td>
<td>Occurrences per million</td>
</tr>
<tr>
<td>p.c.</td>
<td>Personal communication</td>
</tr>
<tr>
<td>PSOLA</td>
<td>Pitch Synchronous Overlap and Add</td>
</tr>
<tr>
<td>PWC</td>
<td>Possible word constraint</td>
</tr>
<tr>
<td>RPT</td>
<td>Rapid Prosody Transcription</td>
</tr>
<tr>
<td>RT</td>
<td>Reaction times</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>st</td>
<td>Semitones</td>
</tr>
<tr>
<td>ToBI</td>
<td>Tones and Break Indices</td>
</tr>
<tr>
<td>VW</td>
<td>Visual-World</td>
</tr>
<tr>
<td>WSW</td>
<td>Weak-strong-weak prosodic structure; also used for all other combinations of S and W</td>
</tr>
</tbody>
</table>
Chapter 1 is the introductory chapter. First, 1.1 sketches the role of the stressed syllable in infant word segmentation and adult word recognition, respectively, and briefly highlights how utterance-level pitch accent type interacts with the stressed syllable (rendering the stressed syllable either high- or low-pitched). This will reveal why it is interesting to study the effect of pitch accent type on infants’ and adults’ processing of lexical stress. Then, in 1.2, we will outline the structure of the thesis.
1.1 Rationale of the thesis

Throughout life, nothing feels more effortless than listening to (native) speech. We listen to caretakers, teachers, or professors, engage in discussions on politics, economy, or sports, watch series and movies, or just overhear random conversations on the bus. Indeed, listening to speech is such a seemingly unspectacular process that in most cases we do not even notice that it is happening – until understanding fails. For instance, when we listen to an unknown language, words seem to be glued together, resulting in a speech stream that appears unbearably fast. We soon realize that, unlike words in a written text, fluent speech lacks clear and unambiguous word boundary cues, a characteristic of spoken language that has become widely known as the word segmentation problem (e.g., Cutler, 2012).

Word segmentation, i.e., the decomposition of fluent speech into discrete units, is especially vital for infant listeners who, unlike adults, cannot rely on a mental lexicon. Infants strongly depend on this mechanism since they rarely hear words in isolation, as recent corpus analyses on infant-directed speech (henceforth IDS) have revealed (e.g., E. K. Johnson, Lahey, Ernestus, & Cutler, 2013; Zahner, Schönhuber, Grijzenhout, & Braun, 2016b). Presumably it goes without saying though that it is part of human development that we are prepared for the task of understanding speech. In fact, the language that surrounds us teaches us a lot about how we can crack the “speech code” (A. M. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), see Kuhl (2004) for a detailed overview. Generally speaking, our experience with language and speech starts long before we utter the first words, even before birth, as soon as the auditory capabilities are in place around the last trimester of pregnancy (Querleu et al., 1988; Richards et al., 1992). The womb filters out much of the spectral characteristics that make up most of the segmental contrasts but preserves the prosodic and rhythmical properties of speech. As a result, the unborn child prosodically tunes in to her native language, showing language-specific preferences and rhythm-based discrimination abilities from the first day of life onwards (e.g., Moon, Cooper, & Fifer, 1993; Nazzi, Bertoncini, & Mehler, 1998). Prosody continues to play an essential role for language acquisition within the first year of life (cf. Prieto & Esteve-Gibert, 2018), with rhythm-based strategies strongly assisting in solving the word segmentation problem (cf. Nazzi et al., 2006). That way, infants who grow up with a stress-timed language, such as German, Dutch, or English, do not only display a preference for the predominant stress pattern, i.e., trochees (Bijeljac-Babic, Höhle, & Nazzi, 2016; Höhle et al., 2017). Besides common objects involving daily routines, e.g., Windel ‘nappy’ or Löffel ‘spoon’, the names of the parents (Mama ‘mum’, Papa ‘dad’) and abbreviations thereof as well as the infants’ own names are very frequent in early conversations (Zahner, Schönhuber, et al., 2016b). These words are particularly often trochaic. To illustrate the point we analysed the word-prosodic structure of the first names of 100 randomly selected infants (51 female, 49 male) tested at the BabySpeech Lab at the...
2009; Jusczyk, Cutler, & Redanz, 1993), but also develop a stress-based segmentation strategy where the stressed syllable is interpreted as a word onset. Applying this Metrical Segmentation Strategy (MSS, Cutler & Norris, 1988), infants in these languages have been shown to extract trochaic units from fluent speech by 9 months of age (e.g., Höhle, 2002; Jusczyk, Houston, & Newsome, 1999; Kooijman, Hagoort, & Cutler, 2005; Männel & Friederici, 2013). This shows that infants have learned to use the cues in the signal to identify stressed syllables and exploit them to localize word boundaries within the second half of the first year of life. The trochaic unit further leaves its traces throughout development in stress-timed languages, such as in children’s early productions, e.g., nana for banana (Demuth, 1996, 2018; Fikkert, 1994; Gwinner, Gaglia, & Grijzenhout, 2012), and still serves as a segmentation unit for adults, as revealed by “slips of the ear” (Cutler & Butterfield, 1992) or – more amusingly – by mondegreens (songs in which listeners missegment foreign lyrics based on native segmentation strategies, e.g., Kentner, 2015). For instance, the original lyrics “Hope of deliverance” [haup ov d’il.værants] (Paul McCartney) have been shown to be mis-perceived as “Hau auf die Leberwurst” [hau auf di: ‘le:be.vurst] ‘hit on the liver sausage’ by German listeners (Kentner, 2015), revealing that German listeners straightforwardly interpreted the stressed syllable ([ItI]) as a word onset in these lyrics.2

Growing up, we develop a mental lexicon, i.e., a memory system in which we store a form of representations of the words, which are considered very abstract by some theories and very detailed (episodic) by others (e.g., Pierrehumbert, 2016, for a recent overview). Once the lexicon begins to emerge around the second half of the first year (Bergelson & Aslin, 2017; Bergelson & Swingley, 2012), infants have the possibility to recognize words in speech not only by signal-derived cues but also by using lexical knowledge. In fact, these lexical cues play a dominating role in adult word segmentation (e.g., see Mattys, White, & Melhorn, 2005, on the integration of lexical and signal-derived cues in adult segmentation). Explicit segmentation, which is the essential process of recognizing words for infants, is in turn assumed to serve an assisting function in adult word recognition (e.g., Cutler, 2012; Hanulíková, 2009; Weber & Scharenborg, 2012). Generally speaking, word recognition is understood as the process in which activated words compete for recognition, with the winner eventually being the word for which the cues in the signal match the stored

---

2 To explain this misperception in more detail, the trisyllabic word deliverance with penultimate stress is parsed as a sequence of an article followed by a trochaic noun (die Leber) in this example. Hence, the stressed second syllable in the English word functions as a word onset in the German 'misperceived' version. Similar tendencies have been observed for Dutch listeners, leading to the so called “mama appelsap”-songs (see radio broadcasts publicly available).
representation best (for overviews, Cutler, 2012; McQueen, 2005; Weber & Scharenborg, 2012). In that sense, processing stress cues correctly is particularly beneficial in languages in which lexical stress distinguishes between words. For instance, in German, the name of the city Konstanz (at the lake of Konstanz) and the noun Konstanz meaning ‘stability’ can be distinguished solely based on the stress status of its syllables (the city name is stressed on the first syllable while the noun is stressed on the second syllable). Similarly, lexical stress distinguishes the English noun object, which is stressed on the first syllable, from the verb object, which is stressed on the second syllable. Beyond these stress minimal pairs, which are in fact infrequent in English and German (Cutler, 2005), lexical stress also distinguishes cohort competitors that overlap in speech sounds but differ in the position of stress (henceforth “stress competitors”). For example, the German stress competitors Libero [li.ba.ro] ‘sweeper’ and Libelle [li.ˈbe.lə] ‘dragonfly’ are segmentally ambiguous until the end of the onset consonant of the second syllable, but differ in the position of the stressed syllable, with the former being stressed on the first and the latter on the second syllable. Accordingly, if listeners take information on lexical stress into account, the number of competitors is reduced and word recognition becomes more efficient. Indeed, there is abundant evidence that listeners – besides using segmental information (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; McQueen & Viebahn, 2007; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) – also make use of lexical stress for word recognition, as soon as the stress cues are available in the signal (e.g., Jesse, Poellmann, & Kong, 2017; Reinisch, Jesse, & McQueen, 2010). Cues to stress in the signal are duration, intensity, spectral cues, and, depending on the type of pitch accent, also f0 (e.g., Cutler, 2005).

This thesis focuses on explicit segmentation in infants and lexical activation in adults – processes of word recognition that are most relevant for the respective target group. Specifically, we examine how pitch accent type affects stress processing in segmentation and lexical activation. As will be shown in detail in 2.1, pitch accents generally mark words and stressed syllables therein as prominent at the utterance level, mostly for information-structural purposes (Ladd, 2008, for an overview). Broadly speaking, pitch accents differ in where – in regard to the stressed syllable – the fundamental frequency (f0) peak is realized, i.e., the f0 peak can precede or follow the stressed syllable, or it can be realized on the stressed syllable. Consequently, utterance-level intonation guides the status of f0 on a stressed syllable (either high-pitched or low-pitched, or rising or falling) and renders f0 an unreliable cue to the position of lexical stress for listeners. Conceivably, pitch accent type might influence stress processing in both infants and adults – especially since
high f0 is used for linguistic grouping (by infants and adults) and in offline stress judgement studies (by adults), see Chapter 3. So far, no study has investigated the perceptual consequences of the interplay between suprasegmental stress cues and pitch accent types in word recognition, lexical activation and explicit segmentation. The current dissertation attempts to fill this gap. To this end, we test how (phonological) alignment differences of the f0 peak in different pitch accent types modulate lexical activation of stress competitors in German adults and the extraction of trochaic units in German infants. In accounting for the effects observed in these mechanisms, from a different viewpoint, we also address the underlying factors to the processing of f0 as a stress cue. One theory posits that high-pitched syllables stand out perceptually (Cho, 2005; Hsu, Evans, & Lee, 2015; Lieberman, 1967), while another suggests that listeners learned to associate high-pitched syllables with metrical stress because of a frequent occurrence of high-pitched accents. Both views could explain why high f0 might be used as a cue to stress. As the former mechanism was more difficult to experimentally manipulate, we here tested whether the use of the f0 cue as a cue to lexical stress is determined by the frequency of different pitch accent types in the immediate input.

Overall, this dissertation sheds light on two core processing mechanisms of human word recognition – lexical activation and explicit segmentation – and at the same time contributes to our understanding of how stress perception (and the cues affecting it) might change throughout life. It hence contributes to current topics in psycholinguistics. Importantly, our study is a first attempt to unravel the mechanisms that guide the processing of f0 as a stress cue. In that regard, it bears implications for the way high f0 is processed as a stress cue, i.e., either directly or via phonological categories and hence also contributes to the phonetics-phonology interface.

1.2 Outline of the thesis

The individual chapters of this thesis are organized as follows: Chapters 2 and 3 represent the background chapters of the thesis. Chapter 2 provides the phonetic and phonological background on the interplay between pitch accent type and lexical stress. In particular, in 2.1, we introduce the two systems (KIM and GToBI) that were used to describe the different pitch accent types under investigation and provide a detailed look at different pitch accent types, discussing theoretical concepts (association and alignment) and their status as phonological categories. In 2.2, we introduce lexical stress as a concrete phenomenon (phonetic correlates) and an abstract property (metrical strength relations). Chapter 3 is devoted to the processing of stress and intonation and thereby focuses on the perspective of the listener. In 3.1, we will discuss how f0 – despite it being an ambiguous cue to the position of
the stressed syllable – functions as a cue to stress in both infants and adults. We will then focus on adult word recognition and infant word segmentation, and the role of stress and intonation cues in these processes in particular (3.2 and 3.3, respectively).

Chapters 4-6 represent the experimental chapters of this thesis: Broadly speaking, Chapters 4 and 5 study stress processing in infants and adults, respectively. Chapter 6 addresses the underlying mechanisms of stress processing in order to explain the effects observed in Chapters 4 and 5. All of the experimental chapters come with their own individual Introduction and Discussion, in which we highlight the relevant aspects and implications in order to guide the reader. Specifically, Chapter 4 presents a series of head turn-preference experiments that address the effect of pitch accent type on German 9-month-olds’ stress-based segmentation strategies. We test whether a misalignment between the f0 peak and lexical stress modulates the ability to extract a trochaic unit from a WSW- (weak-strong-weak) carrier word and whether such misalignment configurations may lead to a re-interpretation of a word’s prosodic structure (e.g., WS to SW when the WS-sequence has a high-low tonal structure). Chapter 5, in turn, investigates the effect of pitch accent type on stress perception and consequently, on the lexical activation of stress competitors in a Visual-World Eye-Tracking study with German adults. To this end, we test whether f0 peaks on unstressed initial syllables trigger the perception of stress and thus lead to the temporary activation of competitors with initial stress. In Chapter 6, we assess one of the underlying factors that could make high f0 a cue to stress: the frequency of occurrence of different pitch accent types. Following previous adaptation studies, we conduct an exposure-test paradigm with adult listeners in which we increase the occurrence frequency of low-pitched accents in the immediate input (exposure phase) and subsequently test the use of f0 as a stress cue in an eye-tracking study (as in Chapter 5). Beyond unravelling the mechanism at play, this further allows us to shed light on the nature of f0 processing.

Chapter 7, the General discussion, first presents a summary of the findings obtained in the experimental studies of this thesis (Chapters 4-6). Then, in 7.2, selective aspects emerging from the results of the individual experiments will be discussed, along with some avenues for future research. We will finally end with some concluding remarks (7.3) in which we highlight the main implications that can be derived from the current thesis and the main contributions of this work to the field of psycholinguistics and the phonetics-phonology interface.
Chapter outline

Chapter 2 provides the phonetic and phonological background on the interplay between pitch accent type and lexical stress, with 2.1 focusing on pitch accents (intonation) and 2.2 on lexical stress. In particular, in 2.1, we broadly define the concept of intonation and explain its correlates and functions. We will then introduce two systems that are used to describe different pitch accent types in German, i.e., the Kieler Intonation Model (KIM) and the German Tone and Break Indices (GToBI) (2.1.1), and outline why pitch accent types are considered phonological categories (2.1.2). In an excursus, we will finally summarize the communicative functions of different pitch accent types (2.1.3). In 2.2, we introduce lexical stress as a concrete phenomenon and an abstract property. Accordingly, in 2.2.1, we will summarize studies providing evidence for different phonetic correlates of lexical stress in German, i.e., duration, spectral balance, and vowel quality. From a phonological perspective, 2.2.2 discusses different typological stress systems and stress placement in three West-Germanic “free-stress” languages (German, English, Dutch), in particular. We will also outline metrical theories and discuss the three West-Germanic languages as stress-timed languages. In 2.3, we summarize the main points of Chapter 2.
2.1 Intonation

In intonation languages, such as German, English, or Dutch, pitch accents and boundary tones are considered to be the building blocks of the intonational contour (Ladd, 2008). We will start out by addressing the concept of intonation in general before we turn to pitch accents in more detail. Intonation in a narrow sense refers to the f0 changes over the course of an utterance ('t Hart, Collier, & Cohen, 1990, p. 10). From a phonetic perspective, the fundamental frequency is the acoustic correlate of intonation, the frequency of the vocal fold vibration the physiological one. Pitch refers to the psychophysical correlate, which is the perceptual impression of listeners. The faster the vocal folds vibrate, the higher the measured f0 (in Hz) and the higher the perceived pitch (for brief overviews, e.g., 't Hart et al., 1990, pp. 10-37; Féry, 2017, pp. 16-34; Gilles, 2005, pp. 3-6). Intonation is a linguistic feature present in all languages around the world (Hirst & Di Cristo, 1998, p. 1, see also Gussenhoven, 2004, for a similar claim). Typologically, different groups of languages can be distinguished with respect to how intonation is exploited (Vaissière, 2004, p. 242, for an overview): tone languages, such as Mandarin Chinese, in which a change in tone on a syllable leads to a difference in lexical meaning (Gussenhoven, 2004); pitch accent languages, such as Japanese, in which f0 movements of a particular shape may distinguish between the meaning of a word depending on the position of the movement (high-low for Japanese, Beckman, 1986; see Hyman, 2009, for a critical discussion on pitch accent languages as a distinct typological category), and intonation languages, such as German, English or Spanish, in which pitch is not used contrastively.

For intonation languages – the focus of the current dissertation – we will adopt a definition of intonation put forward by Ladd (2008, p. 4): “Intonation [...] refers to the use of suprasegmental phonetic features to convey 'postlexical' or sentence-level pragmatic meanings in a linguistically structured way.” Accordingly, we first consider intonation to be a phenomenon that takes place above the segmental level. Second, the meaning that intonation conveys is related to a whole utterance and intonation is hence different from word-level lexical stress and lexical tone, for which the primary function is the distinction between different words. The third component of Ladd’s definition of intonation points to the way linguistic meaning is conveyed through intonation. For Ladd, distinct entities (low vs. high tonal targets or rising vs. falling movements), as opposed to gradient differences, are employed to constitute linguistic meaning. In Ladd’s understanding, f0 movements have a “morpheme-like” character (Ladd, 2008, p. 41): Similar to morphemes, which are the smallest linguistic units carrying meaning, elements of intonation are also thought to carry meaning and as such belong to a system with considerable pragmatics. These intonational morphemes are assumed to convey meaning across
different lexical and linguistic contexts. Similar positions have been advocated by Gussenhoven (1984, p. 27), Ward and Hirschberg (1985), Pierrehumbert and Hirschberg (1990), or Steedman (1991), see Ladd (2008, p. 147ff.) for an overview. For instance, Pierrehumbert and Hirschberg (1990, p. 285) advocate a theory of tune meaning (a tune being composed of a pitch accent type, a phrasal tone, and a boundary tone, see below for details), which speakers use to specify the relation between the propositional content in the intonation phrase and the mutual beliefs of the interlocutors. In their view, a rising-falling-rising contour (described as L+H* L H%, for details on intonational categories see below) is used to signal a correction or contrast, with a speaker placing a “new scalar value for one previously proposed by [the] S[peaker] or [the] H[earer]” (Pierrehumbert & Hirschberg, 1990, p. 296). Similarly, a rising-falling-rising contour with the rise starting later in the accented syllable (as compared to the “correction-contrast” contour) has been assigned a meaning of uncertainty (Ward & Hirschberg, 1985).

Yet, the assumption of intonational morphemes is controversial (e.g., Kohler, 1991c, p. 161). In fact, it has been intensively challenged by recent experimental evidence showing that the relationship between form (intonation) and function (meaning) might not be straightforward in all cases (e.g., Chodroff & Cole, 2018; Mücke & Grice, 2014). Chodroff and Cole (2018), for instance, investigated pitch accent types in non-phrase-final words in read stories in American English. Their intonational analysis did not reveal a consistent use of a particular contour type for predefined communicative categories (here information structure, see below). Studies on the interpretation of intonation draw a similar picture: Perceptual evidence on the interpretation of early-falling contours do not reveal a clear one-to-one relationship between contour and an assigned meaning (Baumann & Grice, 2006) and differences in contours apparent in speakers’ productions are not unambiguously interpreted by the listener (Braun, 2006). Some scholars have even questioned the entire view of tone-meaning relation altogether (e.g., Calhoun, 2010, 2012), based on a semi-spontaneous game task that revealed the tonal distinction between rheme/theme structure to be less straightforward than the prominence distinction. Accordingly, in this thesis, we consider intonation to be able to convey certain communicative functions but refrain from ascribing a one-to-one relationship between form (intonation) and function (meaning).3

3 It needs to be pointed out that Ladd (2008) draws a clear-cut distinction between the phonological and the phonetic nature of intonational analyses, with phonological (categorical) distinctions contributing to linguistic meaning and phonetic (gradient) variation to the paralinguistic meaning of intonation (see also Gussenhoven, 2004, p. 47ff., for a similar position). The paralinguistic function of intonation is traditionally characterized by its universal scope. According to Gussenhoven’s theory on paralinguistic meaning, this language-independence is biologically motivated (Gussenhoven, 2002, 2004). Gussenhoven advocates three physiological conditions, which subsume to a “biological code” (see e.g., Chen, 2009, for an extensive overview). In this dissertation, we will concentrate on the linguistic function of intonation. Note though that it has been controversially discussed in
The distinctive categories of intonation (low vs. high tonal targets or rising vs. falling movements), as Ladd (2008) refers to them, generally serve two purposes – the grouping of speech into chunks (phrasing) and, most relevant to the current thesis, the highlighting of words within an utterance (accentuation); see also Grice and Baumann (2007) for an overview. Linguistic highlighting on the utterance-level, which makes a word (and the stressed syllables therein) acoustically stand out from all the other words (and syllables) in the utterance, is achieved by the use of pitch accents. Linguistic highlighting is not only a matter of accentuation, i.e., the presence or absence of a pitch accent. Pitch accents can be of different types, which show a different alignment of the f0 peak to the stressed syllable. In the following subchapters, we will introduce two systems of intonational description, namely the Kieler Intonation Model (KIM, Kohler, 1991b, 1991c) and German Tone and Break Indices (GToBI, Grice, Baumann, & Benzmüller, 2005), which we used to describe the pitch accent types under investigation. We will then discuss the concepts of alignment and association regarding different pitch accent types and emphasize the phonological status of alignment differences in the pitch accent types under investigation.

2.1.1 Models to describe (German) intonation

Generally speaking, intonation has been described using contour-based and level-based models. Contour-oriented approaches focus on the dynamic movement of the (nuclear) pitch contour – with a strong perceptual emphasis (for German, e.g., Fox, 1984; Kohler, 1991b, 1991c; Pheby, 1975; Von Essen, 1964). These models were developed in the tradition of the British school, which has been very influential in describing English intonation since the 1950s (e.g., Couper-Kuhlen, 1986; Cruttenden, 1986; Crystal, 1969; Halliday, 1967; Kingdon, 1959; O’Connor & Arnold, 1971). Accounts in the framework of the British school described intonational patterns by addressing whole f0 configurations, named tone groups. Each tone group (also referred to as an intonation group or word group) consists of a nucleus (maximally one), which is the only essential element in a tone group. A tone group may further comprise a pre-head, a head, and a tail. The tail consists of the unaccented syllables that can follow the nucleus, while the head comprises all accents up to the nucleus (and thus considerably varies in its length depending on the length of the utterance). The unaccented syllables that precede the head are grouped

---

4 Detailed reviews on different models of intonational description can be found elsewhere (e.g., Braun, 2005, pp. 19-36; Cruttenden, 1986; Féry, 2017, pp. 94-134; Grice & Baumann, 2002; Grice et al., 2005). We will here highlight the main claims that are relevant for the argumentation of the thesis.
into the pre-head (Féry, 2017, p. 125ff.; Mª Luisa García, 1997). All these models have in common that the nucleus is considered to be the most prominent element that carries the meaning (e.g., Halliday, 1967). Hence, this framework explicitly posits a difference between nuclear accent (the last accent) and prenuclear accents (accents preceding the nucleus).

The Kieler intonation model (KIM) – a contour-based approach for German intonation

One well-known and influential contour-based approach for German intonation is the Kieler Intonation model (KIM, Kohler, 1991b, 1991c), developed by Klaus Kohler and colleagues, a phonological model based on experimental work on form-function relations, see Niebuhr (2007a, pp. 33-40) for an overview. KIM's assumptions are thus explicitly based on perception experiments on listeners’ categorical perception of pitch and their semantic judgements for the established categories. In its attempt to model intonation contours as movements rather than as discrete tonal units, the model explicitly situates itself in the tradition of the British School (Kohler, 1991a, p. 20). The model considers a contour to be a sign, crucially only one sign. This sign is confined to the accented syllable, starting before the accented syllable and ending at the start of an adjacent accented syllable or at the end of an intonation phrase. Note that KIM differs from other approaches of the British school in that an intonation unit may start before the accented syllable. In KIM, pitch contours are described as peak or valley contours, based on the form of the f0 movement. Valley contours are falling movements that are most often followed by a concave rise. Peak contours, on the other hand, are rising-falling movements, usually of a convex shape (Kohler, 1991b, 1991c). Peak contours are most important for this thesis and will now be described in more detail.

For peak contours, KIM (Kohler, 1991b, 1991c) generally distinguishes between early-, medial-, and late-peak contours, depending on the position of the f0 peak in regard to the stressed syllable. Figure 1 visualizes two different peak contours on the syllable sequence *Sie hat ja gelogen* ‘She actually lied’ in the notation of the KIM (dot plots), an early peak and a medial peak on the verb *gelogen* [ɡəˈloːɡn].

Following the tradition of the British school, KIM also differentiates between the prenuclear and nuclear position in a contour. Terminal patterns are described as rising, falling, and falling-rising. In addition, four degrees of prominence are annotated, 0 for unaccented, 1 for lexically stressed, 2 for accentuated, and 3 for emphasized syllables.

---

5 Dot plots were traditionally employed by scholars in the British school. In such a dot plot, every syllable is assigned a dot. The size of the dot illustrates prominence relations, with the bigger dot referring to the stressed syllable. The pitch range of the movements lies between the top and the baseline (see e.g., Mª Luisa García, 1997).
German Tones and Break Indices (GToBI) – An autosegmental-metrical model

Currently, the most widespread phonological framework for describing intonation is autosegmental-metrical phonology (AM, Ladd, 2008; Pierrehumbert, 1980). As the name autosegmental-metrical suggests, AM models separate the tonal tier from the metrical tier (Ladd, 2008, p. 48ff., for an overview). Models in this framework position themselves in the tradition of level-based approaches to describe intonation. Early level-based approaches inspired by American Structuralists (e.g., Pike, 1945) described intonation in the form of tonal levels ranging between two and four distinct levels (for German Isačenko & Schädlich, 1966; Moulton, 1962). Isačenko and Schädlich (1966, p. 12), for instance, described intonation as a sequence of monotone pitch levels, high and low, (“zwei[…] und nur zwei[…] Tonstufen” (‘two and only two tone levels’)), but more levels have been assumed in other models. More recent AM models generally postulate two levels of description of intonation (high and low) for the tonal tier. The idea to assume tonal targets instead of contours

---

6 The approaches that can be subsumed under the umbrella term “AM models of intonation” were inspired by three influential PhD theses by Mark Liberman, Gösta Bruce, and Janet Pierrehumbert (Bruce, 1977; M. Liberman, 1975; Pierrehumbert, 1980). Janet Pierrehumbert’s famous dissertation (Pierrehumbert, 1980), has been called the “integrating step” in this approach (cf. Braun, 2005, p. 25). Pierrehumbert eventually described intonation as a finite state model that draws on a finite inventory of different categories. Yet, the name autosegmental metrical itself goes back to Ladd (1996) and has not yet been used by Pierrehumbert (1980) herself.
received experimental support by studies in the late 1990s and 2000s. Specifically, these studies provided evidence for a phenomenon called “segmental anchoring” (see Ladd, 2008, p. 169ff.), which describes the stability of the alignment of tonal targets under different conditions, e.g., different segmental contexts, different speaking rates etc. (Arvaniti, Ladd, & Mennen, 1998; Prieto, 2005; Silverman & Pierrehumbert, 1990).

In sharp contrast to contour-based approaches, AM models describe tonal movements as a linear sequence of tonal targets. Importantly, tonal targets, which are understood as meaningful turning points in the utterance, are the essential parts of the contour rather than the movement itself. The movement itself is considered an epiphenomenon of the interpolation of tonal targets only. Essentially, the tonal tier is linked with the metrical tier at structurally important positions, i.e., at stressed syllables where tones function as pitch accents, and at boundaries where they function as boundary tones. The link between these two independent tiers is called association—an essential claim in this approach and a basic assumption in order to understand the concept of different pitch accent types. Broadly speaking, pitch accents can be high-pitched (H*) or low-pitched (L*), with the star (*) denoting the association between the tone and the stressed syllable. Pitch accents can also be bitonal, e.g., H+L*, in which case the H-tone is referred to as a leading tone (preceding the accentual tone), or L*+H where the H-tone is referred to as a trailing tone (following the accentual tone). In this view, pitch accents and boundary tone can freely combine. To this end, Pierrehumbert (1980) suggested a finite state model for intonation consisting of an initial boundary tone, a pitch accent, a phrase accent, and a final boundary tone, see Figure 2. Note though that the tones in her model do not all possess the same status, but follow a hierarchical structure, similar to the prosodic hierarchy (Nespor & Vogel, 1986; Selkirk, 1984). For instance, the boundary tone terminates intonation phrases, while the phrase accent occurs at the ends of smaller prosodic units; together they form an edge tone. Regarding the status of the accent, however, the distinction between prenuclear and nuclear established in the framework of the British school is not explicit in the model. Being a finite state model, the central assumption of the model is that the described tonal events (initial boundary tone, pitch accent(s), phrase accent, and final boundary tone) can theoretically occur in all logic combinations. By definition, the finite state model does not make claims about the probability of certain accents and their combination. In subsequent

---

7 The notion of starred tones as markers of association has not always been uncontroversial in the literature: Specifically, it has been pointed out that that it is not entirely clear what the star refers to, which has led to different labels for pitch accents in Spanish and English for the phonetically same contour (Prieto, D’Imperio, & Fivela, 2005). Arvaniti, Ladd, and Mennen (2000) further report evidence from Greek rising bitonal pitch accents in which neither of the tonal target is aligned with the stressed syllable, questioning the fact that the tone temporarily aligned with the stressed syllable is the starred one.
distribution analyses though, it has been shown that the probability of certain tonal combinations is higher than for others (e.g., Dainora, 2006, for American English; Peters, Kohler, & Wesener, 2005, for German).

Figure 2: Finite state model of intonation according to and adapted from Pierrehumbert (1980, p. 29). It shows the possible combinations for an initial boundary tone, one or more pitch accents(s), a phrase accent, and a final boundary tone.

The model of intonation that was originally suggested by Pierrehumbert (1980) has been developed into a variety of ToBI transcription systems in different languages, e.g., Mainstream American English ToBI (Beckman & Ayers-Elam, 1997; Beckman, Hirschberg, & Shattuck-Hufnagel, 2005; for earlier descriptions see Silverman et al., 1992), even automatic ToBI systems (AuToBI) have been suggested (Rosenberg, 2010). German ToBI (GToBI) was collectively developed by Martine Grice, Matthias Reyelt, Ralf Benzmüller, Anton Batliner, Jörg Mayer, researchers in Saarbrücken, Stuttgart, München, and Braunschweig (Grice & Baumann, 2002; Grice et al., 2005). For pitch accent types, in most cases direct comparisons between the English and German ToBI system are possible; a fact we will make use of in this dissertation where necessary.

On the tonal tier, GToBI distinguishes six different pitch accent types, two monotonal accents (H* and L*, in which the stressed syllable is perceived as high and low, respectively) and four bitonal accents, containing either leading tones or trailing tones. These include two rising accents that differ in the perception of the stressed syllable as high or low (L+H* is perceived as high and L*+H as low) and two early falls, H+L* with a prominent pitch fall onto the stressed syllable (high to low) and H+!H* where the fall is less steep (high to mid-level). Unlike other AM

---

8 Besides GToBI, there are other AM models of German (Féry, 1993; Grabe, 1998; Uhmann, 1991; Wunderlich, 1988). The interested reader is referred to Grice and Baumann (2002) who describe the individual models in more detail and discuss differences between them.
models for German (Féry, 1993; Grabe, 1998; Uhmann, 1991; Wunderlich, 1988), which postulate only left-headed accents, in analogy to the German foot structure (see below), GToBI allows for pitch accents types with both leading and trailing tones. Regarding phrasing, two levels are annotated: intermediate phrases (ips), indicated by “-”, and superordinate intonational phrases (IPs), indicated by “%”.

To summarize, contour-based approaches describe intonation by focusing on the f0 movement in the area of stressed syllables and at boundaries. That way, KIM differentiates between peak and valley contours. For peak contours, the model distinguishes between three rising-falling movements based on the position of the f0 peak with regard to the lexically stressed syllable: early-, medial-, and late-peak accents. Models in the AM framework describe intonation as a sequence of high and low tonal targets that are associated with structurally important positions. For pitch accents, the starred tone is associated with the stressed syllable; unstared tones are optional. In describing different pitch accent types in this thesis, we will borrow labels from both KIM and GToBI and use them interchangeably, e.g., early-peak accent and H+L*, or medial-peak accent and L+H*.

2.1.2 Pitch accent types as phonological categories

Association and alignment

In this dissertation, we will define pitch accents as “local features in the pitch contour – usually, but not invariably a pitch change, and often involving a local maximum or minimum – which signals that the syllable with which it is associated is prominent in the utterance” (Ladd, 2008, p. 48). As foreshadowed above, the idea that pitch accents and the stressed syllable are associated is one of the central claims in AM phonology. Association, as Ladd uses the term, is an abstract (explicitly theoretical) relation between the tonal and the metrical tier. We also assume that pitch accents share the property of being associated with the stressed syllable, but differ in the alignment of the f0 peak with respect to the lexically stressed syllable: The difference between the concept of association and alignment is most explicitly spelled out in the first edition of Ladd’s book Intonational Phonology:

“I am therefore drawing a distinction […] between alignment and association. Alignment must be defined as a phonetic property of the relative timing of events in the F0 contour and events in the segmental string. Association, on the other hand, is the abstract structural property of ‘belonging together’ in some way. The fact of association entails no specific predictions about the alignment: if an H tone is associated with a given prominent syllable, we may expect to find the peak of F0 somewhere in the vicinity of the syllable, but the peak may be early in the syllable or late, and indeed it may be outside the temporal limits in the syllable altogether.” (Ladd, 1996, p. 55)

9 This dissertation is not aimed at probing the appropriateness of one approach over the other. In many cases, as Ladd points out, for instance, the two approaches can easily be translated in one another (Ladd, 2008, p. 46).
KIM also assumes an association that links the f0 peak and the stressed syllable: Different “f0 peak positions […] represent different phonological categories of intonation associated with the same stressed syllable” (Kohler, 1991c, p. 164). In this thesis, we will refer to association as the abstract structural notion, and to alignment as the temporal link between the tonal target and the stressed syllable.

For our investigation, we will concentrate on three different pitch accent types, which differ in the alignment of their tonal targets in regard to the stressed syllable, but share the association with the stressed syllable: early-peak accents (H+L*), medial-peak accents (L+H*) and late-peak accents (L*+H). These three different pitch accent types are visualized in Figure 3.

![Diagram of pitch accent types](image)

Figure 3: Visualization of the three different pitch accent types under investigation: early-peak accents (H+L*), medial-peak accents (L+H*) and late-peak accents (L*+H). Both the notation in GToBI (Grice et al., 2005) and KIM (Kohler, 1991b, 1991c) is given.

In KIM, these differences represent three distinct contours, depending on the f0 peak in regard to the position of the stressed syllable: In early-peak accents, the f0 peak is realized before the accented syllable, i.e., to its left,10 in medial-peak accents around the centre of the stressed syllable, and in late-peak accents after the accented syllable, i.e., to its right. In GToBI, these accents are described as H+L*, L+H* L*+H, respectively. These three pitch accent types are considered distinct phonological categories, as perception experiments suggest a categorical perception. These experiments will be discussed in the next subsection.

Categorical perception of pitch accent types

The assumption of different pitch accent types as phonological categories rests on the finding that alignment differences of tonal targets have been shown to lead to categorical distinctions in listeners, as indicated in pioneering studies on German and English (Kohler, 1991c; Pierrehumbert & Steele, 1989), but also in subsequent studies including other languages (D’Imperio, 2000, for Neapolitan Italian; Dilley & Heffner, 2013, for American English; Gili-Fivela, 2009, for Pisa Italian; Niebuhr, 2007b, for German). Categorical perception and imitation are two paradigms

---

10Niebuhr and Kohler (2004) showed that the start of the accented vowel may be the critical reference point for the distinction, not necessarily the beginning of the accented syllable.
predominantly applied to establish alignment differences as different phonological categories. Kohler (1991c) used the categorical perception paradigm to study the perception of pitch accent contrasts in German final falling peak contours in sentences, such as *Sie hat ja gelogen* ‘She actually lied’. He systematically moved the f0 peak in the word *gelogen* to the left and right (equidistant steps of 30 ms), taking a medial-peak production, with the f0 peak “around the stressed vowel centre” as a point of departure for the manipulation (Kohler, 1991c, p. 124), see Figure 4 (left panel). From this peak position, there were six equal steps of 30 ms to the left of the basic peak, resulting in an early-peak accent with the f0 peak “well before the onset of the stressed vowel” (Kohler, 1991c, p. 124), see Figure 4 (right panel) for the left-most position of the shifted f0-peak positions. There were also four steps of 30 ms to the right from medial-peak production, resulting in a late-peak accent with the f0 peak “at the end of the stressed vowel” (Kohler, 1991c, p. 124), see Figure 4 for the right-most position of the shifted f0-peak positions (right panel).

Figure 4: Overview of the manipulation procedure in the categorical perception experiments in Kohler (1991c). Left panel: Medial-peak production (point of departure for the manipulation) for the sentence *Sie hat ja gelogen* ‘She actually lied.’ A1 indicates the beginning of the accented vowel ([øː]), A2 the end of the accented vowel ([øː]). Right panel: The left-most and right-most positions of the shifted f0 peak in the alignment continuum (early peak and late peak, respectively). The figures are taken from Kohler (1991c, p. 119).

---

11 The paradigm that was used for categorical speech perception of f0 contours (Kohler, 1987, 1991c; Niebuhr, 2007b) has originally been developed for the discrimination of sound contrasts (Alvin M. Liberman, Harris, Hoffman, & Griffith, 1957). Categorical perception has two principles: discrimination and identification (see Goldstone & Hendrickson, 2010; Repp, 1984, for overviews). Successful discrimination of a contrast does not indicate members of the contrast belonging to distinct categories, as discrimination is also possible within a category. Hence, categorical perception requires identification of categories to be highest, with discrimination across categories being high and discrimination within categories being low. The sharpness of category contrasts can be observed in the steepness of the discrimination and identification curves at stimulus boundaries (see Repp, 1984). In imitation tasks (e.g., Braun, Kochanski, Grabe, & Rosner, 2006; Pierrehumbert & Steele, 1989), listeners are typically presented with a continuum of alignment differences which they imitate. That way, listeners’ imitations reflect their perception via their productions. The rationale of such imitation paradigms is the following: If alignment were a gradient perceptual phenomenon, one would assume speakers’ responses to fall on a continuum. If, on the other hand, alignment differences were perceptually distinct, one would assume speakers’ responses to fall in two categories (the ones which speakers map the stimuli to), i.e., showing bimodal distributions.
In total, 60 listeners were tested in the categorical perception paradigm. Results showed that listeners perceived a categorical distinction as soon as the f0 maximum reached the stressed vowel in ['lɔːt] (at stimulus number 5, which was the first of the 11 step continuum where the f0 peak was realized in the vowel). In the stimuli that preceded stimulus 5 (from left), f0 was falling in the stressed syllable. Kohler (1991c) hence argued that the early-peak contour and the medial-peak contour constitute phonologically different categorical categories. There was a second (but weaker) sensitivity peak (around stimulus 9/10), suggesting a phonological distinction between medial peak vs. late peak. This is when the initial low-pitched stretch was long enough (or the f0 peak late enough) to change the percept. The perception of medial-peak accents compared to late-peak accents is however different from the distinction between early-peak accents and medial-peak accents, pointing to a more gradual rather than a clearly categorical difference. The identification results confirmed this pattern, showing that early- vs. non-early contours are more clearly identified than that of medial- versus late-peak contours. Note that these findings were robust for different syllable structures and across different listener groups (Northern vs. Southern Germany), as Kohler (1991c, p. 149ff.) showed in a control experiment (participants from the Munich area).

Using an imitation procedure with American English listeners, Pierrehumbert and Steele (1989) provided evidence for a phonological (categorical) distinction between medial-peak accents (L+H*) and late-peak accents (L*+H). Note that they did not test the distinction between early-peak accents (H+L*) and medial-peak accents (L+H*). Specifically, Pierrehumbert and Steele (1989) manipulated an alignment continuum of L-tones and H-tones (rising-falling movement) on the syllable sequence only a millionaire, with L+H* and L*+H as the end points of the continuum on either side. The continuum consisted of 20 ms steps (15 in total), with the f0 peak ranging from 35 ms to 315 ms post [m] in millionaire. While the first step in the continuum had the f0 peak in the vowel (35 ms post vowel onset), the last step in the continuum showed the f0 peak “just before the end of the [n]” (Pierrehumbert & Steele, 1989, p. 185). Participants’ task was to imitate the contour they heard after every presentation of a rising-falling movement. In total, five American English speakers were tested (one of them was one of the authors). Overall, imitations of participants showed bimodal distributions of the f0 peak delay. Results for the L-tone delay also broadly grouped in two categories, while duration did not differ in the productions. Furthermore, f0 values of the L-tone did not differ across productions, ensuring that participants produced instances of L+H* instead of a simple H*-accent. Hence, this finding seems to suggest that speakers were unable to produce intermediate stages of the alignment continuum they were presented with. Rather, their productions of the rising-falling movement fell into two broad
categories (which eventually can be described as L+H* vs. L*+H). It needs to be mentioned though that a clear limitation of the study is the small sample size (n = 5 speakers) and the fact that not all productions showed a bimodal distribution (despite the overall tendency for bimodality).

Taken together, there seems to be evidence to consider alignment differences of the f0 peak with respect to the stressed syllable – as present in the different pitch accent types under investigation – as phonological intonational categories. Note though that in the study on German (Kohler, 1991c), the distinction between medial-peak accents and late-peak accents was not as clear as the distinction between early-peak accent and medial-peak accents. A reason why the distinction between early-peak (H+L*) and medial-peak accent (L+H*) was more discrete could be the direction of the accentual movement, which is falling in early-peak accents and rising in medial-peak accents – at least in an on-ramp view (Ritter & Grice, 2015), whereas the accentual movement is rising in both medial-peak and late-peak accents (see also Pierrehumbert & Steele, 1989, for discussion). Also, in Pierrehumbert and Steele (1989), the continuum steps representing late-peak accents (L*+H) clearly peaked after the stressed syllable while in Kohler (1991c) the latest alignment steps (late-peak accents) had its peak earlier – at the end of the stressed vowel, which might also account for the different findings across studies.

Apart from the phonological contrasts induced by tonal alignment, as they will be considered in the current dissertation, there is also “phonetic” alignment differences of tonal targets, which are considered as not implying changes in meaning. Phonetic alignment differences may be induced by different factors, e.g., tonal crowding as a response to limited voicing or an approaching phrase boundary (e.g., Hanssen, 2017; Prieto, van Santen, & Hirschberg, 1995; Rathcke, 2017; Silverman & Pierrehumbert, 1990; Yu & Zahner, 2018), or dialectal variation of speakers (Atterer & Ladd, 2004; Braun, 2007), see Prieto (2011) for an overview. To illustrate the point, testing different accent types in an imitation paradigm, Rathcke (2017) showed that both Russian and German speakers align tonal targets earlier when sonorant material is restricted. For instance, German speakers retracted the f0 peak in rising-falling contours with a medial-peak accent by a proportion of 0.8 (relative to the duration of the accented vowel) in monosyllabic words with a short vowel (Schiff [ʃɪf] ‘ship’ (0.5)) vs. sonorant disyllabic words (Linner [ˈlɪnɐ] ‘nonce word’ (1.3)). Regarding dialectal variation, Atterer and Ladd (2004) found that Southern German speakers (from Bavaria) align prenuclear accentual rises significantly later than Northern speakers (the low tonal target occurred 36 ms later for speakers from the South). We will concentrate on phonological alignment differences in different pitch accent types in this thesis, however.
2.1.3 Excursus: Communicative functions of pitch accent types

Different pitch accent types have been related to different communicative functions, certain notions of information structure in particular. Information structure has originally been defined as the packaging of information in an utterance as a response to the communicative needs of the interlocutors (e.g., Chafe, 1976; Krifka, 2008; Lambrecht, 1996). Generally speaking, three information-structural partitions can be distinguished (three "basic notions" of information structure, Krifka, 2008): the focus – background partition, the topic – comment partition, and the given – new partition (information status). In the following, we will use the term information structure as an umbrella term for all three partitions. Even though the current thesis is not aimed at studying the communicative functions of different pitch accent types, we will briefly outline the communicative differences suggested for early- (H+L*) medial- (L+H*, H*) and late-peak accents (L*+H), respectively, for completeness.

Early- vs. medial peak accents

Early- and medial-peak accents have been shown to evoke different communicative functions in German listeners. In the experiments Kohler (1991c) conducted on the categorical perception of different peak contours (see above), listeners also had to paraphrase the meanings of the contours: Overall, early-peak accents were associated with “established facts” (old information). Early-peak accents were also judged as a concluding remark of an argument. Medial-peak accents, on the other hand, were associated with new facts (new information in the discourse) and the beginning of a new argument. In contrast to early-peak accents, medial-peak accents allowed the discussion to go on, according to Kohler (1991c). Baumann and Grice (2006) confirmed the role of early-peak accents in marking the information status of a referent. They explicitly manipulated different levels of accessibility (textually displaced, part-whole relations etc.), based on the distinction between new-accessible-old information suggested by Chafe (1994, p. 73), in short contexts. Listeners judged the appropriateness of an auditorily presented sentence (e.g., Unsere Tischnachbarn riefen den Kellner 'The people at the next table called the waiter') in which the phrase-final noun, e.g., Kellner, was realized with different PSOLA-resynthesized (nuclear) pitch accent types (H*, H+L*, or deaccented, in which case an H*-accent was realized on the verb). Some relations under investigation, i.e., scenario relation, part-whole as well as textually displaced type, clearly preferred early-peak accents, while other accessible relations preferred de-accentuation.

---

12 The interested reader is referred to Kruijff-Korbayova & Steedman (2003) who trace back the use of information structure terminology to the early 1900s (see references therein).
A communicative difference within medial-peak accents

Within medial-peak accents, i.e., between H* and L+H*, two accent types which differ in the absence / presence of the low-toned leading tone, there seems to be a slight communicative difference: H*-accents have commonly been referred to as the unmarked accent which introduces new information to the discourse, while L+H*-accents have been associated with contrastive information (see, for instance, the description on German ToBI, Grice & Baumann, 2002, p. 19, or the GToBI website; Pierrehumbert & Hirschberg, 1990, on English). In fact, there are a number of studies showing that listeners use pitch accents to interpret information as contrastive or new, with L+H* indicating a bias towards a contrastive reading (Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008; Kurumada et al., 2014; Watson, Tanenhaus, & Gunlogson, 2008). For instance, in the study by Watson et al. (2008), listeners saw displays that contained four objects, a cohort pair, e.g., bed and bell, an object that was contrastively associated with the target (here chair for bed), and a distractor. Listeners were then presented with a short passage, such as *Click on the bed and the chair. Move the chair to the right of the square. Now, move the [be]...*, which introduced a contrast set (bed-chair). Participants’ task was to follow the instructions in the passage while their eye-movements were monitored. Results showed more fixations to bed in the L+H*-accent condition than in the H*-accent condition, suggesting that L+H* is a contrastive accent that makes alternatives salient. Yet, other studies have failed to find a communicative distinction between L+H*- and H*-accents (Krahmer & Swerts, 2001; Welby, 2003). Welby (2003), for instance, tested the ability of rising-falling contours (H* L% vs. L+H* L%) to project focus, i.e., the ability of a pitch accent on a word to mark focus on a larger constituent. Her hypothesis was that the H*-accent condition would have the ability to project focus, while the L+H*-condition, being contrastively used, would not. Hence, an L+H*-accent with a nuclear fall on the sentence-final noun (e.g., in *I read the Dispatch*) was expected to be inappropriate as an answer to a question that triggered broad focus (e.g., *How do you keep up with the news*?). Contrary to what was expected, listeners’ responses indicated that focus projection was not affected by accent type, speaking against L+H* as a contrastive accent. Irrespective of whether or not there is a difference in intonational meaning between H*- and L+H*-accents, in the current study, we will treat these two accents as one category as they share one essential criterion, the alignment of f0 peak and lexical stress, which is critical to the rationale of the study.

Note that authors have challenged the presence of this criterion for distinction between the two accents (Dilley, Ladd, & Schepman, 2005; Ladd & Schepman, 2003). In fact, these two accent types are highly prone to be confused by labellers (Grice et al., 1996).
Medial- vs. late peak accents

The distinction in communicative function between medial-peak accents and late-peak accents was less clear-cut in the task administered by Kohler (1991c) than the distinction between early- and media-peak accents (see above). In the study by Kohler (1991c), late-peak accents were associated with new information (similar to medial-peak accents) – but, in contrast to medial-peak accents, with new information that was additionally uttered with emphasis. Attitudinally, late-peak accents have further been shown to signal sarcasm (Lommel & Michalsky, 2017), and to frequently occur in rhetorical questions, both in wer- and polar interrogatives (Braun, et al., 2019). Late-peak accents in prenuclear position have also been suggested to signal contrastively used sentence-initial topics (for experimental evidence, Braun, 2006; for a theoretically motivated approach, Büring, 1997; Büring, 2006). Braun (2006), for instance, investigated the accent realization and interpretation in sentence-initial topics in German that are either used contrastively or non-contrastively. She employed target sentences with a theme-rheme structure in a short context (e.g., Die Malayen leben von der Landwirtschaft. ‘The Malaysians live from agriculture.’). The theme identified a topic (e.g., Malayen) and the rhyme made a proposition about it (leben von der Landwirtschaft). Importantly, the contexts were either contrastive (e.g., Malaysian vs. Indonesian) or non-contrastive (e.g., Malaysia for Malaysian). Results of the intonational analysis showed that for contrastive contexts, f0 peaks were both significantly later and higher than non-contrastive accents (even though a clear-cut distinction between L+H* vs. L*+H was difficult to entertain based on a perception and annotation experiment).

In sum, different pitch accent types may serve different communicative functions. Table 1 summarizes those communicative functions of pitch accent types that have been pointed out in the literature. In the following section, we will consider lexical stress in German in more depth, addressing both its concrete (phonetic correlates) and its abstract nature (metrical strength relations).

<table>
<thead>
<tr>
<th>Accent type (GToBI label)</th>
<th>Communicative function (as described in the literature)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early peak</td>
<td>H+L* / H+!H*</td>
</tr>
<tr>
<td>Medial peak</td>
<td>H*</td>
</tr>
<tr>
<td>Medial peak</td>
<td>L+H*</td>
</tr>
<tr>
<td>Late peak</td>
<td>L*+H</td>
</tr>
</tbody>
</table>

Inferable / accessible information
New information
Contrastive information
New information, emphasis, sentence-initial topics, sarcasm, “rhetoricity”, self-evidence

14 Note that Braun (2006) makes use of the distinction theme – rheme structure. For the purpose of this thesis, the terms theme-rheme is used interchangeably with topic-comment.
2.2 Lexical stress

Following Ladd (2008), we will draw a distinction between utterance-level pitch accents, which are part of the intonation contour, and lexical stress, which we consider a property of the word. Note that the models for intonational description, KIM and GToBI, also assume this distinction. We will furthermore adopt the view that lexical stress has two notions (Ladd, 2008, p. 58): a concrete phonetic one (dynamic stress), which refers to the acoustic correlates of stress and an abstract one, defined in terms of prominence relations, usually expressed in a metrical foot structure (metrical strength). Based on this assumption, we will describe lexical stress as a property of the word, which is determined by an abstract prominence relation, but surfaces in acoustic correlates in the signal.

2.2.1 Physical correlates of lexical stress

As indicated above, due to utterance-level intonation, the f0 peak is an unreliable indicator of the position of lexical stress. Hence, if a word occurs in accented position, the contribution of the f0 peak to lexical stress depends on the pitch accent type. In unaccented position, f0 information may be irrelevant altogether. Results from various production studies with linguistically diverse speech materials consistently show that f0 is not a direct acoustic correlate of stress (e.g., Dogil, 1995; Kochanski, Grabe, Coleman, & Rosner, 2005; Sluijter & Van Heuven, 1996a, 1996b; Szalontai, Wagner, Mády, & Windmann, 2016). In a recent meta-analysis, Roettger and Gordon (2017) addressed methodological issues in research on the correlates of lexical stress, considering 110 studies in 75 languages. The authors point out that studies on the correlates of stress need to be interpreted with care:

“The literature is rife with studies of the acoustic exponents of what is often referred to as stress but the methodological diversity of this research has created an unclear picture of the properties robustly associated with it.” (Roettger & Gordon, 2017, p. 1)

Roettger and Gordon (2017, p. 1) explain that this “unclear picture” arises mainly due to phonetic factors – that are genuine to lexical stress – being conflated with other factors caused by phrasal phenomena: on the left-edge of a phrase, initial strengthening; on the right-edge of a phrase, final lengthening and boundary movements, such as final lowering and accentual prominence (pitch accents).

For German, several studies have investigated the acoustic correlates of word stress, controlling for the above-mentioned factors that might skew the picture (e.g., Dogil & Williams, 1999; Kleber & Klippahm, 2006; Mooshammer, 2010; Mooshammer & Geng, 2008; Schneider & Möbius, 2007). These studies primarily used stress minimal pairs to directly compare the stress status (stressed vs. unstressed) in segmentally identical contexts, e.g., ['kɔn. stants] vs. [kɔn.'stants], [ˈtɛːnɔʁ]
‘stance’ vs. [teˌnɔː] ‘tenor’, [ˈum.fɑːr.nə] ‘knock sth. down’ vs. [um.ˈfɑː.rən] ‘go around sth.’, but cohort competitors that overlap in segmental material but differ in the position of stress, e.g., [le.na] ‘female first name’ vs. [leˌnɔː] ‘name of a product’ have also been used. Finally, acoustic correlates of stress have also been investigated in reiterated speech. Overall, in studies on acoustic correlates of lexical stress in German, duration, spectral balance, and vowel quality have been analysed.\textsuperscript{15} We will briefly summarize the experimental studies on each of these acoustic correlates.

\textit{Duration}

Duration has been identified as an important correlate of lexical stress in German: Dogil and Williams (1999), for example, studied different acoustic correlates of word stress in the German minimal pair [ˈkɔn. stants] vs. [ˈkɔnˌstants] embedded in sentences. They controlled for accentuation by placing targets in different sentence positions, e.g., \textit{Die Konstanz der Uni Konstanz ist bemerkenswert} ‘The stability of the University of Konstanz is noteworthy’ (position 1) vs. \textit{Die Uni Konstanz zeichnet sich durch eine bemerkenswerte Konstanz aus} ‘The University of Konstanz is characterized by a noteworthy stability’ (position 2). Three speakers produced the stress minimal pair in different sentence positions, repeating each sentence three times. Results revealed that stressed syllables were significantly longer than their unstressed counterparts. In detail, for the member with stress on the first syllable ([ˈkɔn. stants]), the proportional duration for the first syllable in regard to the whole target was 38.11% for position 1 and 40.92% for position 2. For the member with stress on the second syllable ([ˈkɔnˌstants]), the proportional duration for the first syllable in regard to the whole target was 31.29% for position 1 and 30.47% for position 2. Proportional values for the individual sounds of the first syllable ([ˈkɔn]), either stressed or unstressed) revealed that the difference in duration between stressed and unstressed syllable was mainly due to the reduced duration of the vowel in unstressed syllables as well as a reduction of voice-onset time in unstressed conditions; the sonorant syllable coda [n] was not reduced as a function of stress. Dogil and Williams (1999) emphasize, however, that in the minimal pairs employed as experimental materials, acoustic correlates might be hyper-articulated due to the inherent stress contrast. As an alternative to minimal pairs, the authors suggest using reiterated speech, i.e., sequences of recurring nonce syllables, e.g., \textit{dadada}. To counteract this methodological concern, Dogil and Williams (1999) conducted another experiment with reiterated speech. This time, five reiterated \textit{da}-syllables were

\textsuperscript{15} Overall intensity, which has shown to be higher in English stressed syllables compared to unstressed syllables (Fry, 1955), will not be reported as the results are difficult to interpret due to the inherent difficulty of measuring overall intensity (e.g., distance to microphone).
embedded in answers to a question that elicited different focus positions. Results revealed that syllable duration was the only acoustic value that differed significantly as a function of stress status, irrespective of focus position. Based on these findings, the authors conclude that duration is “the primary correlate of German word stress” (Dogil & Williams, 1999, p. 292). Schneider and Möbius (2007) further confirmed the important role of duration in a study that investigated stress correlates in child-directed speech. Utterances of three mothers directed to their children (who were two to six years old) were gathered over a time of at least one year. Nonsense stress minimal pairs were elicited by providing mothers with toys of particular names, e.g., ['bi.mo] for a brown bear and [bi.'mo] for a polar bear. The results showed that the duration of vowels significantly differed as a function of stress. This difference was observed for all three speakers and across all vowels under investigation ([a], [i], [o]), suggesting syllable duration to be a robust acoustic correlate of stress in German. Note that the other acoustic measures they performed (overall intensity, formant measures) were not as consistent as the durational difference.

Spectral balance
Stressed syllables have further been shown to exhibit an increased vocal effort compared to unstressed syllables: For Dutch, Sluijter and van Heuven (1996b) found spectral balance (also known as spectral tilt), i.e., the difference in energy between higher and lower frequencies in the spectrum, to be one of the most robust correlates of lexical stress (stronger than vowel quality and overall intensity, but comparable to duration). Specifically, their production study in which ten Dutch speakers produced disyllabic stress minimal pairs (with and without pitch accent) showed that stressed and unstressed vowels differed in spectral balance, with stressed vowels having more high-frequency emphasis compared to unstressed vowels, i.e., generating a flatter spectrum. Their results indicated that the effect found for spectral balance is due to an increase in physiological effort when producing stressed vowels (Sluijter & van Heuven, 1996b).

For German, Mooshammer (2010) investigated the effect of stress and phrase-level accent on vocal effort, using articulatory and acoustic measures. An increased vocal effort is assumed to cause less symmetric, left-skewed glottal pulses due to the shorter closing phase of the glottis (see Mooshammer, 2010, p. 1052, for a detailed discussion). This change in the pulse of the glottis generates a shift of energy in the spectrum, with more energy in higher frequency regions than in lower frequency regions (also assessed by Sluijter & van Heuven, 1996b, see above). Particularly, Mooshammer (2010) measured glottal pulse and the open quotient of the glottis using a laryngograph. Regarding acoustic measures, H1-H2, i.e., the
difference in energy between the first and the second harmonics (operationalizing the open quotient of the glottis) and H1-A3, i.e., the difference in energy between the first harmonic and the third formant (operationalizing spectral tilt) functioned as dependent variables. Participants produced pairs that differed in stress position (e.g., [ˈle.na] vs. [le.ˈnoʊ]) embedded in sentences, either as new information or as given information (hence triggering accentuation vs. deaccentuation). Loudness was also manipulated, i.e., whether participants produced the targets with a loud vs. a soft voice. The articulatory results produced a clear picture: Stressed syllables were produced with more vocal effort, as indicated by differences in the glottal pulse and the open quotient of the glottis between stressed and unstressed conditions, even in the absence of f0 movements, i.e., in the given information condition. An increased vocal effort created a more asymmetrical glottal pulse with a shortened opening phase of the glottis and hence a steeper edge. These results were similar to loud speech. Yet, contrary to what Sluiter and van Heuven (1996b) found for Dutch, the acoustic results on German for the distinction of stress status were less clear. The open quotient measure H1-H2 was affected by stress, but not by focus or loudness of speech, while spectral tilt (H1-A3) was only affected by loudness, but not by stress (Mooshammer, 2010). The author explains the null results for stress status on spectral tilt for the acoustic measures by changes in the formant frequencies that were caused by vocal tract modifications. From our perspective, it is difficult to judge whether this asymmetry in results between the two studies outlined above is due to cross-linguistic (German vs. Dutch) or methodological differences across studies.

Vowel quality

Stressed syllables show more peripheral vowel quality than unstressed ones. From a phonetic viewpoint, vowel reduction has been described as a phenomenon of vowel target undershoot, due to either coarticulatory processes, a tendency to centralize vowels, or both (see Fourakis, 1990 on English; Lindblom, 1963 on Swedish; Padgett & Tabain, 2005 on Russian). Regarding centralization processes, early on, Delattre (1969) demonstrated that unstressed vowels are reduced compared to stressed vowels in four languages (English, German, Spanish, and French). Specifically, Delattre (1969) measured the first and second formant, as well as the position of the tongue in stress minimal pairs, e.g., Manier [ma.ˈniːə] ‘manner’ vs. manieriert [ma.ˈniːɐ̯.ɐt] ‘affected’ for German and competing [kɔm.ˈpiː.tɪŋ] vs. competition [kɔm.ˈpiː.tʃən] for English. Acoustic results revealed that English exhibited the strongest tendency to reduce unstressed vowels (by a proportion of 18%), as compared to 6% in German, and 7% and 4% in French and Spanish, respectively. Regarding coarticulatory processes, Mooshammer and Geng (2008) showed in an
electromagnetic midsagittal articulography study that unstressed unaccented vowels exhibit a greater degree of coarticulation with the surrounding consonantal context (the preceding and following /t/) than stressed accented vowels. In that study, participants produced words that contained tVt-sequences (V = vowel) embedded in carrier phrases, such as “Ich habe ‘Vt gesagt, nicht tV’tal” ‘I said ‘Vt, not tV’tal’. The authors interpret the increased coarticulation as a result of the decreased effort which leads to a reduced formant space in unstressed unaccented vowels. Yet, the experimental manipulation in Mooshammer and Geng (2008) does not allow for comparisons between stressed (unaccented) and stressed (accented) vowels, as the factors lexical stress and accent were not crossed orthogonally. However, as vocal effort – operationalized as glottal pulse and open quotient of the glottis – was also increased for stressed syllables in unaccented position in the study by Mooshammer (2010), we may tentatively extrapolate that the difference in the degree of coarticulation between stressed and unstressed syllable might hold irrespective of accentuation. In short, from a phonetic perspective, lexical stress is marked by clear acoustic correlates in German: duration, spectral balance, and vowel quality seem to reliably signal lexical stress.

### 2.2.2 Stress as an abstract notion

*Different typological stress systems*

In linguistic typology, there is a general distinction between languages that have stress and languages that do not (Hyman, 1977, 2014). According to Hyman (1977, 2014), most of the world’s languages dispose of stress. An example of a language without stress is Ambonese Malay, a language which is spoken by 250,000 people on an island called Ambon near Maluku Island (Maskikit-Essed & Gussenhoven, 2016). In stress languages, one and only one syllable has the highest degree of prominence. This highest degree of prominence in a (lexical) word has been referred to as the *culminative* function of stress in the phonological literature (e.g., Hyman, 1977). The stress position in a word can either be fixed, as in Hungarian, Slovak, or Czech, or free, as in German, English, or Spanish. In fixed stress languages, the position of stress is typically constrained to a certain syllable in the word. Hungarian, Finnish, Slovak, or Czech, for instance, exhibit word-initial stress by default (Cutler, 2005, for an overview). In Polish, stress typically falls on the penultimate syllable (but there are some exceptions in which stress falls on the antepenultimate syllable, Domahs, Knaus, Orzechowska, & Wiese, 2012, for a summary; Hayes, 1995). Since stress is (highly) predictable in these languages, lexical stress is said to serve a *demarcative* function, serving as a boundary marker that can be used to segment words (Cutler, 2005, for an overview; e.g., Hanulíková, McQueen, & Mitterer, 2010, on Slovak; Suomi, McQueen, & Cutler, 1997, on Finnish). Hanulíková et al.
Chapter 2 (2010), for instance, tested the reliance of Slovak listeners on the possible-word constraint (PWC)\(^\text{16}\) and fixed stress during speech segmentation. In a word-spotting task, listeners had to spot real words such as *ruka* 'hand' in different contexts: a prepositional-consonantal context (e.g., */gruka*/), a non-prepositional-consonantal context (e.g., */truka*), and a syllable context (e.g., */dugruka*/). In the first experiment, listeners were faster in prepositional-consonantal contexts than in non-prepositional-consonantal contexts, in which, in turn, they were faster than in syllabic contexts. Hence, it seemed that listeners did not rely on the PWC. On the contrary, listeners even found it easier to spot words in the consonantal contexts than in the syllable contexts. However, the consonantal and the syllable context differed in regard to stress status in the first experiment: While the target *ruka* was stressed in the consonantal context, it was not in the syllable context. This is why the authors controlled for stress in their second experiment by cross-splicing a stressed version into the syllable context. As a result, the latencies of participants’ responses were reverted: This time, the target *ruka* was recognized faster in the syllable context than in the non-prepositional context. The authors hence concluded that stress, besides the PWC, is an important factor in the segmentation of fixed stress Slovak.

In free stress languages, by contrast, stress can theoretically fall on any syllable in a word and is hence lexically contrastive – even though there are only very few “real” stress minimal pairs in these languages (Cutler, 2005), such as *Konstanz* [kon.stants] ‘the name of a city’ vs. *Konstanz* [kon.'stants] ‘stability’. In these free stress languages, there are language-specific preferences for stress distribution, so stress is typically not entirely unpredictable. The question that arises is whether stress is also an important factor in the segmentation of free stress languages like German, English and Dutch. In the following, we will first concentrate on the distribution of stress placement in free stress languages with a focus on the three West-Germanic languages German, English, and Dutch, and account for the metrical and rhythmic structure in metrical theory.

*Stress placement in three West-Germanic so-called “free stress” languages*

In order to examine regularities in stress placement in English words, Cutler and Carter (1987) analysed two computerized English dictionaries, the *MRC Psycholinguistic Database* on British English (Coltheart, 1981, as cited by Cutler & Carter, 1987) and the *Merriam-Webster Pocket Dictionary* on American English. In their

\(^{16}\) According to the possible-word constraint, PWC (Norris, McQueen, Cutler, & Butterfield, 1997), during segmentation, any lexical parse is dismissed that results in a sequence without vowels (for further experimental evidence, see Cutler, Demuth, & McQueen, 2002; Norris et al., 2001). Testing the use of the PWC in Slovak is particularly interesting since this language allows for single-consonant words and hence provides an interesting test case for the PWC (Hanuliková et al., 2010).
Intonation and lexical stress

Cutler and Carter (1987) coded whether a word was monosyllabic or polysyllabic and whether or not the first syllable was stressed. The concept of word was defined as the lexical and grammatical words (only the coding in the MRC Psycholinguistic Database allowed to differentiate the two though); morphology, i.e., whether a word was monomorphemic or not was not considered (Anne Cutler, p.c.). For the MRC Psycholinguistic Database, the distribution analysis (for lexical and grammatical words) showed that there were 11.7% monosyllables, 50.6% polysyllables with initial primary stress, 10.7% polysyllables with initial secondary stress, and 27.0% polysyllables with a weak initial syllable. Very similar numbers were observed for lexical words only (see their Tables I and III, p. 135-136). Likewise, for the Merriam-Webster Pocket Dictionary, the distribution analysis (for both lexical and grammatical words) showed that monosyllables occurred in 17.3% of the cases, polysyllables with initial primary stress in 48.1%, polysyllables with initial secondary stress in 12.2%, and polysyllables with a weak initial syllable in 22.4% of the cases. Cutler and Carter (1987) further showed that the words with initial stress are more frequent than the words with a weak initial syllable. In a corpus based on spontaneous British English, which contained many high-frequency words, they found 90% of the lexical words being stressed on the first syllable. Hence, even though words with a weak initial syllable make up around a quarter of the English lexicon, these words are underrepresented in actual speech, as Cutler and Carter (1987) argue.

The same tendency can be observed in other West-Germanic languages. Cutler (2012, p. 53) presents a distribution analysis of lexical stress placement in the lexicon of three West Germanic languages: English, German, and Dutch. The distribution of stress placement in English, German, and Dutch is shown from left to right in Figure 5. The graphs for each language are based on data from the CELEX lexical data base (Baayen, Piepenbrock, & Gulikers, 1993) and show monosyllabic words (Mono, purple), and polysyllabic words with stress on the first (Poly1, pink), the second (Poly2, yellow), the third (Poly3, light green), and the fourth or later syllable (Poly4/+, magenta). Again, there is no distinction between numbers of morphemes (Anne Cutler, p.c.). Overall, the distribution reveals a strong tendency towards stress on the word-initial syllable for lexical statistics of English, German, and Dutch. Specifically, Figure 5 shows that for English, monosyllabic words (Mono, purple) occur in approximately 11% of the cases, polysyllabic words with stress on the first (Poly1, pink) in approximately 48%, with

Note that Cutler (2012, p. 53) does not provide the exact percentages of the distribution of stress placement for the three West-Germanic languages in the text, but only visualizes the distribution, see Figure 5, adapted from Cutler (2012, p. 53). The proportions for the different stress placements in different languages given in this thesis were derived from Figure 5 by the author of this thesis and represent hence an approximation only.
stress on the second (Poly2, yellow) in approximately 28%, with stress on the third (Poly3, light green) in approximately 10%, and with stress on the fourth or later syllable (Poly4/+, magenta) in approximately 3% of the cases. For German, monosyllabic words (Mono) occur in around approximately 3% of the cases, polysyllabic words with stress on the first (Poly1) in approximately 59%, with stress on the second (Poly2) in approximately 22%, with stress on the third (Poly3) in approximately 10%, and with stress on the fourth or later syllable (Poly4/+) in approximately 6% of the cases. Finally, for Dutch, monosyllabic words (Mono) occur in around approximately 3% of the cases, polysyllabic words with stress on the first (Poly1) in approximately 59%, with stress on the second (Poly2) in approximately 24%, with stress on the third (Poly3) in approximately 10%, and with stress on the fourth or later syllable (Poly4/+) in approximately 4% of the cases.

As Cutler (2012, p. 53) further outlines – here giving concrete numbers, the tendency towards word-initial stress in these languages becomes even more obvious if secondary stress is included in these calculations. In case secondary stress is included in the counting, then 81%\(^{18}\) of the English lexicon and 89% of the German and Dutch lexicon is stress initial. It needs to be mentioned though that Cutler and Carter (1987) and Cutler (2012) only broadly distinguish between monosyllabic vs. polysyllabic words, but do not further divide polysyllabic words based on the number of syllables. As a whole, there seems to be a strong bias towards word-initial stress in the three West-Germanic languages, but the picture might be more diverse for different polysyllabic words, i.e., disyllabic, trisyllabic or even longer words.

\(^{18}\) Contrary to the numbers regarding primary stress (depicted in Figure 5), Cutler (2012, p. 53) gives exact proportions of different stress placements in regard to counts that include primary and secondary stress. Hence, the proportion of stress-initial words (if primary and secondary stress is included) reported in this thesis are taken from Cutler (2012, p. 53) and are not based on estimation.
A phonologically-oriented distribution analysis on German by Féry (1998), also based on data from the CELEX lexical database (Baayen et al., 1993) differentiated between different types of polysyllabic words. Féry (1998, p. 104) discarded monosyllabic words and proper names from her analysis and further only included monomorphemes, resulting in a corpus of 6100 words in total: 3425 disyllabic words, 1312 trisyllabic words, 991 quadrisyllabic words, and 384 longer words. The main finding of her analysis is that stress placement is highly influenced by whether or not the final syllable contains a schwa. In her analysis of German disyllabic words, Féry (1998) distinguishes words with a final schwa syllable (necessarily stressed on the penultimate) and words which have two full vowels (where both syllables could potentially be stressed). Therefore, if one only considers words with two full vowels, the picture changes and only approximately 40% of the disyllables have penultimate stress, while the other 60% are stressed on the final syllable (penult). According to Féry (1998), similar shifts in lexical statistics can be observed for trisyllabic words. Trisyllabic words with a final schwa are very often stressed on the penultimate syllable (more than 90% of the cases), in 7% of the cases stress falls on the antepenultimate syllable, i.e., the first syllable. This high proportion of penultimate stress is drastically reduced for non-schwa final syllables, occurring in only 18% of the cases. In this group, half of the words are stressed on the final syllable (ultimate), and around one third on the first syllable (antepenultimate).

Frequency-wise, words with a final schwa are in the majority in German disyllabic words: S. Bartels, Darcy, and Höhle (2009) created a list with all disyllabic monomorphemic words, also using the CELEX database. In 83% of the cases, the disyllabic words were trochaic with a final schwa, in 9% they were trochaic with a final full vowel, and in 9% they were iambic. Given the high number of disyllabic words in German in general and the high number of final schwa syllables in disyllabic words in particular (S. Bartels et al., 2009; Féry, 1998), the trochaic pattern seems to be the default in German disyllabic words. It has been argued that trochees may even be more frequent in German than in English (Höhle et al., 2009). Compared to English, German exhibits a lower number of monosyllabic words (see Figure 5). One reason for this asymmetry may be the presence of inflections in German (as has been argued by e.g., Höhle et al., 2009; see Wurzel, 1989, on the German inflectional system). German inflectional endings (marking number, gender, or case) generally add a syllable to a monosyllabic stem, resulting in a trochee with a final schwa syllable. As Höhle et al. (2009) argue, this changes the rhythmic structure of German, as compared to English. The authors further conclude that, as trochees are expected to be more frequent in German than in English, the rhythmic structure in German is assumed to have a more stringent alternation between strong and weak syllables than in English. Extending the argument by Höhle et al. (2009), it seems likely that the number of trochees might even be higher in areas
where the Alemannic dialect prevails (Konstanz, Freiburg, Switzerland). In Alemannic, diminutive suffixes are often added to monosyllabic nouns, e.g., [ˈhauʃ]/[ˈhɔɪʃ.li] for ‘house’/‘small house’ (e.g., Siebenhaar & Wyler, 1997, p. 39, on the number of diminutives in Swiss German). Now that we have established tendencies for primary stress placement in West-Germanic languages, we turn to phonological theories of stress placement.

*Lexical stress and metrical theories*

In their seminal work, M. Liberman (1975) and Liberman & Prince (1977) proposed a metrical theory of word stress. The 1980s and 1990s gave rise to a variety of different metrical stress theories (Alber, 1997; Giegerich, 1985; Hayes, 1995; Kager, 1995; Nespor & Vogel, 1986; Selkirk, 1980). In these theories, stress is considered a relative phenomenon, with one syllable being stronger than the other. This prominence relation has traditionally been expressed via metrical feet, metrical grids, or trees (see Kager, 1995, for a discussion of different notations of prominence relations). In essence, lexical stress is part of a hierarchical metrical structure (Nespor & Vogel, 1986) which determines the abstract prominence relations in a phrase. In this hierarchical structure, the syllable is dominated by the foot, which is in turn dominated by the phonological word, see (1) where \( w \) represents the phonological word, \( \varphi \) the foot, and \( s \) the syllable. The letters appearing as subscript indicate whether a given prosodic element is strong (S) or weak (W).

\[
(1)
\]

![Diagram](image)

Regarding the level of the foot (\( \varphi \)), different foot types have been proposed, depending on the number of syllables and the direction in which feet are constructed (leftwards or rightwards), see van der Hulst (1999) for an overview. Hayes (1995), for instance, assumes strictly binary feet, which either have stress on the initial (SW, trochaic foot) or the final element (WS, iambic foot).\(^{19}\) The prototypical

\(^{19}\) Note that ternary feet also have been proposed (e.g., Halle & Vergnaud, 1987) in order to capture language with ternary rhythm. For theories that assume binary feet, there are two kinds of exceptions, degenerated feet (comprising only one syllable) or extrametricality, a phenomenon which has been used to describe syllables that fall outside the rhythm (Hayes, 1995; Kager, 1995).
foot for Germanic languages is also considered binary, consisting of a sequence of two syllables – one of which is strong, stressed, and the other weak, unstressed (Hayes, 1995; Selkirk, 1980). In the three West-Germanic languages, German, English, and Dutch, stress is typically assigned by building trochaic feet (SW) from right to left, with the rightmost foot receiving word-level prominence (for German, e.g., Féry, 1998; Giegerich, 1985; Jessen, 1999; Wiese, 1996), see (2) and (3) for illustration of a metrical structure in English and German, respectively; the metrical structure for the examples given in (2) and (3) in Dutch would be identical. In this bracket-grid notation, ‘x’ represents the strong element. The first level shows the prominence relations (‘x’ strong, ‘.’ weak) on the foot level, ( ) brackets group the syllables into feet. One level higher, prominence relations between feet are given on the level of the phonological word, with the brackets ( ) framing the phonological word.

(2) English

\[
\begin{align*}
(x) & \quad (x) \quad (x) \\
(x.\) & \quad (x)(x.\) \quad (x)(x.\)
\end{align*}
\]

Tiger September propaganda

(3) German

\[
\begin{align*}
(x) & \quad (x) \quad (x) \\
(x.\) & \quad (x)(x.\) \quad (x)(x.\)
\end{align*}
\]

Tiger September Propaganda

Note that in German stress assignment has not been fully resolved yet. Two opposing theories exist: Representatives of a *quantity insensitive approach* (Eisenberg, 1991; Wiese, 1996) assume the syllable structure not to influence stress assignment in German, while representatives of a *quantity sensitive approach* (Féry, 1998; Giegerich, 1985; Vennemann, 1991) argue that stress assignment is guided by the weight of a syllable. Note that phonological analyses on stress assignment reviewed below (e.g., Féry, 1998; Wiese, 1996) report stress placement from right to left. For Wiese (1996), stress falls on one of the last three syllables in a word, with penultimate stress being the default stress pattern, e.g., Libelle [liˈbɛ.lə] 'dragonfly'.

---

20 The interested reader is referred to Hulst (1999) who provides metrical descriptions of various other European languages.

21 The interested reader is also referred to Janssen (2003, pp. 11-47) who provides a detailed summary of different representatives of the two accounts. We only concentrate on two main representatives here.
other stress positions are considered to be deviations from the norm, which need to be pre-specified in the lexicon. For instance, in *Kimono* [ˈkiːmo.no] ‘kimono’, the final vowel is marked as being extrametrical, whereas in *Samurai* [za.ˈmu.ˈraɪ] ‘samurai’, the final vowel is underlyingly associated with a foot. Wiese (1996) mentions the fact that there is a statistical preference for antepenultimate stress in the case of an open penultimate syllable for borrowed words. Summarizing the stress rules that he proposes, German stress is assigned such that trochaic feet are built, starting at the right edge of the word. Others, most notably Féry (1998), on the other hand, assume syllable weight to be the decisive factor in the assignment of word stress in German. The main argument that Féry (1998) derives from her analysis is that the existence of a schwa-syllable highly influences stress assignment and skews the stress pattern distribution in German. While stress assignment in her approach is based on principles and rules and thus requires fewer lexical pre-specifications, it needs to establish some exceptional rules to account for the stress phenomena in words of Germanic origin. For the purpose of the current thesis, it is important to emphasize that both accounts (quantity sensitive and quantity insensitive) commonly assume trochaic feet, built from right to left, with the rightmost one being the strongest.

Hence, under the assumption that German (as other West-Germanic languages) typically builds trochaic feet from right to left, bisyllabic words typically receive primary stress on the initial syllable (*tiger*/Tiger in (2) and (3) while in poly-syllabic words primary stress occurs in the word-final foot (*September*/September and *propaganda*/Propaganda in (2) and (3)). In that case, initial syllables may bear secondary stress. Clearly, the lexicon of a given language consists of words that differ in the number of syllables they are composed of. As a consequence, the distribution of stress pattern may also vary as a function of word length, as has been confirmed by distribution analyses of stress placement (e.g., Cutler, 2012, p. 53; Cutler & Carter, 1987; Féry, 1998).

*West-Germanic languages as stress-timed languages*

The rhythm of a language is typically assumed to exhibit a tendency for strong and weak syllables to alternate (Hyman, 1977). In this regard, the rhythm of the three West-Germanic languages German, English, and Dutch has traditionally been classified as stress-timed, as opposed to syllable-timed (or morae-timed). The categorization into different rhythm classes rests on the initial assumption that the timing between different units of speech (feet in stressed-timed languages, syllables in syllable-timed languages, and morae in mora-timed languages) is regular in the sense that the respective units are of equal duration, a phenomenon known as *isochrony* (Abercrombie, 1967; Pike, 1945). To this end, since the 1970s, researchers
have grouped languages according to their rhythm in order to account for rhythmical differences between languages (Nespor, Shukla, & Mehler, 2011, for an overview). There are three different groups: stress-timed languages, e.g., German, English, or Dutch, or syllable-timed languages, e.g., Italian or French (Abercrombie, 1967; Pike, 1945), and mora-timed languages, such as Japanese (Warner & Arai, 2001, for a review). From a perception perspective, it is well documented that language-specific differences in rhythm lead to different processing strategies: To illustrate the point, in a seminal study, Nazzi, Bertoncini, et al. (1998) tested French new-borns’ ability to discriminate low-pass filtered sentences from different language pairings in a high-amplitude sucking paradigm. Infants succeeded in discriminating languages that belonged to different rhythmic classes (English vs. Japanese), but failed to do so when the languages belonged into the same rhythm class (Dutch vs. English). Discrimination was robust when languages of the same rhythm class were combined (Dutch/English vs. Italian/Spanish), suggesting that infants group languages in broad rhythmic groups and generalize over differences within one group. As will be shown in Chapter 3, infant and adult listeners further rely on rhythmic units for speech segmentation (Cutler, Mehler, Norris, & Segui, 1983; Nazzi et al., 2006; Otake, Hatano, & Yoneyama, 1996).

In contrast to rhythm-class based differences in perception, the acoustic notion of *isochrony* is controversial. Attempts to find phonetic evidence for the phenomenon of isochrony have largely remained inconclusive (e.g., Arvaniti, 2009; Dauer, 1983; Kohler, 2009; Prieto et al., 2012): Research within the last 25 years, starting with Dauer (1983), has shown that rhythm is based on the language-specific phonotactics, the degree of vowel reduction, or the absence or presence of vowel length contrast rather than isochrony of speech units. Different metrics to quantify rhythm have been developed: Ramus, Nespor, and Mehler (1999), for instance, used the proportion of vocalic intervals (%V), and the average standard deviation of consonantal (ΔC) and vocalic intervals (ΔV), with %V and ΔC best predicting the traditional distinction in different rhythm classes. Generalizing about the metrics proposed in the literature (Nespor et al., 2011, for an overview), stress-timed languages seem to allow for complex consonantal clusters, lengthened stressed, and reduced unstressed vowels. In syllable-timed languages, vowel length is equal over different syllables and clustering is reduced (see Arvaniti, 2009, for a critical discussion on rhythm metrics). West-Germanic languages typologically belong to the stress-timed languages.

Taken together, German is considered to be a stress-timed language, similar to English and Dutch in which stress is free but typically assigned by building trochaic feet from right to left. Given the frequent occurrence of disyllabic words, a substantial portion of all words in German start with a strong syllable, with the richness
of the German inflectional system adding to the high proportion of trochees with a final schwa. Consequently, although stress position in German is not as predictable as in fixed stress languages, there is a strong tendency towards word-initial stress.

2.3 Interim summary

To summarize, in Chapter 2, we discussed the concept of intonation (2.1), different pitch accents in particular, and lexical stress (2.2). Regarding intonation, we saw that in intonation languages, such as German, English or Dutch, different pitch accent types primarily signal pragmatic functions, e.g., information-structural differences in an utterance – even though there seems to be no one-to-one relation between form (accent type) and function (meaning). In intonational phonology, pitch accents are thought to be associated with the stressed syllable (abstract, theoretical relation). Importantly, the actual temporal alignment of the f0 peak with respect to the position of the stressed syllable may vary: In medial-peak accents (L+H*), the f0 peak is aligned with the stressed syllable, while in early-peak accents (H+L*) it is aligned with an unstressed syllable preceding the stressed syllable, and in late-peak accents (L*+H) with an unstressed syllable following the stressed syllable. These alignment differences in different accent types need to be considered phonological distinctions, as categorical perception and imitation studies have revealed. What we can derive from these alignment differences for the relation between pitch accent type and lexical stress is that f0 is an unreliable cue to the position of the stressed syllable because stressed syllables can either be high-pitched or low-pitched.

Indeed, various production experiments with linguistic diverse materials (e.g., target words in different focal positions) have shown that high f0 cannot be regarded as a direct correlate of stress (even though this is the case in many studies that conflate lexical stress correlates with accentuation). For German, we have shown that lexical stress is reliably marked by the acoustic correlates: duration, spectral balance and vowel quality. Specifically, stressed syllables are longer, produced with more vocal effort and more peripheral vowel quality, compared to their unstressed counterparts. Hence, as far as phonetic correlates are concerned, there are clear indicators that mark the position of the stressed syllable in German.

Like English and Dutch, German is a stress-timed language where lexical stress is free, i.e., not restricted to a certain position in the word. In an abstract sense, lexical stress in these languages is integrated in a hierarchical metrical structure that determines the prominence relations in a word (and eventually the phrase). In this prosodic hierarchy, the syllable is dominated by the foot, which is in turn dominated by the phonological word. The prototypical foot for the West-Germanic languages German, English, and Dutch is a sequence of two syllables, one of which is strong
(stressed) and the other weak (unstressed). Stress is typically assigned by building trochaic feet (SW) from right to left, with the rightmost foot receiving primary stress. Bisyllabic words are thus often stress-initial while longer words tend to receive secondary stress on the initial syllable. Indeed, distributional analyses show that the lexicon in West-Germanic languages is strongly dominated by words with initial stress. This suggests that a stressed syllable is a strong indicator for word onsets.

Taken together, we have introduced pitch accents as a prominence phenomenon on the utterance level and lexical stress as a prominence phenomenon at the word level (Gordon, 2014; Ladd, 2008). Within pitch accents, we have distinguished between nuclear (the last accent) and prenuclear accents (accents preceding the last accent), depending on the position in the phrase. Within lexical stress, we have distinguished between its concrete nature (phonetic correlates that render a syllable prominent) and an abstract level of prominence. Table 2 gives an overview of the different notions of prosodic prominence introduced in Chapter 2.

Table 2: Overview of different notions of prosodic prominence (see Grice & Baumann, 2007, p. 28, for a similar distinction).

<table>
<thead>
<tr>
<th>Type of prominence</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical stress</td>
<td>Lexical stress signals out one of the syllables in a word (abstract notion of prominence, see metrical structure); this syllable is acoustically marked by duration, spectral balance, and vowel quality</td>
</tr>
<tr>
<td>Pitch accent</td>
<td>An accented syllable (i.e., a stressed syllable bearing a pitch accent) exhibits additional tonal movement on or in its vicinity (see different pitch accent types)</td>
</tr>
<tr>
<td>Nuclear pitch accent</td>
<td>The nuclear pitch accent is considered to be the last pitch accent in an intonation phrase, and is usually perceived as the most prominent one</td>
</tr>
</tbody>
</table>

In this thesis, we explicitly address the perception of lexical stress (first row in Table 2) as a function of pitch accent type (second row in Table 2). As has been outlined in detail above, the interplay between different pitch accent types and the perception of lexical stress is interesting as it may involve a shift in the cues that listeners can use to identify the stressed syllable. In this vein, the current thesis sets out to investigate the consequences for stress processing if f0 peak and lexical stress are not aligned. In the next chapter (Chapter 3), we will take a close look at how lexical stress and intonation are processed by infants and adults.
3 The processing of stress and intonation

Chapter outline
Chapter 3 provides the psycholinguistic background to the thesis. From the perspective of the listener, this chapter focuses on the processing of stress and intonation. In 3.1, we will discuss how f0 – despite it being an ambiguous cue to the position of the stressed syllable – functions as a cue to stress in both infants and adults. We will then focus on adult word recognition (3.2) and infant word segmentation (3.3), respectively. Regarding adult word recognition (3.2), we will first discuss how different approaches, i.e., episodic and abstractionist views, explain the mapping between the speech signal onto words (3.2.1), before we will present some current models of spoken word recognition (3.2.2). In 3.2.3, we will outline the most common methods to study spoken word recognition in adults, and summarize experimental studies on the use of stress and intonational cues (and their interplay) in adult word recognition. Regarding infant word segmentation (3.3), we will first outline the challenge and the relevance of finding words in speech for infants, before describing some common methods used in infancy research to study word segmentation (3.3.1). Then, 3.3.2 gives an overview of different cues and strategies that infants exploit to solve the segmentation task, while 3.3.3 focuses on stress and intonation (and their interplay) in infant word segmentation. In 3.4, we summarize the main points of Chapter 3.
3.1 F0 as a stress cue?

3.1.1 F0 and the perception of phrasal prominence

Experimental research on the perception of prosodic prominence is flourishing (e.g., Baumann & Winter, 2018, for a recent summary of the prominence literature). Various studies have investigated the effect of different acoustic cues (and their weighting) on the perception of prosodic phrase-level prominence (e.g., Andreeva, Barry, & Steiner, 2007; Andreeva, Barry, & Wolska, 2012; Barry & Andreeva, 2011; Baumann, 2014; Baumann & Röhr, 2015; Baumann & Winter, 2018; J. Cole, Mo, & Hasegawa-Johnson, 2010; J. Cole & Shattuck-Hufnagel, 2016; Wagner, Cwiek, & Samlowski, 2016, among others). Importantly, all these studies focused on the perception of cues that render a word prominent in the phrase, i.e., make it stand out from the other words in the utterance.

In this regard, F0 has been demonstrated to be a relevant cue to the perception of phrasal prominence: For instance, Wagner et al. (2016) showed that untrained listeners rely predominantly on F0 excursions when asked to annotate prominence in a drumming task. In that study, participants heard spoken utterances from three different speakers and had to “repeat” them by drumming beats. The perceived prominence was reflected by the modulation of the drumming velocity. Using the Rapid Prosody Transcription task (RPT, developed by J. Cole et al., 2010; see also J. Cole & Shattuck-Hufnagel, 2016), Baumann (2014) investigated the role of different pitch accent types, duration, and intensity, along with other non-prosodic factors (e.g., word frequency) in the perception of prosodic prominence by untrained German listeners. In RPT, (prosodically untrained) participants are commonly presented with a speech sample and instructed to underline prominent words and boundaries in a transcript that contains no punctuation. Speech is presented as a whole and participants are not allowed to listen to portions of the signal. Specifically, Baumann (2014) presented listeners with 60 sentences of varied length (between 5 to 18 words) produced by 14 native speakers of German. Participants marked the most prominent word in the utterance on a sheet of paper (the German instructions were “betont/hervorgehoben/wichtig” ‘stressed, highlighted, important’, see Baumann (2014, p. 22)). Results revealed that tonal information, rising pitch in particular, was the most relevant factor for the perception of prominent words in German listeners. In a phonetically more controlled study, Baumann and Röhr (2015) corroborated the important role of rising and high pitch for phrasal prominence perception: In that study, 68 German native speakers were asked to judge the prominence of a proper name (e.g., Lana) on a scale from 1 to 100 in sentences, such as *Sie hat mit Lana telefoniert* ‘She was on the phone with Lana’. The proper name was realized with different pitch accent types. Results showed that listeners judged rising pitch accents (L+H*, L*+H) as most prominent, followed
by high-pitched accent types (H*, !H*), which are in turn followed in perceived prominence by early-falling contours (H+L*, H+!H*) and low-pitched accents (L*). Deaccentuation on a word was perceived as least prominent. Taken together, the above reviewed studies seem to suggest that high f0 (in form of rising or high-pitched accented syllables) renders words prominent on the level of the utterance.

3.1.2 F0 and the perception of lexical stress

In addition to the studies on phrase-level prominence perception, considerable research has also been conducted regarding the perception of word-level prominence, which is the focus of the current thesis. There are a number of methods that have been used to study stress perception in adults, such as meta-linguistic same-difference tasks, identification, or odd-one-out tasks (e.g., Egger, 2015; Fry, 1958; Isačenko & Schädlich, 1966; Kohler, 2008; Niebuhr & Winkler, 2017; Schwab & Dellwo, 2017), as well as linguistic grouping tasks (Bhatara et al., 2013; Bion, Benavides-Varela, & Nespor, 2011). Linguistic grouping has also been studied in infants (Abboub et al., 2016; Bion et al., 2011). As diverse the methodological approaches might be, these studies all have in common that they primarily focused on the cues (and their weighting) that are relevant for the perception of lexical stress. In the following, we will review the studies on the role of f0 as a stress cue in both adults and infants.

F0 in adult stress perception

Listeners of West Germanic languages strongly attend to pitch information for stress perception, regardless of whether it is rising f0 movements, high f0 levels or f0 peaks (cf. Fry, 1958; Isačenko & Schädlich, 1966; Kohler, 2008; Niebuhr & Winkler, 2017). In a seminal study, Fry (1958) showed that English listeners primarily relied on a high f0 level to decide between stress minimal pairs, e.g., the noun object and the verb object; even more so than on duration or intensity, which are inherent correlates of metrical stress. He found that the relation between f0 levels on adjacent syllables function in an all-or-none fashion, with a step-up in f0 leading to an iambic perception and a step-down to the perception of a trochee. Fry (1958) also tested the effect of more complex (but non-linguistic) f0 movements on stress perception and showed that rising movements most distinctly triggered the percept of metrical stress. For German, Isačenko and Schädlich (1966) similarly showed that a manipulation of tonal scaling in discrete monotone pitch levels changed the perception of the stressed syllable in a word (from secondary to primary stress and vice versa): übersetzen [ˈyː.ˈbe.ˈzę.tsən] (‘to translate’) was perceived as übersetzen [ˈyː.ˈbe.ˌzę.tsən] (‘to cross over / ferry over’) when the pitch level on über [ˈyː.ˈbe] was...
increased in tonal height (Iščenko & Schädlich, 1966, p. 22). Research on stress perception has been continued up to date. Only recently, Niebuhr and Winkler (2017) showed that an increase of 0.5 semitones (st) in height of the f0 peak counterbalanced a 30% increase in syllable duration (in relation to the original syllable duration) for the distinction between primary and secondary stress. Kohler (2008) demonstrated that listeners judged the second syllable as more prominent in bisyllabic nonce syllable strings baba as soon as an f0 peak (with a preceding 1-2 st rise to the peak and a subsequent fall of 6-7 st) was realized on the second syllable. Conversely, a falling movement of 2-4 st on the second syllable, resulting in a peak-like contour on the first syllable, led to a trochaic percept.

Linguistic grouping studies corroborate the important role of high f0 as a stress cue (Abboub et al., 2016; Bhatara et al., 2013; Bion et al., 2011). These studies investigated how listeners group syllable sequences alternating in three acoustic dimensions (one at a time), i.e., pitch (high vs. low), duration (long vs. short), and intensity (loud vs. soft). They were inspired by the formulation of the iambic-trochaic law, ITL (Hayes, 1995), according to which sound sequences alternating in duration are perceived as iambic patterns, while sequences alternating in intensity and pitch are perceived as trochaic patterns. The ITL – proposed to be a rhythmic universal – goes back to very early investigations by Woodrow (1909, 1911) who was the first to show that alternation in intensity triggered a “group-beginning effect” and alternation in duration a “group-ending effect”; pitch had no effect in his studies (Woodrow, 1911). Experimental support for the ITL mainly comes from German and Italian adults (French adults did not provide evidence for such a grouping behaviour, see Bhatara et al., 2013; Bion et al., 2011).

Bhatara et al. (2013) presented German and French adult listeners with MBROLA-synthesized syllable sequences (Dutoit et al., 1996) alternating in intensity or duration (Experiment 1), or in pitch or duration (Experiment 2). For the intensity manipulation, one syllable was increased in intensity (between 2 to 8 dB) from a baseline of 70 dB. For the duration manipulation, 50, 100, 150 and 200 ms were added to the baseline duration of 260 ms for each syllable. Finally, for the pitch manipulation, an f0 peak was set in a syllable (at 20-50 Hz above 200 Hz baseline pitch). Noise was added to the beginning of the syllable stream in order to avoid order effects. In addition, the stream beginnings were counterbalanced for the acoustic dimensions (for half of the participants the long syllable was the first, for the other half the short syllable was the first, the same procedure was applied to intensity or pitch alternation). In the test phase, listeners’ task was to indicate whether they perceived the sound sequence as a strong sound followed by a weak one (trochaic response) or a weak sound followed by a strong sound (iambic response). Results revealed that German listeners consistently grouped sequences
varying in pitch (high-low-high-low) or intensity (loud-soft-loud-soft) as trochaic units, with the high-pitched syllable (the one with the f0 peak) or the loud syllable forming the strong (stressed) element, respectively. Meanwhile, syllable strings that alternate in duration (short-long-short-long) were grouped as iambs (WS), with the long syllable forming the strong (stressed) element. The reason for the absence of grouping evidence in French listeners might be due to the lack of marking of lexical stress in French (see discussion in Bhata et al., 2013; and other studies on stress deafness in French listeners, e.g., Dupoux, Pallier, Sebastian, & Mehler, 1997; Dupoux, Peperkamp, & Sebastián-Gallés, 2001). In a similar paradigm to the one administered by Bhata et al. (2013), Bion et al. (2011) tested Italian listeners on their grouping behaviour as a function of acoustic alternation. Italian listeners also grouped syllable sequences alternating in pitch as trochees and syllable sequences alternating in duration as iambs (intensity was not tested in that study).

**F0 in infant stress perception**

Sensitivity to pitch information begins from birth, with the auditory system quickly maturing: The ability to perceive pitch in an adult-like fashion has been demonstrated to develop within the first year of life (Montgomery & Clarkson, 1997; Spetner & Olsho, 1990) with some studies demonstrating adult-like processing of pitch intervals even in new-borns (ERP responses, Stefanics et al., 2008). F0 modulations are furthermore used for linguistic tasks, such as discrimination, from early on: For example, French new-borns have been shown to distinguish between rising and falling contours in Japanese bisyllabic (and bimoraic) words (Nazzi, Floccia, & Bertoncini, 1998). The ability to differentiate between rising and falling pitch remains robust, in particular in languages that mark a linguistic contrast via intonation only, e.g., the distinction between polar questions and declaratives in European Portuguese (see Frota, Butler, & Vigário, 2014, on robust discrimination abilities from 5 months onwards; Sundara, Molnar, & Frota, 2015, on Basque vs. English 4-month-olds). Additionally, new-borns are sensitive to the congruency of pitch movements and visual movements, e.g., rising pitch being accompanied by an upward movement of balls (P. Walker et al., 2018). Another recent study furthermore suggests that infants, at 7 months strongly rely on natural f0 modulations to discriminate between languages that are rhythmically similar, i.e., English vs. German (Chong, Vicenik, & Sundara, 2018).

Infants do not only discriminate pitch movements and use f0 modulations to discriminate languages, f0 modulations have furthermore been shown to be one of the most attractive prosodic features infants attend to: After just a few weeks, infants develop a strong preference for so called infant-directed speech (e.g., R. P. Cooper & Aslin, 1990; Fernald, 1985; Werker, Pegg, & McLeod, 1994, among many others, see MetaLab for meta-studies and current ManyBabies initiative,
Many Babies Consortium, in press, a speech style that is predominantly characterized by higher f0 and greater f0 variability (e.g., Fernald & Simon, 1984, for German, but see many others on a variety of different languages). Fernald and Kuhl (1987) suggest that modulations in f0 play a particularly central role in infants’ preference for infant-directed speech (henceforth IDS) over adult-directed speech (henceforth ADS). In sum, infants are highly sensitive to pitch differences from a very young age. This early sensitivity to pitch information might be caused by its perceptual salience and by its frequent use for linguistic contrasts across languages, as Frolta et al. (2014) suggest. Frolta et al. (2014) refer to Gussenhoven (2004)’s claim that some kind of melodic contour (in the form of tone, lexical pitch or intonation) is present in all types of languages. Beyond the sensitivity to pitch information – which needs to be seen as a prerequisite to use this kind of acoustic information – infants have been shown to attend to high f0 as a signal to word-level prominence (e.g., Bion et al., 2011, see more details below). Note though that, compared to research on adults, only little is known about the role of f0 in stress perception in infants (and children), with developmental studies only recently starting to explore this topic (e.g., Abboub et al., 2016; Bhatara, Boll-Avetisyan, Höhle, & Nazzi, 2018; Bion et al., 2011; Quam & Swingley, 2014).

Regarding the role of f0 for linguistic grouping, developmental studies paint a similar picture to the studies with adults: Bion et al. (2011), for instance, exposed Italian 7.5-month-old infants to syllable streams that alternated in either pitch or duration. In a subsequent test phase, infants’ preference for disyllables that had a final prominence during exposure (iambic response) vs. for disyllables that had an initial prominence during exposure (trochaic response) was assessed. Results showed a preference (longer looking times) for words with an initial prominence (trochees) after infants listened to the stimuli that alternated in pitch, indicating that high pitch was interpreted as the strong (stressed) element. No such preference was found after exposure to durational alternation (intensity was not tested). Contrary to the results in Bion et al. (2011), Abboub et al. (2016) found a grouping effect for syllable sequences alternating in duration for French and German 7.5-month-olds (preference for words that had a final prominence during exposure, iambic response). Importantly, Abboub et al. (2016) also replicated the trochaic grouping behaviour after exposure to syllable sequences alternating in pitch in both French

---

22 The MetaLab (Bergmann et al., 2018) is a platform that provides meta-analyses in the field of language acquisition in order to promote replicability of studies. The platform also provides downloadable datasets along with visual analytic tools. More information can be found here: https://langcog.github.io/metalab2/index.html (last accessed: 28.02.2020). The ManyBabies initiative is a large-scale collaborative approach with focus on replication and best practice in research in the field of developmental psychology. As a first project, the preference for infant-directed speech was replicated, with 67 labs in 16 different countries collecting data (ManyBabies Consortium, in press). The BabySpeech Lab at the University of Konstanz also participated in this project. More information can be found on: https://manybabies.github.io (last accessed: 28.02.2020).
and German 7.5-month-olds (recall the lack of such a behaviour in French adults). In the experiment by Abboub et al. (2016), however, the effect surfaced in a novelty preference (shorter looking times to words with an initial prominence, trochees), which the authors interpreted such that German and French infants found it easy to use f0 as a cue for grouping, easier than duration in the same study, but also easier than Italian infants in a comparable study, cf. familiarity preference in Bion et al. (2011).

While the grouping studies reviewed above suggest that high f0 alone can cue lexical stress, preliminary data by Bhatara et al. (2018) reveal that high f0 alone might not be able to trigger the intuitive preference towards the predominant stress pattern (here the trochaic pattern) that has been observed for German and English infants when stress was realized by a conglomerate of cues, i.e., duration, intensity, and high pitch (Höhle et al., 2009; Jusczyk et al., 1993). Specifically, in their overview article, Bhatara et al. (2018) present preliminary data (sample sizes ranging between N = 11-16 infants) in an attempt to dissociate the individual stress correlates (f0, intensity, and duration). They tested the preference for stress patterns (trochaic vs. iambic) in German and French 6-month-old infants when only one cue to stress was present. The stimuli they used were MBROLA-synthesized (Dutoit et al., 1996). Infants showed no preference for a stress pattern in any of the conditions, even when intensity and the pitch cue co-occurred. Possibly, as the authors explained, these data may suggest that individual cues are not strong enough to trigger the percept of lexical stress that infants need to display a trochaic bias while converging cues are (Höhle et al., 2009). Yet, as null effects are impossible to interpret (e.g., Wagenmakers, 2007), the conclusions can only remain vague. Additionally, from this paradigm it is difficult to conclude that infants did not perceive a syllable as the stressed one when cued by f0. In fact, infants might be able to discriminate words with a different stress pattern (exclusively cued by high f0) but simply show no preference for them. More research is needed to resolve the role of f0 in the perception of stress in infants (also for different linguistic tasks, discrimination, preference, grouping etc.).

For older children, the role of individual cues to stress is equally under-researched. To our knowledge, there is only one study by Quam and Swingley (2014) that tested the role of f0 as a stress cue in word recognition in American English pre-schoolers (as compared to adults). In that study, children aged 2.5-5 years made use of converging stress cues (duration, intensity, vowel quality, and high f0) when identifying two visually presented stress competitor objects on screen (*banana* vs. *bunny*), but did not use high f0 alone (PSOLA-superimposed on a stress-ambivalent production of the target). Meanwhile adults reliably distinguished between the presented objects before they were segmentally disambiguated based on the stressed syllable that was solely cued by high f0.
**Summary**

Summarizing the role of f0 in the perception of word-level prominence, both metalinguistic stress judgements as well as linguistic grouping studies with adults clearly demonstrate that adult listeners strongly rely on high pitch as a cue to lexical stress (be it an f0 peak, rising f0, or high f0 in general). For infants, it appears that an f0 peak is a strong indicator of the stressed syllable in linguistic grouping tasks (Abboub et al., 2016; Bion et al., 2011). On the other hand, studies focusing on different tasks (here trochaic bias or word recognition), seem to converge on the view that f0 is strongest as a stress cue when going together with other acoustic cues, such as duration or intensity (Bhatara et al., 2018; Quam & Swingley, 2014).

In sum, from the perspective of the listener, f0 is an attractive cue in the perception of lexical stress – even though, as we discussed in detail in Chapter 2, due to utterance-level intonation this strategy is misleading in cases in which the high-pitched syllable is not the stressed one, i.e., in early-peak or late-peak accents.

In addition to the studies focusing on the individual acoustic cues in the perception of lexical stress (reviewed above), there is a huge (and growing) field that studies the use of stress (considered a conglomerate of cues) in speech processing to which we will turn now. These studies have investigated how adults and infants use stress for word segmentation (as evident in mishearings or via word spotting, Cutler & Butterfield, 1992; Cutler & Norris, 1988; Newman, Sawusch, & Wunnenberg, 2011); for experimental evidence with infants, behavioural looking paradigms as well as neurophysiological paradigms have been employed (e.g., S. Bartels et al., 2009; Junge, Kooijman, Hagoort, & Cutler, 2012; Jusczyk, Houston, et al., 1999; Kuipers, Coolen, Houston, & Cutler, 1998; Männel & Friederici, 2013). Additionally, lexical stress has been demonstrated to play an important role in adult spoken word recognition (as shown in cross-modal priming, fragment completion, or eye-tracking studies, e.g., N. Cooper, Cutler, & Wales, 2002; Donselaar, Koster, & Cutler, 2005; Reinisch et al., 2010), but comparatively little is known on the role of stress in word recognition in toddlers and children. In the next two subchapters, we look at speech processing, focussing on lexical activation in adults and word segmentation in infants. We first discuss the main processes underlying spoken word recognition as well as models thereof and review studies on the use of stress and intonational cues in adult word recognition (3.2). We will then move on to the use of stress in infant word segmentation and review what is known on the role of intonation in this process (3.3).
3.2 Adult spoken word recognition

3.2.1 Mapping the speech signal onto words

The challenge

Understanding how listeners map the acoustic signal onto linguistic units, such as words, is one of the central questions in speech perception research (e.g., Cutler, 2012; Ernestus, 2014; Frauenfelder & Floccia, 1999; Hanulíková, 2009; McQueen, 2005; Pierrehumbert, 2016; Weber & Scharenborg, 2012). In other – very simple – words, psycholinguists have been trying to answer the question of how listeners, for instance, arrive at the notion of cat (as being a carnivorous mammal with fur, four legs, a tail, whiskers and claws, usually kept as a pet) upon hearing an open vowel [æ] preceded by an aspirated voiceless velar plosive [kʰ] and followed by an alveolar plosive [t]. There is consensus that this process – as simple it might appear in the above example – is far from trivial. In fact, the biggest challenge listeners face when recognizing words in speech is posed by the speech signal itself, with the input being highly variable: Words do not only vary as a function of the linguistic context in which they appear, but also differ across speakers, age, and nativeness, among others – a phenomenon of spoken language frequently referred to as the lack of invariance problem (e.g., Cutler, 2012; see Pierrehumbert, 2016, for an overview of sources of input variability). To illustrate the point, words tend to be more reduced if they occur in a linguistic context in which they are highly predictable compared to contexts in which they are less expected (context refers here to the preceding or following words or sounds, see Ernestus, 2014, for an overview). Speakers also differ in the temporal patterns with which they realize individual sounds in words (Remez, 2010). Naturally, words are produced with a different fundamental frequency if uttered by a female or a male speaker; even sexual orientation is encoded in the quality of vowels (Pierrehumbert et al., 2004).

Episodic and abstractionist views

In accounting for the mechanism of the mapping between the acoustic signal and linguistic units, two basic approaches have been proposed: episodic accounts (e.g., Goldinger, 1998; Hintzman, 1986, 1988; K. Johnson, 1997; Klatt, 1979) and abstractionist accounts (e.g., Lahiri & Marslen-Wilson, 1991; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008). In their basic assumptions, these two approaches need to be considered as opposites – even though there have been attempts to reconcile the two theories (see below). Nevertheless, we will take a moment to outline the main ideas in both episodic and abstractionist views (e.g.,

\footnote{Note that beyond these two basic approaches other theories of speech perception have been proposed, e.g., gestural accounts such as the Motor Theory (e.g., A. M. Liberman et al., 1967) or the Direct Realist Theory (Fowler, 1986).}
Ernestus, 2014, for a comparison of basic ideas). Broadly speaking, both views assume a storage system, the mental lexicon, which eventually allows listeners to comprehend the uttered words. Yet, the specific mapping process between signal and linguistic unit tremendously differs in the two views.

Episodic (or exemplar) theory (Goldinger, 1998; K. Johnson, 1997) basically assumes that the mental lexicon contains many (very detailed) episodes or exemplars of words, which together constitute a cloud. These exemplars represent the tokens that a speaker has encountered, either by producing the word herself or listening to words being produced by the interlocutor. As such, exemplars are highly detailed – rich with phonetic and indexical information – as they contain all types of information on the context in which the word was uttered. That way, they perfectly capture the variability present in the input. All newly encountered words are compared to stored exemplars and categorized accordingly, such that the exemplar memory is continuously updated. As K. Schweitzer et al. (2015), for instance, highlight, the memory is very sensitive to frequency effects, as frequently appearing words have many exemplars while infrequent words have only few. In the episodic view, word recognition is thought to be a direct mapping of the acoustic characteristics to the exemplars in the mental storage system, the mental lexicon. An exemplar receives activation, with the degree of the activation depending on how well the signal properties match the exemplar properties. The activated exemplar passes activation to its word node (lexical level), which disposes of all types of information of the word (e.g., Ernestus, 2014, for a summary). Figure 6 (left panel) visualizes the levels of the mapping process between the acoustic signal and the stored representation in the episodic framework.

In sharp contrast to the episodic approaches, a key feature of abstractionist theory (e.g., Lahiri & Marslen-Wilson, 1991; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008) is that the mental lexicon contains only one single representation for every word (or morpheme). Essentially, this representation is composed of abstract units (that could be phonemes or linguistic features, see more details below). Similar to episodic theories, word recognition is understood as the mapping of the acoustic signal onto the lexical representation, but in this view the lexical representation is thought to be abstract. Importantly, abstractionist views assume an intermediate level between signal and (abstract) lexical entry, which is called the pre-lexical level. At the stage of the pre-lexical level, all speaker- or context-specific characteristics of the signal are assumed to be abstracted away from, resulting in an (equally and decidedly) abstract representation at the pre-lexical level. Unlike in the episodic models, in an abstractionist view, the pre-lexical level is

24 We will focus on perception here, but there are also episodic theories of speech production (Pierrehumbert, 2001; Walsh, Schütze, Möbius, & Schweitzer, 2007). The interested reader is referred to Pierrehumbert (2016) or Ernestus (2014) for overviews.
The processing of stress and intonation compared to the stored (abstract) lexical representation, with the goodness of the match determining the degree of activation. As soon as a threshold is reached, the word is recognized (see e.g., Ernestus, 2014, for a summary). Figure 6 (right panel) visualizes the levels of the mapping process between the acoustic signal and the stored representation in the abstractionist framework. In the following, we will outline some effects observed in language processing which can be explained in one approach (but are difficult to account for in the other), calling for hybrid models of speech comprehension.

Figure 6: Simplified overview of the mapping process between the acoustic signal and the lexical level in the episodic framework (left panel) and the abstractionist framework (right panel).

Why episodes? Why abstractions?

There is converging evidence that listeners store word-specific information based on the various contexts a word appeared in, which cannot be explained by abstractionist but by episodic views (e.g., Mendoza-Denton, Hay, & Jannedy, 2003, for discussion; Pierrehumbert, 2016, for a recent review of such effects of indexical information on memory). To illustrate the point, Bradlow, Nygaard, and Pisoni (1999) showed that words are memorized better if they are spoken by the same speaker and at the same speaking rate, suggesting that these characteristics are stored together with the word itself. More recently, studies have provided evidence for stored characteristics that might be less intuitive: A. Walker and Hay (2011), for instance, tested whether words uttered predominately by younger speakers (e.g., *lifestyle*, henceforth young “word age”) or words that are uttered predominantly by older speakers (e.g., *typist*, henceforth old “word age”) are more easily processed when these words are spoken by a person whose age matches the “word age”, i.e., *lifestyle* spoken by a young person (here 22 years), than when “word age” and “speaker age” mismatch, i.e., *lifestyle* spoken by an older person (here 50 years).
Results of a lexical decision task, in which listeners indicated whether a given word is a word of English or not, revealed that the accuracy of the decision increased and the reaction time decreased when age characteristics matched. This finding speaks in favour of age of the voice being stored together with the word and hence in favour of episodic representations. Such context effects even go beyond a mere acoustic similarity between signal and stored representation and extend to non-linguistic factors: In this regard, Hay and Drager (2010) showed that listeners (from New Zealand) shift their vowel categories depending on the association they created with the task (induced by a toy present in the room, either a kangaroo/koala generating associations with Australia or a kiwi generating associations with New Zealand). Specifically, Hay and Drager (2010) asked listeners to match the pronunciation of a short vowel [ɪ], e.g., in *fit* embedded in a sentence, to a 6-step-continuum of [ɪ]-vowels. In this continuum, the first and second formants in [ɪ] were manipulated to resemble Australian-like [ɪ]-vowels (lower F2, higher F1) and New Zealand-like [ɪ]-vowels (more central, higher F1 and lower F2). Listeners judged the pronunciation more Australian-like when the kangaroo or koala was present and more New Zealand-like when the kiwi was present. Hence, this study also calls for detailed representations and hence for an episodic modelling of speech processing.

On the other hand, it has been argued (cf. Cutler, 2012; Hanulíková, 2009) that episodic theories cannot explain how listeners are able to quickly accommodate or adapt to speech characteristics never heard before and, importantly, generalize them to new sounds, as has been shown in various perceptual learning studies or accent adaptation studies (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Hanulíková & Ekström, 2017; McQueen, Cutler, & Norris, 2006; Norris, McQueen, & Cutler, 2003). In a perceptual learning study, McQueen et al. (2006), for instance, trained listeners on words that ended in an ambiguous fricative (spectrally between [f]/[s]). In one training group, the ambiguous fricative replaced a fricative in a Dutch word ending in [f]; in the other training group, the ambiguous fricative replaced a fricative in a Dutch word ending in [s]. During test, Dutch listeners had to decide whether a word was an existing Dutch word or not, with the decision being preceded by the auditory presentation of a priming word with an ambiguous fricative. Importantly, the word for the lexical decision task was not part of the training session. Results indicated that listeners trained with [f]-words were faster in their lexical decision for ambiguous fricatives in [f]-ending words and slower for [s]-ending words, and vice versa. This shows that listeners learned to interpret the fricatives in accordance to the test phase and generalized this knowledge to a new set of words – by shifting their category of a sound. McQueen et al. (2006) interpret this finding as speaking in favour of an abstraction on the pre-lexical level. Regarding accent adaptation, Hanulíková and Ekström (2017), for instance, familiarized German listeners (and
two other non-native groups) with an unfamiliar German dialect (in which [u] was shifted to [o], and [i] to [e]), using a 15 min exposure story. Here again, German listeners adapted to the shift in vowels (particularly for [u]/[o]) and used this knowledge for dialect words not presented during training when they performed lexical decisions in a subsequent test phase. Moreover, in a recent study, E. K. Johnson, Bruggeman, and Cutler (2018) argue that the so called “language familiarity effect” according to which listeners recognize a speaker more accurately when she is speaking the listener’s native language is very likely based on abstract phonological processing. Their claim rests on their finding showing that talker recognition was not modulated by the familiarity to a variety: Canadian and Australian English listeners behaved in the same way, recognizing a talker equally well in one variety of English or the other, but recognition was worse in non-native speech (Dutch). Likewise, Dutch listeners, recognized talkers more accurately when they were speaking Dutch, compared to when they were speaking English, but this disadvantage for English was independent of the variety and the listeners’ familiarity to one or the other variety. Hence, identification seems to depend on an abstract phonological level of processing.

Taken together, on the one hand, abstractionists seem to have good reasons for assuming abstract representations. On the other hand, however, it seems to be the case that listeners also store word-specific information, which facilitates processing. As has been foreshadowed above, there have been attempts to reconcile the two (opposing) theories and the clear-cut distinction has begun to be dissolved, particularly by proposing hybrid models (see discussion in Cutler & Weber, 2007; Ernestus, 2014; Pierrehumbert, 2016; Weber & Scharenborg, 2012). The fact that “such [hybrid] approach[es] may prove necessary to accommodate many linguistic processes” has already been suggested by Goldinger (1998, p. 254), who posits himself in the field of episodic views. In reviewing different models in the different approaches, Weber and Scharenborg (2012, p. 396), for instance, draw a similar conclusion: “[I]t has become obvious that both purely abstract models and purely episodic models are incomplete, and the challenge for the future is to develop a hybrid approach that combines both abstract and episodic representations.” Indeed, such hybrid models have been proposed in the literature (e.g., Connine, Ranbom, & Patterson, 2008; Goldinger, 2007; McLennan, Luce, & Charles-Luce, 2003; Norman & O’Reilly, 2003). They combine both abstract and episodic representations. In the next subsection, we will briefly discuss some of the most influential models of spoken word recognition.
3.2.2 Models of spoken word recognition

MINERVA 2 (Hintzman, 1986, 1988) is an episodic model of speech comprehension that has been successfully employed in simulations (Goldinger, 1998), see Weber and Scharenborg (2012) or Goldinger (1998) for overviews. The memory system modelled in MINERVA 2 consists of various traces of a word, which are highly detailed (and may be hugely redundant). When the model encounters a word, an analog probe (Goldinger, 1998, p. 254) – operationalized as a numeric vector – is sent to all traces. This probe is then compared to the stored traces, whereby activation of the traces depends on how well the traces matched the probe. At this stage, an echo is created based on a weighted aggregation of all activated traces. Essentially, this echo, which may contain information that was not present in the probe, is subsequently sent to the memory. Hence, at this stage, the model includes a type of abstraction process which generalizes over different traces; a fact, often overlooked in the abstractionist literature (see Pierrehumbert, 2016, for discussion).

The most influential abstractionist models are the Cohort Model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994), and Shortlist B (Norris & McQueen, 2008). In these models, the signal is thought to be converted into an abstract representation, which generalizes over the variability in the input and serves to access the mental lexicon. They also share the main assumption that the basic principles accounting for the recognition process are activation, competition, and finally selection: Lexical candidates are assumed to be activated, with the amount of activation depending on the degree of the match between the stored (abstract) representations and the signal. Subsequently, the activated candidates engage in competition and the candidate that receives the highest activation is eventually selected among lexical entries (e.g., Weber & Scharenborg, 2012). There is converging behavioural evidence supporting both parallel activation and competition (e.g., Cutler, 2012; Weber & Scharenborg, 2012): Regarding activation, Zwitserlood (1989), for instance, showed that a word fragment which is compatible with the word beginning of two words (e.g., [kapit] for the Dutch word pair kapitein ‘captain’ and kapitaal ‘capital’) facilitates the recognition of words that are semantically associated with both members of the pair (e.g., schip ‘ship’ associated with kapitein and geld ‘money’ associated with kapitaal). This finding can be interpreted as a parallel activation of both kapitein and kapitaal upon hearing an ambiguous word beginning [kapit]. Furthermore, it has been shown that lexical embedding (e.g., bone in trombone) leads to spurious activation of lexical candidates that are not intended by the speaker (McClelland & Elman,

---

Note that Shortlist B (Norris & McQueen, 2008), as a Bayesian account, does no longer suggest activation of word candidates but probabilities thereof. The probabilities of lexical candidates are independent of one another, amounting in a sum of 1. Hence, a higher probability for one candidate means a lower probability for the other.
The processing of stress and intonation

1986), also being indicative of a parallel activation of different lexical candidates. More recently, activation has been studied using eye-tracking techniques. We will discuss this method in more detail when we outline how suprasegmental information influences the activation of lexical candidates (see 3.2.3). Various other experiments have also provided evidence for competition in word recognition. In particular, McQueen, Norris, and Cutler (1994) showed in a word spotting paradigm that listeners found it harder to spot embedded words such as sack in sacrif (the beginning of sacrifice) than in sacrik, a nonce word onset in English. This asymmetry, as the authors argue, can be taken as a competition effect between sack and sacrifice. Moreover, the number and frequency of similar sounding words in the lexicon have repeatedly been shown to increase competition and consequently slow down word recognition (e.g., Luce, Goldinger, Auer, & Vitevitch, 2000; Luce & Large, 2001; Vitevitch, 2002; Vitevitch & Luce, 1999), see McQueen (2005) for an overview.

However, beyond this common assumption lies some fundamental differences between models: one regards the nature of representations, another the flow of information between different processing levels, and yet another the presence of inter-word competition (see Weber & Scharenborg, 2012, for details). First, even if abstractionist approaches all propose an abstract pre-lexical and lexical level, they highly disagree on the “unit of perception” (Hanulíková, 2009, p. 15), i.e., the composition of the abstract unit. The suggestions range from acoustic phonetic features to syllables or totally underspecified units (for summaries, see Cutler & Clifton, 1999; Frauenfelder & Flocia, 1999; Hanulíková, 2009; Weber & Scharenborg, 2012). While the Cohort Model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978) assumes features at the pre-lexical level and an underspecified phonological structure at the lexical level, Shortlist (Norris, 1994) assumes phonemes (or strings thereof) at the two levels. TRACE (McClelland & Elman, 1986), in turn, models abstract representations as features and phonemes, and Shortlist B (Norris & McQueen, 2008) assumes probabilities of phonemes.

Second, there is no consensus on the direction of information flow between the individual processing levels, i.e., whether there is feedback from the lexical level to the pre-lexical level or not (see Weber & Scharenborg, 2012, for discussion). While all models obviously agree on the flow of information in a bottom-up manner (from the input via the pre-lexical level to the lexical level), TRACE, for instance, additionally allows for top-down spread of activation from the word to the phoneme layer. All other models outlined above (the Cohort Model, Shortlist and Shortlist B) do not. Finally, models differ on whether they allow for lateral competition or not. Serial models of spoken word recognition, e.g., the Cohort Model, assume lexical selection to be a process determined by serial bottom-up activation and bottom-up inhibition. As soon as the signal mismatches a lexical candidate in
the cohort, its activation decreases. By contrast, connectionist models (TRACE or Shortlist) assume lateral inhibitory connections between lexical candidates, such that activated candidates inhibit neighbouring candidates and hence reduce the competitor set more efficiently.

Note that in abstractionist models introduced above, suprasegmental cues are currently not accounted for as the input for these models is most often a string of phonemes which abstracts away from acoustic details (see Connell et al., 2018; Jesse et al., 2017, for discussion) – even though suprasegmental cues have been shown to be used during word recognition (as will be outlined in detail in the next subsection). Shortlist disposes of a (non-acoustic) mechanism of metrical segmentation that aids the model in finding word boundaries in continuous speech (McQueen et al., 1994; Norris, McQueen, & Cutler, 1995). To our knowledge, there is only one model that accounts for tonal information. For Mandarin Chinese, Shuai and Malins (2017) suggested an extension of TRACE, which in addition to phonemes also decodes lexical tones.

Taken together, the mapping from the acoustic signal onto words is far from trivial. One solution to account for this process is provided by episodic theories. They suggest that listeners store all encountered words as detailed exemplars. Word recognition is achieved by a direct mapping between the signal and the stored traces. Abstractionists propose a mediated process of word recognition that proceeds via a pre-lexical level. Both pre-lexical and lexical level are highly abstract and do not contain fine phonetic details. Both views have their strengths and weaknesses. The clear-cut distinction has started to be shifted, with the field already experiencing the advent of hybrid models. Activation and competition have been acknowledged as common principles in word recognition. In the following, we will review the use of suprasegmental stress and intonational cues in adult word recognition and show how activation and competition are affected by these cues.

### 3.2.3 Stress and intonation in spoken word recognition

*Methods to study spoken word recognition in adults*

Psycholinguists have primarily used cross-modal fragment priming and eye-tracking to examine lexical activation and competition in adults. In cross-modal fragment priming (CMFP) experiments, listeners’ reaction times (RTs) to a target are recorded as a function of a preceding prime (see cf. Cutler, 2012, p. 15; Zwitserlood, 1996, on form priming). Usually, the listener’s task is to indicate whether or not a given target is a real word of their language (lexical decision). The cross-modal version of the method presents the prime in one mode (usually auditorily) and the target in another (usually visually); primes are fragments of the target or the competitor (e.g., octo- for octopus). According to the rationale of the method, compared
to the reaction times after a control prime, listeners are faster in their lexical decisions if the prime and the target match in stress pattern (e.g., ['ɔk.to] for the Dutch word ['ɔk.to.pus] 'octopus'). This pattern of results would be interpreted as facilitation. Crucially, if prime and target mismatch (e.g., ['ɔk.to] for the Dutch word ['ɔk.to.bar] 'october'), listeners are expected to be slowed down in their reaction (i.e., showing longer RTs to a given target compared to an unrelated control prime). This pattern of results is in turn indicative of inhibition.

Extending CMFP, the observation of eye-movements provides time-course information, i.e., when activation of word candidates starts and how activation patterns change over time. Eye-tracking technology makes use of the human oculomotor system: fixations and saccades (fast moves between two fixations). Initial attempts go back to R. M. Cooper (1974) who, to our knowledge, was the first to monitor listeners’ eye-movements to scenes when they listened to short narratives (see Huettig, Rommers, & Meyer, 2011, for a historical sketch on the method).

Around 20 years later, Tanenhaus et al. (1995) brought the method back into use by recording listeners’ eye-movements to objects while listeners followed instructions (“Pick up the candy. Now put it above the fork”). The paradigm rests on the assumption that listeners look at the objects they attend to in the visual world (Huettig et al., 2011). Tanenhaus et al. (1995) showed that when listeners hear the fragment cand, they direct the same number of fixations to the object candy and the object candle – hence both are activated; yet, as soon as segmental information disambiguates the two candidates, candy is preferred over candle (if candy is the target) and hereafter recognized. This paradigm became known as the Visual-World (VW) Eye-Tracking Paradigm (Allopenna et al., 1998), which initially used pictures as objects to be presented on screen, but later also printed words were employed (McQueen & Viebahn, 2007).

The usual setup in a VW Eye-Tracking study is the following: Listeners are presented with four objects on screen (visual or written representation of the words). After a certain preview time for the words or objects on screen, participants are then instructed to click on one of the four words or objects (e.g., “Click on the beaker”). The intended referent is called the target (i.e., the object/word that is eventually recognized). Depending on the research question of the study, there are also one or two competitors. In the paradigms that will be reviewed below, competitor and target typically overlap segmentally, i.e., share the onset or rhyme/coda (e.g., beetle for an onset overlap; speaker for a rhyme overlap). The remaining referent(s) are typically unrelated to both the target or competitor(s) and are thus called distractors (example taken from Allopenna et al., 1998; see also Huettig et al., 2011, for a review on the use of the Visual-World paradigm; Tanenhaus & Spivey-Knowlton, 1996, for methodological considerations and fields of application). In the
current dissertation, the VW paradigm is used to study the effect of pitch accent type on the activation of lexical candidates in German adults. We will now review the literature on the use of suprasegmental stress and intonational cues for word recognition. We concentrate on listeners from intonation languages, in which pitch is not lexically contrastive.

Suprasegmental stress cues and word recognition

Processing suprasegmental stress cues correctly supports efficient processing, as these cues can help distinguish between segmentally overlapping words which differ in the position of stress, e.g., museum – musical. That way, the number of competitors is reduced, with one of the cohort competitors mismatching the incoming signal. Fragments that mismatch in stress result in inhibition relative to an unrelated control prime (N. Cooper et al., 2002; Donselaar et al., 2005; Soto-Faraco et al., 2001). Using the CMFP, Donselaar et al. (2005), for instance, presented Dutch listeners with semantically non-constraining sentences ending in a disyllabic word fragment (“Everybody had heard about the [fragment]”, e.g., [ɔ́k.to]). At the end of the fragment, listeners were shown a string of letters on screen (e.g., octopus) and they had to decide whether the given string of letters was an existing word in Dutch. The fragment primes either matched the stress pattern of the visually presented target (e.g., [ɔ́k.to] for [ɔ́k.to.pus] ‘octopus’) or mismatched the stress pattern of the visually presented target (e.g., [ɔ́k.to] for [ɔ́k.’to.bər] ‘October’). Segmentally unrelated fragments served as control primes (e.g., [œy.fo] from [œy.fo.ˈri] ‘euphoria’). As outlined above, reaction times for lexical decisions on the target ([ɔ́k.to.pus] ‘octopus’) are compared for stress-matching primes and control primes, and stress-mismatching primes and control primes. Results showed that Dutch listeners reacted faster to targets when the preceding auditory prime had the same stress pattern as the target, as compared to the reaction time after the control prime (facilitating effect); however, listeners were slower compared to a control condition (inhibitory effect), when there was a mismatch between the stress pattern of the prime and the target. The same finding was found for Castilian Spanish listeners in a very similar study (Soto-Faraco et al., 2001): Identical to the results reported for Dutch listeners (Donselaar et al., 2005), Spanish listeners also reacted faster to targets after stress-matching primes, compared to after control primes; listeners in turn were slower in their reaction to a target, as compared to after a control prime, when there was a mismatch between the stress pattern of the prime and the target, suggesting lexical inhibition.26

26 Note that Soto-Faraco et al. (2001) also tested segmental effects on spoken word recognition using primes that matched or mismatched in vowels or consonants. Overall, segmental mismatch also constrained lexical access. For the purpose of this dissertation, we only report on the use of suprasegmental cues.
More recently, researchers have started to address the time course of the use of lexical stress for spoken word-recognition using eye-tracking to track listeners’ fixations to spoken words (Connell et al., 2018; Jesse et al., 2017; Reinisch et al., 2010; Sulpizio & McQueen, 2012). Using the VW Paradigm, Reinisch et al. (2010), for example, showed that suprasegmental information affects lexical activation immediately. Dutch listeners were presented with four written representations on screen, a critical cohort pair, and an unrelated distractor pair. Members of the cohort pair segmentally overlapped in their first two syllables but differed in the location of primary stress e.g., in the words ([\text{sk.to.pus}] vs. [\text{sk.to.bər}]). On every trial, participants saw a quadruplet of words and followed instructions over headphones (“Klik nog een keer op het woord \text{<Target>}” ‘Click once more on the word \text{<Target>}’) while their eye-movements were monitored. Dutch listeners fixated a target, e.g., [\text{sk.to.pus}], more than its stress competitor, e.g., [\text{sk.to.bər}], before segmental information disambiguated the pair (here the onset phoneme of the third syllable [b]/[p]). Thus, listeners made use of stress information as soon as it became available in the signal. In a subsequent study, it has been shown that seeing a person say fragments of the pairs is sufficient information for Dutch listeners to distinguish between stress minimal pairs (Jesse & McQueen, 2014). Hence, suprasegmental stress cues in visual speech alone can effectively guide word recognition. Witnessing an accompanying random pointing gesture alone, however, did not affect the perceived location of suprasegmental stress cues (Jesse & Mitterer, 2011).

From a cross-linguistic perspective, it has been shown that suprasegmental stress cues are only used for word recognition if this reduces competition between lexical candidates, i.e., only if it is profitable for the listener (Cutler & Pasveer, 2006). In Spanish, the same set of vowels is used for stressed and unstressed syllables and stress is encoded by suprasegmental cues only (Navarro-Tomás, 1968; as cited in Soto-Faraco et al., 2001). German and Dutch tend to reduce vowels in unstressed syllables (Delattre, 1969), but the languages allow for non-schwa vowels in unstressed syllables. In English, more than in Spanish, German and Dutch, many vowels in unstressed syllables are centralized to schwa (Delattre, 1969), making vowel reduction a strong cue to lexical stress on the segmental level. Analysing the vocabulary in German, English and Dutch, Cutler and Pasveer (2006) additionally found that there is more reduction of embedded words (words that are contained in another word; for example bone in trombone) due to stress in Dutch and German than in English (Cutler & Pasveer, 2006). Hence, for Dutch and German listeners, it is more beneficial to rely on suprasegmental stress cues to reduce lexical competition than in English (Cutler & Pasveer, 2006).

In that sense, various early studies on stress perception in English have emphasized that listeners more strongly attend to spectral information than to
suprasegmental information (Bond, 1981; Bond & Small, 1983; Cutler, 1986; Mattys & Samuel, 1997; Small, Simon, & Goldberg, 1988): To illustrate the point, in a phoneme-monitoring task, Small et al. (1988) presented listeners with sentences that contained homographs or non-homographs which were either correctly stressed or mis-stressed, e.g., Mary was a recent convert from Catholicism (convert being an example for a correctly stressed homograph). Mis-stressing in a homograph resulted in an existing word in English (here the iambic verb (to) convert), while mis-stressing in a non-homograph resulted in a non-existing word in English (iambic complain vs. non-existing trochaic complain). Listeners’ task was to monitor the appearance of a given phoneme (e.g., /f/ in from in the current example), which always appeared after the minimal pair. Results showed that mis-stressing did not affect word recognition when the mis-stressing occurred in a homographic word (reaction times to the phoneme were not slower than after correctly stressed words), thus producing the other member of the pair. Yet, mis-stressing affected word recognition when a non-word was created (slower reaction times to target phoneme after mis-stressed words).

The results in Cutler (1986) point in a similar direction: In that study, listeners were presented with sentences that were semantically non-constraining, e.g., Gritting her teeth, she reminded herself / that her forbears had been … / to forebear to mention …, and asked participants to make lexical decisions on visually presented targets that occurred at the end of the auditory prime (e.g., forbear). Crucially, the targets on screen were either semantically related to the primes or not. The results showed that stress minimal pairs, forbear (depending on the stress meaning ancestor or tolerate), primed each other’s associates (ancestor, tolerate). Hence, listeners treated these pairs as homophones, despite their suprasegmental stress difference (Cutler, 1986).

More recently though, it has been shown that English listeners can use suprasegmentals if necessary (Braun, Lemhöfer, & Mani, 2011; Connell et al., 2018; N. Cooper et al., 2002; Jesse et al., 2017; Zahner, Kember, & Braun, 2017), albeit they might do so less efficiently than other listener groups: For instance, N. Cooper et al. (2002) demonstrated that – like in Dutch (and Spanish) – words were recognized faster after a stress-matching prime than after a control prime (e.g., [ædmi] for admiral, facilitating effect). But in contrast to Dutch (and Spanish), there was no inhibitory effect in English listeners, i.e., slower reaction times to words after an inappropriately stressed prime (e.g., [ædmi] for admiration), but rather partial facilitation instead for monosyllabic stress-mis-matching primes (e.g., [mju:] for museum produced even faster reaction times than controls). Furthermore, in a forced-choice identification tasks, English native listeners were outperformed by Dutch non-native listeners in assigning a fragment (e.g., mus-) either stressed or unstressed to its source, e.g., music or museum (N. Cooper et al., 2002). Corroborating these findings, two recent studies investigated the time course of the use of lexical stress for word
recognition in English listeners (Connell et al., 2018; Jesse et al., 2017). Closely following Reinisch et al. (2010), Jesse et al. (2017) investigated when in time English listeners use suprasegmental stress information to distinguish two visually presented stress competitors (e.g. admiral vs. admiration). Comparable to what was observed for Dutch listeners (Reinisch et al., 2010), English listeners showed a target preference in the second syllable (defined as the difference of log-transformed target and competitor fixations) if targets were stressed on the initial syllable (admiral). For targets with primary stress on the third syllable and secondary stress on the initial syllable (admiration) no such preference was found. Furthermore, in contrast to the findings in Dutch, suprasegmental stress cues did not modulate the degree of competition: While, prior to segmental disambiguation, words with stress on the initial syllable were stronger competitors in Dutch than words that had secondary stress on the initial syllable (Reinisch et al., 2010), no such difference was observed for English in the study by Jesse et al. (2017). Given this cross-linguistic difference, Jesse et al. (2017) concluded that English listeners can and do use suprasegmental stress cues, but they are less efficient than Dutch listeners. Jesse et al. (2017) argue that segmental cues to stress might be more relevant. In fact, Connell et al. (2018)’s findings support Jesse et al. (2017)’s conclusion: Connell et al. (2018) presented listeners with four words on screen, a cohort pair consisting of target and competitor and an unrelated distractor pair. The target was stressed on the first syllable (SW) and the competitor on the second (WS). In one condition, the first syllable in cohort pairs segmentally overlapped (no vowel reduction condition, carpet vs. cartoon); in the vowel reduction condition, by contrast, in competitor words the first syllable was reduced while it contained a full vowel in targets (parrot vs. parade). Fixation results revealed an interaction between stress and vowel reduction, indicating that the effect of stress on target fixations was stronger when vowel reduction additionally cued stress on the segmental level. Crucially though, lexical access was also constrained when stress was only cued by suprasegmental cues (in the no vowel reduction condition).

All these psycholinguistic processing studies used naturally produced stress competitors and tested the use of a conglomerate of suprasegmental stress cues – unlike the more psycho-acoustic studies (reviewed in 3.1) that manipulated different correlates of stress individually and in a step-wise fashion (cf. Fry, 1958; Isačenko & Schädlich, 1966; Kohler, 2008; Niebuhr & Winkler, 2017). The descriptions and acoustic analyses of stimuli in these processing studies reveal that the stressed syllable was mostly longer, louder, and higher than their unstressed counterpart (e.g., Cutler, Wales, Cooper, & Janssen, 2007, who analysed the stimuli by Cooper et al., 2002; Donselaar et al., 2005, p. 257; Reinisch et al., 2010, p. 780). Additionally, the embedding of targets in instructions, such as ‘Click on <Target>’ increased the
likelihood of a medial-peak accent \((L+H^*)\) on the target (see Jesse et al., 2017, for discussion). Post-hoc correlation analyses indicated that behavioural responses and fixations were correlated with f0, intensity, and duration, depending on the language and the task (Cutler et al., 2007; Reinisch et al., 2010). English listeners’ behaviour in N. Cooper et al. (2002) correlated with f0 only (Cutler et al., 2007, p. 1914, Table 1), while Dutch listeners’ responses were correlated with duration and intensity in Reinisch et al. (2010).

Taken together, there is substantial evidence from different paradigms and different languages demonstrating that suprasegmental stress cues – as they occur in natural productions – affect spoken word recognition in adults, i.e., lexical activation and lexical inhibition. Unfortunately, analogous experiments to the ones reported in this subsection have not been conducted in German, which is the target language in the current study.\(^{27}\) However, one might expect comparable results to Dutch, given that stress reduction is also less likely in German (Cutler & Pasveer, 2006).\(^{28,29}\) In this thesis, we will build on the finding that listeners use suprasegmental stress cues for efficient word recognition and test whether different pitch accent types affect lexical activation of stress competitors. We will next outline how intonational cues affect word recognition.

The processing of pitch accents for pragmatic purposes

Regarding intonational cues, listeners in intonation languages have been shown to use information on different pitch accent types for pragmatic inferences, such as reference resolution and the prediction of upcoming referents (Braun, Asano, & Dehé, 2018; Chen, Os, & Ruijer, 2007; Dahan et al., 2002; Watson et al., 2008, among others; Weber, Braun, & Crocker, 2006). Dahan et al. (2002), for instance, examined the role of accentenedness on lexical competition using eye-tracking. They monitored the eye-movements to cohort competitors (e.g., candy vs. candle). Listeners received instructions via headphones to move objects around on a screen and place them at a certain position in relation to fixed shapes, e.g., *Put the candle below the triangle*. This first instruction was followed by a second instruction in which accentuation (accented vs. deaccented) and introduction of a new referent (yes/no)

\(^{27}\) A replication of N. Cooper et al. (2002) with German and Australian English listeners is currently on-going (Jenny Yu, Anne Culor, p.c).

\(^{28}\) “If exact analogues of the Dutch experiments to be conducted in German, the results would be as in Dutch. In German, too, exploitation of stress information pays off in competition reduction.” (Cutler & Pasveer, 2006, p. 239).

\(^{29}\) A great deal of what is known on German listeners’ use of suprasegmentals in word recognition comes from work by Claudia Friedrich (Friedrich, 2002; Friedrich, Kotz, Friederici, & Alter, 2004; Friedrich, Kotz, & Gunter, 2001, using electrophysiological methods and behavioural data (cross-modal fragment priming). In her dissertation, Friedrich (2002) investigated the role of prosody in spoken word recognition, explicitly focusing on the f0 contour as a stress cue. We will address these studies in the next subsection in which we discuss the use of f0 contours in word recognition.
was manipulated. Thus, the noun in the second instruction was either accented or deaccented (Now put the candle above the square / Now put the candle above the square), underlining indicates accentuation), it could also introduce a new object or refer to the same object (candy). Listeners looked more to the competitor (candy) when the noun in the second instruction was accented (non-anaphoric reading) and less often when the noun was deaccented (anaphoric reading). Thus, listeners immediately use accentual cues to dissolve the relation between referents. Pointing in a similar direction, Weber et al. (2006) showed that listeners are able to exploit intonational information from contrastively accented (L+H*) adjectives to identify a referent prior to its mention. Similar results have been reported for English adults (Ito & Speer, 2008). Children from six years onwards are similarly able to exploit information from contrastive pitch accent types but their fixation response is delayed compared to adults (Ito, Bibyk, Wagner, & Speer, 2014).

Information on pitch accents (also in combination with boundary tones) is also used in the online interpretation of syntactic structure, as well as speaker meaning (Bibyk, Gunlogson, & Tanenhaus, submitted; Heeren, Bibyk, Gunlogson, & Tanenhaus, 2015a; Kurumada et al., 2014; E.-K. Lee & Watson, 2011, among others). In that regard, Heeren, Bibyk, Gunlogson, and Tanenhaus (2015b) and Bibyk et al. (submitted) investigated the effect of boundary tones on listeners’ interpretation of questions vs. statements. The authors developed an interactive card game (participant against computer) which employs syntactically ambiguous sentences (Gotta candy! vs. Gotta candy?). Participants’ fixations to a blocking card (necessary when the sentence is understood as a statement) vs. to playing cards (necessary when the sentence is understood as a question) are informative on how the ambiguous sentence is interpreted by the listener. The experiment is currently replicated in Australian English, a variety of English that uses uptalk a lot, i.e., in which statements are frequently rising (Asano, Grohe, Kember, & Weber, 2017). In sum, these studies indicate that listeners immediately use intonational cues for sentence processing.

**Pitch accents and lexical processing**

Given that pitch accent types primarily signal communicative purposes, it is obvious that listeners use this kind of information to make pragmatic choices. By contrast, it might seem a more daring venture to expect that pitch accent type could influence lexical processing given that they primarily signal information-structural differences of a referent in intonation languages. Yet, there is some previous research indicating that intonational contours also affect online lexical processing. In a seminal study by Cutler (1976), participants had to detect a phoneme in a target word as quickly as possible, e.g., the phoneme /d/ in dirt. The target was embedded
in a sentence with different continuations, such as “She managed to remove the <Target> from the rug”, triggering either a “high stress version” (contrastive accent on dirt) or a “low stress version” (contrastive accent on rug). The target itself was spliced into these sentences from a neutral version. Hence only the preceding context indicated the intonation condition and not the target itself. Results indicated that participants were faster at detecting a phoneme in a target in the high stress condition, i.e., when the context predicted an accent on the target word than when it predicted no accent, as in the low stress condition (cf. Akker & Cutler, 2003; Cutler, 1976). Hence, prior intonational context affected listeners’ phoneme monitoring latencies for acoustically identical sounds.

Importantly, the intonation contour of an utterance has been shown to affect lexical access. Specifically, Braun, Dainora, and Ernestus (2011) presented Dutch sentences with a broad focus intonation and with an unfamiliar intonation contour, which was an inverted, slightly declining sine contour. The sine contour was imitated by the speaker based on a resynthesized version in order to guarantee natural productions with natural word onset cues. At the end of the auditory sentence, participants saw a visual target word (identical to the last word of the sentence) and performed a lexical decision task. Listeners showed longer RTs when sentences were produced with the sine contour compared to a normal Dutch intonation contour (broad focus). This effect appeared in all tasks administered, i.e., word monitoring and CMFP. The authors conclude that the effect of intonation on lexical processing is robust and long lasting as it affects lexical access as well as processes that follow lexical access and involve the meaning of lexical candidates. Hence, there is evidence from a small amount of studies, suggesting that intonation influences lexical processing.30

The interaction of stress and intonational cues during word recognition

Little is known about how stress and intonational cues interact during word recognition. There is related evidence, however, that distal intonational cues (alternating high and low tonal targets in words preceding targets) influence listeners’ expectations on the stressed syllable of the target and consequently their consideration between segmentally overlapping stress competitors: Brown, Salverda, Dilley, and Tanenhaus (2015), for example, showed that the recognition of a target, either SW (jury) or WS (giraffe), depends on the f0 pattern of the preceding context (e.g., in “Heidi sometimes saw”). Results showed that if stressed syllables preceding the

30 Note that f0 information has been shown to play a role in adult segmentation, with listeners relying on the pitch contour to decide between lexical candidates, such as Norma Nelson and Norman Elson, in Italian, English and French (D’Imperio, Petrone, & Nguyen, 2007; Ladd & Schepman, 2003; Welby, 2007). Yet, these studies focused more on the segmentation of the speech stream as a function of f0 information than on lexical activation per se, rendering them only indirectly relevant for the purpose of our investigation with adults.
target had the same acoustic properties as the first syllable of the target word, the initially stressed candidate (jury) was favoured over the candidate stressed on the penultimate (giraffe), i.e., the first syllable was interpreted as stressed. On the other hand, if unstressed syllables preceding the target had the same acoustic make-up as the first syllable of the target word, then the candidate stressed on the penultimate (giraffe) was favoured over the initially stressed target (jury), i.e., the first syllable was interpreted as unstressed (see Niebuhr (2009) for similar findings on German in an off-line stress judgement task).

Not only distal intonational cues, but also the pitch contour of the target itself has been demonstrated to modulate word recognition based on the evaluation of suprasegmental stress cues. Friedrich, Alter, and Kotz (2001) demonstrated in an identification study that listeners were slower and made more errors when reacting to a SW-target (e.g., Amboss [ˈam.bɔs] ‘anvil’) when the pitch contour in the target was taken from an accented WS-word (e.g., Abtei [ap.ˈtai] ‘abbey’) than when the pitch contour was taken from an accented SW-word. Friedrich, Kotz, Friederici, and Gunter (2004) further revealed that intonation contours not only modulate behavioural responses, but also ERP (event-related potentials) recorded in an electroencephalography (EEG) experiment: In a cross-modal fragment priming experiment with a lexical decision task, ERPs were measured for reactions to target words. Besides faster RTs, listeners showed reduced P350 amplitude to targets (e.g., Regel [ˈreːɡl] ‘rule’) when the monosyllabic prime (e.g., [reː]) was excised from an item with the same f0 contour as the target, suggesting facilitated processing (Friedrich et al., 2004). Judging from the f0 contours in Friedrich et al. (2001), “initially stressed” pitch contours, i.e., the f0 contour for the SW-words, bore an H*-accent on the first syllable. Conversely, “initially unstressed” contours generated by WS-words bore an H*-accent on the second syllable. Similarly, the “stressed” prime in Friedrich et al. (2004) displayed a larger and steeper f0 excursion, compared to the “unstressed” version of the prime. Hence, from these studies by Friedrich and colleagues one may infer that stress processing in German adults is hampered if the stressed syllable is not high-pitched.

In sum, there is abundant evidence that listeners of intonation languages use suprasegmental stress cue to efficiently recognize words in speech – as soon as they become available in the signal. Pitch accent types are also processed immediately, mostly for referent resolution or pragmatic interferences, but lexical processing can also be influenced by intonation. No study has explicitly manipulated utterance-level pitch accent and investigated the effect of pitch accent type on the activation of segmentally overlapping stress competitors. This is what we study in Chapter 5.
3.3 Infant word segmentation

3.3.1 Finding words in speech

The challenge

Spoken language is not only highly variable, but also continuous (e.g., Cutler, 2012). In fact, the speech signal lacks clear and unambiguous word boundary cues (R. A. Cole & Jakimik, 1980; Klatt, 1989), see Figure 7 for a sound pressure wave and a spectrogram of the German utterance *Heute war es schön sonnig* ‘Today it was beautifully sunny.’ Figure 7 suggests that based on the acoustic signal it is challenging for the listener to localize word boundaries in fluent speech as individual words appear to be glued together, without pauses between them (see word segmentation problem, already introduced in Chapter 1).

![Sound pressure wave and spectrogram of the German utterance *Heute war es schön sonnig* ‘Today it was beautifully sunny’, spoken by a female speaker of German.](image)

Word segmentation in adults – A brief sketch

In adult speech processing, word segmentation is usually considered to be a by-product of the competition process during word recognition (see Cutler, 2012; Hanulíková, 2009, 2012; McQueen, 2005; Weber & Scharenborg, 2012, for discussion). As soon as a word is recognized, the continuous speech signal is decomposed into distinct words. Even though the competition process can correctly parse the input into distinct units of words, processing may be taxing in certain situations, e.g., when there are many similar sounding words in the lexicon which compete with the target (e.g., Luce et al., 2000; Luce & Large, 2001; Vitevitch, 2002; Vitevitch & Luce, 1999), see McQueen (2005) for an overview. At the same time, there is abundant experimental evidence that adult listeners additionally use explicit segmentation cues in the signal in order to decompose speech prior to lexical access (e.g., Mattys et al., 2005, for an integrative study). In particular, adult listeners have been demonstrated to use a variety of signal-based cues for explicit segmentation, e.g., prosodic cues, vowel harmony, phonotactic and allophonic cues, as well as pauses (e.g., Christophe et al., 2004; Cutler & Norris, 1988; Hanulíková et al., 2010; Mattys et al., 2005; McQueen, 1998; Norris et al., 1995; Suomi et al., 1997; Tyler &

---

31 MINERVA 2 does currently not provide a solution for the recognition of words in fluent speech (see Weber & Scharenborg, 2012).
Cutler, 2009; Vroomen, Tuomainen, & de Gelder, 1998). Naturally, signal-derived cues are subject to variation between languages, resulting in a language-specific use of cues (Cutler et al., 1983; Hanulíková, 2009; Hanulíková, Mitterer, & McQueen, 2011; Tyler & Cutler, 2009; Vroomen et al., 1998). In this regard, a wealth of studies have confirmed differences in rhythm-based strategies: For listeners from syllable-timed languages, the syllable has been established as the unit of segmentation (Bradley, Sanchez-Casas, & Garcia-Albea, 1993 on Spanish; Content, Kearns, & Frauenfelder, 2001 on French; Dumay, Frauenfelder, & Content, 2002 on French; Floccia, Goslin, Morais, & Kolinsky, 2012 on Italian; Tabossi, Collina, Mazzetti, & Zoppello, 2000 on Italian), while the mora is the unit of segmentation for Japanese listeners (Murty, Otake, & Cutler, 2007; Otake, Hatano, Cutler, & Mehler, 1993; Otake et al., 1996). Evidence for a stress-based segmentation strategy in German, English, and Dutch adults has been reported in word spotting paradigms, and slips of the ear (e.g., Cutler & Butterfield, 1992; Cutler & Norris, 1988; Kentner, 2015; Vroomen & de Gelder, 1995). All these explicit segmentation strategies in adults are considered to aid spoken word recognition and allow the process to be more efficient (see Cutler, 2012; Hanulíková, 2009).

The relevance of finding words in speech for infants

While explicit segmentation assists in adult word recognition, for infants, this process is fundamental. The principles of activation and competition depend on a mental lexicon that infants initially do not possess. If the mental lexicon were a prerequisite to recognize words in fluent speech, it would follow that infants would have to establish one by hearing distinct words in isolation. However, naturalistic studies on infant-directed speech show that this is rarely the case (E. K. Johnson et al., 2013; Van de Weijer, 2002; Zahner, Schönhuber, et al., 2016b), even when mothers are explicitly told to teach their infants new words (Aslin, Woodward, LaMendola, & Bever, 1996; E. K. Johnson et al., 2013). As Junge (2018) summarizes, only 2 to 9% of the utterances infants are exposed to in the second half of the first year of life are one-word utterances (not counting exclamations, fillers, and social expressions which are one-word utterance by nature, counts are based on Brent & Siskind, 2001; Morgan, 1996). Consequently, infants strongly rely on segmentation strategies to decompose the speech stream into discrete units onto which they can map meaning.

It has been argued that speech segmentation is a vital step in first language acquisition: Longitudinal studies have recently started to investigate the relationship – either via correlation analyses or subgroup comparisons – between early

---

Note though that in a small study by Hanulíková (2012) the results by Cutler and Norris (1988) on English were not replicated with German listeners.
segmentation skills and vocabulary size in early childhood, identifying the ability to extract words from fluent speech as a relevant predictor for later language development (Junge, 2018, for an overview): Infants with advanced segmentation skills develop larger vocabularies (e.g., Junge & Cutler, 2014; Kooijman et al., 2013; Newman et al., 2006; Newman, Rowe, & Bernstein Ratner, 2016; Singh, Reznick, & Xuehua, 2012), giving infants a head start into language acquisition (Junge & Cutler, 2014, for a meta-analysis, showing that the effect size of these studies is moderate, 0.33 95% CI [0.17;0.48]).

For instance, Singh et al. (2012) administered two different behavioural segmentation tasks to infants aged 7.5 months, a simple one (as in Jusczyk & Aslin, 1995) and a more complex task, which required abstraction from intonation during familiarization and test (as in Singh, 2008). The vocabulary size of infants was monitored between 8 and 24 months, and a developmental index was taken at 23 months. Results indicated that segmentation abilities in both tasks (simple and complex) correlated with the vocabulary size at 24 months. The cognitive measure correlated with the simple task but not with the complex one. When grouping infants on the basis of their median vocabulary sizes and growth trajectories, considerable differences in segmentation were revealed: Infants with great vocabularies, for instance, recognized words that mismatched in pitch at 7.5 months already, even though this ability was attested only at 9 months of age in earlier studies (see Singh, White, & Morgan, 2008). The neurophysiological results by Junge and Cutler (2014) point in a similar direction: Infants who show a negative and thus more mature response to the recognized words at 10 months perform better in a looking while listening paradigm at 16 months. At the age of 5 years though, Junge and Cutler (2014) did not find a relationship between early segmentation abilities and linguistic development anymore, as was revealed in a battery of normed perception and production tests. Taken together, these data show that word segmentation is vital for forming a mental lexicon and improving linguistic abilities, which makes speech segmentation an essential prerequisite in first language acquisition (even though the relation of the ability and linguistic development in older children is yet to be more thoroughly explored).

Methods to assess word segmentation in infants
The paradigm that has been most widely used to assess infant speech segmentation is the Head-Turn Preference Procedure (HPP, Kemler Nelson et al., 1995). It is employed with children between 4 and 24 months, but most suitable for infants between 6 and 9 months (E. K. Johnson & Zamuner, 2010). Beyond segmentation, discrimination between languages (e.g., Nazzi, Jusczyk, & Johnson, 2000), discrimination of and preferences for stress patterns (e.g., Höhle et al., 2009) as well as
preferences for speech registers (Fernald, 1985) have been successfully assessed in this paradigm.

The usual setup is the following: Infants are tested in a three-sided experimental booth while sitting on a caregiver’s lap. Each trial starts with a green blinking light at the centre of the screen. As soon as the infant orients towards the centre light, signalling that the child is attentive, a red light to the right or left of the child starts blinking. When infants turn their head towards the sidelight, stimuli are presented and looking times are measured. The stimulus is (most often) infant-controlled, such that speech is presented as long as infants orient towards this side. A new trial starts if the infant looks away for more than two seconds. Generally speaking, differences in looking times to different types of stimuli indicate, depending on the research question, whether infants have recognized one type of stimuli and therefore prefer one type of stimuli over the other, or whether they simply and intuitively prefer a certain type of stimuli (Boll-Avetisyan, 2018; Cutler, 2012; Junge, 2018, for recent overviews).

When segmentation is tested, the paradigm typically employs two phases, a familiarization phase and a test phase: Infants are first familiarized with word lists consisting of several repetitions of isolated tokens (e.g., bike, bike, bike …). Then during the test phase, infants’ recognition of the familiarized words is tested by presenting them with passages – half of which contain the familiarized words and half of which contain novel control words. Conversely, the paradigm can also be used in the passage-word order, i.e., familiarizing infants with passages and testing them on word lists. Longer looking times to passages containing one of the word types or lists of one type (familiarized vs. novel) are taken as an indicator for successful segmentation. Hence, successful extraction of the word can surface in a familiarity preference (longer looking times to familiarized words) or a novelty preference (longer looking times to novel words), depending on different factors such as age and task complexity (e.g., Butler, Floccia, Goslin, & Panneton, 2011; DePaolis, Keren-Portnoy, & Vihman, 2016; Houston-Price & Nakai, 2004; Hunter & Ames, 1988), see Chapter 4 for a more detailed discussion of this issue.

More recently, segmentation has also been investigated using electrophysiological methods, i.e., EEG, a non-invasive method that monitors the activity of the brain by recording the voltage fluctuation at the scalp (see Junge, 2018, for an overview of the paradigm). Event-related potentials inform us about brain activity time-locked to a stimulus. An advantage is that this methodology allows for an online index of word segmentation without requiring an active response by infants (e.g., Kidd et al., 2018; Junge, 2018). That is, infants are presented with the experimental stimuli while watching silent movies in order to keep them as still as possible and thus avoid artefacts in the data caused by movement. Similar to the HPP, infants are familiarized with passages that contain critical words and are later tested.
for recognition of these words. The time course of ERPs time-locked to the beginnings of familiar words compared to novel words is indicative of whether and at what point in time infants recognize familiarized words (Junge, 2018). Hence, successful recognition is shown by diverging ERPs (usually of negative polarity), which typically happens around 350-500 ms from word beginnings, i.e., as early as before the offset of the word. This negative brain response to the familiarized words is referred to as the word recognition effect, N200-500 (Junge, 2018). Note though that positive brain responses can also be observed as an effect of recognition. Positive-going brain responses have been interpreted as more immature acoustic processing, usually obtained with infants around 6 to 7.5 months (Kidd et al., 2018; Kooijman et al., 2013; Männel & Friederici, 2013; Rivera-Gaxiola, Klarman, Garcia-Sierra, & Kuhl, 2005), compared to more advanced negative-going brain responses that are usually taken as to reflect mature linguistic processing, which is the direction of fluctuation that is also found in adults (e.g., Kooijman et al., 2013). While the advantage of not requiring an overt response by the infant is quite convincing and segmentation abilities tend to be captured earlier in ERP responses than in behavioural responses (see Kooijman et al., 2013; Teixidó, François, Bosch, & Männel, 2018, for discussion), the high mobility of infants in their first months leads to a high amount of artefacts in the data and seemingly higher attrition rates than in behavioural studies (e.g., 150% in Kooijman, Hagoort, & Cutler, 2009; 78% in Kooijman et al., 2013).

3.3.2 Cues and strategies in infant word segmentation

An overview of cues and strategies in infant word segmentation

Studies using the two paradigms outlined above have revealed that word segmentation abilities in infants are evident from as early as 6 months onwards, with infants exploiting a variety of cues and strategies – some universal, but some highly language-specific (Bergmann & Cristià, 2016, for a current meta-analysis; Boll-Avetisyan, 2018; Junge, 2018; Teixidó et al., 2018, for current reviews). Jusczyk and Aslin (1995) were the first to demonstrate that American-English 7.5-month-old infants were able to extract monosyllabic words from speech and many subsequent studies adding to the picture of infant segmentation followed soon.

Extensive research on infant word segmentation abilities has drawn the following developmental line: One of the earliest cues that infants rely on for segmentation are statistical regularities. Infants as young as 5 months of age have

---

33 Note that a recent study by Kidd et al. (2018) identified three qualitatively different responses to familiarized words within one age group, i.e., 9-month-old Australian English infants. It might hence be the case that the differences in polarity of the effect are not only age-related but point to individual processing differences.
been shown to be able to use transitional probabilities between syllables in order to extract co-occurring syllables from a monotonous speech stream (Thiessen & Erickson, 2013). Similar findings have been reported for slightly older infants (E. K. Johnson & Jusczyk, 2001; Saffran, Newport, & Aslin, 1996). In these studies, infants are typically familiarized with a monotonous syllable stream, such as \( \ldots \text{da-n} \text{pi} \text{golatutudarodaropitudaro}\ldots \) In these streams, three syllables always co-occur with each other, e.g., \text{daropi}, \text{golatu}, which means that the transitional probabilities within such a trisyllabic “word” are high. Transitional probabilities between these trisyllabic “words”, by contrast, are low (as different words can occur after one another). Infants were able to track statistical regularities and recognized the syllable sequences occurring as words in the stream, e.g., \text{daropi} (as opposed to syllable sequences that spanned a word boundary, e.g., \text{pigola}). E. K. Johnson and Tyler (2010) and E. K. Johnson (2012) have argued though that these artificial syllable sequences are highly simplified compared to natural language. E. K. Johnson and Tyler (2010), for instance, tested 5.5- and 8-month-old infants in two conditions, one in which the CV- (consonant-vowel) syllable stream disposed of words of uniform length (four disyllables), and one in which the CV-syllable stream disposed of words with varying length (2 disyllables and 2 four-syllabic words). While both age groups succeeded to segment the “words” in the uniform length condition, both failed when the “words” were of different length, suggesting that the ability to track transitional probabilities might not be equally robust in natural language where words are more variable.

In a natural language environment, segmentation has been observed from 6 month onwards, with factors such as edge alignment or word familiarity promoting early segmentation attempts. Using natural speech, E. K. Johnson, Seidl, and Tyler (2014), for example, showed that (Canadian English) 6-month-olds reliably segmented monosyllabic words (CVC-structure) from fluent speech, when these words were aligned with the edges of the utterance, either in final or initial position. The segmentation effect was robust when the words were presented with different voices during familiarization and test (male vs. female), suggesting generalization. These findings were consistent with the “Edge Alignment Hypothesis” which was postulated by Seidl and Johnson (2006) based on similar findings with 8-month-old infants. Hence, words at the beginning or end of an utterance are particularly felicitous in early segmentation attempts.

Familiar words, such as the baby’s first name and the name of the caretaker have also been shown to bootstrap early segmentation. For instance, Bortfeld, Morgan, Golinkoff, and Rathbun (2005) showed that 6-month-old American English infants extracted words that were preceded by a familiar name, but failed to do so when the word preceding the target was an unknown name, see also Sandoval and Gómez (2016) for a similar finding. Similarly, Altvater-Mackensen and Mani
(2013) showed the familiarity with a word helps German 7.5-month-olds segmenting other words that are segmentally overlapping with the familiar word, e.g., Löffel ['lœ.f] ‘spoon’ promoted the extraction of the nonce word, but similar sounding word Löckel ['lœ.k].

In the second half of the first year of life, infants develop various other segmentation strategies. As shown by Mattys, Jusczyk, Luce, and Morgan (1999), American English 9-month-olds make use of language-specific phonotactic constraints to localize word boundaries. In particular, infants responded differentially to CVC.CVC nonce words in which the C.C-juncture was a very likely word-internal consonant cluster in English (e.g., [ŋ.k]) as opposed to nonce words in which the medial C.C-juncture was rather unlikely to be found in word internal position in English, e.g., [ŋ.t]. At that age, infants further benefit from reduplicated words, e.g., neenee with primary stress on the initial syllable, compared to non-reduplicated words, e.g., neefoo, as has recently been shown for British English 9-month-olds (Ota & Skarabela, 2017). Slightly later, at around 10.5 months of age, infants are also able to use allophonic variation to detect word boundaries. In a study by Jusczyk, Hohne, and Bauman (1999), American English infants exploited the acoustic differences associated with the sounds /t/ and /ɾ/ in the words night rate vs. nitrate for segmentation. In the second half of the first year of life, language-specific rhythmic strategies gain importance. As this kind of strategy is in the focus of this thesis, we will take a moment to discuss these strategies in more detail.

3.3.3 Stress and intonation in infant segmentation

Rhythmic cues

The Metrical Segmentation Strategy (Cutler & Norris, 1988) and the “rhythmic segmentation hypothesis” (Nazzi et al., 2006) posit that infants use these kinds of rhythmic units to segment speech that are most suitable for the language they grow up with. Indeed, there is a wealth of studies providing evidence for the use of language-specific rhythmic strategies to segment the speech stream into word-like units: Specifically, infants growing up with syllable-timed languages (French, Spanish, Catalan, Italian) rely on the syllable as a segmentation unit (e.g., Bosch, Figueras, Teixido, & Ramon-Casas, 2013; Goyet, Nishibayashi, & Nazi, 2013; Nishibayashi, Goyet, & Nazi, 2015, for French). For instance, before French infants learn to extract polysyllabic word, the syllable is the unit of early prosodic segmentation (Nazzi et al., 2006). Monosyllabic words can hence be segmented early: Bosch et al. (2013) showed for Spanish and Catalan infants that monosyllabic words can be extracted from fluent speech, both at 6 and at 8 months of age. Segmentation of monosyllables was recently also confirmed in French 6- and 8-month-olds (Nishibayashi et al., 2015). Regarding bisyllabic words, Nazi et al. (2006) demonstrated
that before the whole bisyllabic word was segmented at 16 months, infants decomposed the bisyllabic word into monosyllabic words and segmented both the final and the initial syllable (the initial syllable was only segmented when it was excised from the target and not when it was produced in isolation). Note though that the segmentation success can be observed earlier, with increased transitional probabilities and a longer familiarization phase (Nazzi et al., 2014; Polka & Sundara, 2012).

Likewise, for infants from stressed-timed languages, the trochaic unit is the essential unit in the segmentation task: In a seminal study that consisted of a series of 15 HPP experiments, Jusczyk, Houston, et al. (1999) pioneered in demonstrating that American English (AmE) 7.5-month-old infants segmented trochaic units (kingdom, hamlet, doctor, candle) from fluent speech, surfacing in longer looking times to familiarized words, for both test orders (word-passage and passage-word). Note that infants similarly showed successful extraction of trochees when the order of the experimental design was reversed. A consecutive control experiment confirmed that infants reacted to the trochee as a whole and not just to the strong syllable of the trochaic unit (ham). However, a subsequent experiment found that 7.5-month-olds failed to extract iambic words from fluent speech. Instead, when presented with iambs they only extracted the final strong syllable, taking the stressed syllable as a word beginning. The tendency to perceive stressed syllables as word beginnings and to group this syllable with a consecutive weak syllable was so strong that iambic words that co-occurred with a fixed weak element (e.g., the guitar is (WS #W), # indicates word boundary), were mis-segmented as taris at 7.5 months. At 10.5 months, infants can also rely on other strategies (e.g., co-occurrences of syllables) and that way extract iambic words from fluent speech. Thus, the stressed syllable is a vital word onset cue for American English 7.5-month-olds. Note that Houston, Santelmann, and Jusczyk (2004) extended the findings in Jusczyk, Houston, et al. (1999), showing that the ability to use the stressed syllable as a marker of a word onset is limited to strong syllables that carry primary stress and does not

---

34 According to the authors, the passage-word order, i.e., passages in familiarization and isolated words during test, might be the more difficult task. Infants seemingly have no difficulty extracting words from fluent speech. The difference in looking times to familiarized and novel words was numerically even bigger in the passage-word order setup (more than 2.04 seconds, compared to 1.12 seconds in the word-passage order; non-significant interaction).

35 A recent study suggests that lexical knowledge promotes segmentation of iambs in American English 7.5-month-olds: Sandoval and Gómez (2016) familiarized infants with iambic words (guitar) in passages. The iambs were always preceded by a familiar name, e.g., mommy. Infants looked longer to familiarized than to novel iambs, suggesting that the metrical bias can be overridden by lexical knowledge. Additionally, Thiessen and Saffran (2007) showed that an exposure phase to iambic words can promote the segmentation of iambic units such that American English 9-month-olds segment both iambs and trochees from an artificial speech stream. Regarding the extraction of iambs in German, Höhle (2002, p. 187ff.) showed that 9-month-olds react to the strong syllable of an iamb (as their English peers). This was not the case for 11-month-olds, which could indicate that they have extracted the whole iambic unit and not just the strong syllable, but this experiment has not been conducted yet.
apply to syllables that carry secondary stress. The use of stress for segmentation seems to be supported by a preference for the typical stress pattern of the native language. German 6-month-olds and English 9-month-olds display a trochaic bias in preference tasks (Höhle et al., 2009, for German; Jusczyk et al., 1993, for English).

Before infants from stress-timed languages learn to integrate several cues, stress cues seem to be the most relevant strategy for them to localize word boundaries: When confronted with ambiguous input or conflicting cues, 5- and 7.5-month-old American English infants have been shown to rely on statistical rather than on phonological (stress) cues for segmentation (Thiessen & Erickson, 2013; Thiessen & Saffran, 2003). These authors have argued that the early cue weighting advantage for statistical probabilities might promote the ability to detect language-specific peculiarities that, in turn, might be informative of word boundaries, such as the stressed syllable for potential word onsets in West-Germanic languages. The cue weights soon change towards a reliance on linguistic cues: At 8 and 11 months, linguistic cues (stress and co-articulation) outweigh statistical cues (E. K. Johnson & Jusczyk, 2001; E. K. Johnson & Seidl, 2009). Moreover, at 9 months, stress is given more weight than phonotactic information when the two types of information conflict (Mattys et al., 1999). Thus, before infants learn to integrate different strategies in their attempts to decompose fluent speech, stress seems to be a powerful cue for determining potential word boundaries.

A closer look at stress-based segmentation in German, English and Dutch
Corroborating the crucial role of stressed syllables in infant word segmentation, almost two decades worth of research has repeatedly shown that infants with a stress-timed language background use the stressed syllable to localize word beginnings and consequently, extract trochaic units from fluent speech: for English infants (Kidd et al., 2018 for Australian English (AusE); Floccia et al., 2016 for British English (BE); Mason-Apps, Stojanovik, Houston-Price, & Buckley, 2018 for BE), for Dutch infants (Houston et al., 2000; Junge et al., 2012; Kooijman et al., 2005; Kooijman et al., 2013; Kuijpers et al., 1998), and for German infants (Altvater-Mackensen & Mani, 2013; S. Bartels et al., 2009; Höhle, 2002; Mani & Pätzold, 2016; Männel & Friederici, 2013). Segmentation of trochees has even been observed cross-linguistically, i.e., with English infants extracting trochees when tested with Dutch stimuli (Houston et al., 2000) or German stimuli (Höhle, 2002, p. 88ff.), see Table 3 for a summary of these studies. Note that in all these studies, “stress” was considered a conglomerate of different suprasegmental cues in the signal.

Generalizing over the findings on stress-based segmentation summarized in Table 3, infants from stress-timed languages seem to reliably extract trochees from
fluent speech from 9 months onwards. Some neurophysiological studies report evidence of successful segmentation of trochees from 6 months. In these studies, early segmentation success mostly surfaces in a positive brain response, which the authors of these studies attribute to a more immature stage of processing (Kooijman et al., 2013; Männel & Friederici, 2013). It appears that some linguistic factors promote stress-based segmentation: For instance, German infants’ ability to extract trochees seems to depend on the quality of the vowel of the final unstressed syllable. While German 9-month-olds failed to extract *Tilsum* [ˈtɪlsʊm], a trochaic nonce word that contained a full vowel in the second syllable, they successfully segmented trochees with final schwa-syllables, e.g., *Balken* [ˈbalkən] ‘beam’. Moreover, intonation-related factors, such as speech register, IDS vs. ADS (Floccia et al., 2016; Schreiner & Mani, 2017), as well as accentuation (Männel & Friederici, 2013) have recently been identified to foster the ability to extract words from fluent speech. These factors will be discussed in the next subsection where we focus on the interplay between intonation and stress-based segmentation.
Table 3: Overview of studies on stress-based segmentation (extraction of trochees) in infants with a stress-timed language background.

<table>
<thead>
<tr>
<th>Study</th>
<th>Lang</th>
<th>Method</th>
<th>Order</th>
<th>N</th>
<th>Mo</th>
<th>Condition</th>
<th>SW?</th>
<th>Dir. of Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jusczyk, Houston, et al. (1999)</td>
<td>AmE</td>
<td>HPP</td>
<td>w-p</td>
<td>24</td>
<td>7.5</td>
<td>NA</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>7.5</td>
<td>NA</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>7.5</td>
<td>WS # W as SW</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td>Kidd et al. (2018)</td>
<td>AusE</td>
<td>EEG</td>
<td>p-w</td>
<td>103</td>
<td>9</td>
<td>NA</td>
<td>Yes</td>
<td>Negativity (1/3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>Positivity (1/3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
<td>None (1/3)</td>
</tr>
<tr>
<td>Mason-Apps et al. (2018)</td>
<td>BE</td>
<td>HPP</td>
<td>w-p</td>
<td>35</td>
<td>10</td>
<td>NA</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td>Floccia et al. (2016)</td>
<td>BE</td>
<td>HPP</td>
<td>p-w</td>
<td>16</td>
<td>10</td>
<td>Exaggerated IDS</td>
<td>Yes</td>
<td>Novelty</td>
</tr>
<tr>
<td>Kuijpers et al. (1998)</td>
<td>Dutch</td>
<td>HPP</td>
<td>w-p</td>
<td>24</td>
<td>7.5</td>
<td>NA</td>
<td>No</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>w-p</td>
<td>24</td>
<td>9</td>
<td>NA</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td>Kooijman et al. (2005)</td>
<td>Dutch</td>
<td>EEG</td>
<td>w-p</td>
<td>28</td>
<td>10</td>
<td>NA</td>
<td>Yes</td>
<td>Negativity</td>
</tr>
<tr>
<td>Kooijman et al. (2013)</td>
<td>Dutch</td>
<td>EEG</td>
<td>w-p</td>
<td>28</td>
<td>7.5</td>
<td>NA</td>
<td>Yes</td>
<td>Positivity</td>
</tr>
<tr>
<td>Junge et al. (2012)</td>
<td>Dutch</td>
<td>EEG</td>
<td>w-p</td>
<td>28</td>
<td>10</td>
<td>NA</td>
<td>Yes</td>
<td>Negativity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>28</td>
<td>10</td>
<td>NA</td>
<td>No</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>w-p</td>
<td>24</td>
<td>9</td>
<td>No Schwa</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Altvater-Mackensen and Mani (2013)</td>
<td>German</td>
<td>Mid Screen</td>
<td>w-p-w</td>
<td>20</td>
<td>7.5</td>
<td>Onset overlap</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>w-p-w</td>
<td>20</td>
<td>7.5</td>
<td>Offset overlap</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td>Mani and Pätzold (2016)</td>
<td>German</td>
<td>Mid Screen</td>
<td>p-w</td>
<td>24</td>
<td>16</td>
<td>IDS</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>31</td>
<td>16</td>
<td>ADS</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
</tbody>
</table>
Table 3 Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Lang</th>
<th>Method</th>
<th>Order</th>
<th>N</th>
<th>Mo</th>
<th>Condition</th>
<th>SW?</th>
<th>Dir. of Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Männel and Friederici (2013)</td>
<td>German</td>
<td>EEG</td>
<td>p-w</td>
<td>24</td>
<td>6</td>
<td>Accented</td>
<td>Yes</td>
<td>Positivity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>6</td>
<td>Unaccented</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>9</td>
<td>Accented</td>
<td>Yes</td>
<td>Negativity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>9</td>
<td>Unaccented</td>
<td>Yes</td>
<td>Negativity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>12</td>
<td>Accented</td>
<td>Yes</td>
<td>Negativity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p-w</td>
<td>24</td>
<td>12</td>
<td>Unaccented</td>
<td>Yes</td>
<td>Negativity</td>
</tr>
<tr>
<td>Houston et al. (2000)</td>
<td>Dutch - Dutch</td>
<td>HPP</td>
<td>w-p</td>
<td>24</td>
<td>9</td>
<td>NA</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td></td>
<td>AmE - Dutch</td>
<td></td>
<td>w-p</td>
<td>30</td>
<td>9</td>
<td>NA</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
<tr>
<td>Höhle (2002, p. 96ff.)</td>
<td>AmE - German</td>
<td>HPP</td>
<td>w-p</td>
<td>30</td>
<td>9</td>
<td>Final schwa</td>
<td>Yes</td>
<td>Familiarity</td>
</tr>
</tbody>
</table>

**Lang**: Language in which infants are tested. Cases with 2 languages refer to cross-linguistic studies: language 1 = native language, language 2 = materials; **Method**: Paradigm used in the study (HPP, EEG or Middle Screen = Mid Screen); **Order**: Order of the procedure (w-p = word-passage, p-w = passage-word); **N**: Number of infants in the sample; **Mo**: Average age of infants in months; **Condition**: Indicates whether there was an additional experimental manipulation to the extraction of SW-units; **SW?**: Indicates whether infants successfully segmented the SW-unit (yes/no); **Dir. of Effect**: Direction of the effect if trochee is segmented: for behavioural studies (familiarity/novelty), for electro-physiological (positivity/negativity).

---

*a* From 10.5 months onwards, infants successfully extracted WS-words, in both orders.

*b* A group of infants with Down syndrome (N = 14) also successfully extracted the SW-words. Here we concentrate on the typically developing group.

*c* Note that all other 12 groups failed in this study (age range between groups: 8-10.5 months, range of sample size: 15-24 N).

*d* Individual differences were identified; the segmentation success correlated with later language development.


Intonation in infant stress-based segmentation

So far, little is known about the use of intonational cues for word segmentation and even less about the interplay between utterance-level intonation and the stress-based segmentation mechanism in infants. There is some related evidence on the use of intonational cues from studies investigating the recognition of phrasal boundaries: These studies show that infants use, among other cues, intonational cues to recognize an intonational phrase boundary (Gerken, Jusczyk, & Mandel, 1994; Holzgrefe-Lang, Wellmann, Höhle, & Wartenburger, 2018; Männel & Friederici, 2009; Seidl, 2007; Wellmann et al., 2012). For instance, Wellmann et al. (2012) demonstrated that German 8-month-olds recognize a phrase boundary when it is collectively marked by pre-boundary lengthening, a pause, and an f0 movement. When only pre-boundary lengthening and f0 movement are present, recognition is still successful, but not when one of these two cues occurs in isolation. Holzgrefe-Lang et al. (2018) recently replicated this pattern of results for German 8-month-olds in a neurophysiological study (the condition including the pause was not tested), and additionally provided evidence for successful recognition of phrase boundaries in German 6-month-olds (in the combined pitch and duration condition, but not in the pitch only condition).

Regarding the recognition of words and intonation, some studies have addressed the effect of infant-directed speech and found different segmentation results than in adult-directed speech (Floccia et al., 2016; Schreiner & Mani, 2017; Thiessen, Hill, & Saffran, 2005). For instance, Schreiner and Mani (2017) familiarized German 7.5- and 9-month-old infants with monosyllabic words embedded in passages using a visual fixation paradigm (slightly different setup as in an HPP study, but similar in its rationale). The passages were recorded in a highly exaggerated IDS register (acoustics were matched to AmE IDS). For test, infants listened to words in isolation, half of which were familiarized and half of which were novel. Results revealed a significant interaction between age group and familiarization, indicating that 9-month-olds extracted the monosyllables from fluent speech while 7.5-month-olds did not. Similarly, Thiessen et al. (2005) found a beneficial effect of register for American English infants – already for 7-month-olds. They tested two groups of American English infants in an HPP experiment; one group was familiarized with nonce sentences with controlled transitional probabilities in the IDS register, one group with the same sentences in the ADS register. In the test phase, infants listened to syllable sequences with high transitional probabilities between syllables (words) and to syllable sequences with low transitional probabilities between syllables (part-words), both recorded with a monotone voice. Importantly, these words were the same for both groups of infants. Their looking times revealed a difference in looking times to words vs. part-words in the IDS condition only, but not in the ADS condition. This indicates that infants extracted the words when
they heard them in IDS sentences with exaggerated contours before but not when they heard them in ADS sentences. This beneficial effect of IDS has been discussed in terms of an “attentional spotlight” mechanism fostering segmentation (cf. Schreiner & Mani, 2017; Thiessen et al., 2005). Alternatively, the effect could be stimuli-driven only since IDS employs intonation contours with large f0 excursions (e.g., Fernald & Simon, 1984, for German). Scaling modifications in pitch could hence be felicitous for early segmentation. However, since intonation was often not explicitly manipulated (Floccia et al., 2016; Schreiner & Mani, 2017), it is difficult to tease apart the effect of intonation and other acoustic cues that coincide with this change in register (duration, intensity). Yet, in the study by Thiessen et al. (2005), overall duration and intensity were controlled such that the beneficial effect of register seems to be due to f0.

There is only one study that has investigated the effect of intonation on speech segmentation by explicitly manipulating intonation in the familiarization phase of the experiment (here presence/absence of a pitch accent). Specifically, Männel and Friederici (2013) used ERPs to demonstrate that word recognition by German infants is influenced by accentuation, i.e., whether or not a trochee (Sirup [ˈziː.rʊp] ‘syrup’) receives a pitch accent: Infants’ ERPs indicated that the recognition of trochees differs with age and is modulated by accentuation: 6-month-olds only recognized trochees that were accented in the familiarization, surfacing in a positive ERP response 500 ms after word onset, i.e., before the end of the trochee. At 9 months, recognition was independent of accentuation and manifested in a (mature) negative response 400 ms after word onset; this effect was followed by a late negativity only for accented familiarized words. At 12 months, infants recognized words independent of accentuation during familiarization; the effect surfaced as a negative response 350 ms after word onset. We may infer from this study that accented trochees are generally felicitous in early segmentation. However, no study has addressed the effect of pitch accent type on the extraction of trochaic units from speech. This dissertation therefore seeks to fill this gap by testing whether pitch accent type and the resulting consequences of (mis)alignment between the stressed syllable and the f0 peak (rather than accentedness in general, Männel & Friederici, 2013) modulate the segmentation success in German infants, see Chapter 4.

3.4 Interim summary
Chapter 3 was devoted to the processing of stress and intonational cues in both infants and adults. We started out by showing that, from the perspective of the listener, various forms of high pitch bear the potential of signalling word-level prominence. In particular, both infants and adults use high f0 for linguistic grouping
while adult listeners of West Germanic languages further strongly attend to different kinds of high pitch for lexical stress perception in meta-linguistic decision tasks.

We then focused on two specific processes, word recognition in adults and word segmentation in infants, and discussed the use of stress and intonation therein. Generally speaking, to understand fluent speech, the listener needs to recognize the words the speaker intends, which is a highly complex process. Different views have been proposed for the mapping between the signal and linguistic units. While episodic or exemplar theory posits a direct mapping between the input and the stored representations, abstractionists agree on a pre-lexical level and hence a mediated mapping. Episodic theories assume that all words encountered are stored as highly detailed traces, while representations are abstract in the other theory. We have reported on effects in language processing which provide evidence for both of these mechanisms and highlighted the emergence of hybrid approaches.

The recognition process itself is commonly understood as a process of activation and competition between lexical candidates. The studies reviewed in this chapter collectively show that lexical activation and competition are affected by suprasegmental information. In order to recognize words efficiently, listeners process these cues as soon as they become available in the signal, quickly distinguishing between segmentally overlapping stress competitors (e.g., [ɔk.to.pus] vs. [ɔk.'to.bar]). Similarly, infants also use stress cues for speech segmentation, a task that is highly relevant for them to develop a mental lexicon. By the age of 9 months, German, Dutch, and English infants have been repeatedly demonstrated to extract trochaic units from fluent speech, interpreting the stressed syllable in the signal as word onset. For both processes (adult word recognition and infant stress-based segmentation), research on the interaction between the use of stress cues and intonational cues is limited.

In light of the background reviewed in Chapters 2-3, this dissertation sets out to investigate how alignment differences in different pitch accent types affect (a) stress-based segmentation in German infants (Chapter 4) and (b) lexical activation of stress competitors in German adults (Chapter 5). Given the perceptual relevance of high f0 for both infants and adults, we predict stress-based segmentation and lexical activation to be influenced in conditions in which the f0 peak and the lexically stressed syllable are misaligned.
Chapter 4

Pitch accent type affects stress-based segmentation in German infants

Chapter outline
In a series of HPP experiments (passage-word order), we investigated the effect of pitch accent type on stress-based segmentation in German infants, i.e., the ability to extract trochaic (SW) units from speech. In Experiment 1, WSW-carrier words – embedded in sentences – were presented in one of three naturally occurring intonation conditions: one in which the f0 peak was aligned with the stressed syllable (medial-peak condition, L+H* L-%) and two misalignment conditions, with the f0 peak preceding vs. following the stressed syllable, early-peak condition (H+L* L-%) or late-peak condition (L*+H L-% / L* H^-H%), respectively. The recognition of the embedded SW-units was tested. Results showed that infants extracted the SW-unit from the WSW-word only when f0 peak and stressed syllable coincided (medial-peak condition), but not in the misalignment conditions. Recognition was observed irrespective of the intonation of test items (similar vs. dissimilar to familiarization), as Experiment 1b confirmed. Experiment 1c excluded the possibility that the tonal alternation in the medial-peak condition (LH*L) rather than the f0 peak makes the stressed syllable stand out (no SW-recognition was observed when infants were familiarized with a horizontally flipped medial-peak condition, HL*H), suggesting high f0 to be a necessary cue to stress in segmentation. Experiment 2 tested whether high f0 is also a sufficient cue to stress, such that infants perceive an embedded iamb (WS) with an HL contour, as in a WWS-word with an early-peak accent, as a trochaic unit (SW). This re-interpretation of the word-prosodic structure due to accent type was only observed with naturally produced early-peak accents (Experiment 2), but not with resynthesized early-peak accents (Experiment 2b) where the f0 cue was isolated from other peak supporting acoustic cues. Hence, pitch accent type guides infants’ stress-based segmentation, with the f0 peak being a necessary and – in naturally produced accents – also a sufficient cue to stress.

Adapted parts of this chapter have been presented in Zahner, Schönhuber, and Braun (2016) and Zahner and Braun (2018), i.e., results of Experiments 1, 1b, and 1c. The results of Experiment 2 have been presented in Zahner, Schönhuber, Grijzenhout, and Braun (2016a).
4.1 Introduction

This chapter addresses the effect of pitch accent type on stress-based segmentation in German infants. As has been shown in 3.3, within the second half of the first year of life, infants have learned to use various segmentation strategies: universal strategies, such as transitional probabilities between syllables (Saffran, Aslin, & Newport, 1996; Thiessen & Erickson, 2013, but see Johnson, 2012, and Johnson & Tylor, 2010), but also language-specific phonotactic constraints (Mattys & Jusczyk, 2001b; Mattys et al., 1999), coarticulatory phonetic cues (E. K. Johnson & Jusczyk, 2001), or position-specific allophonic variants (Jusczyk, Hohne, et al., 1999). Familiarity with words (Altvater-Mackensen & Mani, 2013; Bortfeld et al., 2005; Sandoval & Gómez, 2016), positional factors, such as word occurring at utterance edges (E. K. Johnson et al., 2014; Seidl & Johnson, 2006), may further promote segmentation. Importantly, for infants who grow up with a stress-timed language, stress appears to be the most powerful cue for determining potential word boundaries within the second half of the first year of life. As has been outlined earlier, around 9 months of age, German, English and Dutch infants have repeatedly been shown to employ the MSS and extract trochaic units from speech (e.g., S. Bartels et al., 2009; Jusczyk, Houston, et al., 1999; Kuijpers et al., 1998), see Table 3 for a summary. When stress cues are pitted against statistical cues, infants rely on the stressed syllable for segmentation (E. K. Johnson & Jusczyk, 2001; E. K. Johnson & Seidl, 2009).

Only recently have developmental studies started to investigate the weighting of cues in infant stress perception, as has been discussed earlier. There is evidence from linguistic grouping studies that high f0 plays a crucial role in infants’ perception of stress: Syllable sequences alternating in pitch were grouped as trochaic units, with the high element forming the strong one (Abboub et al., 2016; Bion et al., 2011). Note that other paradigms, i.e., stress discrimination and word recognition, have suggested that the convergence of cues (high f0 and other stress cues) might be more important for stress perception than the occurrence of high f0 alone (see Bhatara et al., 2018; Quam & Swingley, 2014). Crucially, the factors leading to an infant’s perception of a syllable as stressed when segmenting fluent speech has not been exhaustively studied to date (see Gambell & Yang, 2006, for discussion; but see Thiessen & Saffran, 2004, for the role of spectral tilt). Recently, there have

---

37 Thiessen and Saffran (2004) studied the role of spectral tilt for the perception of stress and consequently word segmentation in American English, i.e., spectral tilt as an isolated stress cue or converging with other stress cues (such as duration, amplitude increase across the whole spectrum and an f0 peak). Results revealed that 9-month-olds interpreted a syllable as stressed and consequently as word onset when stress was signalled by the energy distribution in the spectrum alone (more energy in higher frequency areas). Yet, infants at around 12 months of age (similar to adults) did not use the stressed syllable as a word onset cue when only signalled by the energy distribution. Instead, when the energy distribution converged with other stress cues, the syllable was used as a word onset.
been some attempts to unravel the mechanisms underlying metrical segmentation. In this regard, a study by Männel and Friederici (2013) reveals an effect of accentuation on German infants’ ability to extract trochaic units: Infants at the age 6 months already extracted accented trochees from speech but failed when trochees were unaccented. Given the perceptual relevance of high pitch in the perception of stress and the beneficial effect of accentuation for segmentation it appears that trochees realized with a high-pitched accent might be a prime candidate for segmentation.

Only little is known on the intonational realization of the stimuli used in previous studies on infant stress-based segmentation. As stress naturally occurred as a conglomerate of different cues in previous studies, acoustic analyses of the stimuli (most often) remain sparse. Table 4 summarizes the descriptions of the stimuli in the studies that were outlined in Table 3.\textsuperscript{38} As Table 4 reveals, many studies only report the style of the voice with which the words were recorded, mostly a lively voice one would use with an infant (e.g., Kuijpers et al., 1998; Mason-Apps et al., 2018). Others provided average values of acoustic variables (mostly f0 and duration) for syllables (Jusczyk, Houston, et al., 1999), for words (Mani & Pätzold, 2016; Männel & Friederici, 2013), or for whole sound files (Floccia et al., 2016). The acoustic analysis for the trochees in Jusczyk, Houston, et al. (1999) shows that the first (stressed) syllable was higher (and longer) than the second (unstressed) syllable, which seems to suggest that this study might have used (more) medial-peak accents (than other accent types). Except for the measurements with respect to the syllable as a unit of measurement, it is hard to derive the intonational realization from the stimuli in previous studies (as an acoustic analysis based on the word does not allow to derive the direction of the movement, i.e., falling or rising, for instance). Exemplar realizations of the stimuli further provide insights, but do not allow for strong generalizations. Based on the examples that Männel and Friederici (2013), for instance, provide, it seems that for the “accented” condition, medial-peak accents were used in that study, while for the “unaccented” condition, it is not entirely clear whether all trochees were unaccented or realized with a late-peak accent, as the example suggests. Taken together, it is unclear which intonational contours were used in previous studies. There seems to be some (speculative) evidence that medial-peak accents might have been more frequently employed.

The current series of experiments sets out to examine the effect of intonation on stress-based segmentation. Specifically, we test if the position of the f0 peak in regard to the stressed syllable (in different pitch accent types) affects German 9-month-olds’ perception of stress and consequently metrical segmentation. To study

\textsuperscript{38} In Table 4, we do not list studies that used stimuli of previous studies. In these cases, we report the description of the original stimuli. This is why Table 4 has fewer entries than Table 3.
this question, we use the HPP setup with a familiarization phase consisting of passages and a consecutive test phase with words in isolation, i.e., passage-word order (see Floccia et al., 2016, Experiments 1, and 3-5; Jusczyk, Houston, et al., 1999, e.g., Experiment 2; Mani & Pätzold, 2016). Pitch accent type will be manipulated during familiarization and subsequent recognition of trochees is tested. In the remainder of this Introduction, we will outline the rationale of the experiments along with our predictions.

39The facilities in the BabySpeech Lab (BSL) at the University of Konstanz (https://www.ling.uni-konstanz.de/bsl/ - last accessed: 29.02.2020) offer an HPP setup which could be used to assess the research question of this thesis.
Table 4: Overview of description of experimental stimuli, with focus on f0 information, in studies on stress-based segmentation summarized in Table 3.

<table>
<thead>
<tr>
<th>Study</th>
<th>Description of stimuli</th>
</tr>
</thead>
</table>
| **Jusczyk, Houston, et al. (1999)**
AmE, HPP, w-p / p-w, 7.5 and 10.5 mo | Trochees in lists:  
* Repetition of items “in a lively voice, and as if naming the object for an infant.”
* No difference in duration for the first (368 ms) and second syllables (375 ms) > final lengthening
* F0 peaks for first syllables (500 Hz) higher than for second syllables (311 Hz) > 8.2 st
> Suggests medial-peak renditions, with high f0 being the main correlate; considerable f0 excursion
Trochees in passages:
* Passages recorded in “a lively voice, as if reading to a small child”
* First syllable (320 ms) longer than second (191 ms) > no final lengthening, sentence-medial position
* F0 peaks of the first syllables (278 Hz) higher than of the second syllables (220 Hz) > 4.1 st
> Suggests medial-peak rendition, with high f0 and duration signalling stress
* F0 range in lists higher (8.2 st) than in passages (4.1 st) |
| **Kidd et al. (2018)**
AusE, EEG, p-w, 9 mo | * “All target words were disyllabic and initially stressed, as in Jusczyk, Houston, et al. (1999), and the materials were spoken in a clear and lively style but without exaggerated prosody.” > No further information |
| **Mason-Apps et al. (2018)**
BE, HPP, w-p, 10 mo | * “Passages were recorded in a lively motherese voice” > No further information |
| **Floccia et al. (2016)**
BE, HPP, w-p/p-w, 10 mo | Exaggerated IDS vs. IDS
> Acoustic measures for f0 mean, f0 standard deviation (SD), intensity mean, and duration for the whole sound file, making interpretation hard
> Tab. 3 compared exaggerated vs. typical IDS: exaggerated IDS overall higher (on av. 25.9 Hz), more variable in pitch (on av. 12.2 Hz higher f0 SD), longer, and louder
> Fig. 1 shows an example sentence with exaggerated IDS vs. typical IDS (medial peaks, see p.c. Caroline Floccia confirming that mostly medial-peaks were used) |
<table>
<thead>
<tr>
<th>Study</th>
<th>Description of stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kuijpers et al. (1998)</strong></td>
<td>Trochees in lists:&lt;br&gt; * “[S]peaker repeated each target word […] with some variation, 15 times in a row.”&lt;br&gt; Trochees in passages:&lt;br&gt; * “Passages for each of the four target words were read in a lively voice by a phonetically trained adult female, as if talking to a small child.”&lt;br&gt; * Passages recorded with eight filler passages “so that the speaker would not contrastively stress the target words” &gt; Contrastively stressed could mean L+H*, yet realization remains speculative</td>
</tr>
<tr>
<td>Dutch, HPP, w-p, 7.5 and 9 mo</td>
<td></td>
</tr>
<tr>
<td><strong>Kooijman et al. (2005)</strong></td>
<td>Trochees in passages:&lt;br&gt; * “[W]ords and sentences were recorded […] in animated child-directed speech”&lt;br&gt; * Only mean duration of whole target word is given &gt; No further information</td>
</tr>
<tr>
<td>Dutch, EEG, w-p, 10 mo</td>
<td></td>
</tr>
<tr>
<td><strong>Junge et al. (2012)</strong></td>
<td>Trochees in passages:&lt;br&gt; * Passages taken from Kooijman et al. (2005) &gt; Same description&lt;br&gt; Trochees in lists:&lt;br&gt; * Words for word list were excised from sentences to keep acoustic properties constant across familiarization and test</td>
</tr>
<tr>
<td>Dutch, EEG, w-p/p-w, 10 mo</td>
<td></td>
</tr>
<tr>
<td><strong>Höhle (2002, p. 88ff.) &amp; S. Bartels et al. (2009)</strong></td>
<td>Trochees in lists:&lt;br&gt; * Repetition of items “with a slightly different intonation.” &gt; No further information on intonational realization, no information on realization of words in passages, for both conditions (schwa vs. without schwa)</td>
</tr>
<tr>
<td>German, HPP, w-p, 9 mo</td>
<td></td>
</tr>
<tr>
<td><strong>Altvater-Mackensen and Mani (2013)</strong></td>
<td>* “Sentences and words were spoken by a female native speaker of German in moderate infant-directed speech (mean target duration = 723 ms (isolated words) / 376 ms (sentence context); mean target amplitude = 62 dB; mean target pitch = 187 Hz)” &gt; No further information</td>
</tr>
<tr>
<td>German, Mid Screen, w-p-w, 7.5 mo</td>
<td></td>
</tr>
</tbody>
</table>
Table 4 Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Description of stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mani and Pätzold (2016)</td>
<td>Trochees in passages (IDS vs. ADS):</td>
</tr>
<tr>
<td>German, Mid Screen, p-w, 16 mo</td>
<td>* IDS targets had higher maximum f0 (ADS = 235 Hz; IDS = 349 Hz) and minimum f0 (ADS = 204 Hz; and mean f0 (ADS = 202 Hz; IDS = 317 Hz)</td>
</tr>
<tr>
<td>IDS vs. ADS</td>
<td>&gt; F0 excursion in IDS: 9.3 st; f0 excursion in ADS: 5.4 st; direction not clear</td>
</tr>
<tr>
<td></td>
<td>Trochees in lists (same for IDS and ADS):</td>
</tr>
<tr>
<td></td>
<td>* Targets spoken in “a neutral prosody (mean f0 = 218 Hz, max f0 = 298 Hz, min f0 = 157 Hz)</td>
</tr>
<tr>
<td></td>
<td>&gt; Considerable f0 excursion of 11.1 st but not clear whether fall or rise; if citation form, presumably fall</td>
</tr>
</tbody>
</table>

| German, EEG, p-w, 6, 9, and 12 mo, Accented vs. Unaccented | * “[W]ords spoken with accentuation, as compared to those without accentuation, were of longer duration, at higher intensity and pitch levels, and had larger intensity and pitch ranges” |
|                               | * For accented: Mean f0 max: 335 Hz, mean f0 min = 158 Hz, mean f0 mean = 237 Hz; For unaccented: Mean f0 max: 259 Hz, mean f0 min = 147 Hz, mean f0 mean = 197 Hz |
|                               | > F0 excursion (based on above values, see their Tab. 2) considerable in both conditions (Accented: 13.0 st; Unaccented: 9.7 st), but direction of f0 movement unclear clear |
|                               | > Visual inspection of exemplar realization (their Fig. 1) suggests medial-peak accent in condition Accented |
|                               | Trochees in lists (same for Accented and Unaccented):                                   |
|                               | * “[A]coustic comparison of test words with both kinds of familiarization words (accentuated and non-accentuated) suggests that the test words had similarly reduced overall intensity and pitch patterns as the non-accentuated words (Tab. 2)” > Considerable f0 excursion of 10.3 st; direction of f0 movement unclear |

The column Description of stimuli summarizes the description of experimental materials in previous studies, with focus on f0 information. Information marked by * is taken from the methods part of the respective study, either as a summary or as a direct quote. “;” > denotes our own interpretation inferred from the description.
**General overview of experiments and predictions**

**Experiment 1**

In Experiment 1, rare trisyllabic WSW-carrier words (e.g., *Lagune* [la.'gu:na], ‘lagoon’) are embedded in short passages presented in the familiarization phase. In the consecutive test phase, infants are tested on whether they extracted the SW-unit (e.g., [gu:na]) from trisyllabic WSW-carrier words (e.g., [la.'gu:na]). WSW-words are presented in three different intonation conditions (between-subjects): one alignment condition in which the f0 peak is aligned with the stressed syllable (medial-peak accent) and two misalignment conditions in which the f0 peak and the stressed syllable diverge (early-peak and late-peak accents respectively). Two misalignment conditions are used in order to gain a more complete understanding of how the position of the f0 peak and metrical stress interact in infant metrical segmentation. Trisyllabic WSW-carrier words are chosen to allow us to locate the entire pitch accent on the target word in all three intonation conditions. Note that with disyllabic SW-words, as used in many other segmentation studies (see Table 3), the early pitch peak would have been placed on the word preceding the target word. Beyond that, WSW-words represent a strong test case for investigating the role of stressed syllables as word onset cues as these trisyllabic carrier words provide conflicting segmentation cues: On the one hand, metrical stress hints at a word boundary after the first unstressed syllable (W#SW, where “#” again signals a word boundary) while on the other hand, linguistic, phonetic, and statistical regularities (such as the presence of function words, coarticulatory information and transitional probabilities) in fact indicate a word boundary before the first unstressed syllable of WSW-carrier words (#WSW). We do not know the extent to which German infants at this age exploit cues hinting to a boundary before the trisyllabic carrier (i.e., whether there is no support for or whether there is support against the boundary before the SW-unit). Given that American English infants have been shown to extract embedded trochees from a three- or four-syllable sequence from 7.5 months onwards (Echols, Crowhurst, & Childers, 1997; Jusczyk, Houston, et al., 1999), we expect our group to be able to solve the task. Clearly though, infants can solely rely on stress to extract the respective SW-unit in our setting.

Based on the background reviewed above, we predict that infants are able to extract the trochaic units from the WSW-carrier words in the medial-peak condition in which the stressed syllable is high-pitched. On the contrary, early-peak or late-peak contours on the WSW-word in which the f0 peak is not aligned with the second stressed syllable might be less felicitous for the extraction of SW-units.
**Experiment 2**

If the predictions for Experiment 1 are accurate (and infants extract the SW-unit in the medial-peak condition, but fail in the misalignment conditions in Experiment 1), then in a second step (Experiment 2), we test whether infants weigh the high f0 cue as more relevant for the perception of metrical stress and consequently for segmentation than other stress cues (i.e., longer duration, higher intensity, and an increased vocal effort). In other words, we investigate whether high f0 is a sufficient cue to metrical stress such that genuinely unstressed syllables (with an f0 peak) are perceived as stressed and consequently lead to a re-interpretation of the metrical strength relations in a word, such that infants perceive an embedded iamb (WS) with an HL contour as a trochee (SW).

To this end, infants will be familiarized with passages containing nonce WWS-carrier words\(^{40}\) (e.g., *Rumila* [ru. mi.'la:]), which are presented in two intonation conditions (between-subject): an early-peak condition in which the f0 peak precedes the final stressed syllable and a medial-peak-condition with the f0 peak on the final stressed syllable. Under the assumption that high f0 is a sufficient cue for infants to signal stress, we expect infants to perceive the second syllable as the stressed one in a WWS-word ([ru. mi.'la:] with an early-peak rendition, LHL), such that the perceived word-prosodic structure would be WSW. In that case, infants are expected to perceive the last two syllables as the trochee and hence display recognition for a SW-unit ([Imit.la:]). Figure 8 illustrates the expected re-interpretation of the word-prosodic structure under the assumption that high f0 is indeed a sufficient cue to lexical stress for German infants.\(^{41}\)

---

\(^{40}\) In this experiment, we employed nonce words instead of real “meaningful” words to ensure that the words are unknown to all infants (see accumulating evidence that 6-month-olds already know the meaning of many common words, Bergelson & Aslin, 2017; Bergelson & Swingley, 2012).

\(^{41}\) Note that we used WWS-words in this experiment to create a situation in which infants similarly extract embedded units (see Experiment 1), instead of using the first two syllables (WS-part) of the WSW-carrier words (e.g., [la:gu] from *Lagune* [la.'guː.naː]).
4.2 **Experiment 1**

4.2.1 **Methods**

**Participants**

Fifty-four full-term born (at least 37 weeks of gestation) 9-month-olds from monolingual German families who finished the familiarization phase and all twelve test trials were included in the analysis (25 female, 29 male, average age: 0;9.1, age range: 0;8.19-0;9.16). One third of the infants were tested in the medial-peak condition (9 female, 9 male, average age: 0;9.1), one third in the early-peak condition (7 female, 11 male, average age: 0;9.1) and one third in the late-peak condition (9 female, 9 male, average age: 0;9.1). Assignment to the conditions was random. Twenty-six additional infants were tested but excluded from the analysis due to fussiness\(^{42}\) (10), crying (13), or not attending to the blinking lights (3). Parents were reimbursed for public transport or parking fees and received a small present for the child.

**Materials**

**Familiarization stimuli**

Four trisyllabic words with low lexical frequency (less than 0.1 occurrences per million in the CELEX word form dictionary, Baayen et al., 1993) that are not expected to be familiar to 9-month-old infants served as carrier words. All of them consisted of CV-syllables with stress on the penultimate: *Kanone* [ka.ˈnoː.nə], ‘cannon’; *Lagune* [la.ˈɡuː.nə], ‘lagoon’; *Kasino* [ka.ˈsiː.no], ‘casino’; *Tirade* [ti.ˈraː.da], ‘tirade’. Note that the unstressed initial syllables are not reduced to [ə] in German but retain their full vowel quality. For each of the four WSW-words, we constructed six sentences, such that the carrier word appeared in different lexical contexts and different sentence positions (twice in sentence-final position, four times early in the sentence following an article or prenominal adjective). The words preceding and following the targets differed across sentences, see (4) for an exemplar passage (bold face for SW-part in the WSW-carrier words; italics for other accentuated words in the sentence). Appendix A (A.1) shows all passages.


‘Here originated a lagoon. The lagoon was wonderful. The blue lagoon attracts people. A small lagoon is nice. His lagoon was situated in the South. She took a photo of her lagoon.’

---

\(^{42}\) All infants tested received a fussiness score by the experimenter (either the author of dissertation or a trained research assistant) ranging from 1 (for very patient infants that behaved very well during the experiment) to 4 (for very fussy and restless infants that moved a lot or turned around on their caregiver’s lap). Infants with a score higher than 3 were excluded from the analysis.
A twenty-six-year-old female native speaker of Standard German from the southwest of Germany (Baden-Wuerttemberg), who was trained in intonational phonology, recorded the 24 target sentences in the three intonation conditions. In the medial-peak condition, the f0 peak was realized on the stressed syllable (L+H*), in the early-peak condition, the f0 peak preceded the stressed syllable (H+L*), and in the late-peak condition, the f0 peak followed the stressed syllable (L*+H / L* H-%), see Figure 9. As stress is signalled by a variety of acoustic cues that are distributed over the stressed and neighbouring unstressed syllables (Kohler, 2012; Niebuhr, 2007a), we used naturally produced stimuli in order to make all potential cues available for infants. Care was taken that the distribution and type of other pitch accents in the sentences was the same across intonation conditions, see (4). The speaker read the sentences in a natural and lively way. Stimuli were recorded in a sound-attenuated cabin (44.1 kHz, 16 Bit) and analysed in Praat (Boersma & Weenink, 2014). To achieve equally salient f0 movements across conditions, the sentences were recorded several times and the best matching sentences were chosen, so that eventually the average f0 excursion of the fall in the early-peak condition, was matched with the accentual rise in the medial-peak and the late-peak condition. Further acoustic analyses confirmed that the WSW-words in the three intonation conditions were similar with regard to a number of acoustic variables, see Table 5. The average duration of the passages was 16.4 s (SD = 0.4 s) in the medial-peak, 16.7 s (SD = 0.4 s) in the early-peak condition, and 16.0 s (SD = 0.3 s) in the late-peak condition.

Figure 9: Example sound pressure wave, spectrogram, and pitch track of a target sentence in the medial-peak condition (a), the early-peak condition (b), and the late-peak condition (c) in Experiment 1. Tier 1 gives the words in German, tier 2 the English translation; tier 3 displays the GToBI labels.
Table 5: Acoustic realization (mean values (and standard deviations)) of WSW-target words in the familiarization phase for the three intonation conditions in Experiment 1.

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>Peak-stress alignment condition (medial-peak)</th>
<th>Peak-stress misalignment condition (early-peak)</th>
<th>Peak-stress misalignment condition (late-peak)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of movement in st</td>
<td>8.8 (1.7)</td>
<td>8.5 (1.4)</td>
<td>8.4 (1.7)</td>
</tr>
<tr>
<td>Duration of first syllable (unstressed) in ms</td>
<td>182 (36)</td>
<td>193 (35)</td>
<td>191 (32)</td>
</tr>
<tr>
<td>Duration of second syllable (stressed) in ms</td>
<td>253 (25)</td>
<td>256 (23)</td>
<td>258 (21)</td>
</tr>
<tr>
<td>Duration of third syllable (unstressed) in ms</td>
<td>193 (66)</td>
<td>193 (73)</td>
<td>190 (42)</td>
</tr>
<tr>
<td>Duration of onset consonant in stressed syllable in ms</td>
<td>89.9 (35.7)</td>
<td>89.1 (33.4)</td>
<td>93.1 (23.4)</td>
</tr>
<tr>
<td>H1*-A3+ ratio in middle of first vowel in dB</td>
<td>21.5 (6.5)</td>
<td>20.8 (7.3)</td>
<td>26.4 (6.2)</td>
</tr>
<tr>
<td>H1*-A3+ ratio in middle of second vowel in dB</td>
<td>31.5 (14.1)</td>
<td>31.5 (11.9)</td>
<td>37.4 (13.0)</td>
</tr>
<tr>
<td>H1*-A3+ ratio in middle of third vowel in dB</td>
<td>23.2 (5.3)</td>
<td>24.0 (5.0)</td>
<td>29.9 (10.8)</td>
</tr>
</tbody>
</table>

Test stimuli

Each infant was tested on the same set of four test “words”. These four test words consisted of the SW-unit of the WSW-carrier words: [ɡuːnə] taken from Lagune, [ˈraːdə] from Tirade, [ˈnoːnə] from Kanone, and [ˈsiːno] from Kasino. To increase comparability with earlier HPP segmentation studies (see Table 3), the test stimuli were elicited in the same way as in those studies, i.e., produced “as if speaking to a child” (see Table 4) and with varied f0 range. Closely following Jusczyk, Houston, et al. (1999) who report a decrease in f0 between the first and second syllable in SW-units (see Table 4), we chose falling pitch contours for the SW-units.

The same speaker as for the familiarization stimuli recorded each of these SW-units approximately 30 times with a pitch fall and slightly different durations and f0 excursions to increase the phonetic variability of this contour. For the experiment, we chose 15 tokens of each trochee, such that the average f0 excursion of the pitch fall and the average duration of the test word did not differ across test words. The average f0 excursion of the pitch fall was 10.1 st (SD = 1.8 st), ranging from 6.4 st

43 Following Mooshammer (2010), we used the H1*-A3+ ratio as a measure for vocal effort, i.e. the difference between the amplitude of the first harmonic and the third formant (asterisks denote that amplitudes were corrected for formants). In order to perform these acoustic measurements, we adjusted a Praat script downloaded from http://www.seas.ucla.edu/spapl/voicesauce/, last access: 14.03.2018.
Pitch accent type affects stress-based segmentation in German infants

to 13.1 st. The average duration of the test words was 710 ms (SD = 88 ms), ranging from 540 ms to 884 ms. The 15 tokens of each test word were concatenated with an inter-stimulus-interval (ISI) of 800 ms. The lists were on average 22.7 s long (SD = 0.4 s), see Appendix A (A.2) for acoustic details on the individual test lists.

Procedure
Infants were seated on their parent’s lap, facing a three-sided black experimental booth in the BabySpeech Lab at the University of Konstanz, see Figure 10, and tested in the HPP paradigm. Each trial started with a green blinking light at the centre of the booth. As soon as the infant oriented towards the centre, the green centre light was switched off and a red light to the right or left of the child started blinking. When infants turned their heads towards the sidelight, the auditory stimuli started playing. The sound played as long as infants oriented towards this side. If infants looked away for more than 2 s, the next trial started.44

Figure 10: Schematic illustration of the experimental setup (HPP booth) in the BabySpeech Lab.

In the familiarization phase, the two passages were presented semi-randomly from the left or the right side with at most two trials from the same side until infants had listened to each of the two paragraphs for at least 45 s. Then, the test started. In the test phase, infants listened to lists of the SW-units. They were also presented in a semi-random order from the left or the right side, with no more than two trials from either side in a row. Looking times were coded online45 by an experimenter who monitored infants via a video camera and controlled the experiment via button

44 Upon arrival in the lab, parents signed a consent form. The questionnaire on the language background and the infant data was filled out after the experiment in order not to artificially stretch the infant’s attention span. In prior telephone or E-mail conversations, we checked whether the infant was monolingual German and full-term born (necessary inclusion criteria).

45 Prior reliability studies in our lab have shown that the inter-coder reliability between online and offline coding is very high. A trained person re-coded the looking behaviour of four randomly chosen videotapes recorded in the Head-Turn Preference Paradigm. The looking time data for online and offline coding were very strongly correlated ($r = 0.99$, $n = 48$ trials), suggesting that the online coding was reliable.
presses from behind the booth. The experimenter as well as parents wore tight-fitting headphones with masking music in order not to hear the auditory stimuli infants listened to. The experimental session lasted approximately six minutes.

In each of the three intonation conditions, half of the infants were assigned to the Kanone and Tirade familiarization trials, the other half to the Kasino and Lagune trials. In the test phase, all infants listened to test lists consisting of 15 repetitions of the four isolated SW-units, two of which were units of the WSW-carrier words presented in the familiarization phase (e.g., none and rude), and two of which were units of the WSW-carrier words used with the other sub-sample of infants, thus novel syllable sequences (e.g., sino and gune). In total, there were three blocks of four trials, with three pseudo-randomized repetitions of the four lists. Across infants, we counterbalanced the sides from which the test lists were presented (right vs. left) and the list beginnings, such that all four part-words once occurred at a list beginning.

4.2.2 Results and interim discussion

Analysis and results

Looking times in seconds were averaged by familiarity status (novel vs. familiar) for each infant. The average looking times were 10.2 s (SD = 3.1 s) to novel and 8.8 s (SD = 2.7 s) to familiar lists in the medial-peak condition. Thirteen out of 18 infants oriented longer to the novel lists. In the early-peak condition, infants looked on average 8.3 s (SD = 2.3 s) to novel and 8.5 s (SD = 2.2 s) to familiar lists, with 10 out of 18 infants orienting longer to the novel lists. In the late-peak condition, average looking times were 8.5 s (SD = 2.5 s) to novel and 8.4 s (SD = 2.6 s) to familiar items and 8 out of 18 infants oriented longer to the novel lists, see Table 6 (first three columns).

For statistical analysis in R (R version 3.3.3, R Development Core Team, 2015), we first calculated the average looking time difference and the 95% confidence interval (CI) for all three conditions by subtracting the looking time to familiar test lists from the looking time to novel test lists for each infant, see Figure 11 (left panel). The results show a robust difference in looking times in the alignment condition (medial-peak), but not in the two misalignment conditions (early-peak and late-peak condition). The overlap between the CI of the medial-peak and those of the other two intonation conditions is small enough to suggest a robust difference between the alignment condition and the two misalignment conditions (Cumming & Finch, 2005).
Table 6: Overview of average looking times to novel (Nov) and familiar (Fam) test lists in Experiment 1. Standard deviations are indicated in brackets. The third row (Diff) displays the difference in looking time, along with the p-value as indicated by the post-hoc pairwise t-test.

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th></th>
<th>Exp. 1</th>
<th></th>
<th>Exp. 1b</th>
<th></th>
<th>Exp. 1c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Medial-peak</td>
<td>Falling test</td>
<td>Early-peak</td>
<td>Falling test</td>
<td>Late-peak</td>
<td>Falling test</td>
<td>Medial-peak</td>
</tr>
<tr>
<td>Nov</td>
<td>10.21 s</td>
<td>(3.06 s)</td>
<td>8.31 s</td>
<td>(2.34 s)</td>
<td>8.52 s</td>
<td>(2.55 s)</td>
<td>11.11 s</td>
</tr>
<tr>
<td>Fam</td>
<td>8.79 s</td>
<td>(2.71 s)</td>
<td>8.47 s</td>
<td>(2.23 s)</td>
<td>8.45 s</td>
<td>(2.62 s)</td>
<td>9.64 s</td>
</tr>
<tr>
<td>Diff</td>
<td>1.42 s</td>
<td>-0.16 s</td>
<td>0.07 s</td>
<td></td>
<td>1.47 s</td>
<td>0.15 s</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>= 0.01</td>
<td>= 0.70</td>
<td>= 0.88</td>
<td></td>
<td>= 0.01</td>
<td>= 0.78</td>
<td></td>
</tr>
</tbody>
</table>

Figure 11: Means of difference in looking time to novel and familiar items in Experiment 1. Whiskers represent the 95% CI of the difference in looking time. The grey dashed line indicates zero, i.e., no difference. Left panel: Experiment 1 (three intonation conditions); Right panel: Experiments 1b and 1c.

Results of a repeated measures ANOVA (analysis of variance)\footnote{As was suggested by an anonymous reviewer (for Zahner, Schönhuber, & Braun, 2016), ANOVAs, along with post-hoc pairwise t-tests, were preferred over linear mixed effects models (Baayen, 2008; Baayen, Davidson, & Bates, 2008) to increase comparability with previous studies (summarized in Table 3).} with \textit{intonation condition} as between-subject factor and \textit{familiarity status} as within-subject factor showed a statistically significant interaction between the two factors ($F(2,51) = 3.53$, $p = 0.04$). Post-hoc pairwise t-tests for the three intonation conditions separately showed a statistically significant difference between looking times to novel and familiar test lists in the medial-peak condition ($t(17) = 3.2, p = 0.01$) only, but not in the two misalignment conditions (both $p$-values $> 0.7$). A data analysis according
to a Bayesian approach (M. D. Lee & Wagenmakers, 2013)\(^{47}\) shows that in the medial-peak condition the alternative hypothesis is nine times more likely than the null hypothesis \((r = 0.71, \text{bf} = 8.70)\), while the null hypothesis is approximately four times more likely than the alternative hypothesis in the early-peak condition \((r = 0.71, \text{bf} = 0.26)\) and in the late-peak condition \((r = 0.71, \text{bf} = 0.25)\).

**Interim discussion**

In the medial-peak condition where the f0 peak was aligned with the stressed syllable, infants looked significantly longer to the novel than to the familiar test lists. The magnitude of this looking time difference \((1.4 \text{ s})\) is comparable to other segmentation studies summarized in Table 3. In the two misalignment conditions, there was no looking time difference. These findings show that infants extracted the embedded SW-part-words from fluent speech only when the stressed syllable was high-pitched, but not when the pitch peak and stressed syllable were misaligned. In other words, only high-pitched stressed syllables were interpreted as stressed and consequently taken as word onsets while low-pitched stressed syllables did not serve as strong element in the trochee to be used in segmentation in German 9-month-olds. Note though that the recognition effect found in Experiment 1 surfaces in a novelty effect, i.e., longer looking times to the unfamiliarized \((\text{novel})\) test lists. The direction of this effect is different from the direction found in most segmentation studies (see Table 3), but comparable to Floccia et al. (2016, Experiment 4), who used exaggerated IDS contours. Echols et al. (1997) also observed a novelty preference with 9-month-olds when extracting embedded trochees from WSW-carrier words in a slightly different paradigm. They argue that the task was easier than in other segmentation studies (those in Table 3) due to the acoustic match between items in familiarization and test. Such an explanation does not hold in our case (as our stimuli provide phonetic variability even though the similarity between familiarization and test is greatest in the medial-peak condition, see below). From our perspective, the observed novelty preference is due to the fact that the presence of a medial-peak accent made all cues to lexical stress available \((\text{f0 peak, duration, intensity, and spectral information})\), which might have decreased the task complexity for infants at that age (Hunter & Ames, 1988, on the role of task complexity on the direction of preference). German infants have been shown to extract accentuated words at the age of 6 months already (electro-physiological evidence in Männel & Friederici, 2013), making the novelty preference for a high-pitched accent type at 9 months likely.

\(^{47}\) To calculate the Bayes Factor, we used the R package “Bayes Factor” (Morey & Rouder, 2015) applying the function for a paired t-test, ttestBF().
The recognition effect of SW-units in the medial-peak condition allows for two alternative explanations to our finding (calling for two control studies to corroborate the interpretation that the alignment between f0 peak and lexical stress drives the effect). First, the recognition effect in the medial-peak condition could be traced back to the acoustic similarity between familiarization and test stimuli (the trochaic part was falling in both cases) and suffered from the intonational change in the other two conditions (the trochaic part was flat or rising in familiarization in the early- and late-peak contour, respective, and falling in test). In other words, it is plausible that the task was easier for infants in the alignment condition (allowing for a more direct match) than in the two misalignment conditions (which necessitate abstracting away from intonational information). Previous findings on infants’ early representations speak against such a direct matching account though, as infants can generalize over speaker (male vs. female, Houston & Jusczyk, 2000; van Heugten & Johnson, 2012), emotions (neutral vs. happy, Singh, Morgan, & White, 2004), and pitch levels (high vs. low, Singh et al., 2008) by the age of 9 months. It is thus likely that the change in intonation from familiarization to test does not hinder infants’ recognition of the SW-units. Nevertheless, before drawing stronger conclusions about the role of high pitch in infants’ stress perception, this alternative interpretation has to be excluded. For familiarization in Experiment 1b, we hence used the medial-peak condition from Experiment 1, but used SW-units with a rising contour in test. If the looking time difference in the medial-peak condition of Experiment 1 stems from the similarity in intonation contours between familiarization and test alone, infants are not expected to extract the SW-units under these modified conditions. If infants instead rely on the peak-stress alignment as a segmentation cue, they are expected to show a similar recognition effect as in the medial-peak condition of Experiment 1.

A second alternative explanation is that successful recognition of the trochee might not have been triggered by high-pitched stressed syllable, but by tonal alternation. In other words, infants may be particularly sensitive to the stressed syllable when the neighbouring syllables differ in f0, e.g., LHL or HLH, but less sensitive when there is little change, HLL or LLH, as in the misalignment conditions. This explanation is compatible with an early view on prominence perception on the utterance level by Bolinger (1958, p. 112), arguing that it is a “wide departure from a contour” (in any direction) that makes a syllable stand out. It is also compatible with a recent proposal on prominence perception according to which unexpected f0 events lead to increased prominence (Kakouros & Räsänen, 2016; Kakouros, Salminen, & Räsänen, 2018). In Experiment 1c, we hence test a contour, which has the tonal alternation in the opposite direction (HL*H, we call “flipped”-medial peak condition). If the change in f0 level is the necessary cue to trigger metrical segmentation, gene is expected to be extracted from Lagune in Experiment 1c. If, by
contrast, the f0 peak is the necessary cue, *gun* is not expected to be extracted. Another test case would be monotonous contours without any f0 movement. Likely, they would be perceived as unaccented (Männel & Friederici, 2013) which is why they are not useful and discarded here.

### 4.2.3 Experiment 1b: Acoustic matching?

#### Methods

#### Participants

Eighteen full-term infants from monolingual German homes who finished familiarization and all 12 test trials (6 female, 12 male, average age: 0;9.1, age range: 0;8.18–0;9.21) were included in the analysis. Eleven infants were excluded from the analysis due to fussiness (3), crying (2), not attending to blinking lights (3), falling asleep (2), or due to an unusually short overall average looking time (> 2 SD below the average looking time (1)). Criteria for the exclusion score were the same as in Experiment 1.

#### Stimuli

**Familiarization**

We used the medial-peak familiarization stimuli from Experiment 1.

**Test**

The four SW-units for the test phase were the same as in Experiment 1, but recorded with a rising pitch contour, resulting in a low-pitched stressed syllable followed by a high-pitched second syllable. The average f0 excursion of the pitch rise was 12.4 st (SD = 1.6 st), ranging from 9.1 to 15.1 st. The average duration of the test words was 679 ms (SD = 57 ms), ranging from 546 to 814 ms. As before, the 15 tokens were concatenated with an ISI of 800 ms, resulting in lists with an average duration of 21.4 s (SD = 0.3 s), see Appendix A (A.3) for acoustic details on the individual test lists.

#### Procedure

The procedure was the same as in Experiment 1.

#### Results and interpretation

Participants looked on average 11.1 s (SD = 2.1 s) to novel test lists and 9.6 s (SD = 2.2 s) to familiar ones, see Table 6 (fourth column) for a comparison to the other conditions in Experiment 1. Fifteen out of 18 infants looked longer to novel than to familiar items. The average looking time difference and the 95% CI are shown in Figure 11 (right panel). Figure 11 indicates that this difference in looking time is
similar to the medial-peak condition in Experiment 1 (Cumming & Finch, 2005). A pairwise t-test for the medial-peak condition with rising test intonation also showed a statistically significant difference between looking times to novel and familiar test lists ($t(17) = 2.9, p = 0.01$), as in the medial-peak condition in Experiment 1. The alternative hypothesis is five times more likely than the null hypothesis ($r = 0.71, bf = 5.33$). Next, the data of both medial-peak conditions (Experiments 1 and 1b) were pooled and entered into a repeated measures ANOVA with test intonation (between-subject) and familiarity status (within-subject) as factors. The factors did not interact ($p = 0.95$). Hence, as in the medial-peak condition in Experiment 1, infants showed significantly longer looking times to novel than to familiar test lists, despite the fact that the intonation of the familiarization and test stimuli was different. We therefore exclude the alternative explanation that infants in the medial-peak condition of Experiment 1 extracted the trochaic test items only because of the intonational similarity between familiarization and test stimuli. We will discuss infants’ ability to generalize over intonational contours in relation to previous studies below (4.4).

4.2.4 Experiment 1c: Alternation?

Methods

Participants

Another eighteen full-term German monolingual infants who finished the familiarization and all 12 test trials were included in the analyses (9 female, 9 male, average age: 0;9.2, age range: 0;8.18-0;9.17). Twelve further infants had to be tested to arrive at the sample of 18. These twelve infants were excluded due to crying (5), not attending to blinking lights (5), fussiness (1), and interference of a sibling (1). Criteria for exclusion were the same as in Experiment 1.

Familiarization

For familiarization, we used the four passages from Experiment 1 but re-recorded the passages. The same speaker as in Experiment 1 realized the WSW-carrier words (e.g., <em>Lagune</em>) with a flipped-medial peak intonation contour (HL*H), see Figure 12. Note that the (tritonal) contour is not described in the German intonational system (Grice et al., 2005), but it is indeed present in German infant-directed speech, occurring in 7% of the 426 accentual movements in the KIDS (Konstanz prosodically analysed infant-directed speech) Corpus (Zahner, Schönhuber, et al., 2016b). We matched the current WSW-words closely to those in Experiment 1, see Table 7.
Figure 12: Example sound pressure wave, spectrogram, and pitch track of a target sentence in the flipped medial-peak condition in Experiment 1c. Tier 1 gives the words in German, tier 2 the English translation; tier 3 displays the GToBI labels (note that HL*H is not described in GToBI, label is our own description).

Table 7: Acoustic realization (mean values (and standard deviations)) of WSW-target words in the familiarization phase for the flipped medial-peak condition (HL*H) in Experiment 1c. For ease of comparison the acoustic realization of the intonation conditions in Experiment 1 (grey shading) are displayed again.

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>Exp. 1c Flipped medial-peak</th>
<th>Exp. 1 Medial-peak</th>
<th>Exp. 1 Early-peak</th>
<th>Exp. 1 Late-peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of movement in st HL*H:</td>
<td>8.4 (1.0)</td>
<td>8.8 (1.7)</td>
<td>8.5 (1.4)</td>
<td>8.4 (1.7)</td>
</tr>
<tr>
<td></td>
<td>L*H:</td>
<td>L+H*:</td>
<td>H+L*:</td>
<td>L*+H:</td>
</tr>
<tr>
<td></td>
<td>8.5 (0.9)</td>
<td>8.5 (1.4)</td>
<td>8.4 (1.7)</td>
<td></td>
</tr>
<tr>
<td>Duration of first syllable (unstressed) in ms</td>
<td>196 (21)</td>
<td>182 (36)</td>
<td>193 (35)</td>
<td>191 (32)</td>
</tr>
<tr>
<td>Duration of second syllable (stressed) in ms</td>
<td>259 (20)</td>
<td>253 (25)</td>
<td>256 (23)</td>
<td>258 (21)</td>
</tr>
<tr>
<td>Duration of third syllable (unstressed) in ms</td>
<td>187 (50)</td>
<td>193 (66)</td>
<td>193 (73)</td>
<td>190 (42)</td>
</tr>
<tr>
<td>Duration of onset consonant in stressed syllable in ms</td>
<td>91.3 (25.3)</td>
<td>89.9 (35.7)</td>
<td>89.1 (33.4)</td>
<td>93.1 (23.4)</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of first vowel in dB</td>
<td>22.9 (7.4)</td>
<td>21.5 (6.5)</td>
<td>20.8 (7.3)</td>
<td>26.4 (6.2)</td>
</tr>
<tr>
<td></td>
<td>(7.4)</td>
<td>(6.5)</td>
<td>(7.3)</td>
<td>(6.2)</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of second vowel in dB</td>
<td>34.5 (10.7)</td>
<td>31.5 (14.1)</td>
<td>31.5 (11.9)</td>
<td>37.4 (13.0)</td>
</tr>
<tr>
<td></td>
<td>(10.7)</td>
<td>(14.1)</td>
<td>(11.9)</td>
<td>(13.0)</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of third vowel in dB</td>
<td>31.3 (12.4)</td>
<td>23.2 (5.3)</td>
<td>24.0 (5.0)</td>
<td>29.9 (10.8)</td>
</tr>
<tr>
<td></td>
<td>(12.4)</td>
<td>(5.3)</td>
<td>(5.0)</td>
<td>(10.8)</td>
</tr>
</tbody>
</table>

Test

For test, we used the falling test lists from Experiment 1.48

---

48 The falling test lists provided the strongest mismatch between the trochaic part of the carrier in the familiarization and the trochee in the test phase (this decision was based on the effect found in Experiment 1b).
Procedure

The procedure was identical to the one described in Experiment 1.

Results and interpretation

Infants looked 9.6 s (SD = 3.5 s) to familiar and 9.8 s (SD = 3.4 s) to novel test lists, see Table 6 for looking times (last row) in comparison with other conditions and Figure 11 (right panel) for 95% CIs of the looking time differences in Experiment 1c compared to the previous conditions. Nine infants out of 18 looked longer to the novel lists. A pairwise t-test indicated that the difference in looking times was not significant ($t(17) = 0.3, p > 0.78$). The null hypothesis was about four times more likely than the alternative hypothesis ($r = 0.71, bf = 0.25$). To corroborate the difference across conditions, we pooled the data (medial-peak condition in Experiment 1 with data in Experiment 1c) and tested for an interaction between familiarity status and intonation condition in a repeated measures ANOVA with familiarity status as within-subjects factor and intonation condition as between-subjects factor. The interaction between the two factors approached significance ($F(1,34) = 3.32, p = 0.08$). Thus, in contrast to the medial-peak condition in Experiment 1 (LH*H), infants did not extract the SW-units from WSW-carriers with a flipped medial-peak accent (HL*H) in this experiment. A mere alternation of tonal targets of opposite height (HL*H) did not lead to a percept of stress and the use in stress-based segmentation for the syllable with the deviant pitch. Instead, the current pattern of results is similar to the misalignment conditions in Experiment 1 (early-peak and late-peak conditions), speaking in favour of a high-pitched accented syllable guiding the segmentation process in German infants.

Taken together, Experiment 1, together with its control studies (Experiments 1b and 1c), demonstrated that only high-pitched stressed syllables were perceived as stressed and consequently taken as word onsets by German 9-month-olds. An open question is whether for a syllable to be perceived as stressed, the presence of an f0 peak is sufficient. To test this, in Experiment 2, infants are exposed to trisyllabic nonce words (WWS, e.g., Rumi [ru.mi.ˈlaː]) in sentence-contexts. Recall that nonce words were used to safeguard against infants knowing the meaning of the words (Bergelson & Aslin, 2017; Bergelson & Swingley, 2012). The WWS-words were either realized with a medial-peak accent or with an early-peak accent. If the f0 peak on the second unstressed syllable in the WWS-word is strong enough to trigger a percept of lexical stress and thus leads to a metrical re-interpretation of the word-prosodic structure (WS realized with HL, leading to a SW-percept), infants are expected to extract SW [ˈmi.ˈla] from WWS Rumi [ru.mi.ˈlaː] in the early-peak condition (LHL-contour).
4.3 Experiment 2

4.3.1 Methods

Participants
Forty-eight full-term, monolingual German infants were included in the analyses (17 female, 31 male, average age: 0;9.0, age range: 0;8.19-0;9.17). Twenty-four of these infants were tested in the early-peak (10 female, 14 male, average age: 0;9.1) and 24 in the medial-peak condition (7 female, 17 male, average age: 0;8.30), with a random assignment to intonation condition. Thirty-four additional infants were tested but excluded from the analysis due to fussiness (14), crying (14), not turning to the blinking lights (2), or parental interference (1). Additionally, three infants were excluded because of unusual looking behavior: two infants due to a looking time above/below 2 SD of the average looking time, one infant due to a difference in looking times that was > 2 SD above the average difference of all infants. Criteria for the exclusion score were the same as reported in Experiment 1.49

Materials
Familiarization
Four trisyllabic nonce words with stress on the final syllable (WWS-pattern) were created. They only consisted of CV-syllables with tense vowels to avoid influences of syllable weight and vowel quality on stress placement (Féry, 1998; Giegerich, 1985; Vennemann, 1990, 1991): WWS Rumila [ru.mi.'la:], Linuro [li.nu.'ro:], Nalomu [na.lo.'mu:], Morani [mo.ra.'ni:]. All segments were sonorant to allow infants to properly extract f0 in all cases (e.g., Niebuhr, 2017) and at the same to be more suitable for resynthesis in a potential follow-up experiment. The WWS-words were matched for segmental distribution and syllable frequency in the following way: We controlled the distribution of consonants and vowels within each of the four nonce words, such that every vowel or consonant appeared only once in each word and never at the same position in another. The syllables used were under the most frequent 500 syllables in Samlowski, Möbius, and Wagner (2011) and their frequency across experimental nonce words was matched (the frequency of the three individual syllables was summed for every test word; this “overall frequency” of the test word was matched across test words (average “overall frequency” = 1.96 occurrences per billion (o.p.b, SD = 2.0 o.p.b.)). Four native speakers of German assigned grammatical gender on the basis of intuitive ratings: feminine (die) for Rumila and Nalomu, and masculine (der) for Linuro and Morani.50

49 The attrition rate is higher than in Experiment 1 (and its control studies). As all external conditions were identical, we cannot explain why this was the case other than for reasons of sampling.

50 Raters were presented with the nonce words in IPA on a sheet of paper. They were asked to indicate the grammatical gender that they felt most appropriate for the given nonce word. The result of the gender assessment was the following: Rumila: 4 x die (feminine); Linuro 3 x der (masculine), 1
The nonce words were embedded in the familiarization passages used in Experiment 1. The sentences were slightly modified to additionally control the number of syllables across the two different familiarization versions (e.g., by adding or deleting a syllable, see Appendix B (B.1) for all passages). The same speaker as in Experiment 1 recorded the 24 target sentences with the nonce words in two intonation conditions (six sentences for each passage). In the early-peak condition, the pitch peak preceded the stressed syllable (H+L*), resulting in a low-pitched final syllable of the WWS-carrier words (see Figure 13). In the medial-peak condition, the pitch peak was realized within the boundaries of the stressed syllable (L+H* for utterance-medial nonce words or H-^H% for utterance-final nonce words), resulting in a high-pitched final syllable of the WWS-carrier words (see Figure 13). The intonational realization was checked and words were re-recorded until the f0 excursion of the prominence-lending pitch movement (rise in the medial-peak condition and fall in the early-peak condition) and the duration of the nonce word were matched across intonation condition. A pre-selection of experimental materials was then entered into a rating study to ensure that the speaker correctly stressed the intended syllable, i.e., the third syllable in the WWS nonce words. To this end, eight German adult listeners trained in intonational phonology were asked to identify the metrically stressed syllable in all 48 pre-selected WWS-nonce words (and additional 48 WSW-fillers) on a button box with three keys. Overall the correctness was high (> 97%), but lowest in WWS-words with an early-peak accent (95%). Based on these ratings, two early-peak items that were error-prone were replaced and the final sample was considered suitable, for more details on the rating study see Appendix B (B.2). Table 8 summarizes the results of the acoustic analysis for the final set of WWS-words in two intonation conditions. The average f0 excursion was 10.0 st (SD = 0.9 st) in early-peak condition and 9.6 st (SD = 2.1 st) in the medial-peak condition. The average duration of the experimental WWS-words was 570 ms (SD = 54 ms) in the early-peak condition and 550 ms (SD = 51 ms) in the medial-peak condition. Loudness measures across intonation conditions differed slightly, but had a large overlap across intonation condition, see Table 8. As in Experiment 1, the sentences were concatenated with an ISI of 800 ms. The average duration of the whole familiarization passages was 15.6 s (SD = 0.3 s) in the early-peak condition and 15.2 s (SD = 0.2 s) in the medial-peak condition.

x das (neuter); NaNonu: 2 x die (feminine), 1 x das (neuter), 1 x der (masculine); Morani: 4 x der (masculine).

51 We used trained raters to be able to use the terms stress and intonation and be sure that they judged the intended concept.

52 We took more loudness measures in this experiment than in the previous ones, since here the critical condition is a peak-stress-misalignment condition (see below for a detailed discussion).
Figure 13: Example sound pressure wave, spectrogram, and pitch track of a target sentence in the early-peak condition (a) and in the medial-peak condition (b) in Experiment 2. Tier 1 gives the words in German, tier 2 the English translation.

Test

The test stimuli consisted of the last two syllables of the WWS-none words (e.g., mila from Rumila), but with stress on the initial syllable: [ˈmiː.la] for Rumila [ru.mi.ˈlaː], [ˈnuː.ro] for Linuro [li.nu.ˈroː], [ˈloː.mu] for Nalomu [na. lo.ˈmuː], [ˈraː.ni] for Morani [mo.ra.ˈniː]. The same speaker recorded approximately 20 tokens of each test word with falling intonation, 20 with rising intonation, and 20 with a monotonous intonation contour. For the experimental test lists, 15 tokens of each SW-test word were chosen (one third with falling, one third with rising, and one third with monotonous contours).

The average f0 excursion for the test tokens with falling intonation was 8.1 st (SD = 1.0 st), the average f0 excursion for the test tokens with rising intonation 11.7 st (SD = 1.3 st), and the mean f0 for the test tokens with monotonous intonation was 243.6 Hz (SD = 6.5 Hz). The average duration of the test words was 602 ms (SD = 61 ms), ranging from 479 to 710 ms; the first syllable of the test words was on average 299 ms (SD = 30 ms) and the second syllable 303 ms (SD = 55 ms) long, see Appendix B (B.3) for acoustic details on the individual lists. The 15 tokens of each test word were concatenated with an ISI of 800 ms. Across lists, the position in which the respective intonational realization occurred was kept constant. Every list started with two monotonous tokens. A maximum of two items with the same intonational realization occurred in a row. The lists were on average 20.1 s long (SD = 0.1 s).
Table 8: Acoustic realization (mean values (and standard deviations)) of WWS-target words in the familiarization phase for the two intonation conditions in Experiment 2.

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>Early-peak condition</th>
<th>Medial-peak condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of movement in st</td>
<td>10.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Duration of first syllable (unstressed) in ms</td>
<td>137</td>
<td>145</td>
</tr>
<tr>
<td>Duration of second syllable (unstressed) in ms</td>
<td>165</td>
<td>162</td>
</tr>
<tr>
<td>Duration of third syllable (stressed) in ms</td>
<td>267</td>
<td>243</td>
</tr>
<tr>
<td>Duration of onset consonant in second syllable (unstressed) in ms</td>
<td>70</td>
<td>68</td>
</tr>
<tr>
<td>Duration of onset consonant in third syllable (stressed) in ms</td>
<td>109</td>
<td>90</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of first vowel in dB</td>
<td>29.7</td>
<td>31.4</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of second vowel in dB</td>
<td>28.5</td>
<td>26.4</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of third vowel in dB</td>
<td>29.2</td>
<td>27.6</td>
</tr>
<tr>
<td>Mean intensity in first vowel in dB</td>
<td>73.6</td>
<td>70.1</td>
</tr>
<tr>
<td>Mean intensity in second vowel in dB</td>
<td>70.5</td>
<td>70.1</td>
</tr>
<tr>
<td>Mean intensity in third vowel in dB</td>
<td>70.4</td>
<td>71.2</td>
</tr>
<tr>
<td>RMS in first vowel in Pascal</td>
<td>0.099</td>
<td>0.072</td>
</tr>
<tr>
<td>RMS in first second in Pascal</td>
<td>0.070</td>
<td>0.068</td>
</tr>
<tr>
<td>RMS in first third vowel in Pascal</td>
<td>0.059</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Procedure

The experimental procedure was the same as in the previous experiments. Half of the infants were familiarized with passages containing the WWS-nonce words Rumila and Linuro, the other half with passages containing Nalomu and Morani. For test, infants listened to the three repetitions of the four test lists of the SW-words (12 trials), two familiar lists (syllable sequence was part of the experimental WWS-word heard during familiarization, e.g., [ˈmiː.la] for [ru.mi.ˈlaː]) and two novel lists (syllable sequence was not part of the experimental WWS-word heard during familiarization).
4.3.2 Results and interim discussion

As in the previous experiments, looking times in seconds were averaged by familiarity status (familiar vs. novel) for each infant. In the early-peak condition, infants looked on average 8.6 s (SD = 2.2 s) to familiar and 7.4 s (SD = 2.2 s) to novel test lists. Seventeen infants out of 24 infants showed a familiarity preference, i.e., a longer looking time to familiar test lists. In the medial-peak condition, infants looked on average 8.7 s (SD = 2.4 s) to familiar and 8.6 s (SD = 2.4 s) to novel test lists. Twelve infants out of 24 infants oriented longer familiar test lists. Table 9 shows the mean looking times to familiar and to novel items in both intonation conditions (first two columns).

Table 9: Overview of average looking times to novel (Nov) and familiar (Fam) test lists in Experiment 2. Standard deviations are indicated in brackets. The third row (Diff) displays the difference in looking time, along with the p-value as indicated by the post-hoc pairwise t-test.

<table>
<thead>
<tr>
<th></th>
<th>Exp. 2 Early-peak Varied test</th>
<th>Exp. 2 Medial-peak Varied test</th>
<th>Exp. 2b Early-peak resynthesized Varied test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov</td>
<td>7.40 s (2.18 s)</td>
<td>8.56 s (2.41 s)</td>
<td>8.33 s (3.56 s)</td>
</tr>
<tr>
<td>Fam</td>
<td>8.56 s (2.41 s)</td>
<td>8.66 s (2.36 s)</td>
<td>8.75 s (3.41 s)</td>
</tr>
<tr>
<td>Diff</td>
<td>-1.16 s p = 0.003</td>
<td>-0.10 s p = 0.78</td>
<td>0.42 s p = 0.33</td>
</tr>
</tbody>
</table>

A repeated measures ANOVA with intonation condition as between-subject factor and familiarity status as within-subject factor revealed a significant interaction between the two factors (F(1,46) = 4.52, p = 0.04). Post-hoc pairwise t-tests for the two intonation conditions separately indicated that the difference between looking times to familiar and novel test lists was significant in the early-peak condition (t(23) = 3.34, p = 0.003), with the alternative hypothesis being 14 times more likely than the null hypothesis (r = 0.71, bf = 14.13), but not in the medial-peak condition (p = 0.78) where the null hypothesis was five times more likely than the alternative hypothesis (r = 0.71, bf = 0.22). The 95% CI for the two intonation conditions show a small overlap between the two intervals only, suggesting a robust difference in the early-peak condition (zero not included), see Figure 14 (left panel).
Figure 14: Means of difference in looking time to novel and familiar items in Experiment 2. Whiskers represent the 95% CI of the difference in looking time. The grey dashed line indicates zero, i.e., no difference. Left panel: Experiment 2 (two intonation conditions); Right panel: Experiment 2b.

The recognition effect surfaced as a familiarity preference in this experiment, which is the reverse from the medial-peak conditions in Experiments 1 and 1b (where we find a novelty preference). Recall that the direction of the recognition effect has been argued to be driven by different factors such as age or task complexity (e.g., Butler et al., 2011; DePaolis et al., 2016; Houston-Price & Nakai, 2004; Hunter & Ames, 1988), with infants moving on a continuum from a familiarity preference to a novelty preference. These scholars have suggested that an increased duration of the familiarization phase (overall duration or number of trials) makes a novelty preference more likely. Conversely, an increased task difficulty, either due to complex stimuli or the experimental design, favours a familiarity effect. It has repeatedly been mentioned that a novelty preference is also more likely with an increasing age (most studies here refer to Hunter & Ames, 1988), but a recent meta-study demonstrates that age is not an adequate predictor for the direction of the effect (Bergmann & Cristià, 2016). In any case, infants in our experiments had the same age, making age a futile predictor for the shift in preference anyways. Instead, regarding task complexity, we argue that in the medial-peak conditions (Experiments 1 and 1b), the task was easier than in the current experiment (all potential stress cues coincided in Experiment 1 while infants could rely solely on f0 in Experiment 2). Conceivably, this difference in the availability of cues might have increased task complexity and hence resulted in a reversal of the preference (shift from a novelty preference to a familiarity preference).
Interim discussion

Our findings revealed looking time differences between familiarized and novel test words when the test words (last twoyllables of trisyllabic carrier words with reversed stress, e.g., [mi.ə] for [ru.mi.ˈIː]) had a HL pitch pattern in the familiarization phase. The early-peak accent on the WWS-word hence results in a metrical re-interpretation of the word-prosodic structure of the trisyllabic carrier word, with infants interpreting the embedded iambic sequence with an HL pitch pattern as a trochaic sequence (but not with an LH pitch pattern). Hence, high pitch even when realized on unstressed syllables (in form of an H-leading tone) leads to the percept of lexical stress in German infants. Our findings thus seem to invite the interpretation of a pitch-based rather than stress-based segmentation mechanism, with an f0 peak being a sufficient cue to stress, at least in German infants.

We need to bear in mind though that naturally produced stimuli were used in the current experiment, disallowing us to draw stronger conclusions about high f0 being a sufficient cue to stress for the moment. In natural productions, different pitch accent types cause shifts in the distribution of duration and intensity cues on the syllables adjacent to the stressed one. Specifically, the syllable with which the f0 peak is aligned has been shown to be inadvertently realized with higher intensity and longer duration (see "natural parallelism" of f0 and other cues in Kohler, 1991c, p. 144; for experimental evidence, Niebuhr, 2007a, pp. 117-150). The metrical re-interpretation of the word-prosodic structure in the early-peak condition in Experiment 2 might hence be due to peak confounding cues and not high f0 alone. A closer inspection of the acoustic analysis of the familiarization stimuli in both intonation conditions (Table 8) indicates that the acoustic cues other than f0 were closely matched across condition. The acoustic enhancement of the second syllable in the early-peak condition compared to the second syllable in the medial-peak condition seems minor (given the large overlap). Yet, in our stimuli the stressed syllable (last syllable) is clearly marked by intensity (louder than the preceding syllables) in the medial-peak condition, while intensity is in fact higher in the syllables preceding the stressed syllable than in the stressed syllable itself in the early-peak condition. Compared to Niebuhr (2007a, pp. 137-138), who reports a reduced intensity difference between pre-tonic and tonic syllable in early-peak accents (dB pretonic syllable minus dB tonic syllable, -1.4 dB) compared to medial and late-peak accents (dB pretonic syllable minus dB tonic syllable, -3.0 dB and -3.5 dB, respectively), the acoustic difference in the pre-tonic and tonic syllable in our early-peak accents is bigger (dB pretonic syllable minus dB tonic syllable, +0.1 dB), showing even more intensity in the pre-tonic syllable than in the tonic syllable. That way, the other stress cues might have been more ambiguous in the early-peak condition than in the medial-peak condition. As a result, the metrical re-interpretation might have come
about more easily in the early-peak condition compared to the medial-peak condition where the stressed syllable was clearly marked by both duration and intensity (note that this condition was also most error-prone in the rating study). It is thus not entirely clear which acoustic cues (f0 peak vs. other confounding acoustic cues) contributed to the metrical re-interpretation. Consequently, from a psycho-acoustic point of view, the use of resynthesized stimuli is inevitable to isolate the effect of the f0 peak in being a sufficient cue to the perception of stress (and simultaneously eliminate the concern of acoustic confounding). To this end, we re-run the early-peak condition of Experiment 2 under the same conditions but used resynthesized stimuli. Specifically, an early-peak contour was superimposed on the stimuli that were recorded in the medial-peak condition, using the PSOLA technique (Pitch Synchronous Overlap and Add; Moulines & Charpentier, 1990, for an overview on different resynthesis techniques) in Praat (Boersma & van Heuven, 2001). If the f0 peak alone is sufficient to trigger a percept of stress and thus a metrical re-interpretation of the word-prosodic structure of the word (WS-part with HL perceived as SW-unit), we expect to find the same looking time difference to familiarized and novel items as in the early-peak condition of Experiment 2.

4.3.3 Experiment 2b: Isolating the f0 cue

Methods
Participants
Another 24 full-term German monolingual infants who finished familiarization and all 12 test trials were included in the analyses (12 female, 12 male, average age: 0;9.3, age range: 0;8.18-0;9.19). Eleven further infants were tested but not included due to crying (1), fussiness (6), falling asleep (1), not attending to blinking lights (2) and an experimental error (1). Criteria for exclusion scores were the same as before.

Materials
Familiarization stimuli
Two of the four trisyllabic WWS-nonce words from Experiment 2 could be reused: Rumila [ru.mi.'la:] and Morani [mo.ra.'ni:]. Due to quality issues with the result of the resynthesis, two new words were chosen: WWS Nabuti [na.bi.'ti:] replaced Linuro and WWS Lisuka [li.su.'ka:] replaced Nalomu. Nabuti and Rumila were used

53 Note that resynthesized stimuli have been used in other studies investigating infants’ use of different acoustic parameters (Abboub et al., 2016; Bion et al., 2011; Frota et al., 2014; Sundara et al., 2015), consolidating the appropriateness of this type of stimuli for use in experimental research with young infants.

54 We noticed quality issues after piloting with PSOLA-resynthesis for two the four WWS-nonce words of Experiment 2. More specifically, while the resynthesized version of WWS-words Rumila and Morani generally displayed good sound quality, WWS Nalomu and Linuro drastically suffered from the manipulation in terms of overall quality. The deterioration in quality particularly affected
together in one familiarization group, as were Morani and Lisuka. Since the newly chosen nonce words occurred in different familiarizations, the same passage was used for them (the passage originally employed for Linuro). This facilitated the recording for the speaker and made the passages more comparable.

**Resynthesis procedure**

Generally speaking, the stimuli were PSOLA-resynthesized by superimposing the contour of a WWS-target in the early-peak condition on the target that was originally recorded with a medial-peak condition. Note that for the two new test words (Nabuti and Lisuka) recordings were made in both intonation conditions: in the medial-peak condition, which were used as basis for resynthesis, and in the early-peak condition, from which the reference points for resynthesis were calculated. Resynthesis was done as follows: We extracted the proportional alignment of low (L) and high tonal targets (H) in the nonce words originally recorded with an early-peak accent and used these proportions for resynthesis in the medial-peak files, see Figure 15, tier 3. For instance, if a given early-peak WWS-word reached the f0 peak after 70% of the second syllable, an f0 peak was created at 70% of the second syllable in the WWS-file that was originally recorded with a medial-peak accent. Regarding scaling, for the two target words that were used in Experiment 2 (Ramila and Morani), the original Hz-values occurring in the naturally recorded early-peak files were implemented in the medial-peak file. For the two new words (Nabuti and Lisuka), the average Hz-values of tonal targets from the early-peak files used in Experiment 2 were implemented in the medial-peak recordings (f0 peak = 341 Hz; final low tonal targets in stressed syllable \( L = 191 \) Hz). The range of this accentual movement was 10.0 st (341-191 Hz) and thus identical to the original recordings in the early-peak condition (Table 8). Figure 15 shows an exemplar realization of a natural early-peak accent and its resynthesized version. Tonal reference points used for manipulation are displayed on tier 3.

the round back vowel \([o]\). Depending on the position it occurred in, this vowel underwent changes of vowel quality, turning into \([a]\) in medial position (Nalomu became Nalunu after resynthesis) or turning into the lax version \([a]\) in final position (Linuro ended in a lax vowel \([a]\) after resynthesis) – an issue that could not be solved even by highly trained phoneticians. This vowel quality change after resynthesis would lead to a segmental mismatch between the two targets in the familiarization phase (Linuro, Nalunu) and its part-word versions used for test (nuro, lomu), which disqualified these words from the use in Experiment 2b. Further extensive piloting with the PSOLA-resynthesis revealed that the vowel \([a]\) in word-medial position and the vowels \([a]\) and \([i]\) in word-final position do not alter their vowel quality when undergoing resynthesis. We thus generated two new test words that included these robust vowels in the respective positions to be used in Experiment 2b.

\(^{55}\) Given the reduction in quality noticed while piloting for resynthesis, care was taken that the vowel quality was not changed in the pre-selected set of words during resynthesis. To this end, during the resynthesis procedure described above, the Praat script was paused to inspect every file auditorily after the script set the new pitch points automatically. If necessary, these points were manually adjusted to ensure the highest sound quality possible, which eventually led to a slight reduction of the f0 range of the accentual movement (reduction of about 1st), which, however, was still comparable to the f0 excursion in Experiment 1.
Crucially, the $f_0$ contour of the target word was the only factor that was changed by the resynthesis whereas all other acoustic parameters (intensity, duration and spectral information) remained unaltered. That way, potential confounding acoustic cues were inverted, see Appendix B (B.4) for acoustic analysis of the resynthesized stimuli. After resynthesis the sentences were concatenated with an ISI of 800 ms to form 4 passages, consisting of 6 sentences each. The passages were judged to sound natural by three phonetically trained members of the department.

**Test stimuli**

As in Experiment 2, the test words consisted of the last two syllables of the WWS-words but with reverse stress (e.g., mila [ˈmiː.la] for Rumila [ru.mi.ˈlaː]). The two trochees mila and rani (from Experiment 2) were re-used in Experiment 2b. Additionally, [ˈbuː.ti] for Nabuti [na.bu.ˈtiː] and [ˈsuː.kɑ] for Lisuka [li.su.ˈkaː] were used as test words. The same speaker recorded 20 tokens of rising, falling, and monotone realization of these two words. For the experimental test lists, 15 tokens of each SW-word were chosen (one third of falling, rising, monotonous contours each). The newly recorded words (buti and suka) closely matched the SW-units they replaced, see Appendix B (B.5) for acoustic analysis of the test lists. Lists were created in the same way as in Experiment 2 and concatenated with the same ISI (800 ms). The test lists had an average duration of 20.4 s (SD = 0.2 s).

**Procedure**

The procedure was the same as in Experiment 2. Half of the infants were familiarized with passages containing the WWS-nonce words Rumila and Nabuti, the other half with passages containing the WWS-nonce words Lisuka and Morani.
Results and interpretation

Infants looked 8.8 s (SD = 3.4 s) to familiar and 8.3 s (SD = 3.6 s) to novel test lists, see Table 9 for looking times compared to Experiment 2. Thirteen infants out of 24 looked longer to the familiar lists. A pairwise t-test indicated that the difference in looking times was not significant (t(23) = 0.99, p = 0.33). The null hypothesis was three times more likely than the alternative hypothesis (r = 0.71, bf = 0.33). Hence, in contrast to the naturally recorded early-peak condition in Experiment 2, infants did not show a recognition effect for the embedded iamb presented as trochee in the test phase in the current experiment. Different from what was found in Experiment 2, a repeated measures ANOVA with familiarity status as within-subjects factor and intonation condition as between-subjects factor for the pooled data of Experiment 2b (resynthesized early-peak condition) and Experiment 2 (medial-peak) revealed that the interaction between the two factors was not significant (F(1,46) = 0.34, p = 0.56). Comparing the two early-peak conditions (natural vs. resynthesized) in a pooled analysis also revealed a non-significant interaction between with familiarity status as within-subjects factor and intonation condition as between-subjects (F(1,46) = 1.74, p = 0.19). Hence, the effect of intonation condition in the resynthesized early-peak condition statistically differs neither from the effect in the early-peak condition nor the null-effect in the medial-peak condition.

Figure 14 (right panel) shows the 95% CIs of the looking time differences for Experiment 2b, in comparison to the two natural intonation conditions in Experiment 2 (left panel): It becomes obvious that the 95% CI patterns more with the medial-peak condition than with the early-peak condition, speaking against high f0 as a sufficient cue to stress.

Experiment 2b set out to substantiate the claim that high f0 might be a sufficient cue to stress for German infants in metrical segmentation, using resynthesized stimuli. The recognition effect of embedded iambs, presented as trochees in the test phase was not replicated in Experiment 2b, however. How can we explain this difference in results? Methodological explanations seem least likely to us: First, in Experiment 2b, we used the same procedure under the same conditions as in Experiment 2 (except for resynthesized stimuli). Second, judging resynthesized stimuli as inappropriate for the use with infants per se, also appears an inappropriate explanation, given that resynthesized stimuli have been successfully used before in infant research (Abboub et al., 2016; Bion et al., 2011; Frota et al., 2014; Sundara et al., 2015), and quality of the stimuli was carefully checked. Also, the slightly reduced f0 range in resynthesized early-peak accents (9.0 st, SD = 1.2 st) compared to natural early-peak accents (10.0 st, SD = 0.9 st) due the quality constraints (see above for explanation) cannot serve as a potential explanation: In Experiment 1 recognition was observed in a condition where the f0 range is comparable (8.8 st, SD = 1.7 st) to the f0 range in Experiment 2b. Finally, the two new test words used in
Experiment 2b (Nabuti, Lisuka instead of Linuro, Morani) are unlikely to have increased the task difficulty. On the contrary, the embedded iambs presented as trochees in the test phase are arguably more saliently marked by the plosive and fricative respectively (buti, suka) than by the sonorant consonants used as word-onsets in Experiment 2, and infants have been shown to find it easier to segment words with word-initial consonants than less clearly marked word onsets (Kim & Sundara, 2015; Mattys & Jusczyk, 2001a; Nazzi et al., 2005; Seidl & Johnson, 2008).

Instead, we take the finding to mean that the resynthesized isolated f0 cue did not signal stress for infants. Unlike the naturally produced early-peak accents in Experiment 2, the isolated f0 cue in resynthesized stimuli did not lead to the re-interpretation of the metrical structure (WS to SW) in Experiment 2b. Consequently, high f0 cannot be considered a sufficient cue to stress in German infants when segmenting fluent speech. Only naturally produced early-peak accents were able to trigger a shift in the perception of the given word-prosodic structure. As has been discussed in detail above, naturally produced accents that involve a misalignment between the f0 peak and the stressed syllable are arguably more ambiguous in terms of stress cues, compared to resynthesized ones, given the shift in intensity (and duration) on syllables adjacent to the stressed one ("natural parallelism" of f0 and other cues in Kohler, 1991c, p. 144; Niebuhr, 2007a, pp. 137-138, and acoustic analysis of early-peak stimuli in Experiment 2, see Table 7). This "natural parallelism" is very likely the result of the articulatory process, with the acoustic shifts across syllables being unavoidable. Preliminary data from an imitation experiment in which adult speakers of English (BE and AusE) imitated resynthesized pitch accent types corroborate this view by showing that the syllable with the f0 peak was acoustically enhanced – even when the input signal told otherwise (James & Zahner, Accepted). What infants encounter in "real life" though are naturally produced pitch accent types. Hence, our finding – even if the isolated cue was not able to evoke the change in the interpretation of the word-prosodic structure – is astonishing in any case. It offers a novel perspective to account for early speech segmentation, suggesting that pitch accent type might be the driving force in the perception of stress for infant listeners, at least in German.

4.4 Discussion

Taken together, in Experiment 1, we showed that German 9-month-olds' stress-based segmentation behaviour is guided by pitch accent type. In the medial-peak condition where the f0 peak was aligned with lexical stress, infants extracted the SW-unit from trisyllabic WSW-carrier words, but they failed in the early-peak condition and in the late-peak condition. We further demonstrated that the extraction of embedded trochees in the medial-peak condition stems from the alignment of
pitch peak and metrical stress in the medial-peak condition alone and not from the alternation of tonal targets of opposite height, corroborating that high f0 is a necessary cue to the perception of metrical stress in stress-based segmentation. In Experiment 2, we tested whether high f0 is also a sufficient cue for stressed syllables to be perceived as stressed: While in the experiment with natural stimuli the early-peak accent on the WWS-word led to a re-interpretation of the last two syllables, with infants showing recognition for a test unit that was reversed in stress pattern compared to the familiarization phase, this was not the case for the experiment with resynthesized stimuli. Hence, high f0 (together with peak supporting cues) guides infants’ stress-based segmentation behaviour. Importantly, our recognition effects of trochees (Experiments 1 and 1b (WSW, medial-peak condition with both falling and rising test intonation), and 2 (WWS, early-peak condition naturally produced)) were robust against variation of intonation between familiarization and test, i.e., in case of successful segmentation infants recognized the units irrespective of the intonation contour they displayed during test, emphasizing infants’ ability to abstract away from intonational variation.

Infants’ ability to generalize ties in with previous studies showing that infants’ representations become more abstract towards the end of the first year of life (Bortfeld & Morgan, 2010; Houston & Jusczyk, 2000; Singh et al., 2004; Singh et al., 2008). Our data extend these previous findings by showing that infants do not only abstract over different voices and pitch levels (Houston & Jusczyk, 2000; Singh et al., 2008), but also over different pitch contours (falling vs. rising). It seems that 9-month-olds are aware of the fact that pitch is not lexically contrastive in German and they consequently do not store pitch together with the segmental form of the extracted units. This might be interpreted as a move away from episodic storage, which is observed in very young infants. Considering infants’ ability to abstract over intonational patterns and their strong reliance on high-pitched stressed syllables, it becomes clear that pitch plays different roles in the segmentation (extraction of units in familiarization) and recognition (of these units in test) process, respectively. From the current data it appears that, in a first step, high f0 seems to be a necessary (and given natural productions also a sufficient) cue in the perception of stress and consequently needs to be present in order to extract units in fluent speech. In a second step, when the task demands the recognition of the previously embedded SW-units, f0 is no longer considered relevant in the comparison of stored forms to the input. Once infants have established a (be it only temporary) representation of the extracted sound sequence, they seem to generalize over lexically non-contrastive pitch contours. A study by Vihman, Nakai, DePaolis, and Hallé (2004) similarly reports that prosodic cues play a minor role in word recognition at this age: In an experiment with English-learning 11-month-olds, they found that the recognition of (untrained) familiar words was delayed but not inhibited through the misplacement
of stress, whereas segmental mismatches, particularly the mispronunciation of the initial consonant, hampered word recognition. While infants have learned to neglect information on pitch in recognition processes, high f0 seems to be necessary and in natural production also sufficient in German 9-month-olds’ stress perception, thus playing an important role in segmentation processes.

Regarding stress-based segmentation, the present study replicates the findings of previous research that indicated that infants exposed to stress-timed languages are able to extract SW-units from fluent speech (see Table 3), strengthening the crucial role of stressed syllables for segmentation in Germanic languages. Yet, our findings extend these earlier studies that used non-embedded trochees (see Table 3) by showing that German 9-month-olds can segment SW-syllable sequences from fluent speech even if they are embedded in WSW-words (or WWS-words perceived as WSW-words) and thus provide misleading linguistic, phonetic and statistical cues to word onsets (see also Echols et al., 1997, for American English infants in a different paradigm; and Jusczyk, Houston, et al., 1999, in a similar set-up). On the contrary, these cues all hint towards a different word boundary, the boundary of the trisyllabic carriers. Possibly, when infants are able to extract iambic patterns (cf. 10.5 months in Jusczyk, Houston, et al., 1999) and integrate suprasegmental and sequential cues more robustly (Morgan & Saffran, 1995), the extraction of embedded SW-units would become more difficult or impossible (comparable to the difficulty to activate embedded words, such as date from sedate, see Norris, Cutler, McQueen, & Butterfield, 2006). Critics may argue that investigating the ability to extract embedded units effectively tests mis-segmentation, i.e., units not corresponding to real words. We will take up this aspect in detail in the General discussion and argue that the mechanism we identified does not harm real-world segmentation (7.2.2). Third, and most importantly, our findings reveal that utterance-level intonation in form of different pitch accent types modulates German infants’ stress-based segmentation behaviour. That way, we extend earlier research on accentedness and segmentation in German (Männel & Friederici, 2013) as well studies that highlight the importance of exaggerated f0 contours for segmentation (Floccia et al., 2016; Schreiner & Mani, 2017; Thiessen et al., 2005), see a more detailed discussion of our findings in relation to these studies in 7.2.2.

From an (adult) perspective of intonational meaning, one might argue that the manipulation of intonation condition we employed in the current series of experiments was information-structurally not fully appropriate\(^{56}\) and hence placed additional emphasis on the targets (Titia Benders, Pilar Prieto, p.c.). Given that infants have been shown to understand communicative intentions based on prosody

\(^{56}\) Recall that all targets were produced with the same accent type within one familiarization passage even though one would expect medial-peaks in the beginning and early-peaks and de-accentuation in subsequent mentions (cf. Baumann & Grice, 2006; Kohler, 1991c, see Chapter 2 for details).
(and gestures) within the first year of life (for an overview, see Esteve-Gibert & Prieto, 2018; see also Esteve-Gibert, Prieto, & Liszkowski, 2017; Liszkowski, Carpenter, & Tomasello, 2008) and display robust understanding of the category newness at the age of 2 years (e.g., Grassmann, Stracke, & Tomasello, 2009; Grassmann & Tomasello, 2007, 2010), this argument has to be taken. Yet, a recent corpus study on German shows that IDS does not automatically trigger the use of a pragmatically expected pitch accent types, such as early-peak accents for repeated referents (Zahner, Schönhuber, et al., 2016b). This finding is in line with Grünloh, Tomasello, and Lieven (2015), who found that when talking to children, adults did not de-accent given referents as consistently as when talking to adults, but instead used more H*-accents for given referents. The consistent use of medial peaks within one passage might hence be not as inappropriate in IDS as it might seem from an ADS perspective. From the infant’s perspective, it is rather early-peak accents (which only occur in 6% of the cases in German IDS, Zahner, Schönhuber, et al., 2016b) that need to be considered “exceptional.” Yet no recognition of trochees was observed in this special, unexpected condition, which weakens the argument of pragmatic inappropriateness. Future studies need to examine how many accent types of one kind are necessary to trigger successful segmentation.

By using different accent types, our findings contribute to the picture of which cues lead to the percept of a stressed syllable in infants. In this series of experiments, we showed that the f0 peak is a necessary cue to the perception of stress in metrical segmentation in German infants. An f0 peak, when produced with natural peak supporting cues, even seems to be a sufficient cue to stress: F0 can shift the interpretation of relations of metrical strength in a word (at least in naturally produced pitch accents). Why is high f0 such important? Why does its presence or absence determine whether a syllable is taken as stressed word onset in infant stress-based segmentation? Currently, we see two underlying mechanisms that might make f0 a cue to stress: First (Option 1), infants might pick up on high-pitched syllables, which stand out perceptually (Cho, 2005; Hsu et al., 2015; Lieberman, 1967). Given that infants are sensitive to and attracted by f0 from birth (e.g., R. P. Cooper & Aslin, 1990; Nazzi, Floccia, et al., 1998), this option seems plausible. Another factor (Option 2) may be the frequency of occurrence of different pitch accent types: A frequent occurrence of H*-accents in the input could trigger the expectation in infants that stressed syllable necessarily need to be high-pitched. At this point, both explanations (Options 1 and 2) are likely. We will come back to the potential underlying factors in the processing of f0 as a stress cue in Chapter 6. We will now turn to the effect of pitch accent type on the perception of stress and consequently on lexical activation in German adults (Chapter 5).
5 Pitch accent type affects lexical activation in German adults

Chapter outline
In intonation languages, pitch accents are associated with stressed syllables, so that accentuation is a sufficient cue to the position of stress. Acoustically, pitch accent types differ in terms of the alignment of the f0 peak in regard to the position of metrical stress. This chapter investigates how different pitch accent types modulate the online processing of metrical stress in German. In a Visual-World Eye-Tracking study (Experiment 3), listeners were presented with four words on screen, a cohort pair that shared at least the segments of the first syllable but differed in the position of the stressed syllable, along with two unrelated distractors. In experimental trials, WSW-target words were presented either with resynthesized early-peak accents (H+L*, f0 peak preceding the stressed syllable) or resynthesized medial-peak accents (L+H*, f0 peak on the stressed syllable) and fixations to the words on screen were monitored while listeners followed instructions (“Bitte klicke <Target> an” ‘Click on <Target>, please’). Results showed that during the processing of the segmentally ambiguous part listeners fixated the stress competitor (SWW) more when the target word (WSW) was realized with an early-peak accent, compared to a medial-peak accent. Conversely, the target received fewer fixations. Our findings indicate that high-pitched unstressed syllables are temporarily interpreted as stressed. That way, we show that intonation directly affects lexical activation in German adults.

57 Adapted parts of this chapter have been published in Zahner, Kutscheid, and Braun (2019).
5.1 Introduction

This chapter deals with the processing of lexical stress cues and the question of how utterance-level intonation (mainly f0) modulates the perception of lexical stress in German adults. Stress cues differ in how reliably they signal the stressed syllable. In perception, the phonological feature accentuation has been shown to be an unambiguous cue to the position of metrical stress for adults – irrespective of whether it is a high accent (H*) or a low accent (L*) – since in intonation languages pitch accents only associate with stressed syllables but not with unstressed ones (Ladd, 2008). For instance, Shattuck-Hufnagel, Ostendorf, and Ross (1994) demonstrated in a labeling study with trained annotators that stress shift, i.e., perceived initial prominence on multisyllabic words, such as Massachusetts, occurs if the first syllable is pitch-accented (predominantly with an f0 rise), but less often when there was a small rise or a fall only (Shattuck-Hufnagel et al., 1994, p. 371). Dilley and Heffner (2013) also provide evidence for accentuation as a cue to stress. They investigated whether alignment differences in f0 peaks and valleys shift the perception of the accent type category, e.g., from H* to H+L*. Specifically, listeners heard words and nonce words with primary stress on the first syllable and secondary stress on the final syllable, and one or two unstressed syllables in-between, e.g., millionaire (S1WS3), lemonade (S1WS2), lannameraine (S1WWS3); indexing refers to primary (S1) and secondary stress (S2). Millionaire and lannameraine were realized with an H* on S1 and ended in a low boundary tone; lemonade had an L* on S1 and a high final boundary tone. For millionaire and lannameraine, the f0 peak was shifted to the right in various steps so that it ended up on the unstressed syllable; for lemonade, the f0 valley was shifted to the right in various steps so that it was ultimately realized on the unstressed syllable preceding the secondary stressed syllable. Listeners had to indicate whether the first or the last syllable carried primary stress. The results showed that the later the f0 peak and the f0 valley, the more frequently listeners judged stress to be on the last syllable of the word. Dilley and Heffner (2013) interpret this shift in perception as a shift from an H*-accent associated with S1 to an H+L*-accent associated with S2 for the millionaire/lannameraine-series and from an L*-accent associated with S1 to an L+H*-accent associated with S2 in the lemonade-series. These findings indicate that the alignment of f0 peaks and f0 valleys alter the perception of pitch accent type and in turn the interpretation of the position of primary stress. This hence demonstrates the relevance of the phonological cue accentuation for the identification of stress in adult listeners.

The position of high f0, as has been discussed in detail in Chapter 2, is a highly unreliable cue with respect to the position of the stressed syllable. Yet, as has been shown in 3.1 (review of studies providing evidence for high f0 as a stress cue) and in Chapter 4 (experiments with infants), high pitch generally appears to cue lexical
stress in perception despite the ambivalent role of f0 as a stress correlate. For adults, various decision tasks manipulating different correlates of stress in a step-wise fashion have demonstrated that adult listeners of West Germanic languages strongly attend to pitch information for lexical stress perception, irrespective of whether it is rising f0 movements, high f0 levels or f0 peaks (cf. Fry, 1958; Isačenko & Schädlich, 1966; Kohler, 2008; Niebuhr & Winkler, 2017). In all these studies, listeners used f0 information to decide between stress minimal pairs, e.g., the English noun object and verb object, or pairs differing in primary vs. secondary stress, e.g., übersetzen [ˈyːbɐ.ˈzɛtn] (‘to translate’) and übersetzen [ˈyːbɐ.ˌzɛtn] (‘to cross over/ferry over’). Furthermore, recent offline stress judgement tasks show that listeners are worse at indicating the stressed syllable in a target word when intonation varies (rising vs. falling contours in Schwab & Dellwo, 2017) and when the stressed syllable is low-pitched (Egger, 2015).

In sum, these studies indicate that different kinds of high pitch (f0 rises, high f0 levels, and f0 peaks) function as a perceptual cue to metrical strength even though high pitch is unreliable – given utterance-level intonation. However, these studies leave open a number of questions. First, they do not answer the question of how f0-alignment differences as they occur in naturally produced pitch accents affect the perception of stress. Second, it is an open question whether f0 plays an equally important role in words that are not a member of a stress minimal pair. To control for segmental effects, previous studies used minimal pairs (Dilley & Heffner, 2013; Fry, 1958; Isačenko & Schädlich, 1966; Niebuhr & Winkler, 2017) with two separate lexical entries in the mental lexicon (one for each meaning), or nonce words (Kohler, 2008), which lack lexical stress specifications altogether. However, these stress minimal pairs and nonce words are special and may not reflect well how f0 is used in stress perception: Stress minimal pairs are very rare in German non-compounded words, so one is left with minimal pairs in which the members have very different lexical frequencies or with compounds that differ in primary and secondary stress. The asymmetry in lexical frequency, for instance, may lead to strategic responses on part of the listeners as stress placement has been shown to differ in high vs. low frequency words (see e.g., Cutler & Carter, 1987, who report that high frequency words show a greater proportion of initial stress than low frequency words). Moreover, it cannot be ruled out that stress minimal pairs are processed in a special way altogether.58

A recent offline stress judgement study goes beyond the decision tasks reported on above by using target words that only occur with one stress pattern (see

58 Recall that similar methodological criticism applies to studies investigating the acoustic correlates of lexical stress in production (see Dogil & Williams, 1999, Chapter 2).
Experiments 1a and 1b in Zahner et al., 2019, based on Egger, 2015). Furthermore, intonation contours modelled after a naturally produced pitch accent contrast were used in that stress judgement study. Listeners indicated the position of lexical stress in auditorily presented trisyllabic WSW German nouns on a button box with three keys (for stress on the first, the second, or the third syllable). Experimental items (e.g., *Tornado* [tɔr.nəˈdoː] ‘tornado’) were always stressed on the second syllable (fillers had different stress patterns to avoid strategic effects). Crucially, experimental items were presented with three different pitch accent types: PSOLA-resynthesized early-peak accents (H+L* L-%), PSOLA-resynthesized medial-peak accents (L+H* L-%), and PSOLA-resynthesized L*-accents followed by a high boundary tone, i.e., late-peak contours (L* H^-H%). Different resynthesis procedures were used in two different experiments to (a) safeguard against effects of resynthesis quality and (b) to isolate the effect of f0 alignment from other possible confounding acoustic cues (Kohler, 1991c; Niebuhr, 2007a). Results of the stress judgements indicated that German listeners’ stress perception is hampered when f0 peaks and stressed syllables were misaligned. Irrespective of the manipulation procedure, the results showed significantly fewer correct responses in the misalignment conditions, i.e., early- and late-peak contours (on average 59% and 64% correct responses, respectively), compared to the medial-peak condition (on average 75% correct responses). Crucially, erroneous responses (i.e., cases in which the first or third syllable was judged as the stressed one in the WSW-word) revealed a strong response bias towards the syllable with which the f0 peak was aligned, see Figure 16. Specifically, for syllable-1-response errors, 49% occurred in the early-peak condition, i.e., the misalignment condition with the f0 peak on the first syllable (25% and 26% in the medial-peak and late-peak condition, respectively; $\chi^2 = 70.1$, df = 2, $p < 0.0001$). For syllable-3-response errors, 57% occurred in the late-peak condition, i.e., the misalignment condition with the f0 peak on the third syllable (20% and 23% in the medial-peak and the early-peak condition, respectively; $\chi^2 = 65.7$, df = 2, $p < 0.0001$). These patterns of results demonstrate that the syllable with the f0 peak was interpreted as the stressed one even though the pitch accent was phonologically associated with the stressed syllable in all conditions.

---

59 Note that one of the offline stress judgement experiments (Experiment 1b) in Zahner, Kutscheid, and Braun (2019) was first published in Egger (2015), the master thesis by Sophie Kutscheid (née Egger), supervised by Prof. Dr. Bettina Braun. Sophie Kutscheid contributed the idea of studying the role of intonation in stress judgements. I would like to truly acknowledge her efforts.
In the current eye-tracking study, we extend the offline stress judgement task by Zahner et al. (2019) described above in two essential ways: First, and foremost, we address the online processing mode and investigate whether and how different pitch accent types affect stress processing during word recognition, as the utterance unfolds over time – a question that has not been studied before. Second, by using a task that directly taps into lexical activation (Tanenhaus, Magnuson, Dahan, & Chambers, 2000), we can be sure that we address the perception of lexical stress (in contrast to the perception of phrase-level prominence). In the outlined offline stress judgement task that required a conscious meta-linguistic decision it cannot fully be controlled whether or not participants judged the acoustically most prominent syllable in the word – despite precise instructions that explicitly asked for the position of the stressed syllable (betonte Silbe), as in Experiments 1a and 1b in Zahner et al. (2019). This methodological drawback applies to all offline studies but is avoided with the eye-tracking paradigm used in the current study (or any other paradigm tapping into lexical access).

In 3.2, we demonstrated that adult listeners use stress and intonational cues to efficiently recognize words in speech. Regarding the time course of recognition, a number of previous studies have shown that suprasegmental stress cues are immediately exploited by listeners of intonation languages to distinguish between stress competitors (Connell et al., 2018; Jesse et al., 2017; Reinisch et al., 2010). Going beyond these studies, we manipulate pitch accent type to study the lexical activation of stress competitors. In light of the reviewed background, we hypothesize that f0 peaks that are not aligned with the stressed syllable (in early-peak accents) will hamper lexical activation in German because f0 peaks have been shown to cue lexical stress. Specifically, f0 peaks realized on unstressed syllables are predicted to lead to
the temporary activation of lexical competitors with a different stress pattern in German, such that a WSW-word realized with an early-peak (H+L*) will more strongly activate an SWW-competitor than a WSW-word realized with a medial-peak accent (L+H*). As a result of increased competitor activation, the activation of the target is likely to be decreased as a function of pitch accent type.

5.2 Experiment 3

5.2.1 Methods

Participants
Forty-eight German native speakers (39 female, 9 male, average age = 22.5 years, SD = 3.2 years, 28 right eye-dominant) participated for a small fee. They had normal or corrected-to-normal vision and unimpaired hearing. Most of the participants grew up in Southern Germany (Baden-Wuerttemberg or Bavaria, 75%) and were students or staff at the University of Konstanz (recruited via the online platform SONA). None of them had participated in related studies. Data of four additional participants could not be used due to calibration difficulties (due to mirroring issues with glasses).

Materials
Sixty-four trisyllabic cohort pairs were selected: One member of each pair was stressed on the first syllable (SWW, *Libero* ['li baiso] ‘sweeper’), the other member on the second syllable (WSW, *Libelle* ['li ba.lə] ‘dragonfly’). The cohort pairs were segmentally identical up to at least the first consonant of the second syllable. Thirty-two of the 64 cohort pairs were used for cohort trials. Sixteen of the cohort trials were experimental trials (WSW-word as the auditory target and SWW-word as the stress competitor), 16 were distractor trials (SWW as the auditory target and WSW as the stress competitor). The remaining 32 cohort pairs were used as filler trials in which one of the unrelated items served as the auditory target. Appendix C shows a list of quadruplets used in screens for cohort (C.1) and filler trials (C.2).

Cohort members were matched for lexical frequency and number of characters across groups (Dahan & Gaskell, 2007; Lavidor, Ellis, Shillcock, & Bland, 2001; New, Ferrand, Pallier, & Brysbaert, 2006). In cohort trials, the cohort member with initial stress (SWW) had on average 6.5 characters (SD = 1.1) and a lexical frequency of 93 o.p.m. (SD = 99, according to dllexDB (Heister et al., 2011)), the cohort pair with stress on the second syllable had on average 6.9 characters (SD = 1.0) and a lexical frequency of 69 o.p.m. (SD = 119). In filler trials, the SWW cohort member had on average 7.0 characters (SD = 1.3) and a lexical frequency of 201 o.p.m. (SD = 479); the WSW-cohort member had on average 7.0 characters (SD = 1.1) and a
lexical frequency of 115 o.p.m. (SD = 283). For each cohort pair, we selected two
distractors that were semantically and phonologically unrelated to the cohort mem-
bers and which had a comparable length (on average 7.0 characters, SD = 0.9) and
lexical frequency (on average 94 o.p.m., SD = 79). The trisyllabic distractors were
stressed on the first, the second, or the third syllable to increase variability of stress
patterns on screen (20 SWW, 20 WSW, and 24 WWS words in total). Appendix C
(C.3) displays the lexical frequency and number of characters for cohort pairs used
in cohort trials.

A female native speaker of Standard German (aged 25 at the time of recording,
born in the North of Germany, grown up partly in the North and South of Ger-
many), who was trained in intonational phonology, recorded the auditory stimuli
for the eye-racking study, i.e., the instructions “Bitte klicke <Target> an” ‘Please
click on <Target>’, in a sound-attenuated cabin at the Phonlab at the University of
Konstanz (at 44.1 kHz, 16 Bit, mono). Note that this was the same speaker as in
Experiments 1a and 1b in Zahner et al. (2019) but a different one than in the infant
experiments reported on in Chapter 4. The instructions for trials referring to one
of the cohort members were produced in two intonation conditions each: with a
medial-peak accent (L+H*) and an early-peak accent (H+L*) on the auditory target
word; all other words in the instruction were unaccented. The final boundary tone
was always low (L-%). The sentences were re-recorded until we were able to select
pairs that did not differ other than in the relevant acoustic properties. Table 10
displays the results of an acoustic analysis of targets used in experimental trials. For
fillers, half of the sentences were recorded with a medial-peak (L+H*), half with an
early-peak accent (H+L*) on the target, matching the f0 range of their accentual
movement with the f0 range of cohort pairs (rise for medial-peak accents, fall for
early-peak accents).

To avoid effects of distal prosodic context (Brown et al., 2015), all trials were
spliced after the verb (“Bitte klicke | | <Target> an”). For experimental trials
(WSW-word as the auditory target), four different versions of the pre-context “Bitte
klicke” were used to avoid a mismatch between coarticulatory information in the
last syllable of the verb and the onset consonant of the target word (one for WSW-
words starting in [a], [e], [m], respectively, and one for words starting in another
consonant). For each target, the pre-context was the same across intonation condi-
tion.\(^{60}\) On average, the pre-contexts were 575 ms (SD = 25 ms) long. Overall, the
cross-spliced stimuli sounded natural and the splicing was not noticeable, as

\(^{60}\) Note that the four different versions of the pre-context could be used for SWW-trials with an
early-peak accent, i.e., in half of the distractor trials (SWW, early-peak condition) and one third of
the filler trials (SWW, early-peak condition), since in these words the f0 peak had to be aligned with
the last syllable of the pre-context. Hence, an early-peak version of pre-context *Bitte klicke* with a
high-pitched last syllable was used for these trials, again in four different versions depending on
consonantal context of the target word.
indicated in a post-hoc naturalness assessment. In that post-hoc test, four staff members of the linguistics department at the University of Konstanz judged the naturalness of the stimuli. They were presented with ten randomly selected experimental trials, five in each condition, and instructed to indicate whether or not they could detect any kind of manipulation in the signal (the term splicing, or an explanation of it, was not used in the instruction). In total, 95% of the stimuli were judged as not manipulated (the remaining 5% were comments that were unrelated to splicing but speaker peculiarities), corroborating the subtleness of the splicing procedure.

Table 10: Acoustic realization (means (and standard deviations)) of WSW-targets in two intonation conditions (naturally recorded, before PSOLA-resynthesis).

<table>
<thead>
<tr>
<th>Acoustic variable</th>
<th>Medial-peak condition&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Early-peak condition&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(natural)</td>
<td>(natural)</td>
</tr>
<tr>
<td>F0 excursion of accentual movement (st)</td>
<td>Rise: 8.36 (0.60)</td>
<td>Fall: 8.43 (0.67)</td>
</tr>
<tr>
<td>Duration first syllable (ms)</td>
<td>143 (34)</td>
<td>146 (34)</td>
</tr>
<tr>
<td>Duration second syllable (ms)</td>
<td>214 (48)</td>
<td>226 (48)</td>
</tr>
<tr>
<td>Intensity middle of first vowel (dB)</td>
<td>71.7 (2.3)</td>
<td>72.4 (3.3)</td>
</tr>
<tr>
<td>Intensity middle of second vowel (dB)</td>
<td>71.9 (2.5)</td>
<td>71.9 (1.8)</td>
</tr>
<tr>
<td>Mean RMS amplitude first vowel (Pa)</td>
<td>0.075 (0.017)</td>
<td>0.085 (0.026)</td>
</tr>
<tr>
<td>Mean RMS amplitude second vowel (Pa)</td>
<td>0.085 (0.019)</td>
<td>0.081 (0.013)</td>
</tr>
<tr>
<td>H1*-A3&lt;sup&gt;c&lt;/sup&gt; ratio middle first vowel (dB)</td>
<td>27.0 (10.8)</td>
<td>23.2 (10.7)</td>
</tr>
<tr>
<td>H1*-A3&lt;sup&gt;c&lt;/sup&gt; ratio middle second vowel (dB)</td>
<td>30.8 (7.2)</td>
<td>31.8 (6.3)</td>
</tr>
</tbody>
</table>

<sup>a</sup> to be resynthesized to early-peak accents  
<sup>b</sup> to be resynthesized to medial-peak accents  
<sup>c</sup> Following Mooshammer (2010), we used the H1*-A3* ratio as a measure for vocal effort, i.e. the difference between the amplitude of the first harmonic and the third formant (asterisks denote that amplitudes were corrected for formants). In order to perform these acoustic measurements, we adjusted a Praat script downloaded from http://www.seas.ucla.edu/spapl/voicesauce/, last accessed: 14.03.2018.

After splicing, the stimuli were PSOLA-resynthesized by superimposing the contour of a target word in one intonation condition to the target that was originally recorded in the other intonation condition. Thus, the f0 contours of medial-peak accents (L+H*) were superimposed on target words originally recorded with an early-peak accent (H+L*) and vice versa. Figure 17 shows the PSOLA-resynthesized and natural versions of one item in the two intonation conditions. Note that (a) and (b) display the natural and (c) and (d) the PSOLA-resynthesized versions in Figure 17.
Figure 17: Example sound pressure wave, spectrogram, and pitch track for an experimental trial (WSW-target, e.g., *Libelle*). Figure (a) and (b) show the natural productions (used as basis for PSOLA-resynthesis) in the early-peak condition and medial-peak condition, respectively. Figure (c) and (d) are the PSOLA-resynthesized versions of these contours used as experimental materials. Tier 1 gives the words in German, tier 2 the English translation; tier 3 displays the tonal reference points, L and H, used for resynthesis; tier 4 displays the GToBI labels.

Resynthesis for experimental and distractor trials was done as follows (identical to the procedure described for Experiment 2b in Chapter 4): We extracted the scaling and proportional alignment of low (L) and high turning points (H) in the target word and transplanted these values on the recording in the other condition (PSOLA-resynthesis). For instance, if a WSW-target that was recorded with an early-peak accent (H+L*) reached an f0 peak of 315 Hz after 90% of the first syllable, an f0 peak with 315 Hz was created at 90% of the first syllable in the file that was originally recorded with a medial-peak accent (L+H*), see tonal targets (L, H) on tier 3 in Figure 17. For filler trials, we used the resynthesis procedure employed in Experiment 1 in Zahner et al. (2019) as there were no direct lexical competitors and fillers were recorded in one intonation condition only. That is, we calculated the average f0 maximum (314 Hz) and the mean f0 in unaccented syllables in the target words (193 Hz) in the original recordings of the filler items in both intonation conditions, resulting in an average f0 range of the accentual movement of 8.4 st. These values were then used to create a medial-peak accent in the 16 fillers that were originally recorded with an early-peak accent and vice versa to create an early-peak accent in the remaining 16 fillers that were originally recorded with a medial-peak accent. To this end, we placed the tonal targets in the middle of the respective vowels.
Procedure

The experiment consisted of 64 trials in total: 32 cohort trials (16 experimental, 16 distractor trials) and 32 filler trials. Note that experimental trials are of interest for the current hypothesis (more looks to the SWW-competitor when the WSW-target is realized with an early-peak accent compared to a medial-peak accent). Distractor and filler trials served a strategic function only, protecting against an imbalance in clicking responses. That way, participants had to click equally often on cohort words and non-cohort words throughout the experiment and equally often on words with stress on the first and second syllable.

Intonation condition was rotated across trials as follows: For the experimental and distractor trials, intonation condition was distributed in a Latin Square Design. In other words, half of the filler trials were presented with an early-peak accent, half with a medial-peak accent. Each participant was presented with the same fillers. Eight experimental lists were created, pseudo-randomizing the order of trials, such that each experimental half contained the same number of cohort and distractor trials, with the constraint of the experimental item (WSW) being at most the third item of the same condition in a row. Each list started with seven practice trials to accommodate participants to the task and the voice of the speaker: five filler trials, followed by two distractor trials. Participants were randomly assigned to one of eight experimental lists (six participants per list).

Participants were tested individually in a quiet room in the Phonlab at the University of Konstanz using the SR Eyelink 1000 Plus in a desktop mount system at a sampling rate of 500 Hz. The distance between participants and an LCD screen (37.5 cm x 30 cm) was approximately 70 cm, see Figure 18 for an illustration of the setup. Prior to calibration, all participants received written instructions and a test to determine the dominant eye was administered. Their dominant eye was then calibrated in an automatic procedure (pupil and corneal reflection, Eyelink default settings with nine dots on the screen). Every trial of the experiment started with a centred black cross on white background displayed until participants clicked on it. Upon clicking, the four words appeared on screen (Times New Roman Font, size 20). The words were presented in the outer third of the four quadrants of the screen (to avoid peripheral looking) and framed by a rectangular box (6.5 cm x 4 cm), see Figure 19.
Figure 18: Experimental setup in the Phonlab with SR Eyelink 1000 Plus in a desktop mount. Screen shows a fixation cross (between trials).

Figure 19: Exemplar screen for an experimental trial, showing the cohort pair (WSW-target Libelle (top right) and SWW-stress competitor Libero (bottom left)), and two unrelated distractor items (SWW Thymian ‘thyme’ and WSW Safari ‘safari’ (top left and bottom right, respectively)). Screen is depicted to scale.

The position of the items on screen was counterbalanced across intonation conditions, such that the target that participants had to click on occurred equally often in the four possible positions for each intonation condition. The auditory instruction started 2000 ms after the words occurred on screen, leaving a preview of the words for participants of approximately 2575 ms (2000 ms pause + on average 575 ms pre-context). Participants’ task was to click on the word named in the auditory stimulus as fast as possible. Auditory stimuli were presented via headphones at fixed comfortable loudness (Beyerdynamic DT-990 Pro, 250 OHM). Every fifth trial, a drift correction was initiated. After half of the trials, there was an optional pause and participants could initiate the second half of the experiments in a self-paced manner. The duration of the experiment was 15 minutes (including a language background questionnaire).
5.2.2 Analysis and results

Only experimental trials were analysed (WSW-words, e.g., Libelle) as they were of interest for the research question. Recall that we predicted an increased activation of SWW-competitor words (and therefore potentially decreased activation of WSW-targets) in the early-peak condition compared to the medial-peak condition.

Clicks

In experimental trials, participants correctly clicked on the auditory target in 97.1% of the cases (13 mistakes in the early-peak condition; 9 mistakes in the medial-peak condition). Participants had an average clicking response time of 538 ms (SD = 305 ms) to the visually presented target, measured from the acoustic target offset. To investigate whether clicks were affected by intonation condition, we used logistic mixed effects models (glmer) for correctness of clicks (binary response variable) and linear mixed effects models (lmer) for reaction times (continuous response variable). In the respective models, we initially included intonation condition (medial-peak vs. early-peak accent) as fixed factor and participants and items as crossed random effects using the lme4 package (Bates, Maechler, Bolker, & Walker, 2005) in R (Baayen et al., 2008; R version 3.3.3, R Development Core Team, 2015). The random effects structure of the basic model included random intercepts for participants and items (Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek et al., 2017). Random slopes were added if they improved the fit of the model (in terms of LogLikelihood), as indicated by the anova() function in R. Results of the models showed that there was no effect of intonation condition on the correctness of clicks and no effect on the clicking latencies (all p-values > 0.23). Hence, at the end of the auditory target when the segments unambiguously resolved the competition between WSW-target and SWW-stress competitor, participants unsurprisingly identified the correct target without being affected by pitch accent type. For the current research question, it is more interesting to examine whether the processing during segmental ambiguity, i.e., the pattern of activation for target and stress competitor in the segmentally ambiguous window ([lib] for Libelle/Libero), is influenced by intonation condition. Hence, we turn to the analysis of fixations, which are indicative of lexical activation over time.

Fixations

Fixations were extracted in 4 ms-bins but down sampled in 20 ms-bins for subsequent analysis. They were automatically coded as being directed to the target (WSW,

---

61 Specification of final model for correctness of clicks: corr.glmer_resynth = glmer(targetclick ~ cond + (1|participant) + (1|item), data = data, family = "binomial")

62 Specification of final model for clicking latencies: RT.lmer_resynth = lmer(mouselclickrelendTW ~ cond + (1|participant) + (1|item), data = data)
Libelle), the stress competitor (SWW, Libero), or to the unrelated distractors if they fell within a square of 200 x 200 pixels around the respective word. Blinks and saccades were not further processed. The grand average of evolution of fixations to the four words in the two intonation conditions is shown in Figure 20. These figures were created using the VWPre package in R, a package for pre-processing and plotting visual-world data (Porretta, Kyröläinen, van Rij, & Järvikivi, 2017). In Figure 20, the grey vertical dashed lines indicate the segmental reference points, i.e., word boundaries from left to right, shifted by 200 ms, the average time it takes to launch a saccade (Altmann & Kamide, 2004; Fischer, 1992; Matin, Shao, & Boff, 1993). Hence, only after this time fixations can be interpreted as a response to the signal. The segmental uniqueness point (U.P.), i.e., the point in the signal at which acoustic information perceptually distinguishes the cohort pair irrespective of suprasegmental information (e.g., after the release of [b] in Libelle [liˈbe:lə] vs. Libero [ˈli.bə.ro]) is indicated by a black dashed line (at 1077 ms in Figure 20). The window of interest for the analysis of fixations is the time from target word onset until the segmental U.P., both shifted by 200 ms, i.e., from 775 to 1077 ms in Figure 20. This is the shifted time of segmental overlap in which effects of f0 on the interpretation of lexical stress are expected to surface. We will now describe the evolution of fixations (Figure 20) before moving on to the statistical analysis of competitor and target fixations.

Visual inspection of evolution of fixations

Figure 20 shows that as soon as segmental information of the auditory WSW-target became available, fixations to the (segmentally unrelated) distractors decreased in both intonation conditions (at 775 ms in Figure 20), while fixations to the WSW-target and the SWW-stress competitor both further increased, but with clear differences across intonation conditions: In the early-peak condition, competitor fixations increased more quickly than target fixations; in the medial-peak condition, target and competitor fixations increased with an equal slope. In particular, during the segmentally ambiguous part (775-1077 ms in Figure 20, i.e., [lib]), the stress competitor was fixated more than the target in the early-peak condition (Figure 20, upper panel), while target and competitor were fixated almost equally in the medial-peak condition (lower panel). After the segmental U.P. (at 1077 ms in Figure 20), fixations to the SWW-competitor dropped (for both early- and medial-peak condition). Figure 21 visualizes the fixations to the SWW-competitor in the two intonation conditions. We predicted more competitor fixations in the early-peak condition than in the medial-peak condition in the time window from 775 to 1077

63 Note that the decrease in fixations to distractors when segmental information of the target becomes available corroborate the shift by 200 ms (see Figure 21, light blue lines).
ms (segmental overlap). A visual inspection of the competitor fixations suggests that this is the case (see Figure 21).

![Figure 20](image1.png)

**Figure 20**: Evolution of fixations to WSW-target (e.g., *Libelle*, dark blue line), SWW-competitor (e.g., *Libero*, red line) and the two distractors (e.g., *Thymian* ‘thyme’ (SWW), *Safari* ‘safari’ (WSW), light blue lines) in experimental trials for Experiment 3. The early-peak condition is shown in the upper panel, the medial-peak condition in the lower panel. Acoustical landmarks (dashed vertical lines) are shifted by 200 ms.

![Figure 21](image2.png)

**Figure 21**: Competitor fixations across intonation condition (early-peak condition, red vs. medial-peak condition, orange) in experimental trials for Experiment 3. Acoustical landmarks (dashed vertical lines) are shifted by 200 ms.

**Statistical analysis of competitor fixations across intonation condition**

To statistically corroborate the differences in competitor fixations in the two intonation conditions, we used general additive mixed models in R (GAMMs, Baayen, van Rij, de Cat, & Wood, 2018; Baayen, Vasishth, Kliegl, & Bates, 2017; Wieling, 2018; Wood, 2006; Wood, 2017). GAMMs were chosen as they represent a state-of-the-art statistical approach for the analysis of time-varying data with non-linear relationships and auto-correlation (Baayen et al., 2018; Wieling, 2018). The visual representation of GAMMs indicates when in time an effect on a response variable becomes significant. This is an elegant alternative to traditional time-window
analyses, which require fixations to be binned in predefined arbitrary analysis windows (Barr, 2008), or to Growth Curve Analysis (Mirman, Dixon, & Magnuson, 2008), which model differences in shape of the curves by fitting polynomials to the time-series data. Crucially, GAMMs further allow for the correction of auto-correlation, which is inherent in fixation data. GAMMs have successfully been used in other recent eye-tracking studies (e.g., Nixon et al., 2016; Porretta, Tucker, & Järvikivi, 2016; van Rij, Hollebrandse, & Hendriks, 2016), corroborating their adequacy for the use in this study.

Specifically, GAMMs model non-linear dependencies of a response variable and a predictor via smooth functions, which include a pre-specified number of base functions of different shapes, e.g., linear and parabolic functions of different complexity. Fixed effects can be modelled in the same way as in more traditional linear mixed effect regression models (Baayen et al., 2008, as was done for statistical analysis for clicking latencies above). For the GAMM analysis, we used the R package mgcv (Wood, 2011, 2017); the package itsadug (van Rij, Wieling, Baayen, & van Rijn, 2017) was used to plot the model results. Note that GAMM model outputs alone are not sufficient for the interpretation of the results (different from traditional lmer or glmer summary outputs), effects only become obvious through visualization (Wieling, 2018; Wood, 2017).

In our general additive mixed models, we used the following variables: Competitor fixations were taken as the response variable. They were converted to empirical logits (elogs, which is a logit transformed proportion to looks to the competitor, i.e., a ratio of the fixations to the competitor divided by the fixations directed to the three other objects (Barr, 2008)). We included a parametric coefficient for intonation condition, along with a random effect for event (combining item and subject as a unique identifier) which allowed for a random intercept (see e.g., Porretta et al., 2016). For intonation condition, nonlinear functional relations with the response variable over Time were allowed for using the smooth function. In addition, an AR-1 correlation parameter was estimated, using the acf_resid() function implemented in the package itsadug in order to account for the auto-correlation in the fixation time series, see Figure 22. Figure 22 displays the auto-correlation present in our data. The height of the second vertical line indicates the amount of auto-correlation at lag 1 (rho = 0.64). This value is typically used for correction in all subsequent models (see Porretta et al., 2016).
Figure 22: Auto-correlation graph for the basic model for competitor fixations in Experiment 3 (basic.gam). The height of the second line indicates the amount of auto-correlation at lag 1 (\(\rho = 0.64\)).

Following Porretta et al. (2016), we used a backward step-wise elimination procedure to identify the best model. In all our models, *intonation condition* was kept as a parametric coefficient due to the experimental design of the study. As a first criterion for inclusion of smooth terms, we used the estimated \(p\)-value of these smooth functions (see Porretta et al., 2016). A smooth term was considered for inclusion in the model only if it showed a significant \(p\)-value (< .05). We then compared the model including the smooth term to a simpler version of the model (without the smooth term) using the function CompareML(). This comparison revealed whether the inclusion of this predictor significantly improved the model fit (Maximum Likelihood (ML) scores) or whether this predictor had to be removed. It has been argued that the Akaike information criterion (AIC, Akaike, 1974) is not reliable when the model accounts for auto-correlation in the data (Zuur et al., 2009). Therefore, we compared ML scores as a criterion to determine the better model fit (see Porretta et al., 2016).

Model comparisons following the procedure described above indicated that the model with different time-smooth functions for the different intonation conditions was preferred over the model without a smooth function for the factor *intonation condition* over time. Hence, the final model included *intonation condition* as a parametric coefficient and as a non-linear effect (smooth term) over time, as well as random intercepts for the *event* variable. The final model accounted for 67.7% of the deviance, emphasizing that the model captured important features of the fixations over time, see Table 11 for coefficients of the final model.
Table 11: Final general additive mixed model for competitor fixations in Experiment 3 with spelled out variable names. Part A: Estimate, Standard Error (SE), t- and p-value for the parametric coefficients. Part B: Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F- and p-values for the smooth term. Part C: Model specification of the final model (original variable names).

<table>
<thead>
<tr>
<th>Part A. Parametric coefficients</th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (early-peak contour)</td>
<td>-0.404</td>
<td>0.1012</td>
<td>-3.988</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intonation condition (medial-peak)</td>
<td>-0.229</td>
<td>0.1406</td>
<td>-1.626</td>
<td>0.1040</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B. Smooth terms</th>
<th>EDF</th>
<th>Ref.df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random effect (intercept) for Event, s(event)</td>
<td>643.277</td>
<td>730.00</td>
<td>8.655</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time):early-peak</td>
<td>5.416</td>
<td>6.718</td>
<td>30.075</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time):medial-peak</td>
<td>1.688</td>
<td>2.133</td>
<td>44.491</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

| Part C. Model specification | bam(elogC ~ cond + s(TIMESTAMP, by = cond) + s(event, bs = 're'), rho = 0.64, AR.start = df_combined$experiment == "Exp3","start_event, data = df_combined[df_combined$experiment == "Exp3",], nthreads = 4, method = "ML") |

The effect of intonation condition over time is directly visualized in Figure 23. To explain, the plot represents the difference curve in competitor fixations in the two intonation conditions (early-peak condition minus medial-peak condition), as calculated by the statistical model (for specification of the final model, see Table 11). The grey band displays the 95% CI of the mean difference. Values above zero indicate more competitor fixations in the early-peak condition. Conversely, values below zero indicate more competitor fixations in the medial-peak condition. The difference in competitor fixations across conditions is significant if the 95% CI does not include zero; a red line on the x-axis in the plot indicates the significant period(s). Figure 23 reveals that competitor fixations in the two intonation conditions (early-peak condition vs. medial-peak condition) significantly differ in the window 868-1001 ms.\textsuperscript{64}

\textsuperscript{64} Since GAMMs have only recently been started to be applied in eye-tracking research (e.g., Nixon et al., 2016; Porretra et al., 2016; van Rij et al., 2016), we here additionally report the results of a more traditional moving time window analysis (Barr, 2008, p. 464) in which fixations to a target as a function of intonation condition are compared in subsequent 100 ms-windows, starting from target onset, shifted by 200ms processing time (e.g., Alloppena et al., 1998; McQueen & Viebahn, 2007; Mitterer & Reinsch, 2013). For this 100 ms-window analysis, we compared competitor fixations (also computed as elogs) across intonation condition, using lmers (Baayen et al., 2008).
Figure 23: Difference curve in competitor fixations in early-peak condition minus medial-peak condition in Experiment 3. The grey band indicates the 95% CI of the mean difference. Values above zero indicate more competitor fixations in the early-peak condition. Conversely, values below zero indicate more competitor fixations in the medial-peak condition. The difference is significant if the 95% CI does not include zero. Window of significant difference: 868-1001 ms.

Hence, as predicted, in the time period in which information of the segmentally ambiguous part is processed, there are significantly more fixations to the stress competitor when the target is realized with an early-peak compared to a medial-peak accent: The time window of significant differences (868-1001 ms) lies in the time window in which segmental information is ambiguous between the WSW- and SWW-word (775-1077 ms). The time window in which the stress competitor effect appeared lasted 133ms (868-1001 ms), one third of the time window during which the effect can occur. Effects similar in duration to the one we find in our data have been reported in Visual-World Eye-Tracking studies with comparable or even larger acoustic differences (e.g., Jesse et al., 2017; McQueen & Viebahn, 2007).

model fit for each time window was determined individually (Baayen et al., 2008; Bates et al., 2015; Matuschek et al., 2017). Each model initially included intonation condition (medial-peak contour vs. early-peak contour) as a fixed factor and random intercepts for participants and items (Bates et al., 2015; Matuschek et al., 2017). Random slopes were added if they improved the fit of the model (in terms of LogLikelihood), as indicated by the anova() function in R. Results of this analysis revealed a significant effect of intonation condition in the second analysis window (100-200 ms post target onset), i.e., 875-975 ms, with more competitor fixations in the early-peak condition (average elog: -0.66; B = 0.4 [0.05; 0.65], SE = 0.15, df = 732, t = 2.31, p = 0.02). All other 100 ms-windows did not show an effect of intonation condition (all p > 0.18). The detailed results of this analysis are displayed in Appendix C (C.4). Hence, the more traditional analysis reveals very similar results as the more state-of-the-art GAMM analysis (the window of significant differences in competitor fixations indicated by the 100 ms-window analysis (875-975 ms) lies in the window of significance that GAMM indicates (868-1001 ms)). Given the advantages GAMM shows over other kinds of analysis, we consider the use of the more state-of-the-art methods in the following.
Statistical analysis of target fixations across intonation condition

Given the effect of intonation condition on competitor fixations, in a second analysis step, we assessed whether target fixations also differed as a function of intonation condition during the segmentally ambiguous part of the target. Note that differences in fixations to the target across intonation condition, i.e., in the activation of the intended word (auditory target), would be even stronger evidence for interference of pitch accent type during spoken word recognition. As for competitor fixations, a general additive model with empirical logits of target fixations as response variable was fit, initially including intonation condition as a fixed factor and the same smooth terms as described above. Model comparisons followed the same procedure as described for competitor fixations. The final model included intonation condition as parametric coefficient and a non-linear effect (smooth term) of intonation condition over time, as well as a random intercept for the event variable. The final model explained 66.1% of the variance, which is comparable to the model for competitor fixations. Auto-correlation was also very similar to what was observed for competitor fixations (rho = 0.65). Again, the coefficients of the final model are given in Table 12 while the effect of intonation condition over time is directly visualized in Figure 24. Here the plot shows the difference curve of target fixations (instead of competitor fixations) in the two intonation conditions, early-peak condition minus medial-peak condition, as calculated by the statistical model (see Table 12 for specification of the final model). Recall that values above zero indicate more target fixations in the early-peak condition and values below zero more target fixations in the medial-peak condition.

Figure 24 indicates that for 80 ms in the segmentally ambiguous part (844-922 ms), there are significantly fewer fixations to the target in the early-peak condition than in the medial-peak condition. However, compared to the stress competitor effect, the effect of pitch accent type on target fixations is shorter in time (55 ms shorter) and smaller in its extent (smaller elongation on y-axis, cf. Figure 23).

Note that the 100ms-window analysis with the same conventions for modelling as described for competitor fixations, did not capture the small effect of intonation condition on target fixations (all 100ms-time windows $p > 0.4$).
Table 12: Final general additive mixed model for target fixations in Experiment 3 with spelled out variable names. Part A: Estimate, Standard Error (SE), \(t\)- and \(p\)-value for the parametric coefficients. Part B: Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), \(F\)- and \(p\)-values for the smooth term. Part C: Model specification of the final model (original variable names).

### Part A. Parametric coefficients

<table>
<thead>
<tr>
<th>Parameter Description</th>
<th>Estimate</th>
<th>SE</th>
<th>(t)-value</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (early-peak contour)</td>
<td>-0.879</td>
<td>0.094</td>
<td>-9.295</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intonation condition (medial-peak)</td>
<td>-0.093</td>
<td>0.131</td>
<td>0.711</td>
<td>0.4770</td>
</tr>
</tbody>
</table>

### Part B. Smooth terms

<table>
<thead>
<tr>
<th>Parameter Description</th>
<th>EDF</th>
<th>Ref.df</th>
<th>(F)-value</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random effect (intercept) for Event, s(event)</td>
<td>634.440</td>
<td>730.0</td>
<td>7.392</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, (Time):early-peak</td>
<td>4.162</td>
<td>5.293</td>
<td>37.648</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time):medial-peak</td>
<td>1.003</td>
<td>1.101</td>
<td>56.990</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

### Part C. Model specification

```r
bam(\log(T) ~ cond + s(TIMESTAMP, by = cond) + s(event, bs = 're'), rho = 0.65, AR.start = df_combined[df_combined$experiment == "Exp3",]$start_event, data = df_combined[df_combined$experiment == "Exp3",], nthreads = 4, method = "ML")
```

Figure 24: Difference curve in target fixations in early-peak condition minus medial-peak condition in Experiment 3. The grey band indicates the 95% CI of the mean difference. Window of significant difference: 844-922 ms.
5.3 Discussion

The fixation data showed that different pitch accent types affect lexical activation in German adults: While participants were processing the segmentally ambiguous part of the WSW-target word (Libelle), there were more fixations to the SWW-stress competitor (Libero) when the target word was presented with an early-peak (H+L*) than when presented with a medial-peak accent (L+H*). Consequently, participants directed fewer fixations to the intended target during parts of the ambiguous time in the early-peak condition. Hence, early-peak accents, which involve a misalignment between f0 peak and metrical stress, led to a change in the perception of the metrical strength relations in a word: F0 peaks were interpreted as the strong element, which led to the temporary activation of competitors with initial stress. As a result of this increased competitor activation, listeners fixated the intended target word less—an effect that was significant but very short-lived and small in its extent. Taken together, utterance-level pitch accent type, which in intonation languages mainly serves an information-structural function, directly affects lexical processing in German adults such that competition is increased. Additionally, the activation of the intended target seems to be minorly hampered.

The pitch accent types used in the current experiment (medial-peak, L+H* and early-peak accents, H+L*) naturally occur in German and are both pragmatically appropriate renditions for the carrier phrase employed. On the one hand, the objects to be clicked on in each trial represent information-structurally new information, which might favour a medial-peak accent (Kohler, 1991c). On the other hand, the repetition of the pre-context creates a notion of accessibility of the objects as a whole (cf. Baumann & Grice, 2006), which makes an early-peak realization also pragmatically appropriate. To support the felicity of the two intonation conditions, we conducted a post-hoc production study: Another ten participants (6 female, 4 male, average age = 25.8 years, SD = 3.9 years) read the experimental stimuli of the eye-tracking experiment aloud in a sound-attenuated cabin (44.1 kHz, 16 Bit), see Appendix D (D.1) for more details on participants. We used two experimental lists from the eye-tracking study with 64 trials each, resulting N = 640 productions, in total. Specifically, participants were asked to produce the sentences such that sentences would be “appealing” (German: ansprechend)66 instructions in a computer game. An intonational analysis67 showed that medial-peak accents were most frequent (71% of the cases, N = 441). Early-peak accents occurred in 19% (N = 116)

---

66 This adjective was used to avoid declination (M. Liberman & Pierrehumbert, 1984) and a monotone manner of reading.

67 The 640 productions were annotated according to GToBI by the author of this dissertation. Difficult cases were discussed with a peer PhD student in the department of linguistics (Sophie Kutscheid) and/or the author’s first supervisor (Bettina Braun). In total, 15 items of the 640 productions were excluded due to production of an additional word, the production of the wrong target word, and mis-stressing of a target.
and late-peak contours in 5% (N = 32), the remaining 6% (N = 36) were unaccented; see Appendix D (D.2) for example realizations for each category and Appendix D (D.3) for a more detailed analysis of the distribution of accents. Overall, the distribution of pitch accents in the post-hoc reading study is similar to the frequency counts of these accents in German appointment scheduling dialogues (Peters et al., 2005). Therefore, early-peak accents are clearly an option in this setting – albeit not the most frequent one.

Our finding that peak-stress-misalignment affected stress perception sheds new light on previous studies, as it straightforwardly explains a number of recent findings in the literature: Schwab and Dellwo (2017), for instance, showed that intonational variability hampers stress identification in both native and non-native listeners. They investigated the detection of a stress deviant in Spanish trisyllabic words, i.e., stress-minimal triplets, e.g., *numero*, which can be stressed on the first syllable (‘number’), the second syllable (‘I number’), or the third syllable (‘he numbered’). The three listener groups were native listeners of Spanish and two non-native groups with a different background on stress, German listeners (free stress language) and French listeners (fixed stress language). The words in a trial were spoken by the same vs. different speakers and with the same vs. different intonation contours (falling vs. rising). Their results showed that all listener groups were less accurate in identifying a stress deviant when intonation varied across test words (while talker variability only affected French listeners). The authors argue that listeners’ performance was impeded by intonational variability, since f0 could not be relied on in the same way in questions than in declaratives. From our perspective, this processing difficulty is very likely the result of misinterpretations of the stressed syllables due to alignment differences. In the study by Dilley and Heffner (2013), reviewed in the introduction of this chapter (5.1), there were more last syllable stress responses in the rising continua (which rendered the last syllable high-pitched) compared to the falling continua (Dilley & Heffner, 2013, Figure 9; visual inspection; additionally, three participants had a strong preference for the last syllable in the lemonade-series in general). Finally, as outlined in Chapter 3, Friedrich et al. (2001) showed that German listeners were slower and made more errors in reacting to a SW-target in an identification study (e.g., *Amboss* [ˈam.bʊs] ‘anvil’), when the pitch contour in the target was taken from an accented WS-word (e.g., *Abtei* [aˌp.ˈtaɪ̯] ‘abbey’) than when the pitch contour was taken from a SW-word. Given that the contours seem to be realized with H*-accents (see Figures in Friedrich et al., 2001) processing appears to be hampered if the stressed syllable is not high-pitched.

Importantly, the competitor activation effect – along with the decrease in activation for the target – we find in the current study was solely driven by f0-peak alignment. In both intonation conditions the pitch accent was phonologically
associated with the stressed syllable of the WSW-target; non-tonal stress cues pointed to the second syllable as the stressed one. The only cue that differed across intonation condition was f0-peak alignment. Since we used resynthesized materials, we can safely conclude that effects are solely due to the f0 alignment and not caused by other acoustic correlates cohering with the f0 peak. If natural productions of medial-peak accents have additional intensity and duration on the stressed syllable and the natural productions of early-peak accents on the pre-tonic syllable, then the resynthesis reverted these cues. F0 peaks – even when realized on genuinely un-stressed syllables – are clearly interpreted as stressed.

While the claim that f0 is a strong indicator for lexical stress in adults is not new in the literature (cf. Fry, 1958, and other subsequent meta-linguistic decision studies), this study provides the first evidence that naturally occurring f0-alignment differences in different pitch accent types impede stress perception and as such affect lexical activation, see the General Discussion (7.2.2) where we discuss the implications that arise for spoken word recognition and models thereof in more detail. That way, our results extend the findings in the offline stress identification task in Zahner et al. (2019), who also used non-minimal pairs and naturally occurring alignment contrasts, and show that pitch accent type does not only have an effect in meta-linguistic – and thus more artificial – tasks, but that the results directly transfer to online processing. Since we used a method that directly addresses lexical activation (Tanenhaus et al., 2000), our findings firmly suggest that high f0 was interpreted as a cue to lexical stress and not just as an acoustic cue that enhanced prosodic prominence on a given syllable. Meta-linguistic studies, however, remain caught in this methodological trap.

One question remains, i.e., why is high f0 sufficient for German adults to mark the syllable it appears on as the stressed one? Similar to what was suggested in Chapter 4 for German infants, we propose the perceptual salience (Option 1) as well as the distribution of different pitch accent type (Option 2) as potential mechanisms that guide the processing of f0 as a stress cue in German adults. Again, at this point, we are not in the position to give priority to one explanation over the other. We will now change the perspective in Chapter 6 and address the underlying mechanisms to the processing of f0 as a stress cue in order to be able to answer this question.
6 Addressing mechanisms for the processing of f0 as a stress cue

Chapter 6

Chapter outline
Chapters 4 and 5 provided evidence that pitch accent type interferes with the processing of lexical stress both in infants and in adults. Both groups are strongly guided by the position of the f0 peak when perceiving a syllable as stressed, which has consequences for stress-based segmentation in infants and lexical activation in adults, respectively. Possible mechanisms that may render f0 a cue to stress are a) the salience of high-pitched syllables or b) the distribution frequency of different pitch accent types. The current study is an attempt to shed light on these mechanisms, putting the frequency account to test. To this end, we conducted an exposure-test paradigm with adult listeners (Experiment 4) in which we increased the frequency of low-pitched stressed syllables in the immediate input (by an exposure phase of 3 minutes), before we subsequently re-run the eye-tracking experiment (cf. Experiment 3 in Chapter 5). In Experiment 4, fixation results showed a different pattern than in Experiment 3: The effect of intonation condition on competitor and target fixations disappeared. A pooled analysis of the fixation data in Experiment 3 and 4 confirmed that pitch accent type affects competitor fixations differently in the two experiments. This indicates that participants changed their weighting of the f0 cue as a stress cue after exposure to low-pitched accents – at least for the same speaker. Hence, the frequency of occurrence of different pitch accent types is one mechanism that explains why high f0 may function as a cue to lexical stress.

68 Adapted parts of this chapter are included in Zahner, Kutscheid, and Braun (2019).
6.1 Introduction

In Chapters 4 and 5, we observed f0-stress interference in both infant and adult processing, with the f0 peak guiding both German infants’ and adults’ perception of lexical stress. Specifically, in Chapter 4, we established that infants interpret a stressed syllable as stressed only if this syllable is realized with an f0 peak, i.e., if it is high-pitched. In fact, unstressed high-pitched syllables in natural early-peak productions even appear to be erroneously taken as stressed by German infants. In Chapter 5, we demonstrated a comparable pattern of behaviour in adults when tested in a methodologically different paradigm but with an underlyingly similar rationale. In a Visual-World Eye-Tracking study, f0-alignment differences in the form of different pitch accent types affected stress perception and consequently lexical activation: Competitor words with an initial stressed syllable (SWW) received stronger activation if WSW-targets were realized with an early-peak accent than when realized with a medial-peak accent.

Taking a different perspective as in Chapters 4 and 5, we here address one of the mechanisms that might account for why high f0 is interpreted as a cue to stress in both infant and adult processing: As already foreshadowed in the respective discussions in Chapters 4 and 5, we see two mechanisms that can account for the observed f0-stress interference: One mechanism (Option 1) assumes that high-pitched syllables stand out perceptually (Cho, 2005; Hsu et al., 2015; Lieberman, 1967) and that way qualify as markers of lexical stress. We will refer to this account as the “salience account.” The idea is that listeners pick up on this cue as they might expect something important to be conveyed and hence respond to an increased articulatory effort (Lieberman, 1967). It could also be the case that high pitch per se is processed in a special way, as the qualitatively different brain response to raised pitch compared to lower pitch in the study by Hsu et al. (2015) might suggest. Generally – and at the same time very vaguely – the concept of “salience” has been defined as “the property of a linguistic item or feature that makes it in some way perceptually and cognitively prominent” (Kerswill & Williams, 2002, p. 81). It has been pointed out though that perceptual salience is a concept that is difficult to define and even more difficult to quantify (e.g., Hickey, 2000). To clarify, we here refer to a bottom-up salience, i.e., the salience of the stimulus itself due to its physical nature – in contrast to the top-down salience which is based on cognitive pre-activation sometimes observed in semantic or pragmatic studies (for reviews on both types of salience see e.g., Awh, Belopolsky, & Theeuwes, 2012; Blumenthal-Drame, Hanulíková, & Kortmann, 2017; Zarcone, van Schijndel, Vogels, & Demberg, 2016). For the purpose of this thesis, we assume salience to be a psychophysical phenomenon (Cintron-Valentin & Ellis, 2016, p. 2) and define it as a “pop-out effect” (Zarcone et al., 2016, p. 5) that captures the listener’s attention. There
is (yet) no consensus in the literature regarding the reason for this popping-out effect. It has been suggested that the notion of surprise or unexpectedness contribute to salience (Zarcone et al., 2016, p. 5), or that surprisal and frequency together render an event salient, with an initial unexpectedness contributing to salience and the repetitive occurrence of a pattern establishing associations that make an item extra salient (Rácz, 2013; based on salience of a segmental phenomenon in a dialect). Either way, salient events stand out from their surrounding.

The second mechanism we propose (Option 2) posits that listeners learned to associate high-pitched syllables with metrical stress because of a frequent occurrence of H*-accents in the ambient language. Even though the finite state model of intonation by Pierrehumbert (1980) suggests a free combination of tonal events and does not make any predictions about the probability of their occurrence, there are intonational events that are more frequent than others (e.g., Dainora, 2006, for American English; Peters et al., 2005, for German). If high pitch and metrical stress frequently go together, listeners might create a mental link of the two concepts. We will refer to this account as the “frequency account.” We use the term frequency to refer to the number of occurrences in the input.

Importantly, both factors (salience and frequency) have been shown to be beneficial in processing. Regarding salience, based on psychological research, it has been argued that “salient items are more likely to be perceived, to be attended to, and more likely to enter into subsequent cognitive processing and learning” (Cintron-Valentin & Ellis, 2016, p. 2). Indeed, numerous studies with both infants and adults have demonstrated that the perceptual salience of stimuli influences processing (e.g., Bortfeld, Shaw, & Depowski, 2013; Zarcone et al., 2016, for a review). Regarding acquisition processes, Narayan, Werker, and Speeter Beddor (2010), for instance, showed that phonetic acquisition of sound contrasts is affected by the relative salience of the sound contrast, which here is defined as the acoustic difference. While a salient contrast [ma]/[na] was discriminated by (Canadian)-English infants from 4 months onwards, a less salient contrast [na]/[ŋa] could not be discriminated by either English (non-native contrast) nor by Filipino-learning infants (native contrast). A similar observation has been made for the acquisition of the German/Swiss-German contrast of labial plosives (Pohl, 2011; Schönhuber, Czeke, Gampe, & Grijzenhout, 2019), with the voice-onset time contrast (native for Standard German) being more felicitous for discrimination than the seemingly less salient length contrast (closure duration, native for Swiss German). Referential word recognition has also been shown to be facilitated with onomatopoetic words (Laing, 2017). This finding was attributed to an increased acoustic salience (pauses, f0 movements etc.) of these lexical items in the infants’ input (Laing, Vihman, & Ke- ren-Portnoy, 2016).
Similarly, adult processing is also facilitated by perceptual salience (see Bardovi-Harlig, 1987; Burmester, Sauermann, Spalek, & Wartenburger, 2018; Cintron-Valentin & Ellis, 2016; Gong, Lam, & Shuai, 2016; Grohe & Weber, 2016b; Liu, Chen, & Kager, 2017; Liu & Kager, 2012; Zarcone et al., 2016). To give an example, similarly to the studies for infants mentioned above, Liu and Kager (2012) showed that Dutch listeners easily discriminated a tonal contrast in Mandarin if this contrast was very salient (Tone 1 vs. Tone 4), but have difficulty discriminating if the acoustic salience of this contrast is reduced (in a step-wise fashion). Perceptual salience furthermore fosters learning a second language and accent adaptation: In this regard, Grohe and Weber (2016b) recently demonstrated in an eye-tracking study with a pre-fixed exposure phase that acoustically salient accented words promoted learning of an unfamiliar accent. Taken together, the studies on the beneficial effects of salience in both infant and adult processing reviewed above suggest that Option 1 (salience account) can be considered a likely mechanism for the processing of f0 as a stress cue by both infants and adults.

The frequency of pitch accent types in a given language may also very likely influence processing. In fact, effects – for which the frequency of occurrence of a linguistic phenomenon may account for – are well documented in various processing domains, such as language acquisition, language use, and language change (Ambridge, Kidd, Rowland, & Theakston, 2015; Diessel, 2007, for reviews). Regarding first language acquisition, on a very basic level, numerous studies have shown that the more parents speak to and in presence of their infant, the more words children possess in their receptive (and productive) vocabularies as toddlers (e.g., Hurtado, Marchman, & Fernald, 2008; Huttenlocher et al., 1991; Newman et al., 2016; Weisleder & Fernald, 2013). Clearly, frequency effects in language acquisition have been reported beyond the mere quantity of the input. For instance, frequency distributions affect very early acquisition processes, such as the segmentation of fluent speech solely on the basis of co-occurrences of syllable strings at the age of 8 months (cf. Saffran, Aslin, et al., 1996, for a seminal study in this field; see Saffran & Kirkham, 2018, for a very recent review, and references therein) or the use of frequent co-occurrences of syntactic relations for the syntactic

69 Infancy researchers currently display a great interest in the child’s input, as is mirrored in the newly developed repositories, such as the HomeBank system, https://homebank.talkbank.org/, 29.02.2020 (Warlaumont, VanDam, Bergelson, & Cristià, 2017). Certainly technical advances, such as the launch of the LENA System (https://www.lena.org/, last accessed: 29.02.2020) largely contribute to this flourish in IDS data collection.

70 Note though that Saffran and Kirkham (2018) emphasize that, as far as statistical learning is concerned, infants rely more on transitional probabilities, i.e., the co-occurrence of one element with the other, than on the computation of mere occurrence frequencies. Mere frequencies seem to play also a minor role in the discrimination of native vowel contrasts, as has recently been shown by Tsuji et al. (2017): Dutch infants discriminated native frequent ([i]-[e:|]) and infrequent ([u]-[ø:]) vowel contrasts equally well but displayed no discrimination for non-native contrasts.
categorization of novel words at 14-16 months, here the frequent combination of a determiner and a nonce word to classify a noun (Höhle et al., 2004). Furthermore, segmentation of vowel-initial words has recently been demonstrated in German 11-month-olds (Boll-Avetisyan, Fritzshe, & Jäkel, 2017), while the English peers fail in this task until the age of 13 months (Nazzi et al., 2005) – without additional cues such as function words or edge-alignment (Kim & Sundara, 2015; Scidl & Johnson, 2008). Boll-Avetisyan et al. (2017) argue that this age-related difference in the onset of vowel-initial segmentation might be an effect of the frequency of these words in the input, which is higher in German than in English. Frequency effects in processing have also been attested at various levels in adult populations, ranging from sounds, morphemes and words to larger units such as phrases and sentential constructions (Behrens & Pfänder, 2016, p. 1). For instance, on the level of sounds, adults are highly sensitive to the native phonotactics (e.g., Frisch, Large, & Pisoni, 2000; Frisch, Large, Zawaydeh, & Pisoni, 2001; Treiman & Danis, 1988). Based on the theta band rhythm in infant and adult listeners, Bosseler et al. (2013) furthermore show that the sensitivity to frequent syllables observed in very young infants decreases with age, with adults showing more cognitive effort for non-native stimuli than for frequent syllables. Regarding transitional probabilities, similar to infants, adults can use recurring syllable sequences to group syllable strings as cohesive units, both in speech and in music (Neger, Rietveld, & Janse, 2014; Saffran, Johnson, Aslin, & Newport, 1999). Frequent co-occurrences of larger units, i.e., words and constituents also influence syntactic processing (e.g., Bybee & Thompson, 1997). One of the most well-known frequency effects in psycholinguistics is probably the word-frequency effect (cf. Howes, 1957; Howes & Solomon, 1951), where words with a high lexical frequency are recognized faster and words with high-frequent neighbours are recognized slower since high-frequent competitors are difficult to suppress (e.g., Balota & Chumbley, 1984; Dufour, Brunellière, & Frauenfelder, 2013; Goldinger, Luce, & Pisoni, 1989; Yap, Balota, Tse, & Besner, 2008, among others). Not only lexical frequency but also the frequency of a particular realization of a word influences word recognition. In that sense, Hanulíková and Weber (2012) demonstrated that the experience with substitutions of the English interdental fricative [θ] (either [t] for Dutch learners of English and [s] for German learners of English) promotes the recognition of variants in the second language. Taken together, there is plenty of evidence – of which we could only review some selective studies – that frequency effects can obviously account for many findings in the infant and adult speech processing literature.

The salience account (high-pitched syllables stand out perceptually) is clearly hard to manipulate in an experimental paradigm. By contrast, the frequency of occurrence of different pitch accent types can be readily manipulated, at least in the
immediate input. Hence, in the current study, we put the frequency account to test and investigate whether the distribution frequency of different pitch accent types is indeed a mechanism that explains the processing of f0 as a cue to lexical stress. In an exposure-test paradigm, we manipulate the frequency of occurrence of high-pitched stressed syllables in the immediate input such that the occurrence frequency of low-pitched stressed is drastically increased (via an exposure phase of 3 minutes). Subsequently, we test the effect on the use of f0 as a stress cue by re-running the eye-tracking experiment (Experiment 3) presented in Chapter 5.

We focus on adults in this study for practical reasons: The HPP paradigm already consists of two phases (familiarization phase and test phase) and a paradigm with an additional exposure phase of several minutes would very likely exceed the attention span of 9-month-olds. Exposure-test paradigms in the sense of adaptation paradigms are well established in the literature on adult processing (e.g., Eisner & McQueen, 2006; Grohe & Weber, 2016a, 2016b; Hanuliková et al., 2012; Hanuliková & Ekström, 2017; Kraljic & Samuel, 2006; Norris et al., 2003; Reinisch & Weber, 2012; Witteman, Bardhan, Weber, & McQueen, 2015; Witteman, Weber, & McQueen, 2014) and also have been used with toddlers (cf. Schmale, Cristià, & Seidl, 2012; van der Feest & Johnson, 2016; van Heugten & Johnson, 2014; White & Aslin, 2011). For infants under one year, however, to our knowledge, only two studies endeavoured to use more than two phases (see Altvater-Mackensen & Mani, 2013; Thiessen & Saffran, 2007 for studies with three phases with 7 and 9-month-olds, respectively). We hence concentrated on the group for which the paradigm is more firmly established.

For our exposure-test experiment, we used a similar design as in recent accent-adaptation studies (cf. Grohe & Weber, 2016a; Grohe & Weber, 2016b; Reinisch & Weber, 2012; Witteman et al., 2014). Studies on adaptation on the lexical level and perceptual-learning studies have shown that an exposure phase of 3 minutes is sufficient to affect lexical activation (Grohe & Weber, 2016a, 2016b; Kraljic & Samuel, 2006; Norris et al., 2003). Reinisch and Weber (2012) similarly showed for the suprasegmental level that listeners quickly adapt to lexical stress placement errors by non-native speakers (by listening to a 2.3 min-story, 28 critical items) and use this information in word recognition. We hence assume our 3 min-exposure phase to be solid to investigate the effect of occurrence frequency of pitch accent type on the use of high f0 as a stress cue.

The current exposure-test paradigm may result in two possible outcomes, depending on whether or not the frequency of occurrence of different pitch accent types affects the activation of the competitor (and the target). If f0-stress interference is driven by the input frequency of high-pitched stressed syllables in natural speech (Option 2), we expect no difference in competitor and target activation across intonation condition (i.e., a different result to the one found in Experiment
3). If, on the other hand, f0-stress interference is caused by the inherent salience of high f0 (Option 1), we expect to see the same differences in competitor/target fixations activation patterns as in Experiment 3, i.e., more looks to the stress competitor (SWW-word) and conversely fewer looks to the target (WSW-word) when the WSW-target is realized with an early-peak accent compared to a medial-peak accent.

6.2 Experiment 4

6.2.1 Methods

Participants
Another group of 48 German native speakers (32 female, 16 male; average age = 22.2 years, SD = 3.2 years, 34 right-eye dominant) with normal or corrected-to-normal vision and unimpaired hearing participated under the same conditions as in Experiment 3. None of them had participated in Experiment 3 or any related study. Again, most of the participants grew up in Southern Germany (73%) and were also mostly staff or students at the University of Konstanz. Recruiting was the same as in Experiment 3. Data of five additional participants was excluded due to calibration errors (4) and a bilingual background revealed only after testing (1).

Materials
For the exposure phase, we chose 15 alternative-question units and 15 contrastive-topic units since alternative questions and contrastive topics are commonly realized with low-pitched stressed syllables (e.g., C. Bartels, 1999; Braun, 2006; Büring, 1997; Truckenbrodt, 2011). Wh-questions and one-word answers were also part of the units since low-pitched accents are felicitous for them (e.g., Baumann & Grice, 2006; Féry, 1993; Grice et al., 2005; Kohler, 2004). More specifically, the alternative-question units consisted of an alternative question and a one-word answer, which was inferable based on the preceding question (e.g., ‘Do you prefer BologneseL+H or CarbonaraH+L*,L%? – CarbonaraH+L*,L%). The two conjuncts of the alternative question were nominal or verbal alternatives (e.g., ‘walking’ – ‘driving’) with different numbers of syllables and different stress patterns. The contrastive-topic units, in turn, consisted of a declarative in which the subject was realized as the contrastive topic, a wh-question, and another declarative that gave the answer to the question (e.g., ‘The blanketL+H is made of flannelL+L*,L% – What is the pillowL+H made of? – The pillowL+H is made of feathersL+L*,L%.’). The declaratives followed a theme-

71 In seven of the 15 alternative-question units, the first alternative was chosen as the answer, in the other eight, the second alternative formed the answer. This was balanced across alternative-question units involving nominal or verbal alternatives.
rheme structure, in which the theme referred to a topic and the rheme formulated a proposition about the theme (see Braun, 2006, for similar materials). The *wh*-question asked for an alternative theme that formed a contrast to the one given in the first declarative (e.g., ‘blanket’ – ‘pillow’). The second declarative was structurally identical to the first one and gave the answer to the *wh*-question. The word prosodic structure of the contrastive topics was also varied. Appendix E presents the full list of alternative-question units (E.1) and contrastive-topic units (E.2).

The same female speaker as in the previous eye-tracking experiment (Experiment 3, Chapter 5) recorded the exposure stimuli with $L^*+H$, $H+L^*$, and $L^*$-accents, i.e., only low-pitched accents. She was instructed to produce the pitch range of these accents as natural as possible, see Appendix E for the acoustic analyses regarding f0 for both types of units, E.3 for alternative-question units and E.4 for contrastive-topic units. In total, the exposure materials consisted of 120 low-pitched accented syllables (15 alternative-question units x three low-pitched accented syllables + 15 contrastive-topic units x five low-pitched accented syllables). Figure 25 shows an exemplar trial for each of the two units, an alternative-question unit in (a) and a contrastive-topic unit in (b).

Figure 25: Example sound pressure wave, spectrogram, and pitch track for exposure units in Experiment 4. Figure (a) shows an alternative-question unit (alternative question + one-word answer) and Figure (b) a contrastive-topic unit (declarative 1 + *wh*-question + declarative 2), see tier 1 for individual components. Tier 2 gives the words in German, tier 3 the English translation; tier 4 indicates the accents in a phrase (*) and tier 5 displays the GToBI labels.
The individual components of each unit were concatenated with an inter-stimulus interval (ISI) of 400 ms to form one unit. The ISI for the first five trials was 600 ms to acquaint participants with the task. The 30 units were grouped in six blocks with five trials each. The order of trials in a block was pseudo-randomized and trials were separated by an ISI of 1000 ms (the first block had a 200 ms longer ISI to adjust participants to the task). The blocks of stimuli were on average 26.6 sec (SD = 1.7 sec) long.

Procedure
During the exposure phase, participants listened to one block of stimuli at a time while fixating a black cross that was centred on a white screen (the same LCD screen, 37.5 cm x 30 cm as in Experiment 3). Sound presentation occurred right after the presentation of the fixation cross (100 ms delay). After every block of approximately 30 s, participants were asked to rate on a Likert scale from 1 (= trifft gar nicht zu ‘does not apply’) to 5 (= trifft voll und ganz zu ‘fully applies’) whether a given adjective applied to the person they were listening to (e.g., sympathisch ‘likeable’). The adjectives were taken from A. Schweitzer and Lewandowski (2013). In total, there were six ratings, i.e., one rating after each block of stimuli. Figure 26 displays the two screens that participants saw during the rating task.

![Figure 26: Exemplar screens for exposure-test paradigm in Experiment 4. Screen 1 displays a fixation cross during the presentation of the exposure units and screen 2 displays the adjective to be rated (e.g., sympathisch ‘likeable’) along with the 5-point scale.](image)

Participants’ responses during the rating task were not recorded. The test phase immediately followed the exposure phase with the exact same procedure, stimuli, and lists as in Experiment 3. At end of the exposure phase, a screen visually

---

72 The six adjectives were the following: sympathisch ‘likeable’, intelligent ‘intelligent’, freundlich ‘friendly’, kompetent ‘competent’, erfolgreich ‘successful’, selbstbewusst ‘self-confident’ (see A. Schweitzer & Lewandowski, 2013). A question mark was placed after each adjective during the rating. The reason for this was to trigger a low-pitched accented syllable in the silent prosody of participants (L* H-”H% would be the most likely nuclear contour for a neutral question (e.g., Fery, 1993; Grice et al., 2005; Kohler, 2004; Von Essen, 1964)).
displayed the procedure of eye-tracking part of the study (indicated as the second part) in order to remind participants of the change in task. The reason for displaying the instruction visually was to avoid silent prosody that might evoke medial-peak accents. Calibration took place before the exposure phase to have a smooth transition between phases.

6.2.2 Analysis and results

Identical to Experiment 3 (Chapter 5), clicks and fixations were analysed for experimental trials only (WSW-words as auditory targets).

**Clicks**

As in Experiment 3, we used mixers and glmers to statistically analyse clicking latencies and correctness rates, respectively. The modelling procedure (inclusion of variables and model comparisons) was exactly the same as in Experiment 3. Overall, the results of the analysis of clicks (reaction times and correctness rates) were similar to the results obtained in Experiment 3: In Experiment 4, there were 97.2% correct clicks on the target (14 incorrect clicks in the early-peak condition, 8 in the medial-peak condition) in the experimental trials. The average response time to the visual target measured from the offset of the acoustic target was 557 ms (SD = 302 ms). Intonation condition did not have an effect on error rates (p > 0.19) or on reaction times (p > 0.88). An omnibus model for Experiments 3 and 4 that included data of both experiments did not show effects of experiment or intonation condition (all p > 0.12) and importantly no interaction between the two factors (p > 0.47).  

**Fixations**

*Visual inspection of evolution of fixations*

Figure 27 displays the grand average of the evolution of fixations in experimental trials to the four words on screen in the two intonation conditions. Comparable to Experiment 3, from the processing of the beginning of the target word onwards, fixation proportions to the target (Libelle) and the stress competitor (Libero) both further increased with similar slopes in the medial-peak condition (Figure 27, lower panel). In the early-peak condition, competitor fixations increased with slightly steeper slope than target fixations, Figure 27, upper panel). Comparing fixations in the early-peak condition across experiments (Experiment 3 vs. Experiment 4), competitor fixations seem to increase more moderately in the current experiment than

---

73 The best model fit for clicks (both correctness and latencies) was the model with random intercepts only, identical to Experiment 3 (see above for specification).
in the previous eye-tracking experiment (Figure 20, upper panel for Experiment 3; Figure 27, upper panel for Experiment 4).

Figure 27: Evolution of fixations to WSW-target (e.g., Libelle, dark blue line), SWW-competitor (e.g., Libero, red line) and the two distractors (e.g., Thymian ‘thyme’ (SWW), Safari ‘safari’ (WSW), light blue lines) in experimental trials for Experiment 4. The early-peak condition is shown in the upper panel, the medial-peak condition in the lower panel. Acoustical landmarks (dashed vertical lines) are shifted by 200 ms.

**Statistical analysis of competitor fixations across intonation condition**

Identical to Experiment 3, we first tested whether competitor fixations differed as a function of intonation condition during the segmentally ambiguous part of the target. Data analysis and statistical modelling were the same. Identical to Experiment 3, the final GAMM for Experiment 4 also included intonation condition as a parametric coefficient and a non-linear effect (smooth term) of intonation condition over time, as well as a random intercept for the event variable. Auto-correlation (rho = 0.62) was very similar to Experiment 3, see Figure 28.

Figure 28: Auto-correlation graph for the basic model for competitor fixations in Experiment 4 (basicT.gam). The height of the second line indicates the amount of auto-correlation at lag 1 (rho = 0.62).

The final model accounted for a 71.8% deviance, which is even slightly more than in Experiment 3. The coefficients of the final model are given in Table 13 while the
The effect of intonation condition over time is directly visualized in Figure 29 (i.e., the difference curve of competitor fixations, early-peak minus medial-peak condition).

Table 13: Final general additive mixed model for competitor fixations in Experiment 4. Part A: Estimate, Standard Error (SE), t- and p-value for the parametric coefficients. Part B: estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F- and p-values for the smooth term. Part C: Model specification of the final model (original variable names).

<table>
<thead>
<tr>
<th>Part A. Parametric coefficients</th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.484</td>
<td>0.1048</td>
<td>-4.617</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intonation condition: medial-peak</td>
<td>-0.098</td>
<td>0.1466</td>
<td>-0.666</td>
<td>0.5060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B. Smooth terms</th>
<th>EDF</th>
<th>Ref.df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random effect (intercept) for Event, s(event)</td>
<td>642.525</td>
<td>711.000</td>
<td>10.73</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time): early-peak</td>
<td>1.005</td>
<td>1.011</td>
<td>159.68</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time): medial-peak</td>
<td>3.206</td>
<td>4.114</td>
<td>26.13</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

| Part C. Model specification     | bam(e|logC ~ cond + s(TIMESTAMP, by = cond) + s(event, bs = 're'), rho = 0.62, AR.start = df_combined[df_combined$experiment == "Exp4",]$start_event, data = df_combined[df_combined$experiment == "Exp4",]), nthreads = 4, method = "ML") |

Figure 29: Difference curve in competitor fixations in early-peak condition minus medial-peak condition in Experiment 4. The grey band indicates the 95% CI of the mean difference. Values above zero indicate more competitor fixations in the early-peak condition. Conversely, values below zero indicate more competitor fixations in the medial-peak condition. The difference is significant if the 95% CI does not include zero. Window of significant difference: 1077-1080 ms.
Different from what was observed for Experiment 3, competitor fixations in the two intonation conditions did not differ significantly during the segmentally ambiguous part in Experiment 4, i.e., the 95% CI includes zero almost throughout the whole critical window. Note though that at the end of the analysis window (1077-1080 ms), the model indicated that competitor fixations were more frequent in the early-peak condition; an effect we consider to be negligible due its very short duration (only 3 ms).

**Statistical analysis of target fixations across intonation condition**

Next, we examined the effect of intonation condition on target fixations for processing of the segmentally ambiguous part. To this end, we fitted another general additive mixed model for empirical logits for target fixations under the exact same modelling conditions as described above. The final model contained intonation condition as parametric coefficient and a non-linear effect (smooth term) of intonation condition over time, as well as a random intercept for the event variable, and it explained 67.6% of the deviance. For this model, auto-correlation was also similar to the models before (rho = 0.64). We display the coefficients of the final model in Table 14 and visualize the difference curve of target fixations across intonation condition as estimated by the final model in Figure 30.

<table>
<thead>
<tr>
<th>Part A. Parametric coefficients</th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.877</td>
<td>0.0975</td>
<td>-8.993</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intonation condition: medial-peak</td>
<td>-0.008</td>
<td>0.1361</td>
<td>-0.057</td>
<td>0.9550</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B. Smooth terms</th>
<th>EDF</th>
<th>Ref.df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random effect (intercept) for Event, s(event)</td>
<td>625.705</td>
<td>711.000</td>
<td>8.472</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time): early-peak</td>
<td>1.002</td>
<td>1.005</td>
<td>44.871</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time, by condition, s(Time): medial-peak</td>
<td>2.621</td>
<td>3.372</td>
<td>14.493</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part C. Model specification</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bam(elogT ~ cond + s(TIMESTAMP, by = cond) + s(event, bs = ‘re’), rho = 0.62, AR.start = df_combined$df_combined$experiment == “Exp4”,]$start_event, data = df_combined$df_combined$experiment == “Exp4”,), nthreads = 4, method = “ML”)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 30: Difference curve in target fixations in early-peak condition minus medial-peak condition in Experiment 4. The grey band indicates the 95% CI of the mean difference. The difference is significant if the 95% CI does not include zero. Window of significant difference: NA.

**Combined analysis for Experiment 3 and Experiment 4 (Chapters 5-6) – Competitor fixations**

In order to examine whether the difference of competitor fixation across intonation condition and the difference of target fixation across intonation condition differs for the two experiments, we combined the data of Experiments 3 and 4 and conducted a pooled analysis. Statistically speaking, we tested for an interaction between intonation condition and experiment for competitor and target fixations, respectively, closely following the analysis steps for modelling interactions described in Wieling (2018). We first assessed competitor fixations across experiments (Experiments 3 and 4). To this end, we created a new variable (Expcond), which is the interaction of intonation condition and Experiment, resulting in four levels (Exp3-Early-peak, Exp3-Medial-peak, Exp4-Early-peak, Exp4-Medial-peak). This interaction variable was used instead of the two-dimensional intonation condition variable in the model. Similar to the model fitting for individual experiments, for the model of the combined data, we included a parametric coefficient for the interacting variable (Expcond, four levels), along with a random intercept for event. Further, for the interacting factor, nonlinear functional relations with the response variable over time were included using the smooth function. Again, an AR-1 correlation parameter was estimated to account for the auto-correlation in the fixation data (rho = 0.63). Then, we checked whether the model with the interacting factor (Expcond) is better than the model with a smooth term for intonation condition only. This was indeed the case, showing a significantly lower ML score than the model with intonation condition (22407.07 compared to 22432.13, p < 0.0001). Hence, it was necessary to distinguish the competitor activation between experiments.
To formally assess whether the difference in competitor fixations in the early-peak vs. medial-peak condition is different between the two experiments (Experiment 3 vs. Experiment 4), the model was re-specified by implementing binary difference smooths, following the description in Wieling (2018). Specifically, along with *experiment* as a parametric factor, we included a binary difference smooth distinguishing between the early- and medial peak condition irrespective of experiment (*IsEarly*) as well as a binary difference smooth that distinguished early-peak accents in Experiment 3 from all other conditions (*IsExp3Early*), see Table 15 for coefficients of the final model.

Table 15: Final general additive mixed model of combined analysis for competitor fixations in Experiments 3 and 4. Part A: Estimate, Standard Error (SE), *t-* and *p-*value for the parametric coefficients. Part B: Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), *F-* and *p-*values for the smooth term. Part C: Model specification of the final model (original variable names).

<table>
<thead>
<tr>
<th>Part A. Parametric coefficients</th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.632</td>
<td>0.1005</td>
<td>-6.288</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Experiment: Experiment 4 (with exposure)</td>
<td>0.047</td>
<td>0.1433</td>
<td>0.326</td>
<td>0.7440</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B. Smooth terms</th>
<th>EDF</th>
<th>Ref.df</th>
<th>*F-*value</th>
<th>*p-*value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random effect for Event (intercept), s(event)</td>
<td>1285.671</td>
<td>1441.000</td>
<td>9.524</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>s(Time): Exp3</td>
<td>2.383</td>
<td>3.043</td>
<td>32.945</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>s(Time): Exp4</td>
<td>4.029</td>
<td>5.136</td>
<td>19.763</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>s(Time): IsEarly</td>
<td>2.001</td>
<td>2.001</td>
<td>2.826</td>
<td>0.0593</td>
</tr>
<tr>
<td>s(Time): IsExp3Early</td>
<td>6.307</td>
<td>7.580</td>
<td>8.937</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

| Part C. Model specification      | bam(elogC ~ experiment + s(TIMESTAMP, by=experiment) + s(TIMESTAMP, by=IsEarly) + s(TIMESTAMP, by=IsExp2Early) + s(event, bs='re'), data=df_combined, rho=0.63, AR.start= df_combined$start_event) |

To illustrate, s(Time, by = IsExp3Early) represents the difference between the early-peak-vs.-medial-peak contrast in Experiment 3 versus that in Experiment 4, while s(Time, by = IsEarly) represents the difference between the early-peak-vs.-medial-peak contrast for Experiment 3. The difference of the difference in competitor fixations across experiments, s(Time, by = IsExp3Early) see Wieling (2018), is significant. This difference is directly displayed in Figure 31 where the difference in competitor fixations for the early-peak-vs.-medial-peak contrast is larger in Experiment 3 than in Experiment 4. For a short time (~ 920-950 ms), zero is not
included in the 95% CI of the difference curve of the difference, suggesting a significant interaction between experiment and intonation condition for competitor fixations.

![Graph showing the difference curve of the difference in competitor fixations in early-peak vs. medial-peak condition across the two experiments.](image)

Figure 31: Difference curve of the difference in competitor fixations in early-peak vs. medial-peak condition across the two experiments, i.e., difference between fixations in early-peak vs. medial-peak condition for Experiment 3 minus difference between fixations in early-peak vs. medial-peak condition for Experiment 4. The grey band indicates the 95% CI of the mean of the difference (across experiments) of the difference (in intonation condition).

**Combined analysis for Experiment 3 and Experiment 4 (Chapters 5-6) – Target fixations**

Identical to the analysis steps for the effect on competitor fixations across experiments, for target fixations, we also included the interaction variable (Expcond with four levels) along with a random intercept for event and a nonlinear functional relation with the response variable (elogs for target fixations) over time in an initial model (\( \rho = 0.63 \)). This model was then compared to a model that contained a smooth term for intonation condition only. Model comparisons revealed that the model with the interaction variable (Expcond) was preferred over the model with a simple factor for intonation condition, as indicated by a significantly lower ML score for the interaction model (22066.89 compared to 22084.61, \( p < 0.0001 \)). Thus, for target activation it is also necessary to distinguish between experiments.

Next, the interaction between experiment and intonation condition was assessed formally by re-specifying the model with binary difference smooths (Wieling, 2018), based on the modelling procedure described above. The final model included experiment as a parametric factor, a binary difference smooth distinguishing between the early- and medial peak conditions irrespective of experiment (IsEarly) and a binary difference smooth that distinguished early-peak accents in Experiment...
3 from all other conditions (IsExp3Early), see Table 16 for coefficients of the final model (for interpretation thereof see above) and Figure 32 for the difference of the difference in target fixations for the early-peak-vs.-medial-peak condition in Experiment 3 vs. Experiment 4. Different from competitor fixations, zero is included in the 95% CI of the difference curve of the difference throughout the whole analysis window, suggesting no significant interaction between the factors experiment and intonation condition for target fixations.

Table 16: Final general additive mixed model of combined analysis for target fixations in Experiments 3 and 4. Part A: Estimate, Standard Error (SE), t- and p-value for the parametric coefficients. Part B: Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F- and p-values for the smooth term. Part C: Model specification of the final model (original variable names).

<table>
<thead>
<tr>
<th>Part A. Parametric coefficients</th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.788</td>
<td>0.094</td>
<td>-8.407</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Experiment: Exp 4 (with exposure)</td>
<td>0.093</td>
<td>0.134</td>
<td>-0.692</td>
<td>0.489</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B. Smooth terms</th>
<th>EDF</th>
<th>Ref.df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random effect for Event (intercept), s(event)</td>
<td>1266.765</td>
<td>1441.000</td>
<td>8.176</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>s(Time): Exp3</td>
<td>1.026</td>
<td>1.050</td>
<td>58.052</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>s(Time): Exp4</td>
<td>2.941</td>
<td>3.773</td>
<td>13.005</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>s(Time): IsEarly</td>
<td>2.785</td>
<td>3.233</td>
<td>0.130</td>
<td>0.907</td>
</tr>
<tr>
<td>s(Time): IsExp3Early</td>
<td>5.320</td>
<td>6.467</td>
<td>5.620</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

| Part C. Model specification | bam(elogT ~experiment + s(TIMESTAMP, by=experiment) + s(TIMESTAMP, by=IsEarly) + s(TIMESTAMP, by=IsExp3Early) + s(event, bs='re'), data=df_combined, rho=0.63, AR.start= df_combined$start_event) |


Figure 32: Difference curve of the difference in target fixations in early-peak vs. medial peak condition across the two experiments, i.e., difference between fixations in early-peak vs. medial-peak condition for Experiment 3 minus difference between fixations in early-peak vs. medial-peak condition for Experiment 4. The grey band indicates the 95% CI of the mean of the difference (across experiments) of the difference (in intonation condition). If zero is not included, intonation condition and experiment interact.

6.3 Discussion

In the current chapter, we investigated the underlying mechanisms to the use of high f0 as a stress cue, focusing on the occurrence frequency of different pitch accent types. Recall that in Experiment 3 (Chapter 5), listeners fixated a segmentally overlapping SWW-stress competitor more when the auditory WSW-target was produced with an early-peak accent compared to a medial-peak accent. There was also a minor effect on target fixations, with fewer fixations to the target in the early-peak condition as compared to the medial-peak condition. In the current experiment (Experiment 4), listeners were exposed to utterances that only contained low-pitched stressed syllables produced by the same speaker they were presented with during the eye-tracking study. In the subsequent eye-tracking experiment, the stress competitor effect of Experiment 3 was not replicated: There was no effect of intonation condition on competitor fixations in Experiment 4 throughout the greatest part of the window of interest (shifted target word onset to shifted segmental U.P.). The short-lived and tiny difference in competitor fixations at the very end of the analysis window in Experiment 3 is probably due to strategic effects (note that we used an average U.P. across all trials, so some trials have an earlier U.P. than others).

Most importantly, in a direct statistical comparison between experiments, the effect of intonation condition on competitor fixations significantly differed in the two experiments, although the effect size of the difference was small. Different from Experiment 3, target fixations did not differ in Experiment 4 as a function of
intonation condition. A combined analysis revealed that the interaction between experiment and intonation condition for target fixations was not significant. The lack of interaction for target fixations is probably due to the effect on target fixations being small in Experiment 3. Hence, we need to conclude that target activation is not affected differently in the two experiments as a function of different pitch accent types. Only competitor fixations behave differently in the two experiments: While intonation condition affects competitor fixations in Experiment 3, it has no effect on competitor fixation in Experiment 4 after exposure to low-pitched accents.

In Experiment 4, we used a 3 min exposure phase prior to the eye-tracking study. The design followed a procedure that has been successfully employed in previous adaptation studies, i.e., to regional and foreign accents on the segmental level (e.g., Evans & Iverson, 2004; Grohe & Weber, 2016a; Grohe & Weber, 2016b; Witteman et al., 2015), to different speaking styles (e.g., Poellmann, Mitterer, & McQueen, 2014), and to suprasegmental variation in L2 speech (Reinisch & Weber, 2012). In each of these studies, an exposure phase of only a few minutes was sufficient to result in processing differences; the number of critical tokens was often even higher in our study than in these comparable studies. We thus assume that our exposure phase changed the relative frequency of high-pitched stressed syllables in favour of low-pitched stressed syllables, at least for the speaker participants heard in the experiment (cf. Xie & Myers, 2017, on speaker-specific adaptation). The absence of the stress competitor effect in the early-peak condition compared to the medial-peak condition in Experiment 4 shows that the frequency of occurrence of high-pitched stressed syllables affects stress processing. At the same time, it corroborates the plasticity of speech perception and listeners’ ability to generalize accentual realizations to different sentence types. We will address this issue in the General discussion.

Taken together, our finding suggests that it is not directly the acoustic salience of high pitch that makes f0 a stress cue (Option 1) but the frequent exposure to high-pitched stressed syllables in general (Option 2). Otherwise we would have expected the same stress competitor effect as in Experiment 3. Since we employed the same number of participants and stimuli in both experiments, the statistical power was the same in both experiments. Thus, we argue that the use of high f0 as a stress cue is determined by the frequency of occurrence of accent types in the (immediate) input, with listeners’ generating expectations on how stressed syllables are tonally realized.

Extrapolating from our frequency manipulation in the immediate input, we expect the effect of pitch accent type on stress processing (for both infants and adults) to be strongest in languages and/or varieties with a majority of medial-peak accents (H*-accents). For German, there are two studies that have analysed the distribution frequencies of different accent types in German, one on ADS (Peters
et al., 2005) and one on IDS (Zahner, Schönhuber, et al., 2016b). We will take a moment to outline the distribution frequencies of pitch accent types in German IDS and ADS, respectively. Note though that it is generally difficult to estimate the frequency of pitch accent types and compare them across languages as the numbers highly depend on the type of stimuli in the corpus.

Generally, we have two main reasons to assume differences in the distribution frequencies of pitch accent types across register (IDS vs. ADS) – an acoustic and a pragmatic one. First, it is known that IDS highly differs from ADS in its acoustic features. Overall, compared to ADS, IDS is acoustically characterized by a higher mean f0, higher absolute f0, larger f0 ranges, longer pauses, more use of whispering, a reduced speech rate, final syllable lengthening, and more peripheral vowel qualities, as has been shown for many languages (e.g., Fernald & Simon, 1984, for German, but see many others on a variety of different languages). These prosodic differences and the greater variability in f0 in particular (higher f0, more rises etc.), might influence intonational phonology, potentially leading to a different use of intonational categories in IDS compared to ADS. Second, IDS is pragmatically different from ADS. Conversations with young children are predominantly situated in the immediate present and involve daily routines and events, with objects often being visually present (e.g., Dominey & Dodane, 2004; Papoušek, Papoušek, & Symmes, 1991; Werker & McLeod, 1989). Further, IDS is a speech style that makes use of many repetitions (Fernald & Morikawa, 1993; Ferrier, 1978). Given that pitch accent types also cue information status, i.e., the given – new distinction (Baumann & Grice, 2006), these pragmatic peculiarities of IDS might skew the frequency distribution of pitch accent types in IDS. The following section will briefly outline the frequency distributions of different pitch accent types for German ADS and IDS based on the two corpus studies available.

For German ADS, Peters et al. (2005) analysed the distribution frequency of early-, medial-, and late-peak contours in utterances with one or more pitch accents in turn-internal or turn-final position in appointment-scheduling dialogues, in the Kiel Corpus of Spontaneous Speech (Vol. I and II; IPDS 1995, IPDS 1996), see Table 17 for an overview. As Table 17 indicates, for contours with more accents in a phrase, medial-peak contours occurred in 56% of the nuclear accents in turn-medial position (in 60% in turn-final position), late-peak contours occurred in 34% and 13%, respectively, and early-peak contours in 11% and 28%, respectively. For contours with one accent in a phrase, medial-peak contours occurred in 47% in turn-medial position (in 49% in turn-final position), late-peak contours occurred in 23% and 1%, respectively, and early-peak contours in 30% and 49%, respectively.
Table 17: Distribution of contours in German adult-directed speech; for phrases with more than one accent in a phrase (upper panel) and for phrases with only one accent in a phrase (lower panel), adapted from Peters et al. (2005), p. 12.

<table>
<thead>
<tr>
<th>PEAK TYPE</th>
<th>TURN-INTERNAL POSITION</th>
<th>TURN-FINAL POSITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early peak</td>
<td>107 (11.0%)</td>
<td>74 (27.6%)</td>
</tr>
<tr>
<td>Medial peak</td>
<td>543 (55.5%)</td>
<td>160 (59.7%)</td>
</tr>
<tr>
<td>Late peak</td>
<td>327 (33.5%)</td>
<td>34 (12.7%)</td>
</tr>
<tr>
<td>total</td>
<td>977 (100%)</td>
<td>268 (100%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PEAK TYPE</th>
<th>TURN-INTERNAL POSITION</th>
<th>TURN-FINAL POSITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early peak</td>
<td>363 (30.1%)</td>
<td>105 (49.5%)</td>
</tr>
<tr>
<td>Medial peak</td>
<td>562 (46.6%)</td>
<td>104 (49.1%)</td>
</tr>
<tr>
<td>Late peak</td>
<td>281 (23.3%)</td>
<td>3 (1.4%)</td>
</tr>
<tr>
<td>total</td>
<td>1206 (100%)</td>
<td>212 (100%)</td>
</tr>
</tbody>
</table>

Thus, in German ADS, medial-peak accents are most frequent irrespective of turn-position and number of accents in the phrase, followed in frequency by early-peak and late-peak accents. Early-peak accents are more frequent than late-peak accents in phrases with only one accent. In particular, early-peak accents are relatively common in turn-final position. Note that the analysis in Peters et al. (2005) consists of data from speakers from the north of Germany (Kiel area), which could potentially underestimate the number of low-pitched accents, as some authors have suggested L*/L*+H to be a common accent type in Southern Germany, i.e., in the Freiburg and Stuttgart area (Gilles, 2005; Kügler, 2007). To our knowledge however, there is no quantitative study to corroborate this claim.

For German IDS, Zahner, Schönhuber, et al. (2016b) analysed the distribution frequencies of pitch accent types (and boundary tones) using the KIDS Corpus. The multi-layered corpus includes IDS utterances from 16 different mothers directed to their infants younger than one year (in total 524 Intonational Phrases, IPs). Note that KIDS is composed of two main parts of data: a subset of German IDS utterances extracted from the CHILDES database (MacWhinney, 2000; 196 IPs, 675 words, 3min 28sec of speech) and recordings in the Baby Speech Lab (BSL) at the University of Konstanz and in one home environment (328 IPs). The mother-infant-dyads were recorded in typical play situations for around two minutes in the lab. Mothers received a picture book and some other toys, which they could use according to their infant’s interest. They were instructed to play with their infant as they would do at home.74 The distribution frequency of different pitch accent types in KIDS is shown in Table 18. Overall, the most frequent accent types are H* and L+H*, together occurring in 58% of the cases. These high-pitched

74TextGrids of the corpus are publicly available on https://www.ling.uni-konstanz.de/bsl/kids-corpus/ (last accessed: 28.02.2020) - sounds are available on request.
accents are most often followed by a low-boundary tone, as indicated in a combined analysis of the accentual tones and their surrounding in Zahner, Schönhuber, et al. (2016b). The authors argue that the picture book scenario, in which the mothers introduce new words to their infants, might have contributed to the high occurrence frequency of medial-peak accents (see Fernald & Mazzie, 1991, for a similar finding). The low-pitched monotonal accent \((L^*)\) is also common (18%) and is often followed particularly by a high boundary tone, see Zahner, Schönhuber, et al. (2016b). The high occurrence frequency of this low-rising pattern is attributed to the use of this particular pattern in many polar questions (e.g., Féry, 1993; Grice et al., 2005; Kohler, 2004; Von Essen, 1964) which frequently occur in the infants’ input (Cameron-Faulkner, Lieven, & Tomasello, 2003; Fernald & Mazzie, 1991; Newport, Gleitman, & Gleitman, 1977; Soderstrom, Blossom, Foygel, & Morgan, 2008; Toda, Fogel, & Kawai, 1990). Meanwhile, early-peak \((H+L^*)\) and late-peak accents \((L^*+H)\) scarcely occur in the corpus and hence need to be considered infrequent in German IDS.

Table 18: Distribution of GToBI pitch accent types in the KIDS corpus, split by dataset.

<table>
<thead>
<tr>
<th>GToBI Label</th>
<th>KIDS ((N = 832) accents)</th>
<th>CHILDES ((N = 311) accents)</th>
<th>BSL ((N = 521) accents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H^*)</td>
<td>31%</td>
<td>30%</td>
<td>32%</td>
</tr>
<tr>
<td>(!H^*)</td>
<td>8%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>(L^<em>H^</em>)</td>
<td>27%</td>
<td>20%</td>
<td>31%</td>
</tr>
<tr>
<td>(L^*)</td>
<td>18%</td>
<td>25%</td>
<td>13%</td>
</tr>
<tr>
<td>(L^*+H)</td>
<td>8%</td>
<td>11%</td>
<td>6%</td>
</tr>
<tr>
<td>(H+L^*)</td>
<td>6%</td>
<td>7%</td>
<td>5%</td>
</tr>
<tr>
<td>(H+!H^*)</td>
<td>2%</td>
<td>3%</td>
<td>1%</td>
</tr>
</tbody>
</table>

To sum up, \(H^*\)-accents are the most frequent accent type in German IDS and ADS, composing around 50-60% in the input. The given frequency distributions of accent types with many \(H^*\)-accents in German IDS and ADS might hence lead to the expectation that high \(f_0\) and the stressed syllable are linked. Due to the fact that frequency effects in processing are well attested (see introduction of this chapter), these numbers for German indeed qualify as an explanation for why high \(f_0\) was interpreted as a cue to stress, as has been demonstrated in the experimental Chapters 4 and 5 for German infants and adults, respectively. It is very plausible that listeners mentally link the concept of high pitch and lexical stress and use this cue for stress perception. In many cases, it will be a fortunate strategy. As our frequency manipulation shows, upon “counterevidence” for high \(f_0\) and stress belonging together, listeners seem to readily adjust their cue weights, taking the current input frequency of different pitch accent types into account and dismiss high \(f_0\) as a cue to stress.
To conclude, to our knowledge, this study is the first to address underlying mechanisms in the processing of individual stress cues. Results suggest that the use of high f0 as a stress cue can be modulated as a function of the occurrence of L*-accents in the immediate input. We thus argue that the use of f0 as a stress cue is not (or not only) driven by the acoustic salience of high f0, but more strongly by the learned association between high f0 and metrical stress. Even though the effects are small and the resulting conclusions hence need further experimental support, this study helps unravel processing mechanisms that underlie the effects we perceive at the surface. Beyond corroborating the role of frequency of pitch accents as an underlying mechanism, this study also provides important implications for the nature of f0 processing, i.e., whether there is a direct mapping between the acoustic cue f0 to metrical stress or whether the processing of f0 as a stress cue is mediated through phonological categories (accent types). It further bears interesting implications for cross-linguistic processing. We will elaborate on these implications in detail in the General discussion (Chapter 7), to which we will turn now.
7 General discussion

Chapter outline
This thesis investigated the effect of pitch accent type, i.e., phonological alignment differences between the f0 peak and the stressed syllable, on stress perception and consequently on processes in word recognition. Specifically, we tested how different pitch accent types modulate German infants’ stress-based segmentation behaviour and German adults’ lexical activation. Chapter 7 represents the General discussion of the thesis. We first summarize the results obtained in the experimental chapters, i.e., Chapters 4-6 (7.1). Based on these findings we then discuss selective topics and insights gained from the experimental part of the study, along with avenues for future research (7.2). Specifically, 7.2.1 focuses on the question of whether f0 can be seen as a stress cue by tracing the developmental trajectories of this question, its implications for the phonetics-phonology interface, and for cross-linguistic predictions. While 7.2.2 outlines the implications of our findings for infant word segmentation, 7.2.3 pinpoints the implications for adult word recognition. We will finally end with some concluding remarks that highlight the contribution of the thesis for the field of psycholinguistics and the phonetics-phonology interface (7.3).
7.1 Summary of experimental results

The effect of pitch accent type on stress-based segmentation in German infants (Chapter 4)

In a series of HPP experiments, we tested German 9-month-olds’ reliance on pitch and metrical stress for segmentation. In Experiment 1, we examined whether utterance-level pitch accent type affects infants’ ability to extract trochaic units from trisyllabic WSW-carrier words. We tested the extraction of a trochaic part-word (SW from WSW, e.g., *gune from Lagune), since it is known that infants can use probabilistic and other word boundary cues. Using embeddings allowed us to solely focus on the interplay between stress and pitch accent type for this task. During familiarization, the WSW-carrier words (embedded in sentences) were presented in three intonation conditions (manipulated between-subjects): one f0-stress alignment condition (medial-peak accent, (L+)*H) and two misalignment conditions (early-peak accent, H+L* and late-peak accent, L*+H). For test, all infants were presented with the trochaic part of the WSW-carrier word recorded with falling intonation (half of which were novel and half of which were familiarized). Results showed that infants only extracted the trochee in the medial-peak condition, i.e., if the stressed syllable was high-pitched (as indicated by an overall novelty preference). They failed in the two misalignment conditions that involved low-pitched stressed syllables (no preference). A control experiment (Experiment 1b) confirmed that the recognition effect observed in the medial-peak condition was not due to a pattern matching of the intonation contour between familiarization and test. Furthermore, no recognition of SW-units was observed when infants were presented with targets realized with a horizontally flipped medial-peak contour (HL*H) in familiarization (Experiment 1c), confirming that high f0 is a necessary cue for the use of the stressed syllable in stress-based segmentation for German infants.

In Experiment 2, we scrutinized the role of high-pitched syllables in the perception of stress and consequently the stress-based segmentation process, by investigating whether high pitch itself is a sufficient cue to stress for German 9-month-olds. Here, infants were familiarized with trisyllabic nonce WWS-words, e.g., *Rumila [ru.mi.ˈlaː] in sentence-contexts. The nonce words were either presented with an early-peak accent or a medial-peak accent. Test items were the last two syllables of the WWS-carrier word, but with the reverse stress pattern (SW, e.g., [ˈmiː.la] for *Rumila [ru.mi.ˈlaː]), recorded with varied intonation. To isolate the effect of f0 and exclude the possibility of the second syllable to be acoustically enhanced, the experiment was run with naturally produced stimuli in Experiment 2 and PSOLA-resynthesized stimuli in Experiment 2b (see "natural parallelism" of f0 and other cues in Kohler, 1991c, p. 144; for experimental evidence, Niebuhr, 2007a, pp. 117-150). For resynthesis, f0 contours were interchanged between conditions, i.e., the WWS-words recorded in the medial-peak condition received the early-peak
General discussion

For natural stimuli (Experiment 2), results indicated that the f0 peak on the second unstressed syllable in the WWS-word triggered a percept of stress and hence led to a metrical re-interpretation of the word-prosodic structure (WS realized with high-low, leading to a SW-percept): German 9-month-olds treated high-pitched syllables as word onsets even though they were unstressed (overall familiarity preference in Experiment 2). This was not the case with resynthesized stimuli (no preference in Experiment 2b), indicating that the isolated f0 cue was not sufficient to trigger stress.

The effect of pitch accent type on lexical activation in German adults (Chapter 5)

The second experimental chapter in this thesis investigated the effect of pitch accent type on stress perception and consequently on lexical activation in German adults. In Experiment 3, we tested whether f0 peaks on unstressed initial syllables in WSW-words (as in early-peak accents, H+L*) lead to the temporary activation of competitors with initial stress (SWW, similar to the metrical re-interpretation of the word-prosodic structure tested with infants, Experiment 2).

We used the VW Eye-Tracking Paradigm with four printed words on screen. The screen showed cohort pairs with different stress placement (e.g., SWW, Libero [ˈli.bə.ro] vs. WSW, Libelle [li.ˈbe.ʔel]) together with two distractors. In experimental trials, auditory instructions referred to the cohort member with penultimate stress (e.g., “Bitte klicke Libelle an”). Crucially, the WSW-target words (Libelle) were realized either with PSOLA-resynthesized early-peak accents (H+L*, f0 peak preceding the stressed syllable) or PSOLA-resynthesized medial-peak accents (L+H*, f0 peak on the stressed syllable), manipulated in a Latin Square Design. Fixations to the words on screen were extracted and competitor fixations across conditions were consecutively analysed using General Mixed Effects Models (GAMMs), a novel method that can model non-linear time-series data and allows for correction of auto-correlation, which is inherent in fixation data. The GAMM analysis indicated that in the time period in which information of the segmentally ambiguous part was processed, listeners fixated the stress competitor (Libero) more when the WSW-target (Libelle) was presented with an early-peak accent than when presented with a medial-peak accent: Non-intended competitors with a different stress pattern were more strongly activated in the early-peak condition via a metrical re-interpretation of the metrical strength relations in a word. Conversely, target fixations received less activation in the early-peak condition, compared to the medial-peak condition. Hence, our fixation data indicate that high-pitched unstressed syllables in form of H-leading tones in early-peak accents are temporarily interpreted as stressed and directly affect lexical activation. In sum, Chapters 4 and 5 showed f0-stress interference in both infant and adult processing (segmentation and lexical activation,
respectively). Why does high f0 play such an important role in stress processing? This question was addressed in Chapter 6.

Addressing mechanisms for the processing of f0 as a stress cue (Chapter 6)

To account for the observed effects of pitch accent type on stress-based segmentation and lexical activation (Chapters 4 and 5, respectively), we proposed two underlying mechanisms: Option 1 is that high-pitched syllables stand out perceptually (salience account). Option 2 is that listeners learned to associate high-pitched syllables with metrical stress, because of a frequent occurrence of H*-accents in the ambient language (frequency account). Both salience and frequency have been shown to explain many findings in infant and adult processing. Option 1 is clearly hard to manipulate in an experimental setup. Option 2, by contrast, can be influenced in an experiment.

To this end, in an exposure-test paradigm (Experiment 4), we put the frequency account to test and examined whether the weighting of the f0 cue for stress processing is affected by the frequency of occurrence of high- and low-pitched stressed syllables in the immediate input. For practical reasons, we focused on adults. An exposure phase was prefixed to the eye-tracking study by which we increased the occurrence frequency of low-pitched stressed syllables in the immediate input. In total, participants heard 120 L*-accents (in alternative questions, wh-questions, and contrastive topics) while they were engaged in a rating study. The eye-tracking part of the study immediately followed the exposure phase. As expected by the frequency account, the GAMM analysis of competitor fixations revealed that during most parts of the segmentally ambiguous window there was no effect of intonation condition (target fixation did not differ either). Importantly, a pooled analysis of the fixation data in Experiments 3 and 4 showed that there was an interaction between experiment (Experiment 3 vs. Experiment 4) and intonation condition (early- vs. medial-peak accent) for competitor fixations, which was short, but significant. This interaction confirms that intonation condition affected the fixations to the stress competitor differently in the two experiments (this was not the case for target fixations). We can thus infer that the amount of exposure to high-pitched stressed syllables (Option 2) – more so than the perceptual salience of high-pitched syllables (Option 1) – is a mechanism that explains why high f0 is readily interpreted as a stress cue. Taken together, we propose that the frequency with which different pitch accent types occur in spoken communication modulate the cue weights for acoustic cues to stress, here high f0.
7.2 Selected theoretical aspects and future lines of research

As summarized above, we showed that despite high f0 being an unreliable cue to the position of stress, it strongly influences infant and adult lexical processing. We argued that the learned association between high f0 and lexical stress due to a frequent occurrence of H*-accents in the ambient language can account for why listeners rely on this cue. In 7.2.1, we will consider the role of f0 as a stress cue in more detail. We will start out by tracing the developmental trajectory of high f0 as a stress cue, and then discuss the way high f0 is processed (either via a direct interpretation in which high f0 as an acoustical event is interpreted as metrical stress or via a mediation through phonological accent categories). In 7.2.2 and 7.2.3, respectively, we will discuss the implications that arise from our findings for lexical processing in infants and adults.

7.2.1 F0 as a stress cue?

Developmental trajectories

Our findings show that both groups (one at the very beginning and one further up on the developmental scale) are affected by (phonological) f0-stress alignment differences in processing. German infants only treat stressed syllables as stressed in metrical segmentation when high-pitched and additionally take naturally produced high-pitched unstressed syllables as stressed. Similarly, adults also straightforwardly interpret high f0 as stress and temporarily activate non-intended stress competitors. Hence, high f0 holds sway over the metrical relations in a word in both groups of listeners and determines whether a syllable is processed as stressed or not.

There is a developmental difference across age groups (infants vs. adults) regarding the use of the isolated f0 cue, as Chapters 4 and 5 revealed. Unlike for German infants, German adults’ perception of stress was affected by an isolated f0 cue on an unstressed syllable, i.e., independently from all other cues that supported the peak in natural productions. Infants only interpreted unstressed syllables as stressed when the f0 peak was naturally produced. Why would this be the case? A stimuli-based explanation according to which infants had difficulty in processing PSOLA-resynthesized stimuli (in contrast to adults) seems inadequate, given that other studies have shown successful processing of (re)synthesized stimuli, e.g., in grouping studies (Abboub et al., 2016; Bion et al., 2011), or intonation discrimination studies (Frota et al., 2014; Sundara et al., 2015). Accordingly, we need to assume that an isolated f0 cue on a genuinely unstressed syllable is processed differently by infants and adults, with adults making use of this cue as a cue to stress while infants do not (see also Bhatara et al., 2018). Hence, the adult perceptual system seems to be more flexible in adjusting the weight of stress cues to isolated cues, while infants are more strongly fixed to the acoustic composition that they encounter in natural
speech. Our results do not converge with the linguistic grouping studies in which high f0 was interpreted as the strong element by both infants and adults alike (Abboub et al., 2016; Bhatara et al., 2013; Bion et al., 2011), as well as with a study on the use of isolated spectral cues, which signalled stress in infants while adults only used them as a stress cue for segmentation when co-occurring with other cues (Thiessen & Saffran, 2004). Rather, our data show that f0 is most powerful as a stress cue for German 9-month-olds when it converges with other acoustic cues (be it “real” stress cues as in Experiment 1 or weaker peak supporting cues as in Experiment 2). In this regard, related infant studies have also highlighted the important role of f0 when converging with other cues. For instance, German 8-month-olds are able to distinguish different prosodic phrasings only when the prosodic phrase boundary is collectively signalled by phrase-final lengthening and a pitch movement (Holzgrefe-Lang et al., 2018; Wellmann et al., 2012). Similarly, the trochaic bias was also only evoked in 6-month-old German infants when stress cues converged (Höhle et al., 2009), but neither when signalled by a single cue (high pitch or intensity) nor a combination of these two cues, as preliminary data suggest (Bhatara et al., 2018). Our data reveal a difference between infants and adults, compatible with the results by Quam and Swingley (2014), mentioned earlier. In that study, pre-schoolers between the age of 2.5 and 5 years did not consider an isolated f0 cue to signal a stressed syllable, but relied on converging cues instead. This behaviour was different from adult listeners who adapted their cue weights and used both the isolated f0 cue and the converging stress cues to recognize words that differed in the position of stress, bunny / banana (Quam & Swingley, 2014). It needs to be pointed out though that our experimental materials differed from Quam and Swingley (2014): The isolated f0 cue was implemented on an ambiguous syllable (between stressed/unstressed) in the study by Quam and Swingley (2014), while our f0 manipulation (both in the infant and adult studies) resulted in an f0 peak on a genuinely unstressed syllable. Hence, we challenged the f0 cue more strongly in our design than Quam and Swingley (2014) did. In this regard, our finding that infants interpret an unstressed syllable as stressed if the f0 peak is naturally produced and hence acoustically enhanced appears even more astonishing.

We tested two listener groups: infants and adults. Evidently, infant listeners need plenty of time and experience to evolve to the level of adult speech processing. It is beyond the scope of this dissertation to provide a full developmental perspective of the trajectory of stress perception and its interaction with intonational cues. Clearly, future work needs to incorporate other listener groups, e.g., toddlers, school children, but also elderly people, in order to unravel the developmental scope of cue weighting of f0 and other stress cues for stress perception and processes such as speech segmentation and word recognition. Few studies have addressed the role of lexical stress for word recognition in children (e.g., Van
Alphen, De Bree, Fikkert, & Wijnen, 2007; Vihman et al., 2004; C. Wood, 2006), and even fewer the role of stress and intonation in child word recognition (Fikkert & Chen, 2011; Quam & Swingley, 2014). In sum, these studies show that toddlers and pre-schoolers use stress to recognize words and that mis-stressing affects word recognition (De Bree, Van Alphen, Fikkert, & Wijnen, 2008; Fikkert & Chen, 2011; Quam & Swingley, 2014; Vihman et al., 2004; C. Wood, 2006). Yet, the effect of mis-stressing varies: In some studies trochaic words are affected by mis-stressing (e.g., De Bree et al., 2008), while in other studies, iambic words are affected by mis-stressing (e.g., Fikkert & Chen, 2011). These mixed results for the role of lexical stress in child word recognition are indeed astonishing, given that infants at 12 months have been shown to learn new object-word pairings that only differ in the position of lexical stress SWW vs. WSW (Curtin, 2010). More research is definitely needed to clarify the picture on the effect of lexical stress in word recognition in toddlers. Little is known regarding the integration of lexical stress and intonation. Fikkert and Chen (2011) were, to our knowledge, the first and only to shed light on this question. In a visual fixation paradigm, they tested two age groups (14- and 24-month-old Dutch infants) on their word recognition abilities of trochaic (varken ‘pig’) and iambic (ballon ‘balloon’) familiar Dutch words when words were correctly vs. incorrectly stressed and when words were realized with correct vs. incorrect intonation. Their results revealed that the age groups differed in how and when stress and intonation were used: While the younger group did not integrate the two, 24-month-old children processed stress and intonation in an interactive way, making use of the pragmatic appropriateness of the contour. This was the case only for declaratives, but not for questions, however. Quam and Swingley (2014) also tested the role of f0 in word recognition (see details above), suggesting developmental differences in cue weighting of stress cues. Clearly, these studies can only be a first attempt to answer the question on how children integrate stress and intonation for lexical processing. The differences across age groups, e.g., in Fikkert and Chen (2011), as well as the inconsistent findings on the effects of mis-stressing still need to be resolved.

The weighting of intonational and stress cues during processes of word recognition might not differ only between infancy and adulthood (Quam & Swingley,

---

75 Correct intonation referred to the most typical intonation in the context, i.e., rising for polar questions (zie je een .... ‘Do you see a ....’) and falling for statements (kijk naar de/ het .... ‘Look at the …’); conversely, incorrect intonation referred to a falling intonation in questions and a rising intonation in statements. In total, the crossing of the two variables correctness of intonation and correctness of stress resulted in four different conditions: correct stress and correct intonation (CSCI), correct stress and incorrect intonation (CSSI), incorrect stress and correct intonation (ISCI), incorrect stress and incorrect intonation (ISII). The proportion of looking time to the target picture was analysed in two time windows (window 1: shifted target onset plus 1 second; window 2: shifted target onset plus 2 seconds). The authors established this second window as a response to generally longer processing times in younger children, i.e., the 14-month-olds in that study.
2014, for the difference between preschoolers vs. adults in the use of f0 as a stress cue; Thiessen & Saffran, 2004, for the difference between 9-month-olds vs. 12-month-olds and adults in the use of spectral tilt as a stress cue, see also our data on the use of the isolated pitch cue). It may also be subject to developmental changes beyond adulthood – in analogy to developmental differences in accent adaptation studies across the lifespan (see Cristià et al., 2012, for an overview). On the one hand, senior adults have been shown to use prosody for word recognition as young adults do (Wingfield, Lindfield, & Goodglass, 2000). On the other, it has recently been discussed that elderly people do not only have problems with processing fast speech (Janse, 2009) or speech in noise (Ben-David et al., 2011), but they also have difficulty discriminating f0 contrasts (e.g., Souza, Arehart, Miller, & Muralimanohar, 2011). Cue weighting differences in the perception of emotions between younger and older adults have also been reported (Schmidt, Janse, & Scharenborg, 2016): Compared to younger adults, older adults were less sensitive to intensity differences that cued arousal, while they reacted more strongly to f0 differences in cueing valence. The reliance on f0 was related to hearing ability (Schmidt et al., 2016). Conceivably, cue-weighting shifts in stress perception may also occur as a result of ageing. To fully understand the developmental trajectory of the processing of stress and its interplay with intonation, future research optimally needs to address the whole life span.

In this regard, a big methodological challenge will be to find experimental designs that are comparable across various age groups, starting in infancy and ending in late adulthood. One needs to bear in mind that studying stress perception in infants remains notoriously indirect. The two studies on the role of f0 in stress perception (Bhatara et al., 2018; Quam & Swingley, 2014), as well as our own HPP studies, examine the perception of stress via detour – by investigating infants’ (and/or children’s) behaviour in linguistic tasks (here intuitive biases, referential word recognition, or explicit segmentation, respectively). It might be the case that the weighting of the acoustic cues is different for different tasks. For instance, an isolated f0 cue might not be able to evoke the trochaic bias (Bhatara et al., 2018), but might allow for stress pattern discrimination, or to be used for speech segmentation. Similarly, different cues might be of more or less importance in word recognition or discrimination tasks in adults. Hence, it would be wise to find a paradigm that allows for similar tasks across different age groups (see the studies on linguistic grouping, for instance, Abboub et al., 2016; Bhatara et al., 2013; Bion et al., 2011). Another challenge will be to test different age groups with reasonable sample sizes. In response to this aspect, national and intonational multi-lab collaborations, comparable to the on-going ManyBabies initiative (ManyBabies Consortium, in press), might be a promising way to achieve this goal.
Beyond revealing underlying mechanisms for f0 processing, i.e., the frequency of occurrence of different pitch accent types, the findings of Experiment 4 (Chapter 6) also inform us about how high f0 is processed as a cue to stress. There are two theoretical options: first, a direct interpretation in which high f0 as an acoustical event is straightforwardly interpreted as metrical stress or second, a mediation through phonological pitch accent categories. That is, high f0 might lead to an interpretation of a high-pitched intonational category (H*) which would, in the sense of a prominence-cueing characteristic (Ladd, 2008), lead to the percept of metrical stress. In Experiment 3 (Chapter 5) high f0 triggered the percept of stress and temporarily activated stress competitors and shortly decreased target activation. In Experiment 4 (Chapter 6), by contrast, high f0 lost its power in cueing stress after participants received an increased exposure to low-toned phonological categories. It thus seems that high f0 does not directly lead to the perception of stress but to the perception of an H*-accent on the high-pitched syllable, which is in turn interpreted as stressed. If f0 directly led to the perception of stress we would not have seen an effect of the frequency manipulation in Experiment 4.

We presented literature in which various types of high f0 (peak, rise, high plateau) resulted in the perception of stress. It is yet unclear which acoustic factor (rise, peak itself) triggered the percept of stress in our studies: In the eye-tracking study, we only tested rising-falling contours, i.e., the f0 peak was always preceded by a noticeable f0 rise. Similarly, the conditions felicitous for infant segmentation were mostly realized with a rise towards the high-pitched syllable and a subsequent fall (albeit in some cases the rise was less steep). Acoustically, the rising component of a peak contour could be the perceptually most relevant part for the impression of stress in our studies, as already indicated by Fry (1958). At the same time, the f0 peak itself could be relevant, see Baumann and Winter (2018) on the importance of rising and high syllables for prominence perception. Under the assumption that f0 processing is mediated through phonological categories (which we put forward), it is effectively irrelevant which parts of the f0 peak led to the percept of metrical stress. Instead, we claim that the percept of metrical stress was evoked as soon as listeners interpreted a syllable as being realized with an H*-accent. If it holds that the processing of f0 as a stress cue is mediated, we should find f0-stress interference in all pitch configurations that can be interpreted as an H*-accent, i.e., not only in rising-falling movements that are most frequent in German ADS and IDS (Peters et al., 2005; Zahner, Schönhuber, et al., 2016b), but also for other H*-accents, such as H*-accents in hat patterns (C. Bartels, 1999, p. 85, on English; Truckenbrodt, 2011, p. 34, on German). We leave this empirical question for future research.
Cross-linguistic implications of the mediated processing of f0 as a stress cue

Under the assumption that f0 processing is mediated through phonological categories, we predict the tendency of taking high f0 as a stress cue to be strongest in languages with many H*-accents. For instance, in several varieties of English, (L+)H* accents are overwhelmingly frequent (occurring in 90% of the cases in American English, see Dainora, 2006, p. 118; in 72% of the cases in Australian English, see Fletcher & Stirling, 2014, p. 573; and in more than 90% of the cases in British English (except for Belfast region), see Grabe, 2004, p. 14). In this regard, the stress competitor effect observed with German adults has recently also been replicated with Australian English listeners (Zahner et al., 2017). In that study, Australian English listeners also associated high pitch with metrical stress and hence inadvertently activated words with the wrong stress pattern (SWW, e.g., musical) when presented with WSW-words (e.g., museum) with an early-peak accent. In future studies, we also plan to test listeners in related languages which have higher proportions of L*-accents, e.g., Swiss German (Fitzpatrick-Cole, 1999; Leemann, 2012), or Indian English (Pickeing & Wiltshire, 2000) to see whether there is indeed a lack of the stress competitor effect (statistically speaking, we will test for an interaction between language and intonation condition).

Our line of argumentation also offers a new perspective on cross-linguistic differences in the results of correlation analyses on adult stress perception and different acoustic cues in psycholinguistic studies. So far, psycholinguistic studies on the processing of stress have not manipulated intonation contours, i.e., pitch accent types. However, there are a small number of studies that have conducted post-hoc correlation analyses between acoustic cues to stress and the observed lexical-decision responses or fixations (Cutler et al., 2007; Reinisch et al., 2010). The results revealed differences in the use of stress cues across languages: English listeners’ behaviour correlated only with f0, the cue that showed the largest effect size for the difference between the stressed and the corresponding unstressed syllables (Cutler et al., 2007, p. 1914). Dutch listeners’ behaviour, on the other hand, correlated with duration and intensity, but not with f0 in Reinisch et al. (2010), despite a larger effect size for f0 than for duration and intensity in the acoustics. These cross-linguistic differences have largely been explained in terms of language-specific cues to stress (Cutler & Pasveer, 2006): It has been argued that Dutch listeners strongly rely on suprasegmental cues, while English listeners rely on vowel quality (or f0, as the most salient cue, if vowel quality is not informative (e.g., Connell et al., 2018; Cutler, 2012, p. 235)).

Our findings provide an alternative explanation for these cross-linguistic differences: Compared to many English varieties with an abundance of H*-accents (Dainora, 2006; Fletcher & Stirling, 2014; Grabe, 2004), high-pitched stressed syllables might be less frequent in Dutch. Although there are (yet) no quantitative
studies for the distribution of pitch accent types in Dutch (Miriam Ernestus, p.c),
descriptions of Dutch intonation claim that the hat pattern is the most frequent
contour in declaratives (Cohen & t'Hart, 1968). The hat pattern usually consists of
a rise and a downstepped fall on the nuclear syllables, a nuclear contour that is
acoustically and perceptually close to the early-peak contour. This default pattern
leads to a high proportion of low-pitched stressed syllables overall. To corroborate
this claim, we analysed the 88 declarative filler sentences from the experiments re-
ported in Braun, Dainora, et al. (2011), a corpus that was readily available. In nuclear
position, the downstepped fall occurred in more than 90% of the sentences. In
prenuclear accents, the late-peak accent was the most frequent accent type (38%),
followed in frequency by !H* (24%) and H* (15%). The high number of low-
pitched stressed syllables would explain why Dutch listeners do not rely on H* as a
stress cue as strongly as English listeners do. From a learner’s perspective, such
cross-linguistic differences in the distribution of pitch accents may pose challenges.
For instance, given that American English listeners encounter a very high number
of high-pitched stressed syllables (Dainora, 2006), we predict interference in L2-
processing when learning languages with a high proportion of L*-accents, as in
Dutch or Swiss German (Fitzpatrick-Cole, 1999; Leemann, 2012). Again, these pre-
dictions inspire interesting avenues for future research.

After having discussed the developmental trajectories of f0 as a stress cue as
well as the way it might be processed (direct mapping between acoustic cues to
stress or a mediation via phonological categories), we now turn to the processes
that were in the focus of the current dissertation, infant word segmentation and
adult lexical activation. We will highlight the implications that our findings bear for
the respective processes, along with future lines of research.

7.2.2 Implications for infant word segmentation
Knowing about factors that promote word segmentation abilities has become in-
creasingly relevant for acquisition research as early language abilities have recently
been demonstrated to predict later language development (e.g., Junge & Cutler,
2014; Kooijman et al., 2013; Newman et al., 2006; Newman et al., 2016; Singh et
al., 2012): Infants with advanced segmentation skills develop larger vocabularies
(Junge & Cutler, 2014, for a meta-analysis, showing that the effect size of these
studies is moderate, 0.33 95% CI [0.17;0.48]). Hence, segmentation skills are indic-
ative of an infant’s linguistic development and that way may become relevant for
early interventions. It is hence of great importance for developmental and clinical
researchers to identify factors that guide infant word segmentation as well as to
study factors that cause mis-segmentation.
Discovering underlying factors in metrical segmentation

Regarding the factors that guide infant word segmentation, our study sheds light on the interplay between stress-based segmentation and intonation. We extend the finding by Männel & Friederici (2013) by showing that only high-pitched accent types are felicitous for the extraction of trochees, rather than accentedness in general. As mentioned earlier, IDS contours with exaggerated f0 excursions, i.e., modulations in scaling, have recently also been shown to be beneficial for segmentation (Floccia et al., 2016; Schreiner & Mani, 2017; Thiessen et al., 2005). These findings might speak in favour of a salience account that helps infants discover units more easily due to the acoustic enhancement. In these studies, the IDS productions showed larger f0 excursions, i.e., phonetic modifications in pitch, along with durational differences. Crucially, our manipulation involved phonological intonation contrasts. The durational structure and f0 excursions were similar across conditions, but differed in the alignment of the f0 peak with respect to the stressed syllable. In our study, infants were successful in extracting trochees with ADS stimuli (Mani & Pätzold, 2016, on the ability to reliably segment speech in ADS at the age of 16 months) and were strongly driven by the f0 cue. Our data clearly demonstrate that these alignment differences, which lead to different phonological pitch accent types, are relevant for German infants’ stress-based segmentation behaviour. In all our conditions the f0 excursion was comparable in its extent, but segmentation success was only apparent when syllables were high-pitched. Therefore, we argue that it is the f0 modulations on the horizontal axis, i.e., phonological alignment differences, that drive infant speech segmentation rather than f0 modulations on the vertical axis, i.e., phonetic scaling differences (albeit whether or not the f0 movement needs to be of a certain excursion to prompt successful segmentation is an open question left for future research).

Supporting our argument is a very recent corpus analysis that investigated whether word form learning may benefit from the nature of different speech registers (IDS vs. ADS, Guevara-Rukoz et al., 2018, on Japanese). In detail, Guevara-Rukoz et al. (2018) showed that acoustically, words are less distinguishable in IDS than in ADS, but that IDS contains more distinct words due to a large number of onomatopoetic expressions occurring in this register. Combining the metrics of the two levels (acoustic and phonological, i.e., segmental changes to transform one word into another) revealed a small but significant advantage in discriminability for word forms in ADS over IDS (measured with ABX discrimination scores, Guevara-Rukoz et al., 2018). This input-based argument is in line with recent behavioural evidence on the trochaic bias in American English (Wang, Lee, & Houston, 2016). Specifically, Wang et al. (2016) demonstrated that the preference for the trochaic stress pattern is evident in ADS but not in IDS, most likely due to the contrast between the two syllables, stressed vs. unstressed, being acoustically reduced in...
IDS, as compared to ADS (as their post-hoc acoustic analysis of the stimuli revealed). Hence, IDS might not always be beneficial, especially when acoustic distinctiveness is deteriorated. Taken together, our study offers a new perspective on infant stress-based segmentation, highlighting that it is not IDS or accentedness in general that drives segmentation (at least for German infants), but the alignment of f0 in respect to the stressed syllable.

*Ecological validity: Segmentation in the wild*

To scrutinize the role of metrical stress and f0 as a stress cue, we made two compromises to the experimental material. First, we tested for the extraction of embedded units, i.e., units that did not correspond to meaningful words, (e.g., *gune* from *Lagune*). This was necessary in order to not provide infants with other word boundary cues, which German infants can use at that age (e.g., Höhle & Weissenborn, 2003). As mentioned in Chapter 4, critics may argue that the current experimental design (extraction of embedded trochees) effectively tested mis-segmentation. Although the extraction of an embedded trochee in a WSW-carrier (or a perceived WSW-carrier) is considered a “failure” from an adult’s perspective, from the infant’s perspective, however, we argue that these “failures” are unlikely to harm language acquisition in any sense. WSW-words scarcely occur in German IDS, comprising of less than 5% of accented words in the KIDS Corpus (Zahner, Schönhuber, et al., 2016b); hence, instances in which these “failures” occur in real life are rare. On the other hand, trochaic words preceded by a weak syllable, e.g., an article, such as *die Mama* ‘the mummy’, *der Papa* ‘the daddy’, *die Katze* ‘the cat’, are highly frequent; 46% of the accented words followed this prosodic structure in Zahner, Schönhuber, et al. (2016b). The frequency distributions of word-prosodic structure in IDS demonstrate that the extraction of the SW-part of a WSW-carrier does not harm or complicate first language acquisition: On the contrary, it is beneficial in most cases where infants are confronted with this pattern in the real world. In fact, it helps them successfully extract, e.g., *Katze* from a sequence *die Katze*. Thus, relying on the MSS seems to be a useful first heuristic until infants have learned to integrate other segmentation cues. Second, we also tested whether high f0 leads to a re-interpretation of the metrical structure of a word, which can result in the extraction of a unit not corresponding to the intended target. The configurations used in the experiments, however, are very infrequent in German IDS (only 6% of the accents are early-peak accents, Zahner, Schönhuber, et al., 2016b) which is real-world segmentation is very unlikely to be hampered.

In spite of these compromises, using these configurations (embeddings and rare pitch accent types) allowed us to unravel underlying heuristics and show that high f0 guides segmentation, questions that could not have been addressed otherwise. In fact, our study was explicitly designed to shed light on the underlying
processes and heuristics in metrical segmentation as little is known about which cues guide this process (see Gambell & Yang, 2006, p. 20, for discussion). Our findings provide evidence for high f0 as a driving force in stress-based segmentation: Our data clearly speak in favour of a segmentation strategy that is more strongly driven by high f0 than by metrical cues (duration, intensity). In addition to nominating the driving force, we also explain why f0 is interpreted as a stress cue: the frequency of occurrence of H*-accents in the input. We therefore provide empirical support for speculative claims in the literature that “the identification of stresses involves, beyond an obvious perceptual component, certain cognitive/structural representation of speech and subsequent computations, which may be domain-specific phonological knowledge” (Gambell & Yang, 2006, p. 20). By highlighting the power of high f0 and the frequency of H*-accents as an explaining factor, this dissertation contributes to a deeper understanding of stress-based segmentation.

Integrating different strategies and cues

Clearly though, relying exclusively on high f0 (or H*-accents) to localize word onsets is insufficient and can even be unsuccessful in rare instances of real-world segmentation. Instead, infants must learn to extract word forms from speech which are meaningful (e.g., Graf-Estes, 2012; Graf-Estes, Evans, Alibali, & Saffran, 2007; for overviews see E. K. Johnson, 2016; Saffran, 2014). Successful segmentation, in the sense that “meaningful” units are extracted, can only be achieved by the integration of different strategies (Jusczyk, Houston, et al., 1999). In this regard, stress-based segmentation, along with the use of closed class elements (Höhle & Weissenborn, 2003), transitional probabilities (Saffran, Aslin, et al., 1996), allophonic cues (Jusczyk, Hohne, et al., 1999), familiar words (Bortfeld et al., 2005; Sandoval & Gómez, 2016) etc., need to be combined. In the last three decades, a variety of computational models have attempted to formally model infant word segmentation and the integration of different strategies (e.g., Brent, 1999; Cairns, Shilcock, Chater, & Levy, 1997; Çöltekin, 2017; Elman, 1990; Fleck, 2008; Gambell & Yang, 2006; Larsen, Cristià, & Dupoux, 2017). Gambell and Yang (2006), for instance, argue that transitional probabilities and a unique stress constraint (one stress in each word) can bootstrap lexical development. Other computational models have shown that segmentation based on phonotactics alone is also possible (e.g., Cairns et al., 1997). However, Cairns et al. (1997) argued, based on their neural network model, that an early sensitivity to sequential phonological probabilities might help infants to discover the use of more sophisticated word segmentation strategies such as the MSS. Later on, an integrated approach of phonotactics and rhythmical cues seems to be most beneficial (see Morgan & Saffran, 1995, for behavioural evidence showing reliance on rhythmic cues at 6 and integration of cues at 9 months). Using
a different paradigm, namely word onset priming. Becker, Schild, and Friedrich (2018) recently highlighted infants’ sensitivity to different phonological structures in finding word onsets: In their EEG study, they found stress priming (without phoneme priming) at the age of 3 months, the reverse, i.e., phoneme priming (without stress priming) at the age of 6 months, and an integration of the two (stress and phoneme priming) at the age of 9 months. Clearly, the exact trajectories of the integration of different cues need to be identified (for different languages, and different cue combinations).

Based on our findings, we propose that in the stages of word segmentation assessed in the present dissertation, a pitch-guided MSS is the essential mechanism, at least in German. From a broader perspective, we highlight the importance to integrate other strategies and cues in order to master real-world segmentation and to arrive at “real word” components. To conclude, in order to gain a complete picture on the integration of different strategies in infant word segmentation, there is still a long way to go for both infant researchers and computational modellers – a way on which deep learning systems may somewhen provide more insights.

7.2.3 Implications for adult word recognition

Relevance and generalization of our findings

Regarding lexical activation in adults, we find that listeners immediately exploit the pitch cue in the signal for word recognition, in line with previous work on the immediate exploitation of suprasegmental stress cues (Connell et al., 2018; Jesse et al., 2017; Reinisch et al., 2010) or pitch accents for reference resolution (e.g., Dahan et al., 2002; Ito & Speer, 2008; Weber et al., 2006). In our case, pitch accent type influenced lexical processing, such that listeners over-generalized high f0 to be an indicator for stress – even when realized on genuinely unstressed syllables (this interpretation is likely mediated through accent categories, see argument above). Consequently, stress competitors were more strongly activated in the early-peak condition compared to the medial-peak condition, and target fixations also received fewer fixations during small parts of the ambiguous time frame.

Whether or not the observed stress competitor effect – along with the decrease in target fixation – also transfers to other accent types, e.g., L*+H, remains a question that we plan to pursue in future studies. If the stress competitor effect indeed generalizes, as we would predict, then SWW-words realized with a late-peak accent (L*+H) are expected to activate WSW-competitors more strongly than in a control condition. Such cohort pairs require a late uniqueness point, i.e., the end of the second syllable (as this is the syllable where the f0 peak can interfere the earliest in the critical late-peak condition). It will be challenging to generate enough cohort pairs that meet this requirement.
As high f0 is immediately exploited as a stress cue in our eye-tracking study, we would predict that German listeners also exploit stress cues to constrain lexical access in another experimental paradigm, the CMFP (introduced in Chapter 3), such that stress-matching primes would cause facilitating effects while stress-mismatching primes would generate inhibitory effects (see N. Cooper et al., 2002; Donselaar et al., 2005; Soto-Faraco et al., 2001, on English, Dutch, and Spanish, respectively). Experiments allowing for a direct comparison between German listeners and other listeners from West-Germanic languages are currently on-going (Anne Cutler, Jenny Yu, p.c.). Beyond this, given the power of f0 in cueing stress in our paradigm, we would additionally predict that unstressed syllables with an f0 peak (e.g., the first syllable in a WSW-word with an early-peak accent), prime a target with initial stress. Whether or not this priming effect would be dependent on phones is an open issue (Schild, Becker, & Friedrich, 2014a; 2014b, on phoneme-free prosodic representations). In case it would be phoneme-free, we would expect the first syllable in a WSW-word with an early-peak accent (e.g., unstressed [li] excised from [li.'be.lda]) to prime an initially stressed target, both segmentally matching (e.g., Libero ['li.bə.ro]) and segmentally mismatching (e.g., Kaviar ['ka.vi.ə] ‘caviar’, Claudia Friedrich, p.c.). We will leave this issue to future research, too.

Our main prediction was based on the activation of (non-intended) competitors – as a function of pitch accent type, e.g., the activation of Libero when Libelle was presented. Naturally though, the increase in activation of the stress competitor was temporarily restricted, given that segmental information revealed the intended target soon (at the onset consonant of the second syllable at the latest). Critics may argue that the detrimental effect of pitch accent type on word recognition we reveal in our experiment is rather small in its extent. Nevertheless, we find significantly increased competitor activation and potentially higher processing costs based on utterance-level intonation. The fact that German is a language in which pitch does not distinguish between the meaning of words makes the finding more astonishing.

Implications for current models
The finding that pitch accent type affects lexical activation in a language in which f0 is not contrastive poses questions for current models of spoken word recognition (e.g., McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008; see Chapter 3), as they do currently not account for suprasegmentals. It is highly likely though that listeners analyse a combination of spectral (Allopenna et al., 1998; Dahan et al., 2001; McQueen & Viebahn, 2007; Soto-Faraco et al., 2001; Tanenhaus et al., 1995), suprasegmental stress cues (Connell et al., 2018; N. Cooper et al., 2002; Donselaar et al., 2005; Jesse et al., 2017; Reinisch et al., 2010), and intonational cues (see our findings) as soon as they surface in the signal. Lexical activation is
conceivably modulated as a function of the interplay between these components.

How could one account for utterance-level intonation in such a model? Currently, the architecture of these models renders it difficult to integrate suprasegmental cues, as their input does not account for acoustic details. Note though that, as has been mentioned earlier, Shortlist (Norris, 1994) includes a metrical segmentation mechanism, however not in form of acoustic details. Our experimental results call for the inclusion of these cues in current models, suggesting a mechanism that evaluates stress cues depending on pitch accent type. From the perspective of (German) listeners, high-pitched syllables bear a strong potential to signal stress. Adult listeners might take high-pitched syllables (irrespective of whether they are stressed or unstressed) as a cue to stress and hence to word onsets.

Regarding the competition process, if a WSW-word, e.g., *Libelle* [li.'be.la], is realized with an early-peak accent, i.e., if it contains a high-pitched unstressed initial syllable, competition might be increased as SWW-words, e.g., *Libero* ['li. be.ro], will also engage in the competition process (which is not the case when the WSW-word is realized with a medial-peak accent). Hence, the amount of competition in a model would have to be computed based on the intonational configuration with which a word is realized. In this regard, the incremental nature of word recognition might be a challenge for the modelling of suprasegmental, such as duration, intensity, and pitch. They might not be evaluated in a segment-wise manner but based on larger units, e.g., syllables. As mentioned earlier, Shuai and Malins (2017) recently suggested an extension of TRACE for Mandarin Chinese, which at the level of the input decodes Mandarin phonemes and lexical tones. In particular, pitch height and pitch slope are modelled as distinguishing features of lexical tone (Shuai & Malins, 2017, p. 232, for a detailed description). Specifically, the $f_0$ excursion of the four lexical tones in Mandarin Chinese (time-normalized) was divided into five levels of $f_0$ height. Similarly, $f_0$ slope was divided into three levels, resulting in 15 unique combinations of height and slope. Furthermore, Shuai and Malins (2017) assigned a combination of phoneme code and tonal code to syllables, with five time-intervals per syllable. Subsequently, a simulation was run with the model on two psycholinguistic phenomena, one on differences in the timing of resolution of tonal contrasts, and one on the interaction between syllable frequency and tonal probability. The simulation output was comparable to the behavioural data in a Visual-World Eye Tracking experiment with human listeners, suggesting a successful implementation of tonal cues. Clearly though, tonal cues play a different role in a tone language such as Mandarin Chinese than in an intonation language. Yet, this model could function as a role model for the implementation of intonational cues. It is beyond the scope of this thesis, however, to make concrete suggestions for how suprasegmental stress and intonation cues could be implemented in current models of spoken word.
recognition. Future research clearly needs to determine the weighting and temporal integration of segmental, suprasegmental stress, and intonation and model them to even better understand spoken word recognition. In addition, it is likely that there are other relevant cues that yet have to be identified (e.g., visual cues, see Jesse & McQueen, 2014; see Jesse et al., 2017, for discussion).

**Contributing to the debate on the nature of representations**

Our findings may also contribute to the debate on the nature of lexical representations. In Chapter 3, we argued that there is abundant evidence for the storage of detailed representations that may contain information about fine-acoustic details, such as voice, speaking rate, but also about social or contextual cues, such as word age or regional associations (e.g., Bradlow et al., 1999; Hay & Drager, 2010; Mendoza-Denton et al., 2003; Pierrehumbert, 2016; A. Walker & Hay, 2011). While these findings support exemplar-based theories, there is also evidence emphasizing listeners’ ability to adapt and subsequently generalize newly learned associations to words or speakers never heard before (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004; Hanulíková & Ekström, 2017; E. K. Johnson et al., 2018; McQueen et al., 2006; Norris et al., 2003), which supports abstract representations. How can our fixation data contribute to this debate? Generally speaking, we provide evidence for an effect of pitch accent type on lexical activation in German adults. Since pitch accent types do not distinguish lexical meaning in German but serve a communicative function (different from tone languages or pitch accent languages), one might expect efficient lexical processing to abstract away from intonational properties of the word. In Chapter 4, we have argued that on the way to become proficient language users, infants learn to generalize over different voices and different pitch levels (Houston & Jusczyk, 2000; Singh et al., 2008) and also over different pitch contours (falling vs. rising, see findings in Experiment 1b). As a result, infants’ representations become more and more abstract. In an abstractionist view, it seems plausible that intonational properties are not part of adult representations in intonation languages. From an exemplar-based perspective, by contrast, it appears very likely that listeners also store properties of the intonational contour together with the word (even though research on the storage of intonation is limited, but see Calhoun & Schweitzer, 2012; K. Schweitzer et al., 2015; K. Schweitzer, Calhoun, et al., 2010; K. Schweitzer, Walsh, Möbius, & Schütze, 2010). For instance, K. Schweitzer et al. (2015) report on three corpus experiments that provide evidence for frequency-based lexical storage of intonation. Analysing a dataset comprising 2587 falling accents (H*L in their notation) and 5378 rising accents (LH* in their notation) extracted from the DIRNDL corpus (Eckart, Riester, & Schweitzer, 2012), K. Schweitzer et al. (2015) show that the frequency of occurrence of the combination of a pitch accent type and a word influences the f0 excursion of the
respective accent (the more frequent a combination, the larger the f0 excursion). K. Schweitzer et al. (2015, p. 77) provide an exemplar-based explanation for this finding, for which, as the author state, the direction of the effect is rather surprising given that one would assume a reduction (of excursion) with increasing frequency. We will briefly summarize the argument provided in K. Schweitzer et al. (2015, p. 77): According to the authors, before producing a word, the speaker selects the exemplar from the memory that best matches the communicative function she wants to convey. As the utterance the speaker plans requires the general communicative aim “prominence”, it is very likely that a word that carries a pitch accent is selected. This (pitch-accented) token the speaker produces enters the storage system. Some of the subsequent productions will have slightly smaller, some slightly larger f0 excursions, compared to the f0 excursion in the token that had previously been selected. The tokens with a larger f0 excursion are likely to be selected again because they fulfil the communicative aim “prominence.” As a result, with an increased occurrence frequency, the f0 excursion gets larger (K. Schweitzer et al., 2015, p. 77). In sum, this study reveals that listeners might indeed store intonational properties of words – even though the concrete nature of the stored parts is yet to be explored.

In this vein, parts of our results might also be readily interpreted in an exemplar-based view. Given the high frequency of occurrence of high-pitched accent types, i.e., H*-accents (Peters et al., 2005; Zahner, Schönhuber, et al., 2016b), listeners encounter more H*-accented instances of a given word (e.g., Libero [ˈli.fbə.ro]) than productions of this words with other pitch accent types. Accordingly, many H*-accented instances of this word might be stored. Upon hearing a high-pitched syllable [ìi] those exemplars that match the signal will be activated and an echo will be generated (see word recognition process in MINERVA 2 described in Chapter 3). Very likely, the echo, which is a generalization of all activated exemplars, might contain an H*-accent, since this is the most frequent pitch accent type and thus a logical generalization. Hence, instead of interpreting the high-pitched syllable [ìi] as an unstressed syllable with an H-leading tone (which would match Libelle [li.ˈbe.lə] with an early-peak accent, H+L*), the high-pitched syllable [ìi] might be interpreted as bearing an H*-accent and consequently as being stressed (which could explain the activation of Libero [ˈli.fbə.ro]). Hence, our findings might suggest that listeners store words together with their intonational contour.

At the same time, the findings in Experiment 4 (exposure-test paradigm) indicate that listeners are capable of generalizing the high occurrence frequency of low-pitched accented syllable (during exposure) to a new set of words and new sentence types in the eye-tracking study (during test). After exposure to L*.-accents in the immediate input the erroneous activation of the stress competitor is reduced, which
could be a result of many L*-exemplars in the memory which might prevent the echo to dispose of an H*-accent (as an L*-accent would be generalized based on the exemplars) and consequently avoid the activation of a stressed syllable and in turn the activation of a word with initial stress (*Libero*). Yet, the perceptual system of listeners seems to be flexible enough to generalize the newly encountered L*-exemplars to words not part of the exposure phase – even to new sentence types. Under the assumption that each word is stored with its intonational contour, the manipulation of the input frequency would not have been successful in our study (but only if we had used the exact same words and sentence types for exposure and test). Intonational contours and phonemic structure can thus not be inextricably linked in memory, but show flexibility for generalizations (see Schild et al., 2014a; 2014b, on phoneme-free prosodic representations). Hence, an account purely based on exemplar-based storage of a word and its intonational contour is not adequate. Rather, our findings could be taken as speaking in favour of both lexical storage of intonation and, at the same time, in favour of generalization processes. Clearly, future research is needed to obtain more detailed insights in the nature of representations and the role of intonation therein.

*Beyond the lexical level…*

Beyond the processing and storage of spoken words and models thereof, our findings on adult lexical activation also have implications for processing at the utterance level. Specifically, misalignments between f0 peak and the stressed syllable may lead to changes in the interpretation of focus structure, with early-peak accents not just resulting in (temporary) lexical misinterpretations, but in a misinterpretation of information structure. For instance, when an early-peak accent results in an f0 peak on a preceding word, this word may be interpreted as accented (H*), see Dilley and Heffner (2013) for a similar suggestion. An example of such a potential misinterpretation of information structure is shown in (5) and (6) where the f0 peak on the preposition or the adjective may be temporarily interpreted as accented, which may signal a contrast (e.g., *flying from Paris* in (5) or *a small pigeon* in (6), underlining represents accented words).

(5) He flew to *Paris*

(6) They saw a *big pigeon.*
Interestingly, many native speakers of English\textsuperscript{76} often accentuate prepositions, e.g., *I’ll present this to you*, which from an information-structural perspective appears inappropriate. Diachronically, this tendency may have emerged from a misinterpretation of the accented words and transferred into more recent versions of English. Admittedly, this assumption might seem daring at first glance and needs to be further investigated, but based on the findings on lexical processing it is not an inconceivable one.

### 7.3 Conclusion

In conclusion, this dissertation is relevant to both psycholinguistic research and the interface between phonetics and phonology. From a psycholinguistic perspective, it contributes towards unravelling the influence of intonation on lexical processing by showing that high $f_0$ guides the perception of lexical stress in infant metrical segmentation and adult lexical activation. The perception of metrical strength relations in a word can be shifted if the $f_0$ peak and the metrically stressed syllable are (phonologically) not aligned. Hence, pitch accent type influences lexical processing in intonation languages, in which pitch is not contrastively used. In other words, different pitch accent types affect how we understand words – beyond their communicative function. In a broader sense, this thesis hence contributes to a more detailed understanding of human language processing.

At the same time, by looking at two populations (infants and adults), the current dissertation traces the development of the perception of metrical stress and the weighting of corresponding cues. Here, we have demonstrated that processing in both listener groups is guided by the position of the $f_0$ peak, yet more listener groups need to be incorporated to completely trace the developmental trajectories of the use of $f_0$ as a stress cue.

Finally, the manipulation of the occurrence frequency of different pitch accent types, in particular, allows for theoretical conclusions at the phonetics-phonology interface. In this regard, our findings speak in favour of a phonological basis of the association between high $f_0$ and lexical stress. More precisely, we argue that it is not (only) the phonetic cues (here the acoustic cue high $f_0$) that guide the perception of lexical stress. Rather, the learned association between high $f_0$ and lexical stress generates expectations on which cues make a syllable appear to be stressed.

\textsuperscript{76} This claim is based on personal observations based on speakers of American, Australian, and British English.
Zusammenfassung

Einbettung und Problemstellung
Diese Dissertationsschrift befasst sich mit dem Einfluss verschiedener Tonakzenttypen auf die Wahrnehmung von Wortbetonung im Deutschen. Genauer gesagt untersucht sie, wie sich unterschiedliche Tonakzenttypen auf die metrische Segmentierung bei deutschen Kleinkindern und die lexikalische Aktivierung bei deutschen Erwachsenen auswirken. Sie soll außerdem Mechanismen auf den Grund gehen, die die Gewichtung der akustischen Hinweise (cues) für die Wahrnehmung von Betonung bestimmen.


Erwachsenen, hohe Silben zur Gruppierung linguistischer Einheiten verwenden. So werden tonal alternierende Silbensequenzen von Säuglingen und Erwachsenen als Trochäen gruppiert, wobei die hohe Silbe das starke und damit betonte Element bildet (Abboub et al., 2016; Bhatara et al., 2013; Bion et al., 2011). Für Erwachsene wurde, angefangen in den späten 1950ern, wiederholt gezeigt, dass eine hohe Tonhöhe genutzt wird, um Minimalpaare wie das englische Nomen *object* (Betonung auf erster Silbe) und das Verb *(to) object* (Betonung auf zweiter Silbe) zu unterscheiden (an gefangen mit Fry, 1958). Somit ist für beide Gruppen – Kleinkinder und Erwachsene – die hohe Tonhöhe ein attraktiver Hinweis für die Wahrnehmung von Wortbetonung.


Um gesprochene Sprache verstehen zu können, müssen HörerInnen die einzelnen Wörter identifizieren, die SprecherInnen äußern. Das ist nicht einfach, weil es sich bei gesprochener Sprache akustisch gesehen um einen Sprachstrom ohne eindeutige Wortgrenzen handelt. Diese sprachliche Eigenschaft wird mit Blick auf die HörerInnen gemeinhin als Wortsegmentierungsproblem bezeichnet. Zudem variieren Produktionen eines Wortes in Abhängigkeit von Faktoren wie Geschlecht, Alter, benachbarte Laute etc., was als Invarianzproblem bezeichnet wird (z.B. Cutler, 2012). In der Psycholinguistik wird Worterkennung als Prozess verstanden, bei dem verschiedene lexikalische Kandidaten aktiviert werden, wobei der Grad der Aktivierung davon abhängt, wie groß die Übereinstimmung zwischen Information aus dem Signal und der gespeicherten Repräsentation ist. Dieser Abgleich erfolgt über das mentale Lexikon, in dem Informationen über die Form von Wörtern, einschließlich (morpho)-syntaktischer und semantischer Information gespeichert sind. Die aktivierten lexikalischen Kandidaten befinden sich sodann in einem Wettbewerb, den schließlich jener Kandidat für sich entscheidet, der die größte Übereinstimmung zwischen Signal und gespeicherter Information zeigt. Auf diese Weise wird das intendierte Wort schließlich erkannt (z.B. Weber & Scharenborg,
Für eine effiziente Worterkennung verwenden erwachsene HörerInnen bekanntermaßen segmentelle und suprasegmentelle Informationen. In Bezug auf suprasegmentelle cues wurde in verschiedenen Priming Studien gezeigt (N. Cooper et al., 2002; Donselaar et al., 2005; Soto-Faraco et al., 2001), dass Wörter mit einem bestimmten Betonungsmuster, z.B. einem trochäischen Muster, Wörter mit dem gleichen rhythmischen Muster primen, d.h., deren lexikalischen Zugriff beschleunigen, während sie den Zugriff auf Wörter mit dem umgekehrten Betonungsmuster (z.B. einem jambischen) verlangsamen. In Eye-Tracking Studien wurde zudem gezeigt (z.B. Jesse et al., 2017; Reinisch et al., 2010), dass solche suprasegmentelle cues genutzt werden, sobald sie im Signal verfügbar sind. Schon während der Verarbeitung der ersten Silbe unterscheiden HörerInnen zum Beispiel anhand suprasegmentaler cues sogenannte stress competitors, d.h. segmentell überlappende Wörter, die sich in der Position der betonten Silbe unterscheiden, wie z.B. Oktopus und Oktober, mit Betonung auf der ersten bzw. der zweiten Silbe (Reinisch et al., 2010).


Expizite Segmentierungsstrategien, Subroutinen des Worterkennungsprozesses also, sind in manchen Kontexten unumgänglich, vor allem wenn die Hörergruppe das mentale Lexikon, das zum Abgleich verwendet wird, erst noch ausbilden muss. So ist die Zerlegung gesprochener Sprache vor allem für eine Gruppe essentiell: Kleinkinder. Sie müssen sich den Weg in die Sprache vom Signal her erarbeiten. Darauf sind die Gehörorgane bereits im letzten Drittel der Schwangerschaft ausgerichtet und lassen den Fötus im Mutterleib vor allem die rhythmischen Eigenschaften der Sprache wahrnehmen (Querleu et al., 1988;
Zusammenfassung


Zusammenfassung der Ergebnisse

Eine Reihe von Head-Turn Preference Studien beleuchtete die Verwendung von \( f_0 \) und anderen Betonungs-Cues bei der metrischen Segmentierung durch 9 Monate alte deutsche Kleinkinder (Kapitel 4). In Experiment 1 wurde untersucht, ob verschiedene Akzenttypen die kindliche Fähigkeit beeinflussen, Trochäen (SW, für strong-weak, stark-schwach, z.B. \([\text{'gu:nal}]\) aus WSW-Trägerwörtern (z.B. \( \text{Lagune [la.'gu:.nal]} \)) zu extrahieren. Hier wurde die Extraktion von eingebetteten Trochäen getestet, um allein das Zusammenspiel zwischen Akzenttyp und betoner Silbe untersuchen zu können, da bekannt ist, dass Kinder in diesem Alter bereits andere Hinweise auf Wortgrenzen nutzen können. In der Familiarisierungsphase wurden die WSW-Trägerwörter, die in Sätzen eingebettet waren, in drei verschiedenen Intonationssituationen präsentiert: in einer Alignierungs-Bedingung (mittlerer Gipfelakzent, (L+)H*) und zwei Nicht-Alignierungs-Bedingungen (früher Gipfelakzent, H+L*, und später Gipfelakzent, L*+H). In der Testphase hörten alle Kleinkinder die trochäischen Teile der WSW-Wörter, aufgenommen mit fallender Intonation (die Hälfte der Trochäen war neu, die andere Hälfte familiarisiert). In
diesem Paradigma sind Blickdauerunterschiede zu neuen vs. familiarisierten Wörtern ein Hinweis auf eine erfolgreiche Extraktion, ein Wiedererkennungseffekt also (Kemler Nelson et al., 1995). Die Ergebnisse zeigten, dass Kleinkinder die Trochäen nur in der Alignierungs-Bedingung (mittlerer Gipfelakzent, betonte Silbe klingt hoch) extrahierten, was sich in einer novelty preference äußerte. Dahingegen gelang ihnen die Extraktion der eingebetteten Trochäen nicht, wenn die betonte Silbe tief klang (Nicht-Alignierungs-Bedingungen). Ein Kontrollexperiment (Experiment 1b) bestätigte, dass der Wiedererkennungseffekt, der in der Alignierungs-Bedingung beobachtet wurde, nicht auf einen Abgleich der Intonationskonturen zwischen Familiarisierungsphase und Testphase zurückgeht. Darüber hinaus wurde keine Wiedererkennung der Trochäen beobachtet, wenn Kleinkinder mit WSW-Wörtern familiarisiert wurden, die mit einer horizontal gespiegelten mittleren Gipfelkontur (HL*H) realisiert wurden (Experiment 1c). Somit liefert Experiment 1 mit seinen Kontrollstudien Evidenz dafür, dass die hohe Tonhöhe für deutsche Kleinkinder bei der metrischen Segmentierung ein notwendiger Hinweis auf die betonte Silbe ist. Experiment 2 untersuchte die Rolle von hohen Silben genauer, indem getestet wurde, ob die hohe Tonhöhe womöglich ein hinreichender Hinweis für die Betonungswahrnehmung darstellt. Um dieser Frage nachzugehen, wurden Kleinkinder mit WWS-Wörtern familiarisiert (z.B. *Rumila [ˈru.mi.ˈlaː]). Diese Nicht-Wörter wurden entweder mit einem frühen Gipfelakzent (H+L*) oder einem mittleren Gipfelakzent (L+H*) produziert. Als Testwörter wurden die letzten beiden Silben der WWS-Wörter mit umgekehrtem Betonungsmuster verwendet (z.B. SW [ˈmi.ˈla] für WWS *Rumila [ˈru.mi.ˈlaː]); aufgenommen mit variiertem Intonation (wiederum waren die Hälfte der Testwörter neu, die andere Hälfte familiarisiert). Um den Effekt von hoher f0 zu isolieren (und damit die Möglichkeit auszuschließen, dass die zweite Silbe in den WWS-Wörtern in der frühen Gipfelkontur in natürlichen Produktionen akustisch verstärkt ist (Niebuhr, 2007a)) wurde das Experiment mit natürlichen (Experiment 2) und PSOLA-resynthetisierten Stimuli (Experiment 2b) durchgeführt. Für die Resynthese wurden die Intonationskonturen zwischen den Bedingungen getauscht, d.h. die WWS-Wörter, die mit einem mittleren Gipfelakzent aufgenommen wurden, erhielten eine frühe Gipfelkontur. Die Ergebnisse zeigten, dass mit natürlichen Stimuli der Tonhöengipfel auf der zweiten unbetonten Silbe in den WWS-Wörtern stark genug war, um ein Konzept von Betonung zu erzeugen und so zu einer metrischen Re-Interpretation der wort-prosodischen Struktur zu führen (WS realisiert als hoch-tief führte zu einem SW-Perzept): Deutsche 9 Monate alte Kinder behandelten also hohe Silben als Wortanfänge, auch wenn diese unbetont waren, was sich in einer familiarity preference äußerte (Experiment 2). Dies war nicht der Fall mit resynthetisierten Stimuli (keine Präferenz in
Experiment 2b), was zeigt, dass ein isolierter f0 cue allein nicht ausreicht, um Betonung zu erzeugen.


Zusammengenommen zeigen Kapitel 4 und 5, dass der Tonhöhengipfel die Betonungswahrnehmung sowohl bei Kleinkindern als auch bei Erwachsenen...
beeinflusst. Doch hier stellt sich die Frage, warum eine hohe Tonhöhe bei der Betonungswahrnehmung eine so große Rolle spielt, wenn sie – wie eingangs dargelegt – nicht verlässlich ist. Um die beobachteten Effekte von Akzenttyp auf die metrische Segmentierung und die lexikalische Aktivierung (Kapitel 3 und 4) zu erklären, kommen zwei Mechanismen in Frage: Ein Salienz-Mechanismus (Option 1) oder ein Frequenz-Mechanismus (Option 2). Gemäß Option 1 stechen hohe Silben perzeptuell heraus und können deshalb sehr attraktive Signale für die Wahrnehmung von metrischer Prominenz sein. Gemäß Option 2 könnten HörerInnen aufgrund eines häufigen Auftretens von H*-Akzenten gelernt haben, hohe Silben mit lexikalischer Betonung zu assoziieren. Sowohl Salienz als auch Frequenz sind in der Literatur zwei bekannte Faktoren, die viele Prozesse in der kindlichen und erwachsenen Sprachverarbeitung erklären können. Option 1 ist in einem experimentellen Kontext nur schwer zu manipulieren, wohingegen sich ein Exposure-Test Paradigma für die Untersuchung von Option 2 anbietet. Daher wurde mit solch einem Paradigma getestet, ob sich die Gewichtung des f0 cues für die Betonungsverarbeitung durch die Häufigkeit von hohen und tiefen betonten Silben im unmittelbaren Input verschieben lässt. Aus praktischen Gründen wurden erwachsenen HörerInnen als Testgruppe gewählt. Eine dreiminütige Exposure Phase mit ausnahmslos tiefen betonten Silben (in Alternativfragen, Wh-Fragen oder contrastive topics) wurde dem Eye-Tracking Experiment vorgeschaltet (Experiment 4 in Kapitel 6). Der Eye-Tracking Teil der Studie folgte unmittelbar auf die Exposure Phase. Wie durch den Frequenz-Ansatz vorhergesagt, zeigte in diesem Experiment die GAMM-Analyse der Fixationen zum stress competitor für den Großteil des Analysefensters (segmentell ambiger Teil) keinen Effekt von Intonationsbedingung. Auch die Fixationen zum Zielwort unterschieden sich in diesem Experiment nicht zwischen den Intonationsbedingungen. Wichtig ist, dass eine kombinierte GAMM-Analyse der Fixationen aus den Experimenten 3 und 4 (Kapitel 5 und 6) eine kurzezeitige aber signifikante Interaktion zwischen Experiment (Experiment 3 vs. Experiment 4) und Intonationsbedingung (früher vs. mittlerer Gipfel) zeigte. Dies bedeutet, dass der Akzenttyp die Fixationen zum stress competitor in den Experimenten mit und ohne exposure unterschiedlich beeinflusste (nicht aber für TargetFixationen). Die gesteigerte Häufigkeit von L*-Akzenten in der unmittelbar vorausgehenden Phase bedingte somit, dass HörerInnen ihre Gewichtung von Betonungs-Cues dahingehend verschieben, dass eine hohe Silbe nicht mehr notwendigerweise als betont gewertet wird. Sobald genug Gegenevidenz für hohe Tonhöhe als Betonungs-Cue vorliegt, vernachlässigen deutsche HörerInnen diesen Cue. Dar aus lässt sich schließen, dass die Häufigkeit von hohen betonten Silben (Option 2) – mehr als die perzeptuelle Salienz von hohen Silben (Option 1) – ein Mechanismus ist, der die Interpretation von hoher f0 als betont erklärt. Basierend auf diesem Befund ist anzunehmen, dass die Tendenz, hohe Silben als betont wahrzunehmen
(auch wenn diese unbetont sind), in Sprachen mit vielen H*-Akzenten am größten ist. Im Deutschen sind mittlere Gipfel sowohl in kindgerichteter Sprache als auch in Erwachsenensprache häufig, was Option 2 plausibel macht. Insgesamt argumentiert die Arbeit, dass die Verteilung der Akzenttypen in einer Sprache bestimmt, inwiefern ein Tonhöhungsgipfel als Signal zur Wahrnehmung von Betonung verwendet wird.

Zusammengefasst ist diese Arbeit sowohl für die Psycholinguistik als auch für die Schnittstelle Phonetik-Phonologie von großer Bedeutung. Sie leistet einen erheblichen Beitrag zur Ergründung des Einflusses von Intonation auf lexikalische Verarbeitungsmechanismen, indem sie die hohe Tonhöhe als bestimmenden Faktor in der expliziten Segmentierung und der lexikalischen Aktivierung identifiziert, wodurch die Dissertation weitreichende Implikationen für die jeweiligen Prozesse birgt. Gleichzeitig trägt die Arbeit durch den umfassenden Ansatz zu unserem Wissen bei, wie sich Betonungswahrnehmung (und die Gewichtung der dafür verantwortlichen) im Laufe des Lebens verändert. Besonders hervorzuheben sind die theoretischen Implikationen an der Schnittstelle Phonetik-Phonologie, die für eine phonologische Basis der Assoziation zwischen hoher F0 und lexikalischer Betonung sprechen. Anders gesagt bestimmen nicht (nur) phonetisch-akustische Merkmale die Betonungswahrnehmung, vielmehr werden Erwartungen über die Beschaffenheit der betonten Silbe über die gelernte Assoziation zwischen Tonhöhungsgipfel und betoner Silbe generiert.
Bibliography


Table A.1: Familiarization passages in Experiment 1.

**“Lagune” Passage**


‘Here originated a lagoon. The lagoon was wonderful. The blue lagoon attracts people. A small lagoon is nice. His lagoon was situated in the South. She took a photo of her lagoon.’

**“Kasino” Passage**


‘The town was planning a casino. It should become a big casino. His wife wanted to go to the casino. The casino was still very new. A small casino is not nice. The new casino became very popular.’

**“Kanone” Passage**


‘The man wanted a cannon. He had seen a black cannon. The new cannon was expensive. The cannon should last long. The neighbor sold an old cannon. With a cannon one feels safe.’

**“Tirade” Passage**


‘The girl released a tirade. The tirade did not want to end. It was a big tirade. She was familiar with the tirade. He could not take her tirade. This tirade should stop soon.’

Note: Bold face refers to the SW-part in the WSW-carrier word; italics indicate other accentuated words in the sentence. Translations are provided below each passage.
Table A.2: Mean values (and standard deviations) of the individual test lists in Experiment 1 (falling intonation).

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>gune-list</th>
<th>sino-list</th>
<th>none-list</th>
<th>rade-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of pitch fall in st</td>
<td>10.07</td>
<td>10.33</td>
<td>9.70</td>
<td>10.12</td>
</tr>
<tr>
<td>(stressed) in ms</td>
<td>(1.72)</td>
<td>(1.85)</td>
<td>(2.01)</td>
<td>(1.81)</td>
</tr>
<tr>
<td>Duration of first syllable in ms</td>
<td>331</td>
<td>361</td>
<td>406</td>
<td>358</td>
</tr>
<tr>
<td>(stressed)</td>
<td>(48)</td>
<td>(31)</td>
<td>(61)</td>
<td>(62)</td>
</tr>
<tr>
<td>Duration of second syllable in ms</td>
<td>361</td>
<td>358</td>
<td>342</td>
<td>321</td>
</tr>
<tr>
<td>(unstressed)</td>
<td>(48)</td>
<td>(33)</td>
<td>(48)</td>
<td>(31)</td>
</tr>
</tbody>
</table>

Table A.3: Mean values (and standard deviations) of the individual test lists in Experiment 1b (rising intonation).

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>gune-list</th>
<th>sino-list</th>
<th>none-list</th>
<th>rade-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of pitch rise in st</td>
<td>12.18</td>
<td>12.11</td>
<td>12.12</td>
<td>12.54</td>
</tr>
<tr>
<td>(stressed) in ms</td>
<td>(1.32)</td>
<td>(1.64)</td>
<td>(1.46)</td>
<td>(1.67)</td>
</tr>
<tr>
<td>Duration of first syllable in ms</td>
<td>362</td>
<td>385</td>
<td>392</td>
<td>360</td>
</tr>
<tr>
<td>(stressed)</td>
<td>(49)</td>
<td>(90)</td>
<td>(83)</td>
<td>(23)</td>
</tr>
<tr>
<td>Duration of second syllable in ms</td>
<td>361</td>
<td>358</td>
<td>342</td>
<td>321</td>
</tr>
<tr>
<td>(unstressed)</td>
<td>(32)</td>
<td>(79)</td>
<td>(22)</td>
<td>(20)</td>
</tr>
</tbody>
</table>
### Appendix B: Experiment 2 – Materials

Table B.1: Familiarization passages in Experiment 2.

<table>
<thead>
<tr>
<th>“Rumila” Passage</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>“Linuro” Passage</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>“Nalomu” Passage</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>“Morani” Passage</th>
</tr>
</thead>
</table>

Note: **Bold** face refers to the WS-part in the WWS-carrier word; *italics* indicate other accentuated words in the sentence; underlining indicates where syllables were added/removed in comparison to Experiment 1. See Appendix A for translations.
B.2: Rating Study WWS-words

The pre-selected materials (48 WWS-nonce words, i.e., 4 nonce words x 6 sentences x 2 conditions) were used as the experimental stimuli for this rating study. To increase variability in responses, 48 additional filler items were generated. These filler words had lexical stress on the second syllable (WSW-nonce words). For that purpose, the same speaker recorded the experimental passages again, but this time lexical stress was placed on the second syllable of the nonce words (WSW). These words were also recorded in two intonation conditions (early- vs. medial-peak condition). In total, there were 96 trials.

Using Presentation Software (Neurobehavioral Systems, Version 18.0), raters were presented with 96 sentences including the nonce words to be judged; after every sentence, the critical word occurred again in isolation, separated by 800 ms. Raters indicated their judgment on a button box with three keys (1 representing stress on the first syllable, 2 stress on the second syllable, and 3 stress on the third). Figure B.2 shows the average correctness rates. Overall, the correctness rates were very high (97.4%). Specifically, for filler trials (WSW), the correctness rate was 97.4% for both early- and medial-peak condition. For the critical test trials (WWS), the correctness rate was 100% in the medial-peak condition and 94.8% in the early-peak condition. Results of a general linear mixed effect model with correctness rate as the response variable, experimental status and intonation condition as fixed factors and participants and items as random factors did not show an interaction between the two factors \( p > 0.86 \). There was also no main effect of experimental status on correctness rate \( p = 1.0 \) and no main effect of intonation condition on correctness rate \( p > 0.35 \).

![Figure B.2: Correctness rates split by experimental status (filler items (WSW, left) and test items (WWS, right), and intonation condition (early vs. medial-peak condition)).](image)

Errors in WWS-words in the early-peak condition were inspected further \( N = 10 \). Six of the errors were syllable-1-responses, likely due to an initial stress bias in German listeners (e.g., Cutler, 2012); four errors were syllable-2-responses, which are unfortunate as they show the metrical re-interpretation of the word prosodic
structure we tested for. One item ("Nolumu3") was misjudged the most (3 times). The ten misjudged items (plus two randomly chosen files that did not lead to errors) were re-presented to three randomly chosen subjects of the sample (aj, se, ab) using Praat (Play function). All of them correctly judged the 10 files that had initially led to errors in the stress judgement experiment (even the item that was incorrectly judged in three cases in the stress judgement experiment), but two of them misjudged one file (Linuro4), an item that was randomly added to the list and had not been misjudged in the stress judgement experiment. We thus decided to replace these two items (Nolumu3 and Linuro4) altogether. Another three persons from the sample (bb, mp, se) assured that the third syllable was the stressed one in the replacements. On this final sample acoustical analysis was then performed.

Table B.3: Mean values (and standard deviations) of the individual test lists in Experiment 2 (varied intonation).

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>mila-list</th>
<th>nuro-list</th>
<th>lomu-list</th>
<th>rani-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of pitch rise in st</td>
<td>11.88</td>
<td>11.54</td>
<td>11.64</td>
<td>11.61</td>
</tr>
<tr>
<td>(rising items)</td>
<td>(1.43)</td>
<td>(1.87)</td>
<td>(0.99)</td>
<td>(1.08)</td>
</tr>
<tr>
<td>F0 excursion of pitch fall in st</td>
<td>8.15</td>
<td>8.06</td>
<td>8.01</td>
<td>8.17</td>
</tr>
<tr>
<td>(falling items)</td>
<td>(1.01)</td>
<td>(0.55)</td>
<td>(1.65)</td>
<td>(0.59)</td>
</tr>
<tr>
<td>Mean f0 in Hz (monotone)</td>
<td>242.1</td>
<td>244.0</td>
<td>244.6</td>
<td>243.9</td>
</tr>
<tr>
<td>(stressed) in ms</td>
<td>(5.0)</td>
<td>(10.4)</td>
<td>(7.4)</td>
<td>(2.6)</td>
</tr>
<tr>
<td>Duration of first syllable</td>
<td>284</td>
<td>308</td>
<td>307</td>
<td>297</td>
</tr>
<tr>
<td>(stressed in ms)</td>
<td>(39)</td>
<td>(22)</td>
<td>(27)</td>
<td>(25)</td>
</tr>
<tr>
<td>Duration of second syllable</td>
<td>316</td>
<td>297</td>
<td>296</td>
<td>305</td>
</tr>
<tr>
<td>(unstressed in ms)</td>
<td>(62)</td>
<td>(51)</td>
<td>(44)</td>
<td>(64)</td>
</tr>
</tbody>
</table>
Appendix B: Experiment 2 – Materials

Table B.4: Acoustical analysis of resynthesized stimuli (for ease of comparison, column 1 and 2 display the mean values and standard deviations for the early-peak and medial-peak condition in Experiment 2 again).

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>Early-peak condition (natural)</th>
<th>Medial-peak condition (natural)</th>
<th>Early-peak condition (resynthesized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of movement in st</td>
<td>10.0 (0.9)</td>
<td>9.6 (2.1)</td>
<td>9.0 (1.2)</td>
</tr>
<tr>
<td>Duration of first syllable (unstressed) in ms</td>
<td>137 (28)</td>
<td>145 (27)</td>
<td>138 (40)</td>
</tr>
<tr>
<td>Duration of second syllable (unstressed) in ms</td>
<td>165 (26)</td>
<td>162 (25)</td>
<td>160 (21)</td>
</tr>
<tr>
<td>Duration of third syllable (stressed) in ms</td>
<td>267 (41)</td>
<td>243 (38)</td>
<td>260 (44)</td>
</tr>
<tr>
<td>Duration of onset consonant in second syllable in ms</td>
<td>70 (17)</td>
<td>68 (14)</td>
<td>72 (14)</td>
</tr>
<tr>
<td>Duration of onset consonant in third syllable (stressed) in ms</td>
<td>109 (18)</td>
<td>90 (24)</td>
<td>123 (35)</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of first vowel in dB</td>
<td>29.7 (13.9)</td>
<td>31.4 (13.6)</td>
<td>31.1 (12.5)</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of second vowel in dB</td>
<td>28.5 (9.5)</td>
<td>26.4 (9.7)</td>
<td>32.3 (14.0)</td>
</tr>
<tr>
<td>H1*-A3* ratio in middle of third vowel in dB</td>
<td>29.2 (10.7)</td>
<td>27.6 (10.7)</td>
<td>21.5 (11.0)</td>
</tr>
<tr>
<td>Mean intensity in first vowel in dB</td>
<td>73.6 (2.0)</td>
<td>70.1 (2.7)</td>
<td>70.6 (2.4)</td>
</tr>
<tr>
<td>Mean intensity in second vowel in dB</td>
<td>70.5 (2.9)</td>
<td>70.1 (3.3)</td>
<td>68.4 (3.4)</td>
</tr>
<tr>
<td>Mean intensity in third vowel in dB</td>
<td>70.4 (2.1)</td>
<td>71.2 (2.6)</td>
<td>69.5 (3.5)</td>
</tr>
<tr>
<td>RMS in first vowel in Pascal</td>
<td>0.099 (0.022)</td>
<td>0.072 (0.020)</td>
<td>0.071 (0.020)</td>
</tr>
<tr>
<td>RMS in first second in Pascal</td>
<td>0.070 (0.024)</td>
<td>0.068 (0.024)</td>
<td>0.059 (0.022)</td>
</tr>
<tr>
<td>RMS in first third in Pascal</td>
<td>0.059 (0.022)</td>
<td>0.077 (0.026)</td>
<td>0.073 (0.028)</td>
</tr>
</tbody>
</table>

The resynthesis procedure served its purpose in generating a condition in which the acoustic difference between the second and the third syllable is distinct (as indicated by duration, spectral tilt, and intensity measurements; comparable to the natural medial-peak condition in Experiment 2).
Table B.5: Mean values (and standard deviations) of the individual test lists in Experiment 2b (varied intonation).

<table>
<thead>
<tr>
<th>Acoustic Variable</th>
<th>mila-list</th>
<th>suka-list</th>
<th>buti-list</th>
<th>rani-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0 excursion of pitch rise in st (rising items)</td>
<td>11.88</td>
<td>11.15</td>
<td>11.54</td>
<td>11.61</td>
</tr>
<tr>
<td></td>
<td>(1.43)</td>
<td>(0.65)</td>
<td>(1.87)</td>
<td>(1.08)</td>
</tr>
<tr>
<td>F0 excursion of pitch fall in st (falling items)</td>
<td>8.15</td>
<td>9.38</td>
<td>8.06</td>
<td>8.17</td>
</tr>
<tr>
<td></td>
<td>(1.01)</td>
<td>(0.79)</td>
<td>(0.55)</td>
<td>(0.59)</td>
</tr>
<tr>
<td>Mean f0 in Hz (monotone)</td>
<td>242.1</td>
<td>231.9</td>
<td>244.0</td>
<td>243.9</td>
</tr>
<tr>
<td></td>
<td>(5.0)</td>
<td>(3.8)</td>
<td>(10.4)</td>
<td>(2.6)</td>
</tr>
<tr>
<td>Duration of first syllable (stressed) in ms</td>
<td>284</td>
<td>319</td>
<td>308</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>(39)</td>
<td>(35)</td>
<td>(22)</td>
<td>(25)</td>
</tr>
<tr>
<td>Duration of second syllable (unstressed) in ms</td>
<td>316</td>
<td>309</td>
<td>297</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>(62)</td>
<td>(35)</td>
<td>(51)</td>
<td>(64)</td>
</tr>
</tbody>
</table>

As in Experiment 2, the average f0 excursion of the pitch movement (for tokens with falling and rising intonation), mean f0 (for tokens with monotonous intonation) and the average duration of the test words were closely matched across test words. If one of the newly recorded test words differed from the one it replaced in a given acoustic dimension (e.g., mean f0 for monotonous items), care was taken that the difference was also present in the second newly recorded test words, so that it did not affect the overall difference across lists. For instance, mean f0 was on average 15 Hz lower in buti and suka compared to the test words used in Experiment 2. Overall, the average f0 excursion for the test tokens with falling intonation was 8.77 st (SD = 1.03 st; range = 6.51 st - 10.54 st), the average f0 excursion for the tokens with rising intonation was 11.41 st (SD = 0.98 st; range = 9.55 st – 13.07 st), and the mean f0 for the tokens with monotonous intonation was 236.3 Hz (SD = 8.4 Hz; range = 221.0 – 248.4 Hz). The average duration of the test words was 614 ms (SD = 60 ms), ranging from 479 ms to 738 ms; the first syllable of the test words was on average 304 ms (SD = 35 ms) and the second syllable was on average 310 ms (SD = 51 ms) long.
Appendix C: Experiment 3 – Materials and analyses

Table C.1 Quadruplets displayed on screen in cohort trials in Experiment 3 (target or competitor as the auditory target). Translations are provided below each item.

<table>
<thead>
<tr>
<th>Target (all WSW)</th>
<th>Competitor (all SWW)</th>
<th>Distractor 1</th>
<th>Distractor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albaner</td>
<td>Albatros</td>
<td>Galaxy (WWS)</td>
<td>Konfetti (WSW)</td>
</tr>
<tr>
<td>‘Albanian’</td>
<td>‘albatross’</td>
<td>‘galaxy’</td>
<td>‘confetti’</td>
</tr>
<tr>
<td>Alaska</td>
<td>Alibi</td>
<td>Korrelat (WWS)</td>
<td>Laguna (WSW)</td>
</tr>
<tr>
<td>‘Alaska’</td>
<td>‘alibi’</td>
<td>‘correlate’</td>
<td>‘lagoon’</td>
</tr>
<tr>
<td>Anapher</td>
<td>Ananas</td>
<td>Kassette (WSS)</td>
<td>Gremium (SWW)</td>
</tr>
<tr>
<td>‘anaphora’</td>
<td>‘pineapple’</td>
<td>‘cassette’</td>
<td>‘council’</td>
</tr>
<tr>
<td>Anode</td>
<td>Anorak</td>
<td>Fantasie (WSS)</td>
<td>Polizist (WSS)</td>
</tr>
<tr>
<td>‘anode’</td>
<td>‘anorak’</td>
<td>‘fantasy’</td>
<td>‘policeman’</td>
</tr>
<tr>
<td>Aroma</td>
<td>Arie</td>
<td>Parasit (WSS)</td>
<td>Krokoal (WSW)</td>
</tr>
<tr>
<td>‘aroma’</td>
<td>‘aria’</td>
<td>‘parasite’</td>
<td>‘crocodile’</td>
</tr>
<tr>
<td>Embargo</td>
<td>Embryo</td>
<td>Dekade (WSW)</td>
<td>Legion (WSS)</td>
</tr>
<tr>
<td>‘embargo’</td>
<td>‘embryo’</td>
<td>‘decade’</td>
<td>‘legion’</td>
</tr>
<tr>
<td>Enklave</td>
<td>Enkelin</td>
<td>Lexika (SWW)</td>
<td>Aphasia (WSW)</td>
</tr>
<tr>
<td>‘enclave’</td>
<td>‘granddaughter’</td>
<td>‘lexicons’</td>
<td>‘aphasia’</td>
</tr>
<tr>
<td>Eskorte</td>
<td>Eskimo</td>
<td>Alphabet (WSS)</td>
<td>Tamburin (SWW)</td>
</tr>
<tr>
<td>‘escort’</td>
<td>‘Inuit’</td>
<td>‘alphabet’</td>
<td>‘timbre’</td>
</tr>
<tr>
<td>Exotik</td>
<td>Exodus</td>
<td>Zusage (SWW)</td>
<td>Diadem (WSS)</td>
</tr>
<tr>
<td>‘exoticism’</td>
<td>‘exodus’</td>
<td>‘commitment’</td>
<td>‘ciclet’</td>
</tr>
<tr>
<td>Facette</td>
<td>Faserung</td>
<td>Amulett (WWS)</td>
<td>Synonym (WSS)</td>
</tr>
<tr>
<td>‘facet’</td>
<td>‘fibrillation’</td>
<td>‘charm’</td>
<td>‘synonym’</td>
</tr>
<tr>
<td>Furore</td>
<td>Furie</td>
<td>Elefant (WSS)</td>
<td>Opium (SWW)</td>
</tr>
<tr>
<td>‘sensation’</td>
<td>‘fury’</td>
<td>‘elephant’</td>
<td>‘opium’</td>
</tr>
<tr>
<td>Genetik</td>
<td>Genesis</td>
<td>Lilie (SWW)</td>
<td>Nikotin (WSS)</td>
</tr>
<tr>
<td>‘genetics’</td>
<td>‘genesis’</td>
<td>‘lily’</td>
<td>‘nicotine’</td>
</tr>
<tr>
<td>Kabine</td>
<td>Kabeljau</td>
<td>Amateur (WSS)</td>
<td>Zisterne (WSS)</td>
</tr>
<tr>
<td>‘cabin’</td>
<td>‘cod’</td>
<td>‘amateur’</td>
<td>‘cister’</td>
</tr>
<tr>
<td>Kamille</td>
<td>Kamerun</td>
<td>Mystiker (SWW)</td>
<td>Cholera (SWW)</td>
</tr>
<tr>
<td>‘camomile’</td>
<td>‘Cameroon’</td>
<td>‘mystic’</td>
<td>‘cholera’</td>
</tr>
<tr>
<td>Kanone</td>
<td>Kanada</td>
<td>Ethiker (SWW)</td>
<td>Geselle (WSS)</td>
</tr>
<tr>
<td>‘cannon’</td>
<td>‘Canada’</td>
<td>‘ethicist’</td>
<td>‘companion’</td>
</tr>
<tr>
<td>Kannister</td>
<td>Kanapee</td>
<td>Debakel (WSW)</td>
<td>Stadion (SWW)</td>
</tr>
<tr>
<td>‘canister’</td>
<td>‘couch’</td>
<td>‘debacle’</td>
<td>‘stadium’</td>
</tr>
<tr>
<td>Karotte</td>
<td>Karitas</td>
<td>Parabel (WSW)</td>
<td>Salami (WSW)</td>
</tr>
<tr>
<td>‘carot’</td>
<td>‘caritas’</td>
<td>‘parabola’</td>
<td>‘salami’</td>
</tr>
<tr>
<td>Kaverne</td>
<td>Kaviar</td>
<td>Chinese (WSW)</td>
<td>Veranda (WSW)</td>
</tr>
<tr>
<td>‘cavern’</td>
<td>‘caviar’</td>
<td>‘Chinese’</td>
<td>‘veranda’</td>
</tr>
<tr>
<td>Kolumne</td>
<td>Kolibri</td>
<td>Detektiv (WWS)</td>
<td>Adlige</td>
</tr>
<tr>
<td>‘column’</td>
<td>‘colibri’</td>
<td>‘detective’</td>
<td>‘aristocrat’</td>
</tr>
<tr>
<td>Libelle</td>
<td>Libero</td>
<td>Thymian (SWW)</td>
<td>Safari (WSW)</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>‘dragonfly’</td>
<td>‘sweeper’</td>
<td>‘thyme’</td>
<td>‘safari’</td>
</tr>
<tr>
<td>Markise</td>
<td>Magier</td>
<td>Paradies (WWS)</td>
<td>Vitamin (WWS)</td>
</tr>
<tr>
<td>‘awning’</td>
<td>‘magician’</td>
<td>‘parody’</td>
<td>‘vitamin’</td>
</tr>
<tr>
<td>Maneg</td>
<td>Manitu</td>
<td>Atheist (WWS)</td>
<td>Pipette (WSW)</td>
</tr>
<tr>
<td>‘arena’</td>
<td>‘Manitou’</td>
<td>‘atheist’</td>
<td>‘pipet’</td>
</tr>
<tr>
<td>Marille</td>
<td>Marathon</td>
<td>Fusion (WWS)</td>
<td>Retina (SWW)</td>
</tr>
<tr>
<td>‘apricot’</td>
<td>‘marathon’</td>
<td>‘fusion’</td>
<td>‘retina’</td>
</tr>
<tr>
<td>Marotte</td>
<td>Marrakesch</td>
<td>Leopard (WWS)</td>
<td>Tirade (WSW)</td>
</tr>
<tr>
<td>‘quirk’</td>
<td>‘Marrakesch’</td>
<td>‘leopard’</td>
<td>‘tirade’</td>
</tr>
<tr>
<td>Medaille</td>
<td>Medikus</td>
<td>Folio (SWW)</td>
<td>Aktivist (WWS)</td>
</tr>
<tr>
<td>‘medal’</td>
<td>‘medico’</td>
<td>‘folio’</td>
<td>‘activist’</td>
</tr>
<tr>
<td>Monokel</td>
<td>Monitor</td>
<td>Sardine (WWS)</td>
<td>Rämerin (SWW)</td>
</tr>
<tr>
<td>‘monocle’</td>
<td>‘monitor’</td>
<td>‘sardine’</td>
<td>‘Roman’</td>
</tr>
<tr>
<td>Panade</td>
<td>Panama</td>
<td>Vakuum (SWW)</td>
<td>Dissonanz (WWS)</td>
</tr>
<tr>
<td>‘breading’</td>
<td>‘Panama’</td>
<td>‘vacuum’</td>
<td>‘dissonance’</td>
</tr>
<tr>
<td>Posaune</td>
<td>Positiv</td>
<td>Cabrio (SWW)</td>
<td>Gelage (WSW)</td>
</tr>
<tr>
<td>‘trombone’</td>
<td>‘positive’</td>
<td>‘convertible’</td>
<td>‘binge’</td>
</tr>
<tr>
<td>Pärmisse</td>
<td>Prämie</td>
<td>Kapuze (WSS)</td>
<td>Algebra (SWW)</td>
</tr>
<tr>
<td>‘premise’</td>
<td>‘premium’</td>
<td>‘hood’</td>
<td>‘algebra’</td>
</tr>
<tr>
<td>Radieschen</td>
<td>Radius</td>
<td>Zyniker (SWW)</td>
<td>Pazifik (WSW)</td>
</tr>
<tr>
<td>‘radish’</td>
<td>‘radius’</td>
<td>‘cynic’</td>
<td>‘Pacific’</td>
</tr>
<tr>
<td>Spirale</td>
<td>Spiritus</td>
<td>Novize (WSS)</td>
<td>Herberge (SWW)</td>
</tr>
<tr>
<td>‘helix’</td>
<td>‘spiritus’</td>
<td>‘novice’</td>
<td>‘shelter’</td>
</tr>
<tr>
<td>Statistik</td>
<td>Statue</td>
<td>Kreatur (WWS)</td>
<td>Bikini (WSW)</td>
</tr>
<tr>
<td>‘statistics’</td>
<td>‘statue’</td>
<td>‘creature’</td>
<td>‘bikini’</td>
</tr>
</tbody>
</table>
Table C.2: Quadruplets displayed on screen in filler trials in Experiment 3 (one of the distractors as the auditory target). Intonation condition of fillers is given in brackets. Translations are provided below each item.

<table>
<thead>
<tr>
<th>Distractor 1 (Audio, see word)</th>
<th>Distractor 2 (see word)</th>
<th>Cohort member (all WSW)</th>
<th>Cohort member (all SWW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almosen (SWW, e)</td>
<td>Magazin (WWS)</td>
<td>Oktober</td>
<td>Oktopus</td>
</tr>
<tr>
<td>‘alms’</td>
<td>‘magazine’</td>
<td>‘October’</td>
<td>‘octopus’</td>
</tr>
<tr>
<td>Figaro (SWW, e)</td>
<td>Ocean (SWW)</td>
<td>Papaya</td>
<td>Paprika</td>
</tr>
<tr>
<td>‘Figaro’</td>
<td>‘ocean’</td>
<td>‘papaya’</td>
<td>‘pepper’</td>
</tr>
<tr>
<td>Karneval (SWW, e)</td>
<td>Schimpanse</td>
<td>Hellene</td>
<td>Helium</td>
</tr>
<tr>
<td>‘carneval’</td>
<td>‘chimpanzee’</td>
<td>‘Hellene’</td>
<td>‘helium’</td>
</tr>
<tr>
<td>Lyriker (SWW, e)</td>
<td>Satellit (WWS)</td>
<td>Risotto</td>
<td>Risiko</td>
</tr>
<tr>
<td>‘poet’</td>
<td>‘satellite’</td>
<td>‘risotto’</td>
<td>‘risk’</td>
</tr>
<tr>
<td>Pavillon (SWW, e)</td>
<td>Evidenz (WWS)</td>
<td>Kombüse</td>
<td>Komiker</td>
</tr>
<tr>
<td>‘pavilion’</td>
<td>‘evidence’</td>
<td>‘galley’</td>
<td>‘comedian’</td>
</tr>
<tr>
<td>Jaguar (SWW, m)</td>
<td>Lametta (WSW)</td>
<td>Scharniere</td>
<td>Scharlatan</td>
</tr>
<tr>
<td>‘jaguar’</td>
<td>‘tinsel’</td>
<td>‘hinges’</td>
<td>‘quack’</td>
</tr>
<tr>
<td>Kanguru (SWW, m)</td>
<td>Einreise (SWW)</td>
<td>Regatta</td>
<td>Regelung</td>
</tr>
<tr>
<td>‘kangaroo’</td>
<td>‘entry’</td>
<td>‘regatta’</td>
<td>‘regulation’</td>
</tr>
<tr>
<td>Lithium (SWW, m)</td>
<td>Sängerin (SWW)</td>
<td>Tunesier</td>
<td>Tuniq</td>
</tr>
<tr>
<td>‘lithium’</td>
<td>‘singer’</td>
<td>‘Tunesian’</td>
<td>‘tunic’</td>
</tr>
<tr>
<td>Sellerie (SWW, m)</td>
<td>Karussell (WWS)</td>
<td>Hebräer</td>
<td>Hebamme</td>
</tr>
<tr>
<td>‘celery’</td>
<td>‘carousel’</td>
<td>‘Hebrew’</td>
<td>‘midwife’</td>
</tr>
<tr>
<td>Video (SWW, m)</td>
<td>Pupille (WSW)</td>
<td>Karate</td>
<td>Karavan</td>
</tr>
<tr>
<td>‘video’</td>
<td>‘pupil’</td>
<td>‘karate’</td>
<td>‘Caravan’</td>
</tr>
<tr>
<td>Courage (SWW, e)</td>
<td>Fabrikant (WWS)</td>
<td>Schablone</td>
<td>Schabernack</td>
</tr>
<tr>
<td>‘courage’</td>
<td>‘manufacturer’</td>
<td>‘stencil’</td>
<td>‘prank’</td>
</tr>
<tr>
<td>Finale (WSW, e)</td>
<td>Katalog (WWS)</td>
<td>Moräne</td>
<td>Morphium</td>
</tr>
<tr>
<td>‘final’</td>
<td>‘catalogue’</td>
<td>‘moraine’</td>
<td>‘morphine’</td>
</tr>
<tr>
<td>Lavendel (WSW, e)</td>
<td>Hurrikan (SWW)</td>
<td>Pinzette</td>
<td>Pinguin</td>
</tr>
<tr>
<td>‘lavender’</td>
<td>‘hurricane’</td>
<td>‘galaxy’</td>
<td>‘penguin’</td>
</tr>
<tr>
<td>Salpeter (SWW, e)</td>
<td>Gigolo (SWW)</td>
<td>Negierung</td>
<td>Negativ</td>
</tr>
<tr>
<td>‘nitrous’</td>
<td>‘gigolo’</td>
<td>‘denial’</td>
<td>‘negative’</td>
</tr>
<tr>
<td>Zitrone (WSW, e)</td>
<td>Jalonsie (WWS)</td>
<td>Kadaiver</td>
<td>Kadmium</td>
</tr>
<tr>
<td>‘lemon’</td>
<td>‘shutter’</td>
<td>‘carcasses’</td>
<td>‘cadmium’</td>
</tr>
<tr>
<td>Kommode (WSW, m)</td>
<td>Hostie (SWW)</td>
<td>Manöver</td>
<td>Manager</td>
</tr>
<tr>
<td>‘dresser’</td>
<td>‘host’</td>
<td>‘maneuver’</td>
<td>‘manager’</td>
</tr>
<tr>
<td>Kulisse (WSW, m)</td>
<td>Maserung (SWW)</td>
<td>Amöbe</td>
<td>Amrise</td>
</tr>
<tr>
<td>‘setting’</td>
<td>‘texture’</td>
<td>‘amoeba’</td>
<td>‘ant’</td>
</tr>
<tr>
<td>Matratze (WSW, m)</td>
<td>Eleganz (WWS)</td>
<td>Patrone</td>
<td>Patina</td>
</tr>
<tr>
<td>‘mattress’</td>
<td>‘elegance’</td>
<td>‘cartridge’</td>
<td>‘patina’</td>
</tr>
<tr>
<td>Parade (WSW, m)</td>
<td>Antenne (WSW)</td>
<td>Rosine</td>
<td>Rosmarin</td>
</tr>
<tr>
<td>‘parade’</td>
<td>‘arial’</td>
<td>‘raisin’</td>
<td>‘rosemary’</td>
</tr>
<tr>
<td>Table C.2 Continued</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Satire</strong> (WSW, m)</td>
<td><strong>Dialekt</strong> (WWS)</td>
<td><strong>Polenta</strong></td>
<td><strong>Podium</strong></td>
</tr>
<tr>
<td>‘satire’</td>
<td>‘dialect’</td>
<td>‘polenta’</td>
<td>‘dais’</td>
</tr>
<tr>
<td><strong>Abitur</strong> (WWS, e)</td>
<td><strong>Euphorie</strong> (WWS)</td>
<td><strong>Kubismus</strong></td>
<td><strong>Kumulus</strong></td>
</tr>
<tr>
<td>‘A-levels’</td>
<td>‘euphoria’</td>
<td>‘cubism’</td>
<td>‘cumulus’</td>
</tr>
<tr>
<td><strong>Anatom</strong> (WWS, e)</td>
<td><strong>Cousine</strong> (WSW)</td>
<td><strong>Pedale</strong></td>
<td><strong>Pelikan</strong></td>
</tr>
<tr>
<td>‘anatomist’</td>
<td>‘cousine’</td>
<td>‘pedals’</td>
<td>‘pelican’</td>
</tr>
<tr>
<td><strong>Diamant</strong> (WWS, e)</td>
<td><strong>Pharao</strong> (SWW)</td>
<td><strong>Karaffe</strong></td>
<td><strong>Kakadu</strong></td>
</tr>
<tr>
<td>‘diamant’</td>
<td>‘Pharaoh’</td>
<td>‘carafe’</td>
<td>‘cocky’</td>
</tr>
<tr>
<td><strong>Eтикett</strong> (WWS, e)</td>
<td><strong>Botanik</strong> (WSW)</td>
<td><strong>Orakel</strong></td>
<td><strong>Orient</strong></td>
</tr>
<tr>
<td>‘label’</td>
<td>‘botany’</td>
<td>‘oracle’</td>
<td>‘Orient’</td>
</tr>
<tr>
<td><strong>Halogen</strong> (WWS, e)</td>
<td><strong>Mumie</strong> (SWW)</td>
<td><strong>Follikel</strong></td>
<td><strong>Folie</strong></td>
</tr>
<tr>
<td>‘halogen’</td>
<td>‘mummy’</td>
<td>‘follicle’</td>
<td>‘foil’</td>
</tr>
<tr>
<td><strong>Hysterie</strong> (WWS, e)</td>
<td><strong>Razzia</strong> (SWW)</td>
<td><strong>Askese</strong></td>
<td><strong>Asien</strong></td>
</tr>
<tr>
<td>‘hysteria’</td>
<td>‘crackdown’</td>
<td>‘asceticism’</td>
<td>‘Asia’</td>
</tr>
<tr>
<td><strong>Jubilar</strong> (WWS, m)</td>
<td><strong>Aquator</strong> (WSW)</td>
<td><strong>Tomate</strong></td>
<td><strong>Tombola</strong></td>
</tr>
<tr>
<td>‘jubilarian’</td>
<td>‘equator’</td>
<td>‘tomato’</td>
<td>‘tombola’</td>
</tr>
<tr>
<td><strong>Monolog</strong> (WWS, m)</td>
<td><strong>Giraffe</strong> (WSW)</td>
<td><strong>Vanille</strong></td>
<td><strong>Valium</strong></td>
</tr>
<tr>
<td>‘monologue’</td>
<td>‘giraffe’</td>
<td>‘vanilla’</td>
<td>‘Valium’</td>
</tr>
<tr>
<td><strong>Ornament</strong> (WWS, m)</td>
<td><strong>Kamerad</strong> (WWS)</td>
<td><strong>Praline</strong></td>
<td><strong>Praktikum</strong></td>
</tr>
<tr>
<td>‘ornament’</td>
<td>‘comrade’</td>
<td>‘chocolate’</td>
<td>‘internship’</td>
</tr>
<tr>
<td><strong>Papagei</strong> (WWS, m)</td>
<td><strong>Melone</strong> (WSW)</td>
<td><strong>Senator</strong></td>
<td><strong>Senior</strong></td>
</tr>
<tr>
<td>‘parrot’</td>
<td>‘melon’</td>
<td>‘senator’</td>
<td>‘senior’</td>
</tr>
<tr>
<td><strong>Rituell</strong> (WWS, m)</td>
<td><strong>Granate</strong> (WSW)</td>
<td><strong>Malaie</strong></td>
<td><strong>Malerin</strong></td>
</tr>
<tr>
<td>‘ritual’</td>
<td>‘shell’</td>
<td>‘Malayan’</td>
<td>‘painter’</td>
</tr>
<tr>
<td><strong>Sinfonie</strong> (WWS, m)</td>
<td><strong>Akrobat</strong> (WWS)</td>
<td><strong>Koralle</strong></td>
<td><strong>Korridor</strong></td>
</tr>
<tr>
<td>‘symphony’</td>
<td>‘acrobat’</td>
<td>‘coral’</td>
<td>‘corridor’</td>
</tr>
<tr>
<td>Target (T): WSW</td>
<td>Competitor (C): SWW</td>
<td>Num char</td>
<td>Lexical Frequency (o.p.m.)</td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------</td>
<td>----------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Alaska</td>
<td>[a.'las.ka]</td>
<td>6 5</td>
<td>80 71</td>
</tr>
<tr>
<td>Albaner</td>
<td>[al.'ba:nə]</td>
<td>7 8</td>
<td>45 19</td>
</tr>
<tr>
<td>Anapher</td>
<td>[a.'na:fə]</td>
<td>7 6</td>
<td>4 63</td>
</tr>
<tr>
<td>Anode</td>
<td>[a.'no:də]</td>
<td>5 6</td>
<td>10 11</td>
</tr>
<tr>
<td>Aroma</td>
<td>[a.'ro:ma]</td>
<td>5 4</td>
<td>102 293</td>
</tr>
<tr>
<td>Embargo</td>
<td>[em.'ba:go]</td>
<td>7 6</td>
<td>41 97</td>
</tr>
<tr>
<td>Enklave</td>
<td>[en.'klə:və]</td>
<td>7 7</td>
<td>39 97</td>
</tr>
<tr>
<td>Eskorte</td>
<td>[es.'kɔ:ɾ.tə]</td>
<td>7 6</td>
<td>26 24</td>
</tr>
<tr>
<td>Exotik</td>
<td>[ɛk.'sɔ:tək]</td>
<td>6 6</td>
<td>14 66</td>
</tr>
<tr>
<td>Facette</td>
<td>[fa.'se:ta]</td>
<td>7 8</td>
<td>13 1</td>
</tr>
<tr>
<td>Furore</td>
<td>[fu.'ro:ra]</td>
<td>6 5</td>
<td>28 20</td>
</tr>
<tr>
<td>Genetik</td>
<td>[ɡe.'ne:tək]</td>
<td>7 7</td>
<td>42 158</td>
</tr>
<tr>
<td>Kabine</td>
<td>[ka.'bi:nə]</td>
<td>6 8</td>
<td>174 27</td>
</tr>
<tr>
<td>Kamille</td>
<td>[ka.'mɛ:la]</td>
<td>7 7</td>
<td>30 168</td>
</tr>
<tr>
<td>Kanister</td>
<td>[ka.'nɛs.te]</td>
<td>8 7</td>
<td>32 70</td>
</tr>
<tr>
<td>Kanone</td>
<td>[ka.'no:nə]</td>
<td>6 6</td>
<td>145 442</td>
</tr>
<tr>
<td>Karotte</td>
<td>[ka.'ɾɔ:tə]</td>
<td>7 7</td>
<td>10 2</td>
</tr>
<tr>
<td>Kaverne</td>
<td>[ka.'vɛ:ɾə]</td>
<td>7 6</td>
<td>17 100</td>
</tr>
<tr>
<td>Kolumne</td>
<td>[ko.'lu:mənə]</td>
<td>7 7</td>
<td>18 10</td>
</tr>
<tr>
<td>Libelle</td>
<td>[li.'be:la]</td>
<td>7 6</td>
<td>40 8</td>
</tr>
<tr>
<td>Manege</td>
<td>[ma.'ne:ʒə]</td>
<td>6 6</td>
<td>55 2</td>
</tr>
<tr>
<td>Marille</td>
<td>[ma.'ɾi:la]</td>
<td>7 8</td>
<td>3 47</td>
</tr>
<tr>
<td>Markise</td>
<td>[mar.'ki:zə]</td>
<td>7 6</td>
<td>8 98</td>
</tr>
<tr>
<td>Marotte</td>
<td>[ma.'ɾɔ:tə]</td>
<td>7 10</td>
<td>33 23</td>
</tr>
<tr>
<td>Medaille</td>
<td>[me.'dalja]</td>
<td>8 7</td>
<td>200 6</td>
</tr>
<tr>
<td>Monokel</td>
<td>[mo.'nɔ:kl]</td>
<td>7 7</td>
<td>87 124</td>
</tr>
<tr>
<td>Panade</td>
<td>[pa.'na:da]</td>
<td>6 6</td>
<td>5 184</td>
</tr>
<tr>
<td>Posaune</td>
<td>[po.'zɔ:u.nə]</td>
<td>7 7</td>
<td>60 47</td>
</tr>
<tr>
<td>Prämisse</td>
<td>[pr.'rä:si]</td>
<td>8 6</td>
<td>64 143</td>
</tr>
<tr>
<td>Radieschen</td>
<td>[ra.'dʒi:sçən]</td>
<td>10 6</td>
<td>59 113</td>
</tr>
<tr>
<td>Spirale</td>
<td>Spiritus</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Statistik</td>
<td>Statue</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>
Table C.4: Overview of lmer results of time-window analysis in 6 subsequent 100 ms-windows in Experiment 3. Intonation was modelled as a fixed factor, participants and items as crossed-random factors. Best model fit is indicated (it appeared to be the same for all windows).

<table>
<thead>
<tr>
<th>Exp. 3</th>
<th>Win 0a (575-675 ms)</th>
<th>Win 0b (675-775 ms)</th>
<th>Win 1 (775-875 ms)</th>
<th>Win 2 (875-975 ms)</th>
<th>Win 3 (975-1075 ms)</th>
<th>Win 4 (1075-1175 ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Into condition</td>
<td>$p=0.99$</td>
<td>$p=0.75$</td>
<td>$p=0.60$</td>
<td>$\textcolor{red}{p=0.02^*}$</td>
<td>$p=0.18$</td>
<td>$p=0.59$</td>
</tr>
<tr>
<td><strong>Best model fit</strong></td>
<td>lmer(elogC ~ cond + (1</td>
<td>participant) + (1</td>
<td>item), data = data)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that window 0 refers to the time before information about the target words was present in order to check for initial biases. Window 1 starts at the onset of the target shifted by 200 ms.

$^*$ $\beta = 0.4 \ [0.05; 0.65], \ SE = 0.15, \ df = 732, \ t = 2.31, \ p = 0.02$
Appendix D: Experiment 3 – Post-hoc reading study

Table D.1: Summary of meta-data of participants.

<table>
<thead>
<tr>
<th>ID</th>
<th>List</th>
<th>Age</th>
<th>Gender</th>
<th>State (born)</th>
<th>State (grown up)</th>
<th>Used to reading</th>
<th>Like reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>p02</td>
<td>2c</td>
<td>26</td>
<td>Male</td>
<td>BW</td>
<td>BW</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>p03</td>
<td>1a</td>
<td>32</td>
<td>Male</td>
<td>RLP</td>
<td>RLP</td>
<td>No</td>
<td>Neutral</td>
</tr>
<tr>
<td>p04</td>
<td>2c</td>
<td>22</td>
<td>Female</td>
<td>BW</td>
<td>BW</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>p06</td>
<td>2c</td>
<td>33</td>
<td>Female</td>
<td>LS</td>
<td>LS</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>p07</td>
<td>1a</td>
<td>23</td>
<td>Female</td>
<td>BW</td>
<td>BW</td>
<td>Yes</td>
<td>Neutral</td>
</tr>
<tr>
<td>p08</td>
<td>2c</td>
<td>23</td>
<td>Male</td>
<td>BY</td>
<td>BY</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>p09</td>
<td>1a</td>
<td>24</td>
<td>Female</td>
<td>BW</td>
<td>BW</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>p10</td>
<td>2c</td>
<td>23</td>
<td>Female</td>
<td>BW</td>
<td>BW</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>p11</td>
<td>1a</td>
<td>28</td>
<td>Male</td>
<td>NRW</td>
<td>NRW/BW</td>
<td>No</td>
<td>Neutral</td>
</tr>
<tr>
<td>p12</td>
<td>1a</td>
<td>24</td>
<td>Female</td>
<td>HE</td>
<td>BW</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

BW = Baden-Wuerttemberg
BY = Bavaria
RLP = Rhineland-Palatinate
NRW = North Rhine-Westphalia
LS = Lower Saxony

Note that two participants were excluded (p01 and p05), one due to a technical error with the recording device and one for having a Swiss-German background.
Table D.2: Exemplar realizations for all four nuclear categories produced in the reading: a) early-peak accent, b) medial-peak accent, c) and d) late-peak contours, and e) a stress accent. Tier 1 gives the word in German, tier 2 the English translation. Tier 3 show accented words and Tier 4 provide GToBI labels.

**a) Early-peak accent (H+L* L-%), produced by p03**

**b) Medial-peak accent (L+H* L-%), produced by p08**

**c) Late-peak accent (L*+H L-%), produced by p03**

**d) Late-peak contour (L* H-^H%), produced by p04**

**e) Stress accent, produced by subject 04, female, BW**
D.3: Distribution of accent types in post-hoc reading study

Figure D.3a Distribution of accents in post-hoc reading study, split by experimental quarters.

Figure D.3a suggest that early-peak accents became more frequent towards the end of the experiment. To illustrate, in the first experimental quarter, early-peak accents occurred in 17% of the cases (n = 26), medial-peak accents in 74% of the cases (n = 112), late-peak accents in 4% of the cases (n = 6), and stress accent in 5% of the cases (n = 8). In the second experimental quarter, early-peak accents occurred in 17% of the cases (n = 26), medial-peak accents in 70% of the cases (n = 110), late-peak accents in 7% of the cases (n = 11), and stress accent in 6% of the cases (n = 10). In the third experimental quarter, early-peak accents occurred in 14% of the cases (n = 22), medial-peak accents in 78% of the cases (n = 123), late-peak accents in 2% of the cases (n = 3), and stress accent in 6% of the cases (n = 10). Finally, in the fourth experimental quarter, early-peak accents occurred in 27% of the cases (n = 42), medial-peak accents in 61% of the cases (n = 96), late-peak accents in 8% of the cases (n = 12), and stress accents in 5% of the cases (n = 8). The very last word was realized with an early-peak accent by six of the ten participants.
Figure D.3b shows the distribution of accents in post-hoc reading study, split by word-prosodic structure. Medial-peak accents seemed to be less frequent in WSW-words than in other stress patterns. To illustrate, for SWW-words (present in cohort and filler trials in the eye-tracking experiments), 18% were realized with an early-peak accent ($n = 44$), 75% with a medial-peak accent ($n = 186$), 7% with a late-peak accent ($n = 17$), and 0.8% with a stress accent ($n = 2$). For WSW-words (present in cohort and filler trials), 20% were realized with an early-peak accent ($n = 51$), 66% with a medial-peak accent ($n = 172$), 5% with a late-peak accent ($n = 12$), and 10% with a stress accent ($n = 25$). For WWS-words (only present in fillers), 18% were realized with an early-peak accent ($n = 21$), 72% with a medial-peak accent ($n = 83$), 3% with a late-peak accent ($n = 3$), and 8% with a stress accent ($n = 9$).
## Appendix E: Experiment 4 – Materials in exposure-phase

Table E.1: List of alternative-question units used in the exposure phase of Experiment 4. Translations are provided for each item.

<table>
<thead>
<tr>
<th>Type</th>
<th>Alternative Question*</th>
<th>Answerb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal</td>
<td><em>Grillen wir oder bestellen wir was?</em>&lt;br&gt;‘Are we having barbecue or do we ordering something?’</td>
<td><em>Wir grillen.</em>&lt;br&gt;‘We’re having a barbecue.’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Laufen wir in die Stadt oder fahren wir?</em>&lt;br&gt;‘Are we walking to the city or are we driving?’</td>
<td><em>Wir fahren.</em>&lt;br&gt;‘We’re driving.’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Verschenken wir den Kuchen oder verkaufen wir ihn?</em>&lt;br&gt;‘Are we giving away the cake or are we selling it?’</td>
<td><em>Wir verkaufen ihn.</em>&lt;br&gt;‘We’re selling it.’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Bepflanzen wir den Garten oder lassen wir ihn so?</em>&lt;br&gt;‘Are we planting the garden or do we leave it as it is?’</td>
<td><em>Wir bepflanzen ihn.</em>&lt;br&gt;‘We’re planting it.’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Schreiben wir oder telefonieren wir?</em>&lt;br&gt;‘Are we writing or are you calling?’</td>
<td><em>Wir telefonieren.</em>&lt;br&gt;‘I’m calling’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Feilen oder sägen wir das Stahlstück?</em>&lt;br&gt;‘Are we filing or sawing the steel piece?’</td>
<td><em>Wir feilen es.</em>&lt;br&gt;‘We’re filing it.’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Ist das gestrickt oder gehäkelt?</em>&lt;br&gt;‘Did you knit or crochet this?’</td>
<td><em>Gestrickt.</em>&lt;br&gt;‘I knitted it.’</td>
</tr>
<tr>
<td>Verbal</td>
<td><em>Bist du dir sicher oder vermutest du das?</em>&lt;br&gt;‘Are you sure or are you guessing?’</td>
<td><em>Ich vermute es.</em>&lt;br&gt;‘I’m guessing.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Hat Italien oder Brasilien gewonnen?</em>&lt;br&gt;‘Did Italy or Brazil win?’</td>
<td><em>Brasilien.</em>&lt;br&gt;‘Brazil.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Hast du eine Fliege oder eine Mücke gesehen?</em>&lt;br&gt;‘Did you see a fly or a mosquito?’</td>
<td><em>Eine Fliege.</em>&lt;br&gt;‘A fly.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Willst du die Schatulle oder die Tasche verschenken?</em>&lt;br&gt;‘Do you want to give away the casket or the bag?’</td>
<td><em>Die Schatulle.</em>&lt;br&gt;‘The casket.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Hast du dir eine Bürste oder einen Kamm gekauft?</em>&lt;br&gt;‘Did you buy a comb or a hair brush?’</td>
<td><em>Einen Kamm.</em>&lt;br&gt;‘A comb.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Willst du Pommes oder eine Bratwurst essen?</em>&lt;br&gt;‘Would you like fries or a sausage?’</td>
<td><em>Eine Bratwurst.</em>&lt;br&gt;‘A sausage.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Hat dich deine Schwester oder dein Vater besucht?</em>&lt;br&gt;‘Did your sister or your father come visit?’</td>
<td><em>Meine Schwester.</em>&lt;br&gt;‘My sister.’</td>
</tr>
<tr>
<td>Nominal</td>
<td><em>Magst du Bolognese oder Carbonara lieber?</em>&lt;br&gt;‘Do you prefer Bolognese or Carbonara?’</td>
<td><em>Carbonara.</em>&lt;br&gt;‘Carbonara’</td>
</tr>
</tbody>
</table>

The accented words are **underlined**. The intonational contour realized on the accent words is:
* L*+H H+L* L-%
* H+L* L-%
Table E.2: List of contrastive-topic units used in the exposure phase of Experiment 4. Translations are provided for each item.

<table>
<thead>
<tr>
<th>Contrastive-topic units</th>
<th>Declarative 1(^a)</th>
<th>Declarative 2(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wb-question</strong>(^b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Womit fährt die Bäuerin?</em></td>
<td>‘What does the farmer.fem drive?’</td>
<td><em>Die Bäuerin fährt mit dem Sportwagen.</em></td>
</tr>
<tr>
<td><em>Was serviert der Kellner?</em></td>
<td>‘What does the waiter serve?’</td>
<td><em>Der Kellner serviert das Dessert.</em></td>
</tr>
<tr>
<td><em>Woraus macht man Seide?</em></td>
<td>‘What is silk made of?’</td>
<td><em>Seide wird aus Seidenraupenkokons hergestellt.</em></td>
</tr>
<tr>
<td><em>Was wird in der Weihnachtszeit gemacht?</em></td>
<td>‘What is done for Christmas?’</td>
<td><em>In der Weihnachtszeit werden Plätzchen gebacken.</em></td>
</tr>
<tr>
<td><em>Welche Nummer trägt der Stürmer?</em></td>
<td>‘Which number does the striker wear?’</td>
<td><em>Das Kissen ist aus Federn.</em></td>
</tr>
<tr>
<td><em>Woraus trinkt man Bier?</em></td>
<td>‘What glasses are used for beer?’</td>
<td><em>Für einen Apfelkuchen braucht man Apfel.</em></td>
</tr>
<tr>
<td><em>Was braucht man für einen Apfelkuchen?</em></td>
<td>‘What do we need for an apple cake?’</td>
<td><em>Im Sommer kann man in den See hüpfen.</em></td>
</tr>
<tr>
<td><em>Was kann man im Sommer unternehmen?</em></td>
<td>‘What can one do in summer?’</td>
<td><em>Nachts wird meist geschlafen.</em></td>
</tr>
<tr>
<td><em>Was macht man nachts?</em></td>
<td>‘What does one do in the night?’</td>
<td><em>Bei den Alemannen sagt man Grüßgott.</em></td>
</tr>
<tr>
<td><em>Was sagt man bei den Alemannen?</em></td>
<td>‘What do the Alemanns [in the south] say?’</td>
<td></td>
</tr>
<tr>
<td><em>Wo wachsen Tannenbäume?</em></td>
<td>‘Where do fir trees grow?’</td>
<td></td>
</tr>
<tr>
<td><em>Welche Nummer trägt der Torwart?</em></td>
<td>‘Which number does the keeper wear?’</td>
<td></td>
</tr>
<tr>
<td><em>Wann blühen Narzissen?</em></td>
<td>‘When do daffodils blossom?’</td>
<td></td>
</tr>
<tr>
<td><em>Woraus macht man Seide?</em></td>
<td>‘What is silk made of?’</td>
<td></td>
</tr>
<tr>
<td><em>Wann schneidet man die Weihnachtskugeln?</em></td>
<td>‘When do one cut the Christmas balls?’</td>
<td></td>
</tr>
<tr>
<td><em>Wo wachsen Tannenbäume?</em></td>
<td>‘Where do fir trees grow?’</td>
<td></td>
</tr>
<tr>
<td><em>Woraus macht man Seide?</em></td>
<td>‘What is silk made of?’</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Declarative 1
\(^b\) Wb-question
\(^c\) Declarative 2

For further information, see the original text in the reference.
### Table E.2 Continued

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>'It is easy to eat sorbet with a spoon.'</td>
<td>'What is meat easy to eat with?'</td>
<td>'It is easy to eat meat with a fork.'</td>
</tr>
</tbody>
</table>

The accented words are underlined. The intonational contour realized on the accent words is:

a $L^*+H H+L^* L\%$

b $L^* H\H L\%$

c $L^*+H H+L^* L\%$
Table E.3: Acoustic analysis in regard to f0 for alternative-question units, means (and standard deviations).

<table>
<thead>
<tr>
<th>Acoustic variable</th>
<th>Alternative question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch range of accentual movement in L*+H in st</td>
<td>10.6 (1.7)</td>
<td>NA</td>
</tr>
<tr>
<td>Pitch range of accentual movement in H+L* in st</td>
<td>12.7 (2.4)</td>
<td>13.9 (2.0)</td>
</tr>
<tr>
<td>Mean f0 of low-pitched stressed syllable in L*+H in Hz</td>
<td>168.7 (64.1)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean f0 of high-pitched unstressed syllable carrying in L*+H in Hz</td>
<td>283.8 (61.4)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean f0 of low-pitched stressed syllable in H+L* in Hz</td>
<td>193.8 (28.7)</td>
<td>192.9 (47.0)</td>
</tr>
<tr>
<td>Mean f0 of high-pitched unstressed syllable in H+L* in Hz</td>
<td>332.9 (49.6)</td>
<td>344.7 (44.0)</td>
</tr>
</tbody>
</table>

Table E.4: Acoustic analysis in regard to f0 for contrastive-topic units, means (and standard deviations).

<table>
<thead>
<tr>
<th>Acoustic variable</th>
<th>Declarative 1</th>
<th>Declarative 2</th>
<th>Wh-question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch range of accentual Movement in L*+H in st</td>
<td>8.5 (1.9)</td>
<td>9.6 (1.5)</td>
<td>NA</td>
</tr>
<tr>
<td>Pitch range of accentual Movement in H+L* in st</td>
<td>12.1 (1.7)</td>
<td>12.7 (1.4)</td>
<td>NA</td>
</tr>
<tr>
<td>Pitch range of accentual movement in L* H-^H% in st</td>
<td>NA</td>
<td>NA</td>
<td>11.8 (4.2)</td>
</tr>
<tr>
<td>Mean f0 of low-pitched stressed syllable in L*+H in Hz</td>
<td>151.8 (10.7)</td>
<td>160.2 (23.6)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean f0 of high-pitched unstressed syllable in L*+H in Hz</td>
<td>228.6 (28.1)</td>
<td>253.5 (53.1)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean f0 of low-pitched stressed syllable in H+L* in Hz</td>
<td>190.2 (46.0)</td>
<td>179.3 (38.5)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean f0 of high-pitched unstressed syllable in H+L* in Hz</td>
<td>294.7 (32.0)</td>
<td>294.1 (57.0)</td>
<td>NA</td>
</tr>
<tr>
<td>Mean f0 of low-pitched stressed syllable in L* in Hz</td>
<td>NA</td>
<td>NA</td>
<td>151.5 (30.1)</td>
</tr>
<tr>
<td>Mean f0 of high-pitched unstressed syllable of H-^H% in Hz</td>
<td>NA</td>
<td>NA</td>
<td>243.3 (45.0)</td>
</tr>
</tbody>
</table>